How Cultures Converge: An Empirical Investigation of Linguistic Exchange

Arthur Blouin
University of Toronto

Julian Dyer
University of Exeter

January 6, 2023*

This paper aims to address the question of how languages evolve. To do this we construct a data-set that measures linguistic influence, and that covers most languages in the world. These data reveal that while linguistic influence is relatively common, languages are influenced by the language groups they interact with heavily, so much so that cases of distant influence are rare. When we focus on local linguistic influence, we find that economic trade incentives shape both the extent, and direction of linguistic exchange. We estimate gains from local agricultural trade based on plausibly exogenous complementarity in nutritional endowments. This variation allows us to empirically document three main points. First, linguistic convergence is largest for groups with the greatest incentives to trade. Second, the linguistic influence that results from trade goes well beyond what would be strictly required for trade, and instead impacts domains such as technology, philosophy, and politics. Third, economic leverage determines the direction of convergence within a pair of language groups. The role of economic leverage in asymmetric convergence suggests power dynamics as one possible mechanism. We pursue this idea further in an ancillary empirical exercise by exploring the role of colonialism in shaping language across the globe. We find that colonial relationships, which are characterized by extreme power asymmetries, also exhibit significant and asymmetric linguistic influence.

Keywords: Language Evolution; Linguistic Distance; Linguistic Exchange; Loanwords; Trade Incentives.

JEL Codes: O10, Z10, Z13.

[*] We are grateful to Francesco Amodio, Elliott Ash, Gustavo Bobonis, Loren Brandt, Andrew Dickens, Shari Eli, Benjamin Enke, James Fenske, Per Fredriksson, Leander Heldring, Julian Jamison, Nicholas Li, Rocco Macchiavello, Stelios Michalopoulos, Martina Miotto, Peter Morrow, Sharun Mukand, Naomi Nagy, Nathan Nunn, Hilton Root, Jordan Roper, Dominik Stammbach, and Joachim Voth for helpful comments. Frederick Gietz, Dina O'Brien, and Matthew Schwartzman provided outstanding research assistance. We also thank seminar and workshop participants at ASREC, BSE Summer Forum, CAGE Summer School, Econometric Society Meetings, ETH Zurich Economics + Data Science Seminar, University of Exeter, H2D2, Columbia University, New York University, University of British Colombia, University of Toronto, and University of Warwick. We gratefully acknowledge financial support from SSHRC and the Connaught fund.

1. Introduction

Questions surrounding the evolution of languages have been central in the social sciences dating back at least to the Enlightenment, when it was argued that language evolution could be attributed to economic forces. The predominant view was that people "...naturally begin to form that language by which they would endeavour to make their mutual wants intelligible to each other" (Smith 1762, pp. 203). Since then, the study of language development, and cross-cultural variation in language (or lack thereof), has been prominently undertaken by anthropologists (Lévi-Strauss 1951), philosophers (Foucault 1971), biologists, (L. L. Cavalli-Sforza and F. Cavalli-Sforza 1994), psychologists (Pinker 2003), and of course, linguists (Chomsky et al. 2006). Economists have also taken a keen interest in language differences, with a focus on how they constrain economic trade and development,[1] or how they are established in the first place (Galor, Özak, and Sarid 2018; Dickens et al. 2022).[2] However, given the prominence of the question in other disciplines, there has been surprisingly little attention given, within economics, to how languages evolve and are influenced once they are established.[3] Accordingly, "[b]eyond case studies, the empirical literature on... economic integration, and culture is yet inconclusive" (Bisin and Verdier 2014).

One reason that economists have often shied away from this question is a relatively stronger preference for empirical analysis, combined with the fact that suitable data on changes in language have not been available. In this paper we overcome that challenge by using machine-learning techniques inspired by computational linguistics to compile a global, sub-national and consistently measured data-set on loanwords. Loanwords are words that have been adopted from other languages, and represent about 20% of a typical language.[4] They are interesting in their own right since they directly measure language influence but they have also generally been interpreted as an indicator of cross-cultural transmission more broadly for nearly 100 years (Bloomfield 1933; Scotton and Okeju 1973; Frankopan 2016).[5] Notably, societies concerned with cultural erosion are often primarily

[1] This literature is very large. A few notable contributions on the development side are (Alesina, Devleeschauwer, et al. 2003; Desmet, Ortuño-Ortín, and Wacziarg 2012; Alesina, Michalopoulos, and Papaioannou 2016). On the cultural barriers to trade, contributions include (Dasgupta and Serageldin 1999; Glaeser, Laibson, and Sacerdote 2002; Durlauf and Fafchamps 2005; Putnam 2007; Guiso, Sapienza, and Zingales 2008; Hall and Jones 1999; Rauch and Trindade 2002; Giuliano, Spilimbergo, and Tonon 2014; Melitz 2008; Guiso, Sapienza, and Zingales 2009; Felbermayr and Toubal 2010; Gokmen 2017).

[2] There is also a literature that uses language as a proxy measure of cultural characteristics (Chen 2013; Jakiela and Ozier 2021)

[3] Naidu, Hwang, and Bowles (2017) is among the exceptions, showing how power and social conditions within a language-group can influence the evolution of unequal linguistic conventions.

[4] There is a 150 year-old literature in linguistics that aims to distinguish between cognate words, which are inherited from parent languages; loanwords, which are adopted from neighbouring languages; and native words.

[5] e.g. "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.

concerned with language erosion (Schmid, Zepa, and Snipe 2004), and conversely, fears of cultural erosion are a key factor limiting loanword adoption (Haspelmath and Tadmor 2009).

We leverage the loanwords data to empirically investigate whether it is appropriate to think about language "...as an exogenous variable determined by the static linguistic fabric of society" (Ginsburgh and Weber 2016). The computational linguistics approach that we employ to generate the data systematizes the process of determining loanwords that is outlined in *Loanwords of the World's Languages: A Handbook* (Haspelmath and Tadmor 2009). We follow the guidelines as closely as possible. This entails using only information on language ancestry and how words sound, how they are spelled, and what they mean, to estimate loanwords. Using this framework, we are able to estimate the loanword status of nearly all documented words in nearly all documented languages. While linguists have never (to our knowledge) estimated the loanword status of this many words, the use of machine learning to estimate loanwords has been demonstrated to be a valid approach. For example, we use the same training data as J. E. Miller et al. 2020, make predictions on the same types of features, and achieve similar performance.

The result is a data-set that measures directed linguistic influence and adoption: (1) for over 16 million language-pairs (covering over 4,000 ethnolinguistic groups) across the globe; (2) in an objective and consistently measured way; and (3) across a wide range of dimensions. These data can be aggregated to the language-pair level; the language group level; or the country or country-pair level. They can also be disaggregated to the concept-language-pair level, for all concepts/ideas that can be represented by a list of keywords. We validate these data for use by economists by showing that greater adoption of loanwords between a pair is associated with reduced cultural distance, using a variety of data sources commonly used to measure cultural traits.[6]

Our main goal is to explore the determinants of language convergence between societies.[7] We start with a basic hypothesis in the linguistics literature, that inter-group contact is a core driver of language exchange. We show that the distance between groups - a proxy for the extent of contact - is a very strong predictor of linguistic exchange. This holds globally, within countries, and even within relatively small regional neighbourhoods. While it is not particularly surprising that language exchange is common locally, what is perhaps more surprising is that among distant language-group pairs, language exchange is nearly non-existent. It is surprising because the absence of almost any distant inter-group influence stands in contrast to a linguistics literature that views globalized trade as a core threat to linguistic diversity.

To explore the role of trade more concretely, we need exogenous variation in local

[6] In a similar vein, it is also true that more loanwords are associated with reduced barriers to the diffusion of development.

[7] Here we define societies as language groups included in the Ethnologue. For ease of exposition, we will refer to language groups and societies interchangeably.

economic interaction.[8] To this end, we estimate a simple model of local agricultural Ricardian trade. The intuition of the model draws on insights from Galor and Özak 2015 and Costinot and Donaldson 2012. We focus on the core mechanism that agricultural trade is driven by exogenous variation in soil characteristic complementarity, assuming that societies exchange nutritionally complementary crops to increase their population. To estimate the model we combine data on the agro-ecologically determined potential production of most crops - which we interpret as nutritional endowments - with information on human nutritional requirements. This simple model with multiple goods - crops - and a single factor - agricultural land - accurately predicts contemporary production for each locally traded crop, among other validation checks.[9]

We use this model to estimate exogenous measures of trade incentives between any two language groups within the same geographic region.[10] To generate these parameters we estimate a society's trade utility both under (a) free trade with all groups, and (b) the counterfactual scenario(s) where each society can trade with all but one group (for each group in turn). For each language-group pair in a region, we interpret the percentage welfare gained under free trade as *gains from trade*, and also construct *trade influence*, which is gains from trade from the reversed perspective.[11]

We then use these two exogenously determined parameters in a reduced-form analysis to explore trade and linguistic convergence. First, the trade incentives data allow us to replicate the result that inter-societal contact generates linguistic convergence. For instance, a trading partner's quality strongly predicts the intensity of language adoption. In terms of magnitude, the gains from trade with a typical society's best local agricultural trade partner accounts for about 10% of the local loanwords for a typical society. Even within this relatively narrow focus on local agricultural trade, the evidence confirms that inter-group contact can homogenize cultures.[12] Furthermore, a society's gains from trade are *only* associated with linguistic borrowing, and *not* with linguistic lending.

This latter result suggests that economic leverage may be important in determining how languages evolve. If economic leverage did not matter - and inter-group contact was all that mattered - the expectation might be that trade incentives going in either direction

[8] The variation should be local given the absence of variation in linguistic influence among distant partners.

[9] See Feenstra 2004 for a discussion of this classic model, first introduced in Ricardo 1817. See also Costinot and Donaldson 2012 for empirical support for this type of model applied to agricultural production.

[10] Going forward, for ease of exposition, we will refer to language groups and societies interchangeably, and we refer to groups within the same region as neighbours.

[11] To be more precise. For neighbours i and j, utility under free trade is U_i^{FT}; and under the counterfactual where they can trade with all but one neighbour it is U_i^{FT-j}. *Gains from trade* is $\frac{U_i^{FT}-U_i^{FT-j}}{U_i^{FT}}$, and conversely *trade influence* is $\frac{U_j^{FT}-U_j^{FT-i}}{U_j^{FT}}$.

[12] These results also hold with a data-set we constructed of contemporary bilingualism, reinforcing the assertion that our loanwords measure accurately captures linguistic adoption.

would generate symmetric adoption. To pursue this idea further, we empirically confirm the importance of trade leverage by looking *within* a language-pair to fix the extent of inter-group contact. On average, within a society-pair, the party who linguistically converges the most towards the other is the one who gains the most from trade.[13] In fact, it typically takes only an 8% difference in gains from trade to induce one society to be the only one to converge.

The causal argument for our main result is predicated on the assumption that gains from trade do not influence language adoption other than through actual trade.[14] To demonstrate this, we show that there is no correlation between gains from trade and borrowing for society-pairs that are not viable trade partners. Further, we investigate several alternative mechanisms. For instance, a separate exercise highlights that migration cannot explain the patterns in our data. Finally, each empirical specification controls for land diversity, which has been shown to influence ethnic diversity (Michalopoulos 2012).

While the asymmetry in language exchange suggests that trade leverage, or bargaining power, determines the direction of convergence, there could be other mechanisms at play as well. In an ancillary empirical exercise, we take a suggestive but more direct look at the role of power. We explore the impact of colonialism - one of the primary drivers of recent language change across the globe and an example of relationships with significant asymmetry in power. Consistent with power asymmetries playing a prominent role, in our data colonies adopt about 50 times more from their colonizers than from another distant partner. This result holds both using a descriptive approach, or if we implement an identification strategy inspired by Michalopoulos and Papaioannou 2014 that exploits the arbitrary borders established by colonists. However, the reverse is not true. At least on average, colonial influence is nearly entirely one-sided - colonists adopt very little from their colonies. This strong asymmetry suggests that power dynamics could play an important role in how languages evolve, though there are other potential mechanisms generating asymmetric interaction. Power and status asymmetries have been hypothesized by linguists to be important for the evolution of languages (Labov 1964; Labov and Harris 1994), as those with less status and power tend to adopt from those with more.

This analysis contributes to the literature by highlighting that as societies interact, either economically or politically, they influence each other's language, reducing linguistic distance in systematic ways. A related point - that economic exchange can shape language-group ancestry - has been demonstrated by Dickens 2019.[15] A second contribu-

[13] This is consistent with recent archaeological evidence suggesting that societies in Southern Africa maintained persistent exchange networks over hundreds of kilometres, well beyond the range of a forager. They did this in order to mitigate subsistence risks and the societies who gained the most from exchange were the ones making social investments to maintain these networks (Stewart et al. 2020).

[14] A second identifying assumption is that language exchange does not influence the degree of complementarity in soil characteristics among neighbours, which seems plausible.

[15] Dickens 2019 is particularly relevant, as his work also focuses on trade. He relies on Swadesh lists,

tion is to empirically establish that economic leverage shapes the direction of language influence, and perhaps more generally, cultural adoption.[16,17] We also show suggestive evidence that power asymmetries defined more broadly shape cultural evolution. There is a large literature on the economic consequences of culture (Guiso, Sapienza, and Zingales 2004; Becker and Woessmann 2009; Atkin 2016),[18] as well as on how cultural traits evolve (Nunn and Wantchekon 2011; Alesina, Giuliano, and Nunn 2013; Becker, Boeckh, et al. 2016; Lowes, Nunn, et al. 2016; Blouin 2022). Our analysis differs by directly exploring language adoption. To our knowledge this is among the first studies to examine global patterns in the horizontal transmission of either language or culture, rather than how various cultural traits are influenced by their surroundings.[19] The closest work in this vein shows that historical trade routes predict contemporary Muslim adherence (Michalopoulos, Naghavi, and Prarolo 2018).

2. Loanwords Background

We measure cultural exchange by studying *loanwords*. A loanword is a word in one language whose sound and meaning enter the language's lexicon because it was adopted from another language (i.e. horizontally transmitted). Loanwords are distinct from *cognate* words, which are inherited from a parent language (i.e. vertically transmitted). Linguists typically take considerable effort to first distinguish between loanwords and cognates; and then conditional on identifying a loanword, to identify the direction of transmission.

The field of linguistics uses the terms *borrowing* and *lending* to describe the transfer of words between languages in either direction. It is fully understood that the metaphor is not particularly apt, yet it persists, in part, because it has become so entrenched. For instance,

> ...the borrowing takes place without the lender's consent or even awareness, and the borrower is under no obligation to repay the loan...The real advantage of the term 'borrowing' is the fact that it is not applied to language by laymen. It has therefore remained comparatively unambiguous in linguistic discussion (Haugen 1950, pp. 211–212).

which are words that are most likely to be vertically transmitted. Changes in these words are likely due to linguistic drift, after subgroups break away from the main group. The broader literature also includes related work on the origins of language groups themselves (Ahlerup and Olsson 2012; Ashraf and Galor 2013; Michalopoulos 2012; Dickens 2019).

[16] Directional cultural change has been explored theoretically (Kónya 2006), but not empirically that we know of.

[17] There is far more work on culture's influence on trade than on the reverse. This may be due to the lack of data measuring cultural convergence, as the corresponding theoretical literature is more active (Olivier, Thoenig, and Verdier 2008; Gabszewicz, Ginsburgh, and Weber 2011).

[18] See Nunn 2012 for a review.

[19] There is of course a large literature in labour economics on assimilation, e.g. Bleakley and Chin 2010; Ward et al. 2015; Algan, Mayer, Thoenig, et al. 2021. Our work differs in its focus on societal-level cultural change.

For similar reasons, throughout this paper we will also use the terms linguistic *borrowing* and *lending* to denote the process of directional transmission of words from one language to another. Furthermore, we use *language adoption* and *bilingualism* interchangeably, and we refer to *diffusion* as the process of a potential loanword spreading, and taking hold in a particular language.

When linguists study language evolution, both their focus and methodology are quite different from what would be typical in economics. For example, economists have often focused on the downsides of linguistic diversity. In linguistics, the predominant view has been that homogenization is not desirable, and a considerable amount of attention has been dedicated to how best to stave off language 'extinction.' In this vein, linguists typically view language exchange as taking place because it is imposed on a culture. For instance, "[i]t has everything to do with colonization and globalization" (Mufwene 2004).

Methodologically, linguists also differ quite substantially. They typically study loanwords in the context of detailed case studies of individual languages. These studies provide evidence that loanwords among neighbouring languages is often asymmetric and closely linked to "the nature and extent of cultural contacts" (Scotton and Okeju 1973). Loanwords are often thought to be heavily influenced by the socio-cultural context of a particular society in relation to their neighbours. Even borrowings into the core vocabulary of a language can be prevalent with enough contact with another society. The intensity of loanword adoption should therefore be thought of as the result of a socio-cultural process involving the interactions of individual speakers of languages. Accordingly, more loanwords are often viewed as a proxy for reduced cultural distance.[20]

While the adoption of foreign words or cultural traits happens by individuals, in order to become entrenched as part of a society's language or culture, it must diffuse throughout the society and take hold at the societal level. There is a large literature in socio-linguistics that focuses on diffusion. This literature attempts to identify the linguistic conditions (i.e. grammatical structure, phonetic similarity, etc.) that influence within-language diffusion of a word with foreign origins. Additionally, linguists have focused on the class and prestige of adopters as factors influencing diffusion (Labov 1964; Labov and Harris 1994), but factors like age (Sankoff and Blondeau 2007), gender (Cameron and Kulick 2003), ethnicity (Cukor-Avila and Bailey 2001), and social structure (Paolillo 2001) have also been considered.

[20] Research on loanwords has traditionally been more descriptive. However, recently linguists have begun asking less descriptive questions such as 'why are words for body parts rarely borrowed but words for objects are?' The answers are often complex. For instance, the English word *window* was adopted from Old Norse even though English had previously used the word *eagpyrel* in the same manner (Haspelmath and Tadmor 2009).

3. Language Data

The loanwords data-set that we construct is at the level of a pair of language groups. The data are directional, so we measure borrowing by a particular borrower from a particular lender separately from the borrowing by that same lender from that borrower, which is a separate observation. We define a language-group[21] according to the digitized Ethnologue map of ethnolinguistic societies (Lewis 2009). The Ethnologue includes all contemporary languages and provides the borders of each group's location in the data, from which we identify the centroids used for computing distances.[22] This is the base of the data to which everything else is matched.[23]

We proceed first by describing the sources of data we used and how they were processed. This is followed by a description and validation of the classification algorithm used to classify word pairs as a loanwords (or not). To arrive at the final data-set, we aggregate predicted loanwords to the directed language-pair level (how much each group borrows from each other group) or the society level (how much of a group's language is borrowed) as appropriate. Since the data is built from the word-level, we also aggregate to the group-pair-topic level, measuring the share of words relating to a given topic that are loanwords.[24]

3.A. Language Data Sources

In order to construct data on linguistic exchange, we need word lists, or *lexicons*, from as many societies as possible. For this, we draw on the PanLex database, which takes thousands of translation dictionaries and converts them to a single common structure, covering millions of words.[25] The data represent all known words in all known languages, as closely as we believe is possible.[26] The coverage of this data-set goes far beyond

[21] We will use the terms 'language-group' and 'society' interchangeably for ease of exposition.

[22] The Ethnologue is also the source of our data on group-level bilingualism. The standard Ethnologue data-set does not have complete bilingualism information, but the online version is more complete. We scraped the Ethnologue website, and matched the bilingualism information that exists online to our data. See appendix A.1 for further detail on the scraping and text parsing used to extract bilingualism information.

[23] To see how observations change from this base data-set as we add in information on words, language families and societal similarity, see table G1.

[24] We explain the process of identifying words related to specific topics in more detail in appendix D. We adopt a seed word approach to do this, tying our hands by starting from the Library of Congress classification system as the basis for defining word lists for each of the following concepts: technology, geography, military/warfare, science, politics, and philosophy/religion. We then use a multilingual semantic similarity routine to identify similar meanings in a way that is not reliant on English or Indo-European word associations.

[25] PanLex is a non-profit organization with a mandate to build the largest possible lexical translation database with the aim of improving resources available to under-served languages: see https://panlex.org. The database is constantly being updated, in this paper we use the SQL database from October 1, 2018.

[26] PanLex is based on expressions which can be made up of multiple words. This is particularly common in the lexicons of heavily-resourced languages, so for languages with more than 100,000 words we keep only single-word expressions for comparability. English, being the global *lingua franca*, is documented

the coverage possible with sources based on textual and archival resources, which are restricted to languages with a significant body of written history. Figure 1 shows a map of the language groups in PanLex, and table G2 shows that the coverage in PanLex is uncorrelated with a wide variety of country-level characteristics. For each word in the data-set, PanLex provides the language, spelling, and a meaning identifier. We convert each word in PanLex into the International Phonetic Alphabet (IPA) using the models and data from Ager 2019 and David R Mortensen, Dalmia, and Littell 2018.

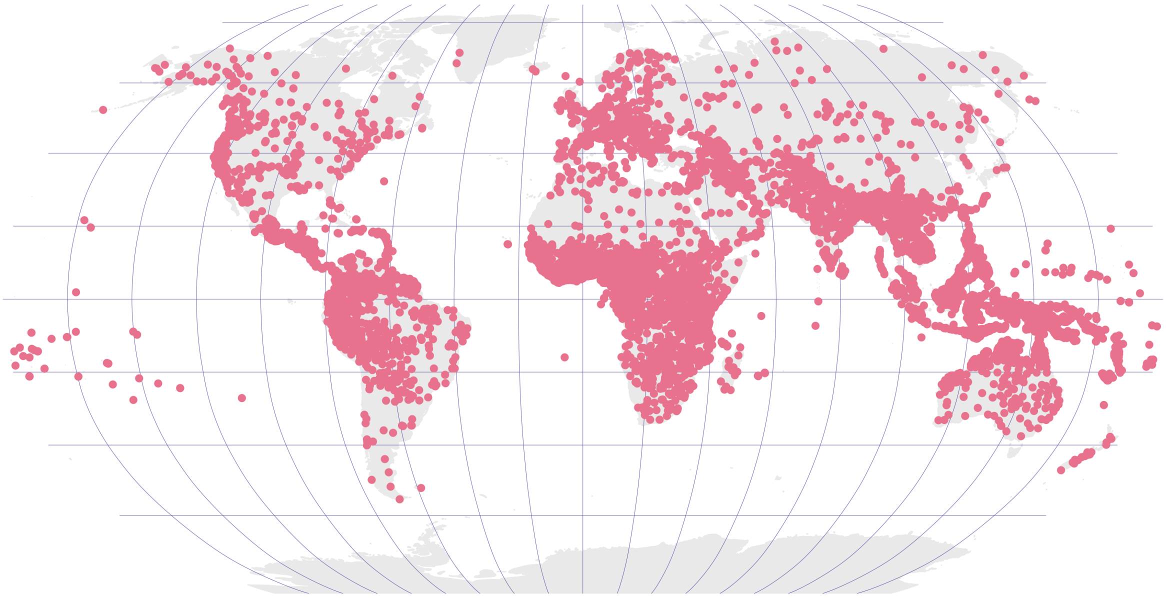


Figure 1: Map of PanLex language groups

Note: This map shows each of the borrower and lender languages in the PanLex data-set.

English is a particular outlier in the PanLex data-set. It includes many expressions that do not represent words (e.g. "A++" or "A10009"), and many words that represent proper nouns (such as "Xerox"). For this reason, when extracting the English PanLex lexicon, a number of additional data cleaning steps are necessary. First, we remove any expressions that include any non-alphabetic characters, and any expressions including a capitalized letter. Furthermore, in all of our analysis tables we do not include English as a borrowing language. We provide further details on how we clean the lexical data in appendix A.

PanLex does not include information on loanwords, so it must be matched with a data source that does. For this we use the World Loanword Database (WoLD), which is a scientific publication by the Max Planck Institute for Evolutionary Anthropology. WoLD is the first aggregated data-set of rigorously-identified loanwords under a consistent set of criteria. It provides "...vocabularies (mini-dictionaries of about 1000-2000 entries) with

especially differently from the rest, for instance, it includes almost every expression on the internet. So we do not include English as a borrowing language in our analysis. For further detail on, and steps taken to ensure the quality of, the lexical data, see appendix A comprehensive information about the loanword status of each word" (Haspelmath and Tadmor 2009) for forty-one languages across the world's language families.[27] WoLD includes words borrowed from three hundred and sixty-nine source languages, and is also a standard training data-set used in the computational linguistics literature for loanword identification (J. E. Miller et al. 2020).[28] Figure 2 presents a map of the spatial distribution of each language in WoLD, and table G3 shows that the coverage in WoLD is uncorrelated with a wide variety of country-level characteristics.[29]

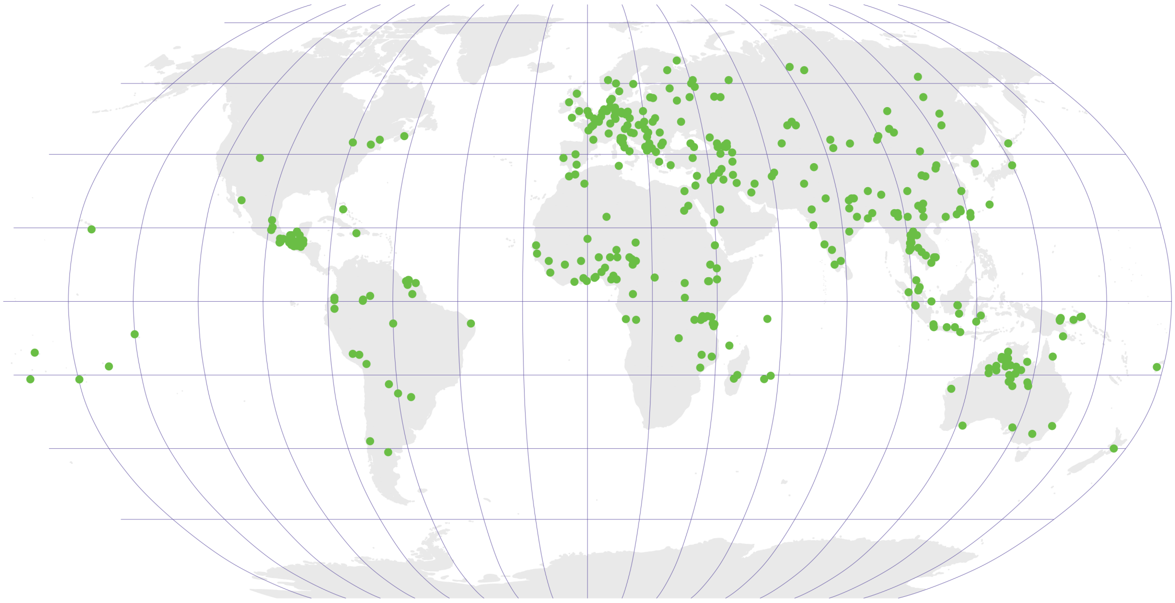


Figure 2: Map of WoLD language groups

Note: This map shows each of the borrower and lender languages in the WoLD data-set.
Source: Author constructed using data from WoLD: https://wold.clld.org/language. Last Accessed December 22, 2020 2:00pm EST.

The main reason why we focus exclusively on loanwords and not cognates or neologisms,[30] is that the monumental effort that has been put into compiling WoLD has not (yet) been made to collect analogous unified lists of known neologisms or cognates. However, loanwords are of interest primarily because it is difficult to measure cross-societal influence in other ways. This is quite different from ancestral relationships, which are also clearly an important source of cross-societal differences. Ancestral relationships could, in theory, be measured with cognates, but they can also be measured in a more direct way

[27] The WoLD data for Swahili, for example, is based on thirty-three academic publications by twenty-seven separate authors, published between 1861 and 2001 (table B3).

[28] We discuss the validity of WoLD as a training set for the broader PanLex data-set in greater detail in appendix A.

[29] As WoLD appears have a small sample of languages in the Americas, we later show (in table G9) that the results of the main specification are robust to dropping the New World. To further account for any under-representativeness of the languages in WoLD, we designed our algorithm to include only features at the word-level and not include measures like loanword share of the language or lexicon size.

[30] Cognates are words that share a sound and meaning because they share a common ancestral language. Neologisms are words that were invented within a language.

with language family trees.

3.B. Machine Learning Algorithm

WoLD is quite a large data-set, however it covers only a small fraction of PanLex. Ideally, we would like to understand, for every word in every language, whether it is a loanword, and, if so, where in the world it was borrowed from. To make progress in this direction we trained a machine learning prediction algorithm, which is the only currently feasible way to accomplish this at the scale required. Our approach closely follows and simplifies the typical method used in computational linguistics. The simplification was necessary in order to feasibly apply the algorithm to every language in PanLex, and incorporates additional data sources on ancestral languages.

The standard process used by linguists to identify loanwords is described in the reference *Loanwords in the World's Languages: A Comparative Handbook*, by Haspelmath and Tadmor 2009.[31] Similar to Blair and Ingram 1998; Blair and Ingram 2003, we generate a computational analogue to the methods that linguists use. In this subsection we summarize the steps that we took to develop the algorithm, some of the important challenges and considerations, and how we dealt with them. All of the data processing steps are described in detail in appendix B. In that section we provide a more detailed description of each step that we took to generate the data, as well as how each of those steps compares to the traditional method of loanword identification. All of that is summarized in table B1, which lists: (i) each step in the standard process that linguists use; (ii) the description of how linguists approach each step, quoted from Haspelmath and Tadmor 2009; and (iii) our computational approximation.

i) Structure and logic of the algorithm: A flow-chart depicting the logic of the algorithm is in figure B1. From PanLex we started by creating a word-pair level database of semantically similar words.[32] For the subset of word-pairs that appear in WoLD we had a good understanding of whether one word was borrowed from the other, and the direction of transfer (nodes 1a and 1b in figure B1). We drew a random sample from this subset of word-pairs with known loanword status to generate a training set for the machine learning algorithm (node 2a of figure B1).

One challenge in generating a training set was that the data are heavily unbalanced. The number of loanword pairs is minuscule relative to the number of non-loanword pairs. This is a potential problem, because the classifier could estimate that there has never, in any language ever been a loanword, and achieve very high accuracy, which is clearly not the goal. We used a combination of two methods to deal with this issue. The simplest is to under-sample the heavily-represented group, the second is synthetic minority

[31] In particular, see section *Recognizing Loanwords*.

[32] For a more detailed description of the semantic similarity routine, see appendix B.2.2.

oversampling (Chawla et al. 2002; Lemaitre, Nogueira, and Aridas 2017). This approach is similar to other methods of data augmentation used for training algorithms in similar applications in computational linguistics (Mi, Zhu, and Nie 2021).[33]

Both of these methods involve drawing a stratified random sample of different types of observations. In our case we draw from three types of non-loanword word pairs and a fourth category is the correct loanword pairs. The first type of non-loanword is word-pairs where one word actually was borrowed from the other, but the direction of transfer is wrong. Second, is word-pairs where one word is a loanword, but the other word is not the correct source. The last type, which is by far the most common, is word-pairs where neither word is a loanword.

We found, not surprisingly, that the algorithm could do extremely well at avoiding classifying obvious non-loanwords even with minimally deep trees, but that similar sounding words that were not loanwords (e.g. distinguishing loanwords from cognates) required deeper trees that considered more features. To account for this, we implemented the algorithm in two stages. The first was a coarse pass, and was accomplished using a training-set sample of (1) 2,000 actual loanword pairs (representing 8% of all known loanwords); (2) 2,000 word-pairs where the direction is flipped; (3) 30,000 loanwords matched to the wrong source; (4) 30,000 non-loanword pairs.[34]

We implemented a second pass on a sample of suspected loanwords, which were determined based on the prediction weights from the first pass. These suspected loanwords were drawn from all of the first stage observations, plus an additional 540,000 randomly drawn observations (node 4 in figure B1). We then applied the algorithm to this second stage training set, generating another set of prediction weights that are relevant specifically for suspected loanwords. This same two-step process was used both to determine the accuracy of the classifier on the test set, and to estimate the loanword status of all word-pairs in PanLex.

The selection of sample size is discussed in more depth in section C.1. To arrive at the sample sizes, we bootstrapped the training set at different sample-size intervals, getting progressively larger each time. We stopped increasing the sample at the point where, as more observations were added to the sample, the accuracy of the classifier was no longer increasing. A plot of the accuracy at each sample size we tested - for both the first and second pass - is in figure C1. The important thing to note is that in both cases test-set accuracy was increasing as we added additional observations at first, but had tapered off by the time we approached the training set sizes that we actually used in the analysis.

[33] 'Synthetic' examples of the under-represented group were re-sampled with replacement. These synthetic examples were constructed as a convex combination of nearest neighbours of the same type within feature space.

[34] There are about 25,000 loanwords in WoLD, and to match WoLD to PanLex we relied on exact spelling matches. A majority of the words did not have exact matches, so 2,000 loanwords starts to get close to all of the accurately matchable loanwords. We also leave a side a number of loanwords to be able to cross-validate the algorithm.

Table 1: Evaluating Classifier Performance

Classifier Type	Classifier Performance Measures (on Test-Set)				
	Accuracy Score	F1 Score	Balanced Accuracy Score	Precision Score	Recall Score
	(1)	(2)	(3)	(4)	(5)
Panel A: Overall:					
Random Forest	0.9825	0.8394	0.9214	0.8261	0.8532
Extremely Random Forest	0.9742	0.7598	0.8797	0.7461	0.7741
Voting Classifier	0.9826	0.8393	0.9185	0.8318	0.8471
Panel B: First-Stage:					
Random Forest	0.9835	0.8386	0.9215	0.8251	0.8526
Extremely Random Forest	0.9751	0.7580	0.8794	0.7437	0.7729
Voting Classifier	0.9836	0.8386	0.9187	0.8308	0.8466
Panel C: Second-Stage:					
Random Forest	0.9125	0.8979	0.9108	0.8968	0.8990
Extremely Random Forest	0.9067	0.8872	0.9005	0.9186	0.8579
Voting Classifier	0.9088	0.8922	0.9055	0.9019	0.8828

Note: Each column presents algorithm performance using a different metric, each weights false-positives relative to false-negatives differently. We show each for both our first and second stage classifiers. All performance measures are performed on the test-set, none of the training-set data is included in this table. Accuracy is simply the share of word-pairs classified as a loanword, are actually loanwords, with the correct direction of borrowing. The precision score is the share of predicted positives that are true positives. The recall score is the share of actual positives that are predicted. Balanced accuracy is the mean of the true positive rate and the true negative rate. The F1-score is the harmonic mean of precision and recall. On a random sample of approximately 8 million word-pairs, the share of word-pairs passed from the first-stage classifier to the second-stage was 0.01369, so we use this to construct a weighted average we use for the overall scores in Panel A.

The main takeaway of the classifier performance metrics in table 1 is that the algorithm is about 98% accurate on the test-set that was not used for training, with an F1 score (another performance metric that takes into account both false positives and false negatives) of 0.8383. This can can also be see from figure C1, as the overall accuracy is a combination of the accuracy in the first (over 98%) and second ($\approx$ 92%) stages.[35]

ii) Estimation Details: The estimation itself was done using Random Forest classifiers, as well as an Extremely Randomized Forest, which is conceptually similar but further improves out-of-sample fit.[36] We selected these classifiers for a number of reasons. First, our priority in this project was to use an intuitive method that is conceptually easily understood by most applied economists. We considered the machine-learning review article by Mullainathan and Spiess 2017 as a guide to the methods that would be conceptually

[35] In our main results, we use the standard classification threshold, where we include a word-pair as a predicted loanword when more than 50% of the classifiers in the ensemble identify it as a loanword. Given that the data in our application is unbalanced, with many more non-loanword pairs than loanword pairs, we later show that our main results are robust to using higher thresholds of 60% and 70% for identifying loanwords (therefore further reducing false positives) in tables G7 and G13.

[36] These ensemble classifiers improve out-of-sample fit by decreasing over-fitting to the training set (Varian 2014; Mullainathan and Spiess 2017). To choose hyper-parameters we used a grid search method over the number of features available at each split of the decision tree; the maximum depth of the decision tree; and the minimum number of observations in each final leaf. We select the parameters that performed best on different folds of the training set.

accessible for a broad set of applied economists. Since random forests are covered in depth in this reference and neural networks are not, we decided to prioritize methodological transparency by using random forest ensembles. Furthermore, given the scale of the project, and the associated computational demands, it was infeasible to implement the more computationally intensive neural networks. Though, given that our model's performance is similar to other work that does use a neural network (Blair and Ingram 1998; Mi, Yang, et al. 2016; J. E. Miller et al. 2020; Mi, Zhu, and Nie 2021), there is no reason to believe that such an exercise would produce substantially different results.

The first step to implement the classifier was to generate the characteristics, or features, of each word-pair that it would use as the basis of its predictions.[37] For each word-pair, we generated a variety of features to measure the factors that linguists consider when assessing if a word is a loanword. One of those factors is the ancestral distance between the languages, which is crucial to help to rule out cognates. We use the extent of overlap in the language trees to measure this, as in Blouin 2021. Next, linguists typically consider a variety of factors related to how similar a potential loanword (called a *target word*) is to the the potential source of the loanword (i.e. the *source word*), and how similar the source and target words are to the typical words in each language. This is helps to determine how likely it is that one word originated from the other, and how likely it is that each word originated within their own language. Accordingly, we include a number of measures that indicate the linguistic similarity of both the target word and the source word to their own respective languages; and features that measure the similarity of the potential target and source words to each other. Lastly, we include the *difference* between these two measures - i.e. between the own-language similarity of the source and target words.

Each of the similarity measures that we include is based either on either how words are spelled (orthographic similarity), or how they are pronounced (phonetic similarity).[38] The orthographic similarity measures are based on work by Jaro 1989 and Winkler 1990. We include the Jaro-Winkler similarity - which has been used in other applications in economics (e.g. Eli, Salisbury, and Shertzer 2018), as well as some other related measures (see the appendix for details). These measures typically take values ranging from 0 to 1, with 1 indicating identical spellings.

Properly accounting for phonetic similarity is slightly more complicated because not all phonetic differences are considered equal signals of loanword status. For example, in English moving from the sounds for *b* (ɓ) to *p* (p) is a more natural and common slip

[37] These features are listed and explained in detail in appendix B, including a description of how orthographic and phonetic measures were implemented and a table summarizing how these measures approximate the techniques used by linguists in practice.

[38] A *phoneme* is a sound associated with characters in a language's alphabet. So, *phonetic similarity* refers to the similarity in how two words sound.

than moving from *b* (ɓ) to *ng* (ŋ).[39] So the presence of the latter is more likely to rule out that a word pair represents a loanword pair than the former. To account for this issue, we follow David R. Mortensen et al. 2016. The general approach is to construct a 'weighted-distance' between sounds, where larger weights are assigned to differences in sounds that are less likely to evolve over time. This weighted distance can then account for the likelihood that one word was indeed borrowed from another, but now sounds slightly different because it has evolved independently since the time that it was adopted. For example, a sound's *sonority* - the amplitude of a sound, or how constricted the vocal tract is - is unlikely to drift over time and is given the highest weight of 1.0. On the other hand, the *length* of a vowel is more likely to drift and is given the lowest weight of 0.125. The full list of phonological features, along with their weights and a short description, are listed in Table B2.

Complicating matters even further is that some phonemes are more common in some languages than others. For instance, the English sound for *h* (h) is rare in French, so a French word using this sound may be likely to have been borrowed, but the same cannot be said in English. To address this issue we constructed all 2- and 3-gram phonemes that exist in a language, based on the IPA transcriptions. An *n-gram* is a sequence of *n* linguistic items. The items can be words, or syllables, or sounds; in our case they are phonemes. From here it is straightforward to compute the likelihood that a particular word - represented by an n-gram phoneme - occurs natively in a particular language. This provides a measure for the likelihood that the sound of any given word is native to any given language. This approach is analogous to other work that uses WoLD to estimate loanwords in a machine learning framework. The idea is based on the concept of *word expectedness* "which reflect unique characteristics of phonological and phonotactic [likelihood of a combination of phonemes] clues which can be used to identify borrowings" (J. E. Miller et al. 2020, p.5).[40]

Using these classifiers and features, we built an ensemble Voting Classifier that predicts the likelihood that any word-pair in our data represents a loanword in a particular direction.[41] Whenever two source words were identified for the same loanword, we kept the source word with the highest probability from the second stage classifier.[42] We report

[39] In brackets are the IPA representations of the phonemes we are referring to, which is more precise than listing the English letters. We primarily refer to the (less precise) English letters simply because many economists are unlikely to be familiar with the IPA representations.

[40] The authors of that study apply this in order to predict loanword borrowings of contemporary German words that are not in WoLD.

[41] This was the most computationally challenging step of the process. We relied heavily on the Niagara supercomputer at SciNet, which is owned by the University of Toronto and Compute Canada, to manage the whole process. Niagara includes a homogeneous cluster of 61,200 cores, of which we were allocated 13.5 core-years. The whole algorithm ran for approximately 43,760 core-hours, which took about one week to execute, using 300 cores.

[42] We also drop loanwords where the source word was itself identified as a loanword, so our final measure of language exchange only includes unambiguously identified loanword pairs.

descriptive statistics for language exchange measures generated by this methodology in table 2.[43]

Importantly, we relied only on the features described above to make loanword predictions. The classifier did not observe variables that directly indicate the identities of the languages themselves, such as language family, lexicon size, population, region, etc.[44] This is important because the discussion and study of loanwords can be an emotional topic, bringing in other opinions or biases (Schadeberg 2009, p. 93). It is important to note however, that these biases may be introduced inadvertently if correlated with other features, as discussed in Kleinberg et al. 2018.

Finally, we return to the estimation implications of the issue relating to special characters, mentioned above. While the vast majority of these characters arise in English, which was removed as a borrower entirely, there may be some special characters that are difficult to detect in other languages that may not use Latin script. Should any expressions containing non-alphabetic characters remain included in the lexicons after data cleaning, this will not impact the pair-wise results that form the basis of the main analysis. This is because any such characters will be harshly penalized by the phonetic and the orthographic distance measures. In the orthographic case, a non-alphabetic character will be treated as a different character than any it is being matched to. Similarly, in the phonetic distance, a non-alphabetic case (such as '+') that cannot be transcribed into IPA (international phonetic alphabet) phonemes will be assigned the maximum possible phonetic distance from any other phoneme it is compared to.

Any non-word expressions that exist in Panlex would therefore appear to be extremely different from any other expression they are compared to, and hence unlikely to be identified as loanwords. In this case, when we compute the shares of each language that is borrowed from each other language, these special cases will show up in the denominator but not the numerator. Since this will be true for loanword borrowing from every other language, it will be captured by language fixed effects, which are included in each regression.

The data that result from the algorithm are summarized in table 2. The main thing to note for now is that while a small share of a typical language is made up of loanwords from a particular group (panel A row 1), a large share of a typical group's language is made up of loanwords overall (panel B row 1). In other words, while there has typically been quite a lot of borrowing - making up about 20% of an average language - this borrowing is relatively targeted. That is, the small share at the pairwise-level is partially driven by a large share of pairs with no borrowing. Consistent with this, the borrowing from a group's best agricultural trade-partner is about triple the mean pairwise borrowing.

[43] See figure G1 for a plot of the distribution of aggregate language-group level language adoption.

[44] This, of course, does not mean that predictions might not be correlated with these things. For example, both borrowing and lending are significantly correlated with GDP per capita, as is cultural openness (the minimum of borrowing and lending). See for example figure G2.

Table 2: Descriptive Statistics

	Mean (1)	Variance (2)	Min (3)	Max (4)	N (5)
Panel A: Society Pair Level					
Linguistic borrowing	0.26	1.92	0	100	9,564
Bilingualism	0.07	0.25	0	1	9,564
Gains from trade (percentage change)	0.076	1.24	-2.84	3.34	9,564
Gains from trade (percentile rank)	0.500	0.289	0	1	9,564
Trade utility	2.61	1.72	0.003	14.27	9,564
Population (1,000)	9,429	72,625	0	871,558	9,564
Share of Arable Land	0.96	0.137	0.003	1	9,564
Land diversity	29,545	39,880	0	343,933	9,564
log Distance to neighbour	4.63	1.29	0	8.64	9,564
Panel B: Society Level					
Linguistic borrowing (total)	20.01	16.9	0	100	2,606
Bilingualism	0.33	0.47	0	1	2,606
Linguistic borrowing (best neighbour)	0.79	3.54	0	100	2,606
Linguistic lending (best neighbour)	0.63	3.19	0	100	2,606
Gains from trade (best neighbour, pct change)	0.96	1.45	-2.83	3.34	2,606
Trade influence (best neighbour, pct change)	0.78	1.31	-2.83	3.34	2,606
Gains from trade (best neighbour, pctl rank)	0.73	0.25	0	1	2,606
Trade influence (best neighbour, pctl rank)	0.69	0.24	0	1	2,606
Trade utility	2.50	1.70	0.003	14.27	2,606
Population (1,000)	1,389	18,518	0	871,558	2,606
Share of Arable Land	0.97	0.129	0.003	1	2,606
Land diversity	24,628	37,913	0	343,933	2,606
log Distance to neighbour	4.29	1.12	1.57	8.534	2,606
Share of viable trading relationships	0.61	0.34	0	1	2,606

Note: The table shows descriptive statistics for the main variables used throughout the empirical analysis. We have word-level data for 4,257 societies. The population data comes directly from the Ethnologue. Distance variables are all author constructed based on the Ethnologue centroids.

The most important thing for our exercise is that overall, loanwords are quite a common occurrence, and therefore they do represent a meaningful outcome of interest. This would be especially true if they are indicative of broader cultural change. We tackle that issue next.

3.C. Validation of Language Data

We undertake a number of exercises to validate the use of these language data to measure cultural adoption and reduced cultural barriers to interaction. First, we show in table C2 that greater loanwords borrowing is associated with greater cultural similarity both in

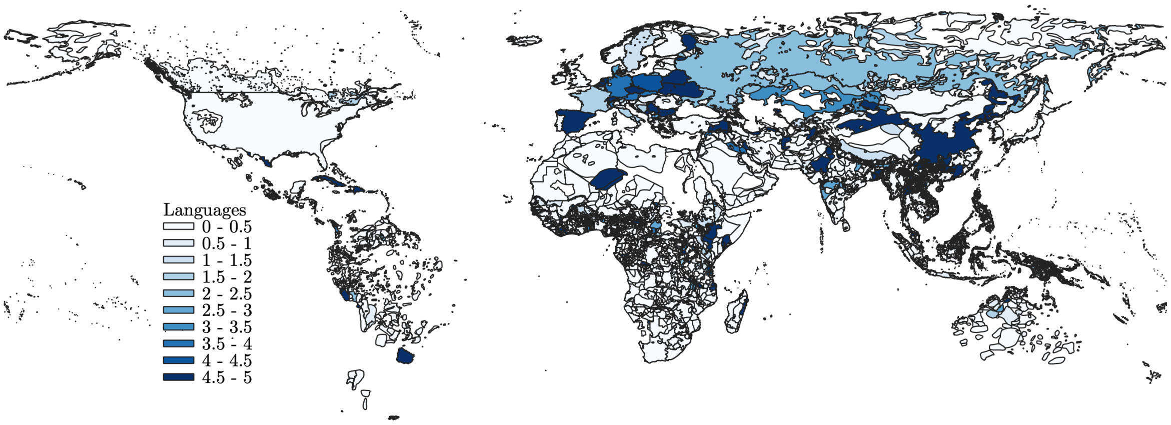


Figure 3: Mapping Loanword Intensity

Note: This map illustrates the global variation in loanword intensity. Loanword intensity is measured as the share of a language lexicon that was borrowed from another language. Darker shades represent a higher intensity of language borrowing.

the ethnographic atlas (Murdock 1959) and in the Folklore data-set (Michalopoulos and Xue 2021).[45] We also show in table C4 that greater loanword adoption is associated with reduced barriers to the diffusion of development, using the data and specification from Spolaore and Wacziarg (2009). Importantly, this is robust to inclusion of a standard measure of cognate words, which shows our loanwords data is not simply identifying cognates. For details on these validation exercises, please see section C.2.

4. Is the linguistic fabric of society static?

We begin the analysis by exploring one of the most basic and fundamental questions of language evolution. On the one hand, economists have typically been content to assume that, once established, language is relatively static. On the other, linguists and sociologists have more deeply considered the idea that as groups interact with each other they influence each other's languages. While even most economists would likely concede that language does endogenously change in the long run, the role of cross-cultural influence has not been explored empirically. Investigating this question reveals that the process of linguistic convergence may be more complex and subtle than one might think.

In particular, the question of how much contact is required for convergence has not been systematically explored and is *a priori* not obvious. This question is crucial to understanding key issues across the social sciences. For example, within linguistics it seems fairly uncontroversial that globalization is undermining linguistic diversity (Kubota 2001; Okwudishu 2019). That hypothesis, however, generally rests on the presumption that even weak societal ties can generate significant cultural convergence.

[45] We also show that greater loanwords borrowing related to a specific topic is associated with greater similarity in traits relating to that topic, holding borrowing of other topics constant.

4.A. Inter-group contact and language exchange

To explore when and how linguistic convergence occurs, we begin by asking whether loanword adoption is more likely between groups that are more likely to have experienced direct and extensive contact between their respective populations. The simplest factor in contact is proximity - interaction is more likely among groups that are closer to each other. It is less clear, however, how much influence takes place among groups who are farther apart. The relationship between language exchange and distance between groups is, therefore, the starting point for our analysis.

The pair-wise empirical set-up is very simple. We include on the right hand side distance between the groups and on the left hand side the share of words that were borrowed from the other language-group:

$$
\begin{aligned}
[\frac{100 \cdot \#\text{ loanwords}}{\#\text{ words in lexicon}}]_{ijcl} &= \alpha_{ci} + \alpha_{cj} + \alpha_{li} + \alpha_{lj} + \beta_1 \cdot \text{CentroidDist}_{ij} \\
&+ \beta_4 LanguageFamilySimilarity_{ijcl} + \epsilon_{icl}
\end{aligned} \tag{1}
$$

We want to focus on interaction between groups rather than the impact of shared ancestry or the fact that nearby groups may share similar shocks within the same country. Accordingly, we include control variables for these factors as follows. Fixed effects at the country (c) or language family (l) - for each of society i and j - are represented by α. $LanguageFamilySimilarity_{ijcl}$ represents the similarity of the language family trees for the two languages in a pair. $CentroidDist_{ij}$ is the distance between the geographic centroids of group i and group j.

The results from estimating these specifications are in table 3. Unsurprisingly, geographic proximity is a very strong predictor of linguistic exchange. The estimate suggests that for every 1,000km of distance between two groups, language exchange drops by about 8% of the language exchange between a typical language-pair (column 1). The result is not driven by cognates, since when we include language tree distance the estimate does not change at all (column 2). This further helps to reinforce that the machine learning estimates were fairly accurate. Likewise, we wanted to make sure that the estimates were not driven by group size. The concern is that since distance is determined by centroids, two small nearby groups are likely to have centroids nearer to each other than two larger groups.[46] However, in column 2 we also add the area occupied by both the lender and borrower as well as the interaction between the two. Group size does not seem to be driving the effect at all, since once again, the estimates in columns 1 and 2 are very similar. The same is true for population of both groups and their interaction. Language

[46] For example, suppose a large group occupied a circle with radius 1,000km. Then, by definition, there could be no groups within 1,000km of any neighbours, and it would appear as if all other groups were far away

family and country fixed effects also do not meaningfully change the estimate (columns 3 and 4, respectively).

Table 3: Inter-group contact strongly predicts linguistic influence

Dependent Variable	Share of borrower's language borrowed from lender (%)						
Sample	All				< 500km	< 100km	> 2500km
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Distance (1,000km)	-0.000291***	-0.000211***	-0.000219***	-0.000231***	-0.112***	-1.462***	1.94e-05
	(1.83e-05)	(1.96e-05)	(1.67e-05)	(2.07e-05)	(0.00821)	(0.192)	(2.77e-05)
Similarity in family tree between borrower and lender		0.499***	0.498***	0.497***	0.353***	0.302***	3.479***
		(0.0465)	(0.0464)	(0.0464)	(0.0301)	(0.0390)	(1.163)
Area Borrower		✓	✓	✓	✓	✓	✓
Area Lender		✓	✓	✓	✓	✓	✓
Area Borrower x Lender		✓	✓	✓	✓	✓	✓
Population of Borrower		✓	✓	✓	✓	✓	✓
Population of Lender		✓	✓	✓	✓	✓	✓
Population of Borrower x Lender		✓	✓	✓	✓	✓	✓
Lexicon Size		✓	✓	✓	✓	✓	✓
Language Family FE			✓	✓	✓	✓	✓
Country FE (for borrower & lender)				✓	✓	✓	✓
Obs.	17,686,230	15,884,210	15,828,462	15,828,462	462,100	51,320	13,232,286
R sq.	0.000	0.007	0.013	0.017	0.040	0.095	0.022
Dependent Variable Mean	0.00347	0.00359	0.00359	0.00359	0.0227	0.0923	0.00270

Note: The regression is run at the society-pair level. Standard errors are two-way clustered by the two societies in a society-pair. *, **, *** denotes 10%, 5%. 1% significance respectively. Language Family FE are FE for the 4-level from root language family based on language trees for both the borrower and the lender. Country fixed effects are included for both the borrower and the lender.

The results in columns 1-4 are consistent with the hypothesis that intense inter-group contact is necessary for language exchange. In fact, it appears that there is nearly no language exchange outside of very small geographic regions. For example, figure 4 plots a binscatter with language exchange on the y-axis and distance on the x-axis. The figure shows how essentially all borrowing takes place within 1,000km. In columns 5 and 6 we restrict the sample to groups within 500km and 100km, and consistent with this hypothesis, the estimate gets much stronger in each case. In column 7 we restrict to groups farther than 2500km, and for those groups there is no relationship between distance and borrowing.

Even though economists have typically assumed that language is unchanging, the fact that contact induces language exchange is not particularly surprising. It is the first time (to our knowledge) that the result has been documented empirically at scale, but is consistent with broader scholarship on linguistic influence. The almost extreme degree to which language exchange is local, however, is a surprise. One ramification of this finding is that it casts some doubt on the often assumed inevitability of global linguistic homogenization.

Expectations of mass linguistic homogenization have historically been motivated by the fact that language differences are a major trade barrier, and that language exchange

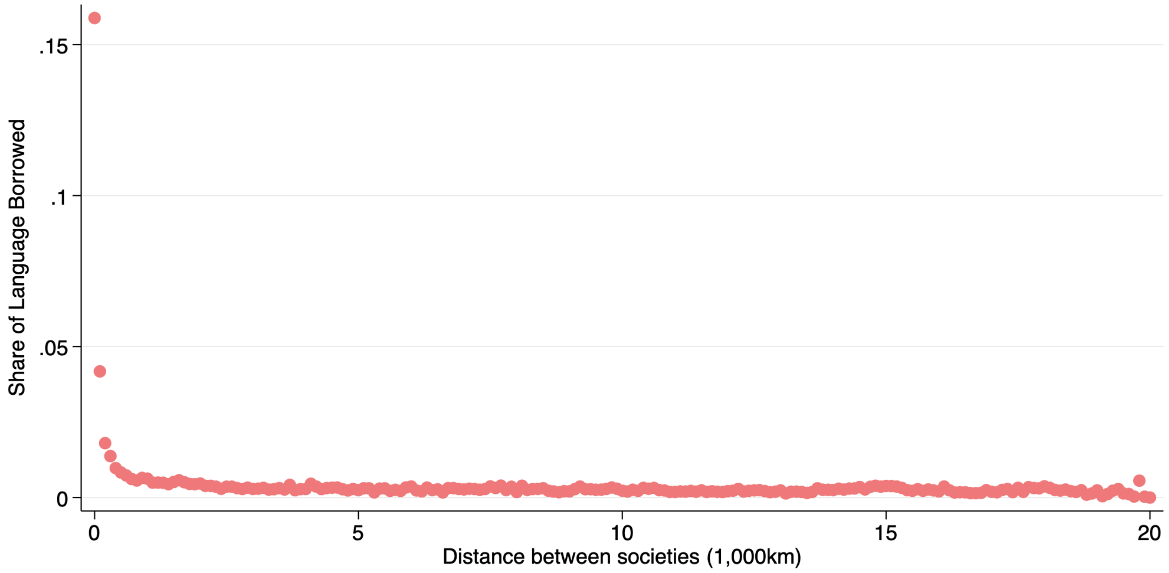


Figure 4: Relationship between distance and language exchange.

Note: The figure is a binscatter plot with language borrowing on the y-axis and 100km distance bins on the x-axis. It shows that language exchange is overwhelmingly local.

is associated with the diffusion of economic development.[47] It dates back to Gottfried Leibniz in the 17th century, who contemplated the idea of linguistic universalism (Leibniz 1916, translation date). This concept was prominent enough to appear in Encyclopedia Brittanica since 1911 (Chisholm 1911) and, more recently, has been discussed in both popular writing (Ettinger 2014) and academia (Bennett 2018). Indeed, the linguistics literature in the past few decades has intensely focused on globalization's impact on language diversity. Linguists have argued that the increased interaction with the west would imminently generate linguistic homogenization, with only 10% of currently spoken languages surviving (Reagan 2005).

Leaving aside the desirability of such homogenization, our evidence suggests that this may be unlikely in light of the extensive amount of contact required for convergence. The idea that very strong cross-societal relationships are necessary for language convergence is also consistent with recent work in sociology. It has become more common among sociologists to argue that while Granovetter's 'weak ties' (Granovetter 1983) are sufficient for information flow, they are not enough to profoundly influence culture (Centola 2021).

Crucially, even if *globalized* trade does not appear to have generated universal long-distance linguistic convergence, this does not imply that economic incentives are not important. However, how those incentives shape the evolution of language remains unclear. To better understand this question we need exogenous variation in the incentive for local groups to interact.

[47] See appendix C4 for descriptive evidence of this association.

5. Economic Incentives and Language: Empirical Strategy

The results above support the assertion that the linguistic makeup of societies is not static. Contact with other nearby groups, and intense interactions with the closest groups, were especially significant sources of cross-group linguistic influence. The economic interpretation of this, however, is less clear. The linguistic influence of nearby groups may be driven by economic incentives for interaction. It could also be driven by non-purposeful accidental interaction, or one of the numerous other potential reasons for interaction, such as shared shocks or complementary technology.

We would therefore like to more closely explore the role of economic incentives and relationship-leverage on language exchange. To do so, we need a plausibly exogenous and consistently observable source of economic incentives. For this reason we turn to agricultural trade incentives. We can estimate trade incentives with a straightforward Ricardian model of local agricultural trade, drawing upon Costinot and Donaldson 2012. This framework models productivity in different crops determined by geographic factors. Inspired by Galor and Özak 2015, we augment the model with a utility function defined by objective nutritional requirements. This allows us to identify incentives arising from complementarity in exogenously determined and globally observable nutritional endowments.

Focusing on local agricultural incentives necessarily involves a trade-off between breadth and empirical identification. We significantly narrow the conceptual focus to consider only local interactions and a narrow range of incentives for interaction, but with more precise identification of incentives for a group to interact with their neighbours. Indeed, the advantage of the approach is that it allows us to estimate pairwise gains from trade for each partner within a trading relationship.

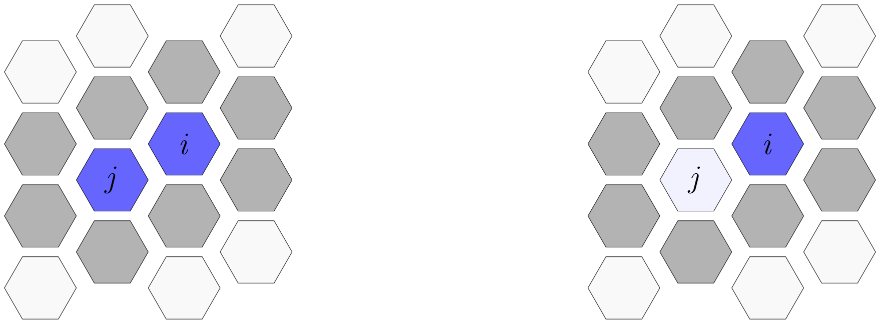
We can then empirically examine whether *within* a trading relationship the group that benefits more from trade than their partner (i.e., has less trade leverage) is more amenable to adopting elements of the trading partner's language, in order to facilitate economic transactions. If so, this implies that economic leverage is an important predictor of how languages evolve.

5.A. Local agricultural trade

In this section we describe a simple trade model, with a single factor, in our case agricultural land, and many goods, in our case different crops.[48]

The model generates a welfare gains from trade - based solely on land characteristic complementarity, for each language group pair within a trading region. This requires that we define a region that sets the scope of possible historical agricultural trade. We will refer to these regions as neighbourhoods, which we define as the union of the immediate

[48] See Feenstra 2004 for a discussion of this classic model, first introduced in Ricardo 1817.



(a) Full Trade **(b)** Counterfactual without j

Figure 5: Neighbourhoods Used for Constructing Gains from Trade

Note: This figure illustrates the counterfactual neighbourhoods used for our structural estimates of gains from trade at the language-pair level. A dark shaded polygon indicates a society that is included in the given counterfactual neighbourhood. In panel a) we show the neighbourhood used for our full trade counterfactual between group i and j, made up of the union of immediate neighbours of i and j. In panel b) we show the counterfactual neighbourhood where j is dropped from the neighbourhood that i can trade with.

Ethnologue neighbours for a society-pair. See figure 5 for a graphical representation of the neighbourhoods of interest.

i) Production We model crop production similarly to Costinot and Donaldson 2012, where production depends on land quantity and the productivity of the land for producing various crops. Data on agricultural productivity comes from the Global Agro-Ecological Zones (GAEZ) data-set (IIASA/FAO 2012), which includes measures of potential production for 49 crops at the 5 arc-minute grid-cell level for the entire world.[49] We convert the GAEZ data into group-level productivity in each crop by taking the mean of cells that fall within the borders of a group's homeland on Ethnologue map.

Societies ($i \in \{1, 2, ..., I\}$) allocate land to different crops ($c \in \{1, 2, ..., C\}$), denoted by the I by C matrix Λ. They do so given the productivity of their land for each crop, denoted by the I by C matrix Ω. The Hadamard product of these two matrices produces output $Y = \Lambda \odot \Omega$, also an I by C matrix:

$$Y = \begin{pmatrix} \Lambda_{1,1}\Omega_{1,1} & \cdots & \Lambda_{1,C}\Omega_{1,C} \\ \Lambda_{2,1}\Omega_{2,1} & \cdots & \Lambda_{2,C}\Omega_{2,C} \\ \vdots & \ddots & \vdots \\ \Lambda_{i,1}\Omega_{i,1} & \cdots & \Lambda_{i,C}\Omega_{i,C} \end{pmatrix} \tag{2}$$

ii) Demand On the demand side, we would ideally have continued to follow Costinot and Donaldson 2012 by assuming price-taking and directly using price data. However, the differences in the context make this strategy more difficult for us, for a number of reasons. First, Costinot and Donaldson 2012 show that Ricardian trade does a good

[49] To avoid concerns regarding endogenous irrigation or other agricultural inputs, we use the potential yields for low-input, rain-fed agriculture. This is similar to the methodology used for generating the measures of crop productivity in Galor and Özak 2016.

job of explaining crop production heterogeneity, while we need some notion of utility to measure the welfare generated by each neighbour. Second, we narrowly model local trading relationships, while their focus is more global. We do this because it helps our empirical identification to exploit only the plausibly exogenous complementarity of land endowments of a society and its neighbours.[50] The trade-offs with this focus include that in very localized trading networks the price-taking assumption employed by Costinot and Donaldson 2012 seems less reasonable, and in any case, we do not observe local prices of all agricultural goods around the world.

For these reasons we model demand to recover relative prices. We build upon Galor and Özak 2015 by treating societies as having the incentive to increase the population they can support, where each adult requires a subsistence bundle of calories and essential nutrients. Galor and Özak 2015 show that caloric potential dominates agricultural suitability. We go one step further by considering the full range of nutritional requirements. This is necessary in our context because it is what gives rise to the incentive to exchange crops. We use data on the nutritional content of each crop, which come from FAO databases (FAO 2017b; FAO 2017a) and are matched to the crops in the GAEZ data.[51] To measure the required essential nutrients to sustain the average adult human,[52] we use the Dietary Reference Intakes (DRI) tables produced by the National Academy of Sciences (NAS) Institute of Medicine (2006). In the context of primary agricultural production, nutritional diversification is known to be an important driver of local trade (Gray and Birmingham 1970).

Accordingly, gains from trade are based on a Cobb-Douglas utility function over calories as well as all essential nutrients.

$$U(x_0, x_1, \cdots, x_{16}) = x_0^{\alpha_0} x_1^{\alpha_1} \cdots x_N^{\alpha_{16}} \tag{3}$$

where x_0 represents daily calories, and x_1 through x_{16} are the sixteen essential micronutrients. α_0 through α_{16} represent the preference weights on calories and each essential nutrient.

[50] This helps to side-step the issues previously discussed, as well as issues like colonialism that likely impacted both trade and language, but not necessarily in a causal manner. It also helps to avoid considering endogenous heterogeneity in other types of trade. We focus on sources of gains from trade where we do not have to worry about the potential endogenous heterogeneity in access to globalized markets, long-distance trade routes, or productivity in industrial products.

[51] Specifically, these data cover twenty-three micronutrients for forty-one of the forty-nine crops included in our agricultural productivity data.

[52] There are sixteen nutrients in our crop content data that are also identified as essential nutrients by feeding experiments in Chipponi et al. 1982, where "[t]he dietary *essentiality* of an organic compound signifies that it serves an indispensable physiological function, but cannot be synthesized endogenously."

The preference weights are constructed as follows:

$$\alpha_n = \frac{\gamma_n}{\sum\limits_n \gamma_n}, \tag{4}$$

where the DRI amounts in the NAS data define γ_n, for $n \in \{1, 2, \cdots, 16\}$. We normalize the weights so that the exponents sum to one, and so that they capture the relative importance of nutrients in the diet.[53]

iii) Trade To get production and utility under full trade we proceed in two steps. The first step is to determine the optimal utility for the region as a whole, subject to a constraint on each society's crop productivity. The second step is to determine how to distribute the aggregate output to compute society-specific utilities, based on their own production possibilities. This requires that we estimate prices, to be able to compare, for instance, how well off are societies that produce similar amounts of different crops. This approach implicitly assumes efficient, frictionless trade within each local region.

More formally, the first step is to compute the matrix Λ^*, denoting the land allocation shares that maximize the *regional* utility function (U^R), based on Equation 3.[54] The optimization problem is:

$$\max_{\Lambda} U^R\left(YN'\right) \text{ s.t. } \Lambda J_{C,1} = J_{I,1} \tag{5}$$

where N is a 17 by C matrix that includes the nutritional content of each crop (and N' is its transpose). This allows for the conversion of units of crop production in Y to the nutritional units that are demanded in U. $J_{C,I}$ is a C by I matrix of ones, and likewise $J_{I,1}$ is a column vector of ones. The constraint here simply ensures that the sum of land allocation shares is equal to 1 for each group.

To estimate society-level consumption, and therefore welfare, we need to estimate prices. We describe our solution method intuitively here, with a full detail of the numerical optimization problem, constraints, and procedure in appendix E.2. The solution relies

[53] For γ_0, the weight for calories, we calibrate using observed population figures (details are in Appendix E.1). This is because the DRI figures we use are derived from modern North American diets, and it is not reasonable to assume that the implied tradeoff in macro- and micro-nutrients can be generalized to the preindustrial local trading systems we are trying to approximate.

[54] Before optimizing production and solving the nutritional trade model, we first simplify the neighbourhoods. These neighbourhoods are constructed for each pair of neighbouring language groups where we observe linguistic data as the pairwise union of neighbours. Some neighbourhoods from regions with high linguistic diversity contain many, many neighbour groups. This is an issue for the optimization procedure as each additional group means an additional 41 crop choice variables, which massively increases computation time and increases the likelihood of being trapped at a local optimum rather than the global optimum. To deal with this issue, we aggregate all very small language groups (individually being $< 0.5\%$ neighbourhood land share, with a cumulative maximum of 5% of neighbourhoods total land) into one synthetic group of small groups who act as price takers, and whose aggregate production has little effect on the neighbourhood's total production.

heavily on the fact that all crops a group chooses to grow must yield at least as much income per unit of land as any crop that the society chooses not to grow. This generates one set of constraints on prices. Similarly, land allocation must be such that all crops a group grows must provide the same income per unit of land. This generates a second set of constraints on prices. Finally, since Cobb-Douglas utility features identical, homothetic preferences, the ratio of consumption of different goods will be the same at all income levels (Feenstra 2004). Each society will therefore consume in the same proportions as the aggregated regional land shares, since they all face the same prices.

These constraints allow us to solve numerically for the price vector that equalizes the ratio of prices to marginal utilities (at the optimal consumption bundle) across crops at the aggregate level. Using these prices and the land allocation and crop productivity, we can compute the share of total neighbourhood income earned by each group. This generates the share of total neighbourhood output consumed by each group, and therefore the utility for each group for a given trading scenario.[55,56] These utilities form the basis of the gains from trade measures.

5.B. Defining gains from agricultural trade:

To make things concrete, consider group i, for whom we can define full-trade utility U_i^{FT}. Since we are interested in gains from trade, we construct a variable that measures the welfare improvement that arises from adding a neighbouring group to each group's trading network.[57] In addition to the utility under full trade, we therefore also estimate utility for group i under the counterfactual where another group in the neighbourhood (call them group j) does not exist. To do this, we use the exact same process as described above, but simply exclude j from the feasible trading partners in the neighbourhood.[58] This generates U_i^{FT-j}, which we use to measure i's gain from trading with j.

[55] We solve our trade model by implementing the Byrd-Omojokun Trust-Region SQP method (see Lalee, Nocedal, and Plantenga 1998; Nocedal and Wright 2006). This method smooths the objective function (to avoid getting 'stuck' on local minima) by making a linear approximation to the function over a 'trust-region,' where the size of this trust-region is adjusted at each iteration. Intuitively, this means the algorithm makes large approximations at first to identify the region where the global (and not local) minimum will be found, and then successively makes smaller approximations until it finds the minimum.

[56] Solving over a forty-one dimensional or higher input vector is computationally intensive, so the algorithm first solves, under coarse optimality tolerance, considering all forty one crops. It then restricts to only those crops allocated at least 1% of land and solves again with much finer optimality tolerance over this filtered set of crops, and use this optimization result for land-use, consumption and utility.

[57] Gains from agricultural trade is our primary variable of interest, the *level* of agricultural income is of some independent interest. We control for the model estimated agricultural incomes of each society under full trade throughout the analysis.

[58] Here we want to focus on gains from trade arising from group i interacting with group j, and not from changes to i's utility that come from some third group k changing their production/prices in response to i's inclusion. We therefore fix all prices to be those from the full trade scenario other than for crops grown by at least one of i or j. See appendix E.3 for more details.

Gains from trade can now be constructed as:

$$c_{ij} = \frac{U_i^{FT} - U_i^{FT-j}}{U_i^{FT-j}} \tag{6}$$

Which can be interpreted as the contribution of j (via trade) to the utility of i. Note that $c_{ij} < 0$ if j is a competitor to i and $c_{ij} > 0$ if i and j make natural trade partners from the perspective of i.

We plot a histogram of c_{ij} in figure G5(a), and find that it is centred approximately around zero. The figure only displays the range {-10, 10} to avoid severe x-axis distortion.[59] The maximum value of the variable is over 3,000, and there are more than a handful of society-pairs with values over 100.[60] To avoid estimates that are driven by societies in the tail of the distribution, and that may have undue influence on a regression, we will show results with the percent change variable, as in equation 6,[61] and a percentile rank version of the same variable.

At the society-level we focus on the neighbour with the biggest gains from trade. This is because of the possibility of crowding-out of language adoption.[62]

$$c_i = \max_j \{ \frac{U_i^{FT} - U_i^{FT-j}}{U_i^{FT-j}} \} \tag{7}$$

Note that $c_i = 0$ if i is, at best, indifferent towards trade, and in practice most societies have at least one partner that they are at least indifferent towards. $c_i > 0$ is therefore typical, which can be seen in figure G5(b), which shows the c_i histogram.[63] As with the pairwise measure, we show results using both the percent change (as in equation 7) and a percentile rank version of the same variable.

In addition to exploring how much a society's gains from trade is related to language borrowing, we should also expect trade incentives to be associated with loanwords lending. Accordingly, the influence of society i on the agricultural trade of neighbour j is ι_{ij}, while ι_i represents the same, but at the society level:

[59] We do this only for the purpose of the visualization, not the broader analysis. This eliminates about 2% of the sample, typically ones with very large values rather than very small. Even a value of 10 is very large, it suggests a single neighbour improves the welfare of a society by 10-fold.

[60] These are typically societies that would not be able to survive without a particular neighbour, and thus have a near-zero denominator.

[61] Here we deal with the outliers by winsorizing at 5%.

[62] One illustrative example of crowd-out would be societal adoption requiring a tipping point or critical mass of adopters Centola (2021). In this case, increasing trade with a less-popular partner could draw away adopters from a more popular partner which would lead to increased crowd-out and reduced overall borrowing.

[63] We also show, in Figure 6(b) how this measure is distributed spatially.

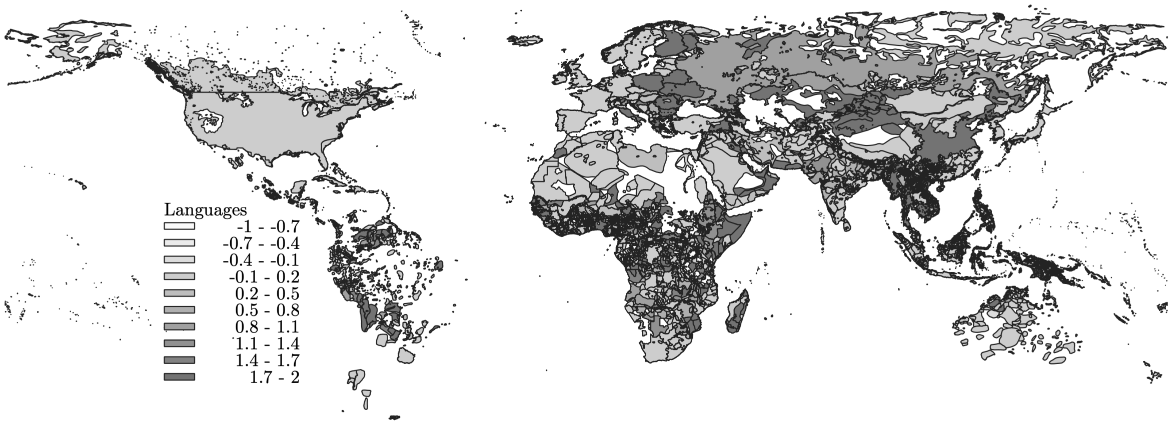


Figure 6: Mapping Gains From Trade

Note: This map illustrates the global variation in local agricultural gains from trade. Gains from trade is defined as in equation 7. Darker shades represent more gains from trade.

$$\iota_{ij} = \frac{U_j^{FT} - U_j^{FT-i}}{U_j^{FT-i}} \tag{8}$$

$$\iota_i = \frac{\sum_{j \in \mathcal{J}} \iota_{ij}}{J} \tag{9}$$

Where J is the total number of neighbours of i.

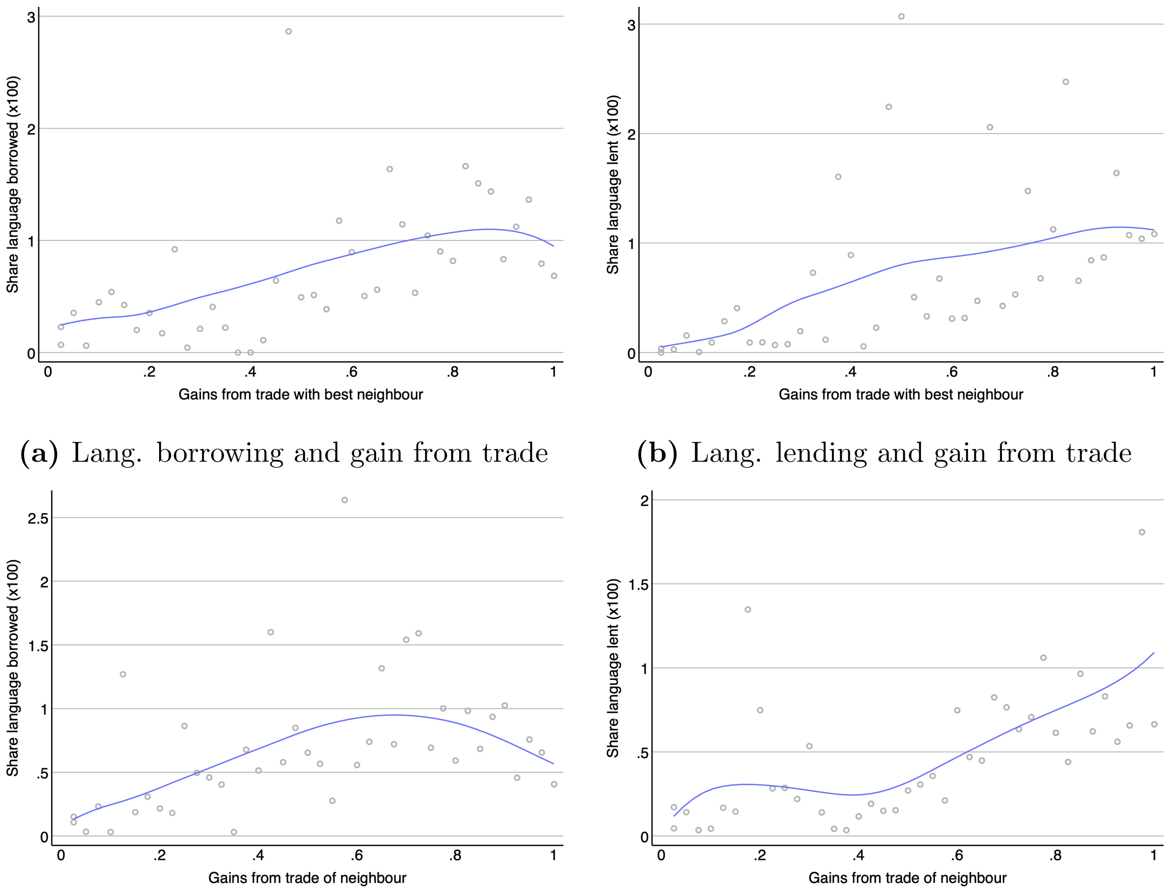
5.C. Validity of Trade Measures

The validity of these measures are discussed in appendix section F. We take three approaches to validating these measures. The first shows that the model-estimated production of each crop that is typically locally traded significantly predicts the actual production of crops, even controlling for the suitability of all crops (section F.1). As anticipated given our local focus, the same is not true for the crops that are typically globally traded. Second, we examine population (section F.2). The population of each group that is estimated by the model is a significant predictor of the actual population of each group. Finally, we look at prices (section F.3). The idea is that when trade is high, price differences get arbitraged away, and prices become more similar. The empirical exercise demonstrates that, as expected, the gains from trade generated by the model is a significant predictor of regional correlation in crop-prices.

6. Economic Incentives and Language: Results

6.A. Language-level Analysis

Based on the previous findings, we expect a positive correlation between the gains from trade with a society's best trading partner and the total loanwords in the society's lan-



(a) Lang. borrowing and gain from trade

(b) Lang. lending and gain from trade

(c) Lang. borrowing and trade influence

(d) Lang. lending and trade influence

Figure 7: Gains from trade and language exchange

Note: The figure shows the correlation between language exchange and trade gains/influence. The figure plots the share of words borrowed, as well as the analogous measure for lending. These are plotted against gains from trade with their best neighbour (as defined in equation 7) and trade influence (equation 9). All of these measures are then winsorized at the 1% level to regulate the axis-scale (this is not done in the analogous tables). The scatterplot groups observations into 0.01 lending bins. The fit line in each graph is based on a biweight kernal of degree 1, with a bandwidth of 0.035.

guage.[64] Figure 7 plots the raw data, and does reveal a positive correlation between gains from agricultural trade and linguistic exchange. Notably, both lending and borrowing are associated with both gains from trade and trade influence. This should be expected if inter-group contact is a primary driver of language exchange.

However, slightly complicating matters is the extremely strong correlation between gains from trade and trade influence (figure G7), which indicates that some of these raw correlations may be spurious. Accordingly, we employ the following horse-race style regression between gains from trade and trade influence:

$$(10) \qquad [\frac{100 \cdot \#\text{ loanwords}}{\#\text{ words in lexicon}}]_i = \alpha_{colonizer} + \alpha_{continent} + \alpha_{li} + \beta_1 c_i + \beta_2 \iota_i + X_i'\Gamma + \epsilon_i$$

The outcome variable represents the total share of words in a society (denoted i) borrowed

[64] This is because gains from trade is an incentive for interaction, and because for any partner other than the best trade partner crowd-out may be an issue.

from neighbours. We include fixed effects for colonizer, continent, and language family denoted $\alpha_{colonizer}$, $\alpha_{continent}$, and α_{li} respectively. X' is a vector of controls, including the following: agricultural wealth (see section 5.A); the share of arable land in the region held by the group; distance to neighbours; and linguistic distance. Asymmetries in diffusion are commonly thought to be a function of the population of the society, its neighbours, and their ratio, so we also include each of those in X'. We also include two measures of land diversity to account for the ethnolinguistic convergence mechanism proposed by Michalopoulos 2012.[65] The two variables we are interested in are c_i and ι_i. c_i is the gains from trade variable, defined in equation 7, while ι_i is the analogous trade influence measure (equation 9).

Our focus is on the gains from trade on language adoption, but another possible source of heterogeneity in economic incentives is heterogeneity in the cost of adoption. For example, it might be easier for a French speaker to adopt a more similar language, like Italian, than to adopt a word from German. In order for this to be a confounding factor, however, would require that language similarity is systematically correlated with pairwise complementarity in nutritional endowment, which seems implausible. Furthermore, we control for language family similarity and shared ancestry in a number of different and flexible ways without any substantive change in the estimated coefficients.[66]

The specification outlined in equation 10 may be able to shed some light on the mechanisms that cause language exchange. However, to fully disentangle mechanisms we need to make use of the bilingualism data described in section 3. We will describe the expectated coefficients under the different hypotheses for borrowing, keeping in mind that we expect the reverse (i.e., swapping β_1 and β_2) for the lending outcomes. The first possibility is that *any* inter-group contact generates some natural cultural exchange, whether intentionally or not. In that case we expect $\beta_1 > 0$ and $\beta_2 > 0$ for both loanwords and bilingualism. Second is that economic leverage matters. One way this could arise is if trade incentives lead one party to make purposeful cultural investments in a foreign culture to facilitate trade. In that case we expect $\beta_1 > 0$; $\beta_2 = 0$ - again - for both loanwords and bilingualism.

Third relates to asymmetries in diffusion. It could be that people learn languages that they come into contact with, more or less at random. So bilingualism essentially follows the logic of the first mechanism outlined above. However, the linguistics literature is quite clear that diffusion need not be the same for both groups. For example, large

[65] The first is the mean variance in suitability of each crop, and the second takes the sum of the absolute difference in crop suitability for all crops.

[66] There are a number of other reasons that this is not a concern for our analysis. For instance, even if similar languages have reduced *costs* of borrowing, it is also equally likely that the *benefits* of borrowing are also reduced. If linguistic similarity is associated with cultural and other ancestral similarities, the resulting costs of cross-group interaction may also be lower. In this case, the incentive for adoption as an investment in reducing cultural distance to facilitate trade will accordingly be lower. Therefore, the overall bias is minimized by countervailing mechanisms.

differences in population size can generate significantly different rates of diffusion, which may also be driven by gains from trade. This would generate differences in loanwords that were asymmetric, as in the second mechanism (i.e., $\beta_1 > 0$; $\beta_2 = 0$), making them more difficult to disentangle. However, in that case we would expect bilingualism to react differently than loanwords to trade incentives. This is based on an interpretation of bilingualism as language adoption that has not necessarily diffused broadly, generating loanwords. In particular, if asymmetries in diffusion were the mechanism underlying asymmetries in loanwords, we would expect $\beta_1 > 0$ and $\beta_2 > 0$ in specifications where bilingualism is the outcome.

The results are in table 4.[67] As in figure 7, gains from trade are associated with linguistic borrowing, and estimates using the percentage of welfare gain (column 1) and the percentile rank of gains from trade (column 2) are consistent with each other. Similarly, trade influence is associated with linguistic lending using either the percentage welfare measure (column 3) or the percentile rank (column 4). Third, the results on bilingualism are completely consistent with the loanwords results for both the percent change (column 5) and percentile rank (column 6) measures of trade.[68] So across the various specifications, $\beta_1 > 0$ for both loanwords and bilingualism.

Table 4: Agricultural trade and language exchange

Dependent Variable:	Language Borrowed		Language Loaned		Bilingualism	
Utility measure	percent change	percentile rank	percent change	percentile rank	percent change	percentile rank
	(1)	(2)	(3)	(4)	(5)	(6)
Gains from trade with neighbours	0.0989***	0.924***	0.0674	0.192	0.0152**	0.0821**
	(0.0355)	(0.321)	(0.0723)	(0.306)	(0.00609)	(0.0396)
Influence on trade with neighbours	-0.0357	0.0721	0.246**	1.634***	0.00909	0.0690
	(0.0387)	(0.357)	(0.118)	(0.523)	(0.00682)	(0.0437)
Trade wealth (structurally estimated)	✓	✓	✓	✓	✓	✓
Population	✓	✓	✓	✓	✓	✓
Land Share	✓	✓	✓	✓	✓	✓
Land diversity	✓	✓	✓	✓	✓	✓
Distance to Neighbour(s)	✓	✓	✓	✓	✓	✓
Linguistic Distance	✓	✓	✓	✓	✓	✓
Language Family FE	✓	✓	✓	✓	✓	✓
Colonizer FE	✓	✓	✓	✓	✓	✓
Continent FE	✓	✓	✓	✓	✓	✓
N	2,606	2,606	2,606	2,606	2,606	2,606
R^2	0.118	0.120	0.151	0.152	0.196	0.197
Dependent Variable Mean	0.951	0.951	0.951	0.951	0.331	0.331

Note: The unit of observation is a society as defined by the Ethnologue. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Gains from trade with neighbours is the c_i measure defined in equation 7, and analogously, influence on trade with neighbours is ι_i (equation 9). Trade wealth (structurally estimated) refers to the level of utility under full trade (i.e. when all neighbours are included) from the same model used to estimate gains from trade. In each case, in order to aggregate to the societal level we take the maximum value from the society's neighbours. Distance to neighbours is a mean distance to neighbours, and in this case captures the density of the neighbourhood. Bilingualism is a binary variable denoting whether the society is heavily bilingual in any of its neighbours' languages.

The percentage change measure may be the more helpful to interpret the magnitudes

[67] Robustness to different rules to define a loanword appears in table G7.

[68] We lose precision when we use the bilingualism data, with the statistical significance dropping to the 10% level, but this is to be expected given how noisy those data are.

(0.085 in column 1 and 0.29 in column 3). On average, a society's best neighbour improves their welfare by about 96% (table 2), which corresponds to a roughly 0.09 percentage point increase in loanwords. This means that gains from trade with a typical society's best trade partner contributes to about 10% of that society's regional linguistic borrowing. If we look at the percentile rank measure, going from the bottom to the top of the distribution represents about 1 percentage point change in borrowing (columns 3 & 4), which is about the mean of linguistic borrowing. Keeping in mind that the estimates capture only local trade in agricultural goods, these estimates seem plausible and appropriate.[69]

Overall, the table highlights the robustness of the correlations in figure 7 (a) and (d) to the various controls and fixed effects in the regression. Importantly, gains from trade remains strongly correlated with borrowing controlling for influence, and vice versa.[70] Also important is that once we control for gains from trade, trade influence no longer positively covaries with linguistic borrowing. In column 1 the estimate is actually negative, while in column 2 it is nearly an order of magnitude smaller than the gains from trade estimate. We find similar differences for the bilingualism outcome in columns 5 and 6. Likewise in columns 3 and 4 we find that trade influence matters and not gains from trade. So, consistently $\beta_2 = 0$. These null results are all a first step towards understanding mechanisms. In particular, the evidence so far supports the idea that inter-group contact is only important in shaping language for the group with more to gain from the relationship.

We expect that results would be strongest on the gains from trade with the best neighbour and total language borrowing due to the possibility of crowd-out. However, we also explore borrowing from only the best neighbour, as well as the - perhaps *a priori* more obvious - specification regressing average linguistic borrowing on average gains from trade. For the first, as anticipated, results are nearly identical in a specification that regresses the gains from trade with the best neighbour on borrowing from the best neighbour (table G10, columns 1-2). Moreover, columns 3 and 4 report the estimates based on gains from trade with an average neighbour and average borrowing. In this case, as expected, the results are much weaker - over six times smaller - albeit in the same direction. Likewise, in columns 5 and 6 we show a much weaker, and even negative correlation between gains from trade with the best neighbour and language exchange with the worst. These results in columns 3-6 are consistent with crowd-out, and help to justify the focus on the best trade partners.

A final consideration is about identification. How do we know that economic trade is the reason why trade incentives are associated with language exchange? There could, after all, be some unobserved variable that is correlated with both gains from trade and

[69] We show that this is robust to focusing only on the Old World in table G9.

[70] The results for gains from trade and trade influence on their own (i.e. not in the horserace specification) are in appendix table G8.

language exchange. One way to investigate this possibility is to examine unviable trading relationships. For these society-pairs we can still observe a continuous gains from trade variable, however we do not expect that this variable will impact actual trade for these societies. Once a relationship is not viable there will not be trade regardless of how unviable it is. We show this falsification test in table G14. Indeed, we find no evidence that gains from trade influences language exchange for unviable trading relationships, reinforcing our belief that the main results are in fact driven by local agricultural trade.

6.B. Language-pair Analysis

i) Empirical Specification The sharpest empirical test of whether trade leverage shapes language influence is to look *within* a language-group pair. This allows us to fix the extent of inter-group interaction, and see whether both groups are equally likely to bear the cost of convergence.

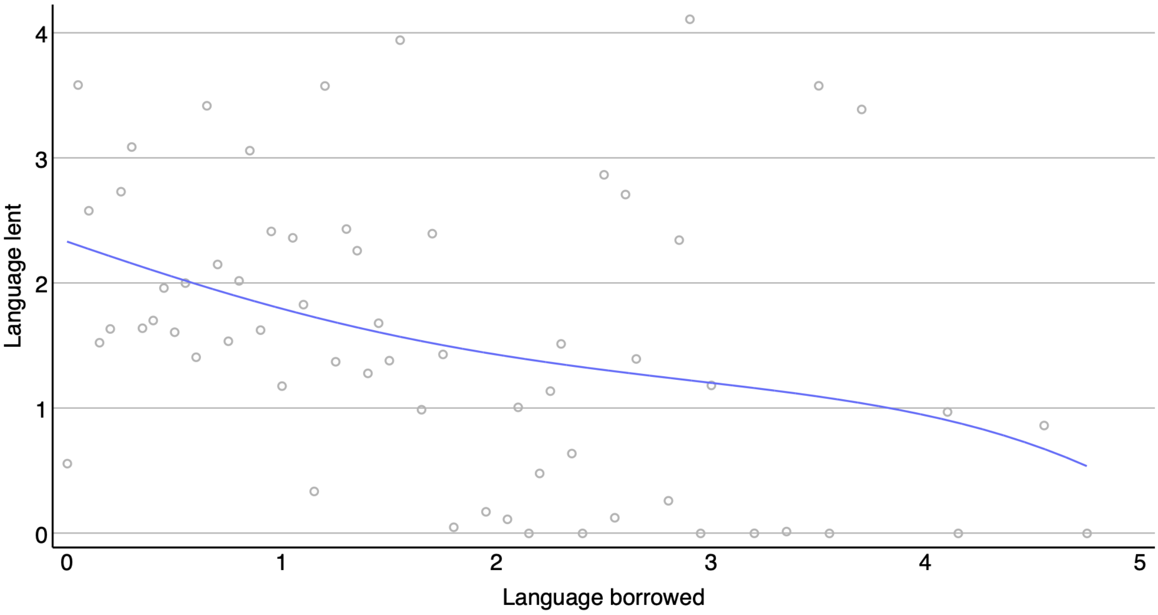


Figure 8: Correlation between language borrowing and lending

Note: The figure shows the correlation between language borrowing and language lending. The figure plots the share of the words in a lexicon that are borrowed (i.e. the sum of the dependent variable in equations 1 and 12 over all neighbours) as well as the analogous measure for lending. That is, for a given language, we define lending as the mean share of the language that was borrowed by neighbours, while borrowing is defined as the share of words in a language that were borrowed from neighbours. Shares are constructed so that they range (theoretically) from 0-100. In each case we condition the sample on the share being less than 5 to regulate the axis-scale. The scatterplot groups observations into equidistant 0.05 lending bins.

We continue our study of these relationship dynamics by documenting a negative correlation between linguistic borrowing and lending (figure 8), which already suggests that local language exchange is asymmetric. This is notable because if all of language exchange was driven by inter-group contact, we would expect that contact would generate both more borrowing and more lending. Instead, the negative correlation is consistent with the idea that leverage or status within a relationship is important. For instance, in the case of economic trade, only one common language is required to carry out a transaction, so only one party needs to bear the cost of convergence.

To more systematically isolate the role of relationship-level incentives, we look *within* potential trading relationships. Investigating this prediction empirically requires us to move from a society-level analysis to a society-pair-level analysis. We can now include society-pair fixed-effects to assess whether the society that benefits more, borrows more (or equivalently the one who benefits less, lends more). The introduction of the society-pair fixed-effects additionally allows us to control for much more than we previously could, and serves as a robustness check to the analysis previously discussed. We test specifications with society-pair-level fixed effects as follows:

$$[\frac{100 \cdot \#\text{ loanwords}}{\#\text{ words in lexicon}}]_{ij} = \alpha_{pair} + \beta_1 c_{ij} + X'_{ij}\Gamma + \epsilon_{ij} \tag{11}$$

Everything is as previously described, with the exception of α_{pair} which denotes a language-pair fixed-effect.[71] We also omit some of the controls that we had in the society-level analysis since they are made redundant by the inclusion of the society-pair fixed-effects.

The estimates generated by the pair-level regressions can be seen in table 5.[72] We examine both borrowing (columns 1 & 2) and bilingualism (columns 3 & 4), as we did before, and the results of both are once again similar. Just as with the language-level estimates, this narrows our focus to the two adoption based mechanisms rather than the diffusion based mechanism. The estimates on loanwords (column 1) suggest that if a society gains 7.6% more from trade than their partner, they borrow about double the typical loanwords of a viable relationship.[73] So this implies that this 7.6% gains from trade is enough to have one party take on all of the linguistic borrowing in a typical viable trading relationship.

We therefore find that societies do borrow more when trade is more profitable, *for the same level of interaction* - in this local trading context, whenever i interacts with j, j also interacts with i. The result highlights in the clearest way so far, that cultural exchange evolves according to within-relationship economic leverage, rather than solely as an unintentional by-product of contact.[74] Investigating the relationship using the percentile rank - which is less prone to influence from outliers - generates no change in interpretation (column 2).

[71] In the data, ij and ji are different observations, in the first case i is the borrower and j the lender, and in the second case these roles are flipped. The α_{pair} variable accounts for this pair of observations, allowing us to identify off of differences in gains from trade within a trading relationship.

[72] Robustness to various loanwords thresholds is in table G13.

[73] This comes from the estimate: 0.0731, and the mean of the dependent variable: 0.278. Combining these, we see that 2 x 0.278 / 0.0731 = 7.6

[74] We also explore heterogeneity in this relationship by within-pair differences in population in table G11, and in linguistic distance between the pair in table G12.

Table 5: Loanwords and trade incentives at the relationship level

Dependent Variable:	Language Borrowed		Bilingualism	
Utility measure:	percent change	percentile rank	percent change	percentile rank
	(1)	(2)	(3)	(4)
Gains from trade with neighbours	0.0731***	0.725***	0.0330***	0.215***
	(0.0272)	(0.264)	(0.00835)	(0.0408)
Relationship Fixed Effects	✓	✓	✓	✓
Baseline controls	✓	✓	✓	✓
N	5,693	5,693	5,693	5,693
R^2	0.518	0.519	0.553	0.556
Dependent Variable Mean	0.278	0.278	0.0833	0.0833

Note: The unit of observation is a society-pair. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Gains from trade with neighbours is the c_{ij} measure defined in equation 7, and analogously, influence on trade with neighbours is ι_{ij} (equation 9). In each case we aggregate to the society level by taking the maximum value from the society's neighbours. Viable trading relationships are any relationships where at least one of the two parties can gain from trade. Bilingualism is a binary variable denoting whether the society is heavily bilingual in the partner's language. 'Relationship Fixed Effects' means we include a fixed effect for a specific pair. Controls are as follows: trade wealth (estimated); population; land share; land diversity.

Finally, we again examine our main robustness check relating to the gains from trade between the unviable trade partnerships. In table G14 we again find that for societies with no benefit from trade, differences in land characteristic complementarity have no influence on either linguistic exchange or bilingualism. Once again, this reinforces our belief that the reason that we find that land complementarity is important because it represents gains from agricultural trade.

We can also explore the adoption of words by word-type to assess whether this linguistic adoption induced by economic incentives is indicative of trade-specific language adoption or broader linguistic convergence. Since our data-set is aggregated from the word-pair level, it can also be aggregated by subsets of words that correspond to various concepts. We focus on the following: technology, geography, military/warfare, science, politics, and philosophy/religion. We do this in order to automate the process of word categorization. These categories form the core categories that might be of interest to economists, that appear in library cataloguing systems. We explain the empirical routine that assigns words to specific topics in more detail in appendix D.[75]

In table 6 we once again find very little heterogeneity in the treatment effect by word-type. In other words, gains from trade is associated with a broad range of linguistic adoption, covering many concepts and aspects of culture.

The results thus far show that economic incentives shape patterns of cross-cultural language influence. In particular, we see that language influence is often asymmetric, with groups converging towards their partners with greater trade leverage. This suggests

[75] In brief, we tie our hands by starting from keywords in the Library of Congress classification system to define our starting lists of seed words. We then use a multilingual semantic similarity routine to identify similar meanings in a way that is not reliant on English or Indo-European word associations.

Table 6: Loanwords and trade incentives by word type

Dependent Variable:	Language Borrowed					
Word Type	Technology (1)	Geography (2)	Military (3)	Science (4)	Politics (5)	Religion / Philosophy (6)
Gains from trade with neighbours	0.0608** (0.0264)	0.0514** (0.0238)	0.0779* (0.0404)	0.0636** (0.0268)	0.0699* (0.0389)	0.133* (0.0731)
Relationship Fixed Effects	✓	✓	✓	✓	✓	✓
Baseline controls	✓	✓	✓	✓	✓	✓
N	5,526	5,521	5,366	5,502	5,434	2,709
R sq.	0.528	0.533	0.509	0.519	0.525	0.548
Dependent Variable Mean	0.197	0.178	0.230	0.158	0.190	0.313

Note: The unit of observation is a society-pair. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Gains from trade with neighbours is the c_{ij} measure defined in equation 7, and analogously, influence on trade with neighbours is ι_{ij} (equation 9). In each case we aggregate to the society level by taking the maximum value from the society's neighbours. Viable trading relationships are any relationships where at least one of the two parties can gain from trade. Bilingualism is a binary variable denoting whether the society is heavily bilingual in the partner's language. 'Relationship Fixed Effects' means we include a fixed effect for a specific pair. Controls are as follows: trade wealth (estimated); population; land share; land diversity.

that bargaining power may play a role in how cultures converge, though there are other mechanisms by which asymmetric trade might generate asymmetric convergence. To explore the role of power, we now explore cultural convergence that resulted from intense interactions and an overwhelming asymmetry of power: colonization.

7. Colonialism and Language Exchange

Colonialism is one obvious example from which to learn about language influence, especially in light of the evidence above that suggests that economic leverage influences the direction of language exchange. In the context of power imbalances, colonists had a dramatic, and often catastrophic impact on many aspects of colonial societies (e.g. Acemoglu, Johnson, and Robinson (2001) and Lowes and Montero (2018)). It is therefore natural to consider what this experience of domination had on languages and societies more broadly. Colonists also played an active role in establishing national institutions - like schools. Schooling has been shown to have had a clear and deliberate influence on language homogenization (Blanc and Kubo 2021).

In order to systematically explore the role of colonial influence on language adoption, we regress loanword intensity on distance, as well as an indicator variable signifying a colonial relationship. This produces a regression such as:

$$(12)\qquad [\frac{100 \cdot \# \text{ loanwords}}{\# \text{ words in lexicon}}]_{ijcl} = \alpha_{ci} + \alpha_{cj} + \alpha_{li} + \alpha_{lj} + \beta_1 \text{CentroidDist}_{ijcl} + \beta_2 Colonized_{ijcl} + X'\Gamma\epsilon_{icl}$$

As before, α represents fixed effects at the country (c) or language family (l) levels for each of society i and j. The potential endogeneity of the $Colonized_{ijcl}$ variable warrants some discussion, since a large literature has taken a variety of strategies to resolve this issue. In this case, the identifying assumption needed is different from the typical case, since in this case an observation is a society-pair rather than a country or a society as is more common in the literature.

This small difference has big identification implications because the finer unit of observation allows us to control for both colonizer and colonized fixed effects. This accounts for anything about a country that would have made it likely to be colonized, and likewise it controls for the disproportional global power and influence that colonists have experienced. With these factors accounted for, any endogeneity concern would arise from the thought that a specific place was endogenously colonized by a specific colonist - that is, the colonist-colony match. However, this type of strategic colonial selection is inconsistent with the largely haphazard and uninformed way in which colonist-colony matches were determined both in Africa (Michalopoulos and Papaioannou 2014) and Latin America (R. J. Miller, Lesage, and Escarcena 2010).

We do, nevertheless, take steps to further resolve such identification concerns. The ideal way to do this is to include country-pair fixed effects to control for the colony-colonist match. This then requires within-colony variation in colonial intensity. To this end, we build upon Michalopoulos and Papaioannou (2014), who argue that "Europeans' presence in Africa was (with some exceptions) limited to the coastline and the capital cities." This follows the now classic argument that to understand colonialism we must explore "the broadcasting of European power by examining the cost structure facing white leaders attempting to broadcast power" (Herbst 2000, pp. 134). Colonists were often interested in establishing a presence in a cost-efficient manner, with only a few central or strategically located outposts.

We therefore test whether groups who were originally located nearby a future capital city were more likely to interact with colonists, and in turn more likely to have been influenced by them. One concern, however, with relying on distance to a capital could be that capitals could have been selected endogenously. As an additional robustness check we proxy for colonial presence using the centroids of contiguous habitable regions colonised by a particular coloniser.[76] The centroids of these colonial holdings are plausibly exogenous since the borders themselves were arbitrary (Herbst 2000; Michalopoulos and Papaioannou 2014).[77] The results of this robustness exercise are in table G6 and closely

[76] Table G4 shows that the distance of the societies to the capital is very precisely correlated with their distance to the centroid of the contiguous colonial holdings

[77] To identify the centroids for each functionally contiguous colonial cluster, we start from a map of colonial boundaries (figure G3) using data on colonial history from Hensel and Mitchell 2007. We restrict to populated regions by only considering areas with an estimated potential caloric yield above 1,000 kcal, in the spirit of Galor and Özak (2015). We further split clusters connected by narrow 'bridges' using small buffer zones to avoid overlap. Based on these regions, we identify the cluster centroids using GIS software, and construct the distance from the centroid of each society's homeland boundaries to the

follow the results of the main specification.[78]

This empirical strategy generates the following regression equation in the case of distance to the capitals. In the following specification, α_{cp} represents country-pair fixed-effects, and everything else is as before.

$$(13) \qquad [\frac{100 \cdot \#\text{ loanwords}}{\#\text{ words in lexicon}}]_{ijcl} = \alpha_{cp} + \beta_1 distance_{ijcl} + \\ \beta_4 Colonized_{ijcl} \cdot DistanceCapital_{icl} + \\ \beta_5 DistanceCapital_{icl} + X'\Gamma + \epsilon_{ijcl}$$

Columns 1-3 in table 7 report the results from equation 12. In column 1 we look at any group in a colonial relationship on either side. So, the regressor considers whether colonists were influenced by colonies as well as whether colonies were influenced by their colonizers. We see a positive but not statistically significant relationship. In column 2 the focus is on groups that were colonized, ignoring the effect of colonies on their colonizers. In that specification the estimate strengthens dramatically and becomes statistically significant. Column 3, which only considers the impact of colonized on their colonizers, yields small and insignificant results. The estimate underscores that power asymmetries are typically associated with asymmetries in cultural exchange. Columns 4-6 report results from equation 13. Columns 4 and 5 mirror columns 1 and 2 but with the interaction, and the results there are consistent with columns 1 and 2. Column 6 also confirms that colonies had essentially no influence on the language of their colonizers, in the within country-pair specification.

In table 8, we present results analogous to column 5 in table 7 but for each word-type. In each column, the outcome is the share of all words in a given topic category that are loanwords. There does not appear to be substantial heterogeneity in the types of words adopted. Colonial relationships led to significant adoption across all word-types, with the largest magnitudes for politics and philosophy/religion and the smallest effects for geography. We conclude from this that language exchange with colonists was quite broad.

Overall, the colonialism findings suggest that there is substantial asymmetry in language influence. This asymmetry could indicate that for a given level of contact some groups are either more influential or are more susceptible to influence. However, we should stress that these results remain suggestive due to one important caveat. To interpret the effects as causal requires that we assume symmetry in cross societal interaction. This builds on the premise that if one group interacts with another, then necessarily that same level of interaction is reciprocated. While that seems sensible generally or in the trade context, in the case of colonialism it is far less clear that this was true. After all,

centroid of the relevant colonists contiguous land holdings.

[78] Figure G4 also reveals a correlation between colonial centrality and colonial language adoption.

Table 7: Colonial Language Adoption

Dependent Variables	Share of borrower's language borrowed from lender (%)					
Definition of Colonial Relationship	All (1)	Colonized only (2)	Colonizer only (3)	All (4)	Colonized only (5)	Colonizer only (6)
Colonial Relationship	0.140	0.280*	-0.000986	0.124	0.406***	0.000190
	(0.107)	(0.153)	(0.000840)	(0.0811)	(0.108)	(0.00134)
Distance to Capital (1,000km)				-0.000471***	-0.000461***	-0.000467***
				(0.000113)	(0.000113)	(0.000113)
Colonial Relationship x Distance to Capital (1,000km)				-0.0368	-0.341***	-0.00581
				(0.0940)	(0.0780)	(0.0118)
Distance between lender and Borrower	✓	✓	✓	✓	✓	✓
Similarity in family tree between borrower and lender	✓	✓	✓	✓	✓	✓
Lexicon Size	✓	✓	✓	✓	✓	✓
Language Family FE	✓	✓	✓	✓	✓	✓
Area of Borrower	✓	✓	✓	✓	✓	✓
Area of Lender	✓	✓	✓	✓	✓	✓
Area of Borrower x Lender	✓	✓	✓	✓	✓	✓
Population of Borrower	✓	✓	✓	✓	✓	✓
Population of Lender	✓	✓	✓	✓	✓	✓
Population of Borrower x Lender	✓	✓	✓	✓	✓	✓
Country FE	✓	✓	✓			
Country-pair FE				✓	✓	✓
Obs.	15,884,210	15,884,210	15,884,210	15,854,423	15,854,423	15,854,423
R sq.	0.007	0.007	0.007	0.026	0.026	0.026
Dependent Variable Mean	0.00359	0.00359	0.00359	0.00358	0.00358	0.00358

Note: The regression is run at the society-pair level. Standard errors are two-way clustered by the two societies in a society-pair. *, **, *** denotes 10%, 5%. 1% significance respectively. Colonial relationship codes the pair as 1 if the adopting society was colonized by the lending society, but not vice-versa. Lexicon size accounts for the PanLex coverage of the borrowing group. Language Family FE are FE for the 4-level from root language family based on language trees for both the borrower and the lender. Country fixed effects are included for both the borrower and the lender.

Table 8: Colonial Language Adoption by Word Type

Dependent Variables	Share of borrower's language borrowed from lender (%)					
Word Type	Technology (1)	Geography (2)	Military (3)	Science (4)	Politics (5)	Philosophy / Religion (6)
Colonial Relationship	0.285***	0.238***	0.273***	0.367***	0.417***	0.413**
	(0.0711)	(0.0590)	(0.0787)	(0.107)	(0.127)	(0.160)
Distance to Capital (1,000km)	-0.000883***	-0.000770***	-0.000922***	-0.000725***	-0.000932***	-0.000555**
	(0.000206)	(0.000188)	(0.000295)	(0.000186)	(0.000210)	(0.000236)
Colonial Relationship x Distance to Capital (1,000km)	-0.226***	-0.179***	-0.238***	-0.268***	-0.302***	-0.336***
	(0.0576)	(0.0550)	(0.0701)	(0.0829)	(0.102)	(0.111)
Distance between lender and Borrower	✓	✓	✓	✓	✓	✓
Similarity in family tree between borrower and lender	✓	✓	✓	✓	✓	✓
Lexicon Size	✓	✓	✓	✓	✓	✓
Language Family FE	✓	✓	✓	✓	✓	✓
Area of Borrower	✓	✓	✓	✓	✓	✓
Area of Lender	✓	✓	✓	✓	✓	✓
Area of Borrower x Lender	✓	✓	✓	✓	✓	✓
Population of Borrower	✓	✓	✓	✓	✓	✓
Population of Lender	✓	✓	✓	✓	✓	✓
Population of Borrower x Lender	✓	✓	✓	✓	✓	✓
Country-pair FE	✓	✓	✓	✓	✓	✓
Obs.	15,710,963	15,699,008	15,312,549	15,667,171	15,575,559	8,003,076
R sq.	0.022	0.021	0.018	0.014	0.015	0.016
Dependent Variable Mean	0.00434	0.00415	0.00465	0.00382	0.00448	0.00487

Note: The regression is run at the society-pair level. Standard errors are two-way clustered by the two societies in a society-pair. *, **, *** denotes 10%, 5%. 1% significance respectively. Colonial relationship codes the pair as 1 if the adopting society was colonized by the lending society, but not vice-versa. Lexicon size accounts for the PanLex coverage of the borrowing group. Language Family FE are FE for the 4-level from root language family based on language trees for both the borrower and the lender. Country fixed effects are included for both the borrower and the lender.

we could interpret "interaction" either as individual-level direct interaction, or as indirect interaction through, for instance, education, or legal institutions. Framed in that light, clearly the level of interaction between colonists and their colonies was not symmetric.[79] Nevertheless, combined with the results above on economic leverage, these results suggest that asymmetric power plays a prominent role in shaping asymmetric cultural convergence.

8. Conclusion

How do languages develop, evolve, and influence each other? Our analysis suggests that intense cross-societal interaction is crucial for convergence, and that language adoption is heavily asymmetric when there is an imbalance in leverage between groups. We support these conjectures using plausibly exogenous variation in incentives for inter-group economic interaction. This analysis shows that greater gains from trade is associated with greater language adoption and that the group with less leverage in a trading relationship will adopt more of their partner's language. We complement this with an analysis of colonialism and show that historical relationships with asymmetric power also generate asymmetric language exchange. Language is therefore not a static feature of a society's culture, and instead responds to economic incentives and power.

References

Acemoglu, Daron, Simon Johnson, and James A. Robinson (2001). "The Colonial Origins of Comparative Development: An Empirical Investigation". In: *The American Economic Review* 91.5, pp. 1369–1401.

Ager, Simon (2019). *Omniglot.*

Ahlerup, Pelle and Ola Olsson (2012). "The roots of ethnic diversity". In: *Journal of Economic Growth* 17.2, pp. 71–102.

Alesina, Alberto, Arnaud Devleeschauwer, et al. (2003). "Fractionalization". In: *Journal of Economic growth* 8.2, pp. 155–194.

Alesina, Alberto, Paola Giuliano, and Nathan Nunn (2013). "On the Origins of Gender Roles: Women and the Plough". In: *The Quarterly Journal of Economics* 128.2, pp. 469–530.

Alesina, Alberto, Stelios Michalopoulos, and Elias Papaioannou (2016). "Ethnic inequality". In: *Journal of Political Economy* 124.2, pp. 428–488.

[79] Even in the case of individual contact, it is likely that in most colonial contexts there are few colonists who influence a large share of a community - this difference in breadth vs intensity of contact can have implications for diffusion.

Algan, Yann, Thierry Mayer, Mathias Thoenig, et al. (2021). “The Economic Incentives of Cultural Transmission: Spatial: Spatial Evidence from Naming Patterns across France”. In: *The Economic Journal* Forthcoming.

Ashraf, Quamrul and Oded Galor (2013). “Genetic diversity and the origins of cultural fragmentation”. In: *American Economic Review* 103.3, pp. 528–33.

Atkin, David (Apr. 2016). “The Caloric Costs of Culture: Evidence from Indian Migrants”. In: *American Economic Review* 106.4, pp. 1144–81.

Becker, Sascha O., Katrin Boeckh, et al. (2016). “The Empire Is Dead, Long Live the Empire! Long-Run Persistence of Trust and Corruption in the Bureaucracy”. In: *The Economic Journal* 126.590, pp. 40–74.

Becker, Sascha O. and Ludger Woessmann (May 2009). “Was Weber Wrong? A Human Capital Theory of Protestant Economic History”. In: *The Quarterly Journal of Economics* 124.2, pp. 531–596.

Bennett, Karen (2018). “Universal languages”. In: *A History of Modern Translation Knowledge: Sources, Concepts, Effects*, pp. 195–201.

Bisin, Alberto and Thierry Verdier (2014). “Trade and cultural diversity”. In: *Handbook of the Economics of Art and Culture*. Vol. 2. Elsevier, pp. 439–484.

Blair, Alan D and John Ingram (1998). “Loanword formation: A neural network approach”. In: *SIGPHON’98 The Computation of Phonological Constraints*.

— (2003). “Learning to predict the phonological structure of English loanwords in Japanese”. In: *Applied Intelligence* 19.1, pp. 101–108.

Blanc, Guillaume and Masahiro Kubo (2021). *French*. Tech. rep.

Bleakley, Hoyt and Aimee Chin (Jan. 2010). “Age at Arrival, English Proficiency, and Social Assimilation Among US Immigrants”. In: *American Economic Journal: Applied Economics* 2.1, pp. 165–192.

Bloomfield, Leonard (1933). *Language*.

Blouin, Arthur (2021). “Axis-orientation and knowledge transmission: Evidence from the Bantu expansion”. In: *The Journal of Economic Growth* 26.4, pp. 359–384.

— (2022). “Culture and Contracts: The Historical Legacy of Forced Labour”. In: *The Economic Journal* 132.641.

Cameron, Deborah and Don Kulick (2003). *Language and sexuality*. Cambridge University Press.

Cavalli-Sforza, Luigi L and Francesco Cavalli-Sforza (1994). “The great human diasporas: a history of diversity and evolution”. In:

Centola, Damon (2021). *Change: How to make big things happen*. Hachette UK.

Chawla, N. V. et al. (2002). “SMOTE: Synthetic Minority Over-sampling Technique”. In: *Journal of Artificial Intelligence Research* 16, pp. 321–357.

Chen, M. Keith (Apr. 2013). “The Effect of Language on Economic Behavior: Evidence from Savings Rates, Health Behaviors, and Retirement Assets”. en. In: *American Economic Review* 103.2, pp. 690–731.

Chipponi, J X et al. (May 1, 1982). “Deficiencies of essential and conditionally essential nutrients”. In: *The American Journal of Clinical Nutrition* 35.5, pp. 1112–1116.

Chisholm, Hugh (1911). *The Encyclopædia Britannica: Tonalite-Vesuvius.* Vol. 27. At the University Press.

Chomsky, Noam et al. (2006). *Language and mind.* Cambridge University Press.

Costinot, Arnaud and Dave Donaldson (2012). “Ricardo’s theory of comparative advantage: old idea, new evidence”. In: *American Economic Review* 102.3, pp. 453–58.

Cukor-Avila, Patricia and Guy Bailey (2001). “The effects of the race of the interviewer on sociolinguistic fieldwork”. In: *Journal of Sociolinguistics* 5.2, pp. 252–270.

Dasgupta, Partha and Ismail Serageldin (1999). *Social capital: a multifaceted perspective.* The World Bank.

Desmet, Klaus, Ignacio Ortuño-Ortín, and Romain Wacziarg (2012). “The political economy of linguistic cleavages”. In: *Journal of Development Economics* 97.2, pp. 322–338.

Dickens, Andrew (June 2019). *The Historical Roots of Ethnic Differences: The Role of Geography and Trade.* Working Paper 1901. Brock University, Department of Economics.

Dickens, Andrew et al. (2022). “Understanding ethnolinguistic differences: The roles of geography and trade”. In: *The Economic Journal* Forthcoming.

Durlauf, Steven N. and Marcel Fafchamps (2005). “Social Capital”. In: ed. by P. Aghion and S. N. Durlauf. Amsterdam: North Holland.

Eli, Shari, Laura Salisbury, and Allison Shertzer (2018). “Ideology and migration after the American Civil War”. In: *The Journal of Economic History* 78.3, pp. 822–861.

Ettinger, Mark (2014). “Here’s Why The World Can Never Have One Universal Language”. In: *Business Insider.*

FAO (2017a). *FAO/INFOODS Analytical food composition database version 2.0, AnFooD2.0.* FAO. Rome, Italy.

— (2017b). *FAO/INFOODS Food Composition Database for Biodiversity Version 4.0, BioFoodComp4.0.* FAO. Rome, Italy.

Feenstra, Robert C. (2004). *Advanced international trade: theory and evidence.* Princeton, N.J: Princeton University Press.

Felbermayr, Gabriel J and Farid Toubal (2010). “Cultural proximity and trade”. In: *European Economic Review* 54.2, pp. 279–293.

Foucault, Michel (1971). *L’ordre du discours.*

Frankopan, P. (2016). *The Silk Roads: A New History of the World.* Knopf Doubleday Publishing Group.

Gabszewicz, Jean, Victor Ginsburgh, and Shlomo Weber (2011). “Bilingualism and communicative benefits”. In: *Annals of Economics and Statistics/Annales d'Économie et de Statistique*, pp. 271–286.

Galor, Oded and Ömer Özak (2015). “Land Productivity and Economic Development: Caloric Suitability vs. Agricultural Suitability”. In: *Working Papers - Brown University, Department of Economics* 2015-5.

— (2016). “The Agricultural Origins of Time Preference”. In: *American Economic Review* 106.10, pp. 3064–3103.

Galor, Oded, Ömer Özak, and Assaf Sarid (2018). *Geographical roots of the coevolution of cultural and linguistic traits*. Tech. rep. National Bureau of Economic Research.

Ginsburgh, Victor and Shlomo Weber (2016). “Linguistic distances and ethnolinguistic fractionalization and disenfranchisement indices”. In: *The Palgrave handbook of economics and language*. Springer, pp. 137–173.

Giuliano, Paola, Antonio Spilimbergo, and Giovanni Tonon (2014). “Genetic distance, transportation costs, and trade.” In: *Journal of Economic Geography* 14.1, pp. 179–198.

Glaeser, Edward L, David Laibson, and Bruce Sacerdote (2002). “An economic approach to social capital”. In: *The economic journal* 112.483, F437–F458.

Gokmen, Gunes (2017). “Clash of civilizations and the impact of cultural differences on trade”. In: *Journal of Development Economics* 127, pp. 449–458.

Granovetter, Mark (1983). “The strength of weak ties: A network theory revisited”. In: *Sociological theory*, pp. 201–233.

Gray, R. and D. Birmingham (1970). *Pre-Colonial African Trade: essays on trade in Central and Eastern Africa before 1900*. Oxford U.P.

Guiso, Luigi, Paola Sapienza, and Luigi Zingales (May 2004). “The Role of Social Capital in Financial Development”. In: *American Economic Review* 94.3, pp. 526–556.

— (2008). “Social capital as good culture”. In: *Journal of the European Economic Association* 6.2-3, pp. 295–320.

— (2009). “Cultural biases in economic exchange?” In: *The quarterly journal of economics* 124.3, pp. 1095–1131.

Hall, Robert E and Charles I Jones (1999). “Why do some countries produce so much more output per worker than others?” In: *The quarterly journal of economics* 114.1, pp. 83–116.

Haspelmath, M. and U. Tadmor (2009). *Loanwords in the World's Languages: A Comparative Handbook*. De Gruyter Mouton.

Haugen, Einar (1950). “The analysis of linguistic borrowing”. In: *Language* 26.2, pp. 210–231.

Hensel, Paul R. and Sara M. Mitchell (2007). *The Issue Correlates of War (ICOW) Project Supplementary Data Set: Colonial History Data Set.* Edition: V2 Section: 2007-11-28.

Herbst, Jeffrey (2000). *States and power in Africa.* Princeton University Press.

IIASA/FAO (2012). *Global Agroecological Zones (GAEZ v3.0).* IIASA, Laxenburg, Austria and FAO, Rome, Italy.

Institute of Medicine (2006). *Dietary Reference Intakes: The Essential Guide to Nutrient Requirements.* Red. by Jennifer J. Otten, Jennifer Pitzi Hellwig, and Linda D. Meyers. Washington, DC: The National Academies Press.

Jakiela, Pamela and Owen Ozier (Sept. 2021). *Gendered Language.* preprint. Economics Working Papers.

Jaro, Matthew A. (1989). "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida". In: *Journal of the American Statistical Association* 84.406, pp. 414–420.

Kleinberg, Jon et al. (2018). "Human decisions and machine predictions". In: *The quarterly journal of economics* 133.1, pp. 237–293.

Kónya, István (2006). "Modeling cultural barriers in international trade". In: *Review of International Economics* 14.3, pp. 494–507.

Kubota, Ryuko (2001). "Teaching world Englishes to native speakers of English in the USA". In: *World Englishes* 20.1, pp. 47–64.

Labov, William (1964). "Phonological correlates of social stratification". In: *American Anthropologist* 66.6, pp. 164–176.

Labov, William and Wendell A Harris (1994). "Addressing social issues through linguistic evidence". In: *Language and the law*, pp. 265–305.

Lalee, M., J. Nocedal, and T. Plantenga (Aug. 1, 1998). "On the Implementation of an Algorithm for Large-Scale Equality Constrained Optimization". In: *Journal on Optimization* 8.3, pp. 682–706.

Leibniz, Gottfried Wilhelm (1916). *New Essays Concerning Human Understanding with an Appealing...: Transtated from the Original Latin, French and German Writeen.* Open Court Publishing Company.

Lemaitre, Guillaume, Fernando Nogueira, and Christos K. Aridas (2017). "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning". In: *Journal of Machine Learning Research* 18.17, pp. 1–5.

Lévi-Strauss, Claude (1951). "Language and the analysis of social laws". In: *American Anthropologist* 53.2, pp. 155–163.

Lewis, Paul M. (2009). *Ethnologue : languages of the world.* Texas: SIL International.

Lowes, Sara and Eduardo Montero (2018). "Concessions, Violence, and Indirect Rule: Evidence from the Congo Free State". In: *Unpublished manuscript.*

Lowes, Sara, Nathan Nunn, et al. (2016). “The Evolution of Culture and Institutions: Evidence from the Kuba Kingdom”. In: *Econometrica*.

Melitz, Jacques (2008). “Language and foreign trade”. In: *European Economic Review* 52.4, pp. 667–699.

Mi, Chenggang, Yating Yang, et al. (Oct. 2016). “Recurrent Neural Network Based Loanwords Identification in Uyghur”. In: *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation: Oral Papers*. Seoul, South Korea, pp. 209–217.

Mi, Chenggang, Shaolin Zhu, and Rui Nie (Apr. 2021). “Improving Loanword Identification in Low-Resource Language with Data Augmentation and Multiple Feature Fusion”. en. In: *Computational Intelligence and Neuroscience* 2021. Ed. by Nian Zhang, pp. 1–9. (Visited on 06/28/2022).

Michalopoulos, Stelios (2012). “The origins of ethnolinguistic diversity”. In: *American Economic Review* 102.4, pp. 1508–39.

Michalopoulos, Stelios, Alireza Naghavi, and Giovanni Prarolo (2018). “Trade and Geography in the Spread of Islam”. In: *The Economic Journal* 128.616, pp. 3210–3241.

Michalopoulos, Stelios and Elias Papaioannou (2014). “National Institutions and Subnational Development in Africa”. In: *The Quarterly Journal of Economics* 129.1, pp. 151–213.

Michalopoulos, Stelios and Melanie Meng Xue (2021). “Folklore”. In: *The Quarterly Journal of Economics* 1, p. 54.

Miller, John E. et al. (Dec. 2020). “Using lexical language models to detect borrowings in monolingual wordlists”. en. In: *PLOS ONE* 15.12. Ed. by Søren Wichmann, e0242709.

Miller, Robert J, Lisa Lesage, and Sebastián López Escarcena (2010). “The international law of discovery, indigenous peoples, and Chile”. In: *Neb. L. Rev.* 89, p. 819.

Mortensen, David R. et al. (2016). “PanPhon: A Resource for Mapping IPA Segments to Articulatory Feature Vectors”. In: *COLING*.

Mortensen, David R, Siddharth Dalmia, and Patrick Littell (2018). “Epitran: Precision G2P for many languages”. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Mufwene, Salikoko S. (2004). “Language Birth and Death”. In: *Annual Review of Anthropology* 33, pp. 201–222. (Visited on 07/12/2022).

Mullainathan, Sendhil and Jann Spiess (2017). “Machine Learning: An Applied Econometric Approach”. In: *Journal of Economic Perspectives* 31.2, pp. 87–106.

Murdock, George Peter (1959). *Africa : its peoples and their culture history*. New York [u.a.]: McGraw-Hill.

Naidu, Suresh, Sung-Ha Hwang, and Samuel Bowles (May 2017). “The Evolution of Egalitarian Sociolinguistic Conventions”. en. In: *American Economic Review* 107.5, pp. 572–577. (Visited on 07/14/2022).

Nocedal, J. and S. Wright (2006). *Numerical Optimization.* Springer New York.

Nunn, Nathan (2012). “Culture and the Historical Process”. In: *Economic History of Developing Regions* 27.S1, pp. 108–126.

Nunn, Nathan and Leonard Wantchekon (2011). “The Slave Trade and the Origins of Mistrust in Africa”. In: *American Economic Review* 101.7, pp. 3221–3252.

Okwudishu, Appolonia U (2019). “Globalization, multilingualism and the new information and communication technologies”. In: *In the Linguistic Paradise: A Festschrift for E. Nolue Emenanjo* 2, p. 1.

Olivier, Jacques, Mathias Thoenig, and Thierry Verdier (2008). “Globalization and the dynamics of cultural identity”. In: *Journal of international Economics* 76.2, pp. 356–370.

Paolillo, John C (2001). “Language variation on Internet Relay Chat: A social network approach”. In: *Journal of sociolinguistics* 5.2, pp. 180–213.

Pinker, Steven (2003). *The language instinct: How the mind creates language.* Penguin UK.

Putnam, Robert D (2007). “E pluribus unum: Diversity and community in the twenty-first century.” In: *Scandinavian political studies* 30.2, pp. 137–174.

Rauch, James E and Vitor Trindade (2002). “Ethnic Chinese networks in international trade”. In: *Review of Economics and Statistics* 84.1, pp. 116–130.

Reagan, Timothy (2005). “Mark Janse and Sijmen Tol (eds.). Language Death and Language Maintenance: Theoretical, Practical and Descriptive Approaches.” In: *Language Problems and Language Planning* 29.1, pp. 83–86.

Ricardo, David (1817). *On the Principles of Political Economy and Taxation.* London: John Murray. URL: https://www.econlib.org/library/Ricardo/ricP.html.

Sankoff, Gillian and Hélène Blondeau (2007). “Language change across the lifespan:/r/in Montreal French”. In: *Language*, pp. 560–588.

Schadeberg, Thilo C. (2009). “Loanwords in Swahili”. In: *Loanwords in the World’s Languages: A Comparative Handbook.* Ed. by M. Haspelmath and U. Tadmor. De Gruyter Mouton, pp. 77–102.

Schmid, Carol, Brigita Zepa, and Arta Snipe (2004). “Language policy and ethnic tensions in Quebec and Latvia”. In: *International Journal of Comparative Sociology* 45.3-4, pp. 231–252.

Scotton, Carol Myers and John Okeju (1973). “Neighbors and Lexical Borrowings”. In: *Language* 49.4, pp. 871–889.

Smith, Adam (1762). “Lectures on rhetoric and belles lettres”. In:

Spolaore, Enrico and Romain Wacziarg (2009). “The diffusion of development”. In: *The Quarterly journal of economics* 124.2, pp. 469–529.

Stewart, Brian A. et al. (Mar. 2020). “Ostrich eggshell bead strontium isotopes reveal persistent macroscale social networking across late Quaternary southern Africa”. en. In: *Proceedings of the National Academy of Sciences* 117.12, pp. 6453–6462.

Varian, Hal R. (2014). “Big Data: New Tricks for Econometrics”. In: *Journal of Economic Perspectives* 28.2, pp. 3–28.

Ward, Zachary et al. (2015). *The Role of English Fluency in Migrant Assimilation: Evidence from United States History.*

Winkler, William (Jan. 1, 1990). *String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage.*

Appendix A. Language Data Appendix

We construct data on linguistic exchange, starting from the word lists, or *lexicons*, included in the PanLex database, which takes thousands of translation dictionaries and converts them to a single common structure, including millions of words. PanLex includes the largest collection of living languages in a single data-set, and our final data-set includes all languages that can be matched from PanLex into the Ethnologue and other supplementary linguistic data-sets. While PanLex is an exceptionally rich data-set, among the millions of expressions it includes there are expressions that are not strictly words, and we outline here a number of approaches we take to deal with this.

For large languages[80] the lexicons include a non-negligible share of multi-word expressions, that should not themselves be considered words. To deal with this, we remove all expressions including multiple words, which are more common in heavily-resourced languages, and keep only single-word expressions. In order to remove named entities, we use all twenty languages covered by the `polyglot` package's Named Entity Recognition module. We identify all expressions that are identified as a Named Entity in any of the twenty languages, and then designate all meanings associated with those expressions as named entities. With this list, we then exclude any expressions in any language where one of the meaning identifiers is on this list of named entity meanings. While the PanLex database is not perfect, it does represent the largest collection of lexical data available. More importantly, the primary analysis we intend for this data is at the group-pair level, allowing the inclusion of language fixed-effects, which would capture variation in likelihood of identifying a loanword on average for a given language. We therefore caution that other researchers using this data should follow suit and use a research design primarily at the language pair-level and aggregated without drawing strong conclusion from small subsamples.

In order to train a machine learning classification algorithm that will identify likely loanwords and their sources in PanLex, we use the World Loanword Database (WoLD). We chose to use the World Loanwords Database (WoLD) as the training data for this model, as it is the most complete and authoritative data-set on the topic compiled by expert linguists. Despite the great care that went in to creating this resource, there may be some potential limitations, though these will not plausibly impact the main results in this paper. While the list of items included in the vocabularies for WoLD may be based on synonyms from Indo-European languages, the final list was selected to include meanings that are commonly included across most languages in the sample to reduce the likelihood of heterogeneous rates of missing data across the sample. In our algorithm, we ensure as much as possible that we do not introduce bias through definition of synonyms, and use a broader set of semantically similar meanings. In order

[80] we set the threshold at 100,000 expressions in the PanLex lexicon

to not favour Indo-European languages, we do this by using word embedding models trained on two hundred and ninety-six languages to identify similar concepts. Therefore, any bias that comes from an Indo-European basis for synonyms is minimised within the training set by the selection of commonly included concepts to reduce missing data, and bias in applying the algorithm *outside* the training set is minimised as much as possible by considering semantic embeddings from a very broad range of languages. Finally, the semantic similarity measure is used only to restrict to a set of feasible potential loanwords, and is not included as a feature used by the classifier, meaning that the potential for bias is further reduced. While WoLD may include a different number of words for different languages in the sample, our analysis is at the expression-level rather than at the language-level. This means that any bias would need to be introduced from the word-pair information as we excluded language-level information such as the loanword share or size of WoLD vocabulary as features. More importantly, any bias introduced by these language-level concerns would have to be correlated with pair-level characteristics[81] which is unlikely to be the case. We also show that our results are robust to concerns about WoLD appearing to under-sample language in the Americas, and show in table G9 that the results of the main pairwise horse-race specification are robust to dropping the New World. While WoLD may have potential drawbacks and may over-sample parts of the distribution of groups in PanLex, it is without a doubt the most consistent and reliable global source for validated loanwords, data. Where possible, the algorithm has been designed to reduce the risk of bias, and any limitations are highly unlikely to impact the pair-level analysis that is the core of this paper.

A.1. Bilingualism Data Extraction

The Ethnologue online stores additional information on Language Use. We used an automated web scraper to access the Ethnologue page for each language in the sample, and searched for a field labelled "Language Use" in the table of information. The information in this table is stored as text data, such as *"Also use Kâte [kmg]. Also use Tok Pisin [tpi]. Also use Yabem [jae]. Used as L2 by Musom [msu]. Shifting to [bub"* with references to other languages used by a given group, and other languages that use a group's language as a second language (or L2). From this we used text parsing to extract the associated three-letter Ethnologue codes, and use this to build a data-set recording whether a group has become bilingual or also uses another group's language for all the pairs in our analysis.

[81] Specifically, this bias introduced by language-level information would have to be correlated with complementarity in nutritional characteristics at the pair level.

Appendix B. Machine Learning Appendix

The preparation of the data-set follows the following rough order. We start with a data-set containing nearly all words in nearly all languages, and convert the original *orthographic representations* (i.e. the words as written in their original language, in the most natural script for the language) into a standardized phonetic representation so that we can compare across languages using different scripts. We then compute own-language dissimilarity measures to identify words that look like outliers. This is all matched to data on contextual similarity, which we use to construct a data-set containing candidate word-pairs that are in the same semantic space. For these word-pairs, we then compute additional pairwise distance measures between words in a candidate word-pair.

We train a classifier algorithm using all of these features. This generates a set of predictions of whether a word-pair is a loanword in a particular direction, and we apply these predictions to the full PanLex database. Our algorithm is based on the standard approach used by linguists, as documented in Haspelmath and Tadmor 2009. The mapping between this standard approach and our automation of it is summarized in table B1, and discussed in detail below.

B.1. Data Extraction and Phonetic Transcription

The first task in creating this data-set was extracting data on words (PanLex calls them *expressions*) from the PanLex data-set, after which we excluded named entities, prepared the necessary features for each expression, and converted orthographic representations into phonetic representations (i.e. the International Phonetic Alphabet (IPA)) using the methodology and data outlined in David R Mortensen, Dalmia, and Littell 2018.[82] Some language families were not represented. We therefore coded orthographic-phonetic mappings using orthography tables from Ager 2019 for 15 further languages, to give full coverage of the major language families included in our sample. We then use Ethnologue data on language families to match all languages in our sample to the nearest language sharing the same script included in our augmented list.

For each language, we build a data-frame including all expressions and extract the following information for each expression: *Unique Expression ID*, *Raw Text*, *Degraded text (no accents, etc.)*, *Language code*, and *IPA-converted text.* Meaning identifiers in the PanLex data-set refer to abstract meanings, that may be associated with one or more expressions. If two expressions are assigned the same meaning identifier, they can be thought of as translations.

[82] This method relies on mappings between orthographic and phonetic units. The data includes 64 language-script pairs. For example 'eng-Latn', for English in Latin script, and 'tir-Ethi' for Tigrinya in Ethiopic script.

Table B1: Computational Approximation of Linguistics Best Practice

Step	**Quote from Handbook**	**Computational Approximation**
Identify words with similar *shape*	*Linguists identify words as loanwords if they have a shape and meaning that is very similar to the shape and meaning of a word from another language* p 43-44	We use standard measures of differences in spelling as well as phonological differences, taking into account the likelihood of sound drift over time
Identify words with similar *meaning*	*Linguists identify words as loanwords if they have a shape and meaning that is very similar to the shape and meaning of a word from another language* p 43-44	We use the encoding of meanings in the PanLex data-set to identify words that share the exact same meaning. We expand this using a trained model of semantic similarity to expand beyond this and allow for matching of words from a similar context, but not the exact same meaning.
Consider likelihood of word being inherited from ancestor	*[. . .] of course, we need to exclude the possibility of descent from a common ancestor,* p 43-44	We create features for the classifier that measure similarity to the *Swadesh* words, which are likely to have been inherited from an ancestor, as a way to take into account the likelihood of a word being inherited. We also use language distance, measured by splits in the language tree, so that the classifier can take into account shared ancestry and set thresholds for other variables accordingly. We also restrict to loanword pairs where there is no ambiguity in borrowing, meaning it only looks like a loanword pair in one direction.
Identify plausible source word and donor language	*In general, a word can only be recognized with certainty as a loanword if both a plausible source word and a donor language can be identified*	Our algorithm operates purely at the word-pair level, so it never identifies a loanword without considering the paired source word in a donor language.
Use morphological and phonological criteria to determine direction of borrowing	*First, if the word is morphologically analyzable in one language but unanalyzable in another one, then it must come from the first language. For instance, German Grenze 'border' must have been borrowed from Polish granica 'border' [. . .] because -ica is a well recognized suffix in Polish, and the stem gran- occurs elsewhere, whereas German Grenze is not analyzable in this way. [. . .] This is the case, in particular, if the word is phonologically aberrant in a way that would be explicable by a borrowing history of the word. For example, Thurgood (1999: 11) notes that many loanwords from Mon-Khmer languages into Chamic languages (of the Austronesian family) can be recognized by their loan phonemes, sounds which occur only in borrowed words (e.g. implosives; thus, Chamic *ia+ 'little' seems to have a Mon-Khmer origin, Thurgood 1999: 313).*	We measure the similarity of a word to its own language, as well as measuring how often the bigrams and trigrams in a word appears in the other words in the given language. We then take the difference between these for the source and target word, to identify which word looks most like the outlier.
Consider other semantically similar words in the same language	*[. . .] the meaning often helps: Sanskrit nakra- 'crocodile' is likely to be a loanword from Dravidian (e.g. Kannada negar), because Indo-Aryan speakers coming from northern India would not have brought a word for crocodile with them (Burrow 1946: 9)*	We use the same relevant meaning procedure as with our cross-language comparisons to identify semantically similar words in the same language, and consider lexical, phonetic and bigram/trigram dissimilarity to these words. This allows the classifier to identify if a word in the language is similar to words with relevant meanings.

Note: This table summarizes how our computational methodology, and the features used to construct our classification algorithm, approximate best practice used by linguists.

Table B2: Phonological Features With Weights

Phonological Feature	**Feature Weight**	**Feature Description**
syllabic	1.0	Is the segment the nucleus of a syllable?
sonorant	1.0	Is the segment produced with a relatively unobstructed vocal tract?
consonantal	1.0	Is the segment consonantal (not a vowel or glide, or laryngeal consonant)?
continuant	0.5	Is the segment produced with continuous oral airflow?
delayed release	0.25	Is the segment an affricate?
lateral	0.25	Is the segment produced with a lateral constriction?
nasal	0.25	Is the segment produced with nasal airflow?
strident	0.125	Is the segment produced with noisy friction?
voice	0.125	Are the vocal folds vibrating during the production of the segment?
spread glottis	0.125	Are the vocal folds abducted during the production of the segment?
constricted glottis	0.125	Are the vocal folds adducted during the production of the segment?
anterior	0.25	Is a constriction made in the front of the vocal tract?
coronal	0.25	Is the tip or blade of the tongue used to make a constriction?
distributed	0.125	Is a coronal constriction distributed laterally?
labial	0.25	Does the segment involve constrictions with or of the lips?
high	0.25	Is the segment produced with the tongue body raised?
low	0.25	Is the segment produced with the tongue body lowered?
back	0.25	Is the segment produced with the tongue body in a posterior position?
round	0.25	Is the segment produced with the lips rounded?
tense	0.25	Is the segment produced with an advanced tongue root.
long	0.125	Length of sound. “For instance, Classical Latin had five contrasting long and five contrasting short vowels. This feature applies as well to consonants; long consonants are often also called geminates. ” (Hayes 2009, p.83)
velaric	0.25	Refers to the position of the velum, e.g. “The velum, or soft palate. This is a flap of soft tissue that separates the mouth from the nasal passages. [. . .] When the velum is high, then the velar port is closed, and air is confined to the oral passage.” (Hayes 2009, p.5)

Note: Phonological features used to compute phonetic similarity. In the unweighted version, all features are given the same weight. In the weighted distance, features are weighted by the inverse likelihood of these features drifting over time. Features with the highest weights are those that are least likely to drift and the most indicative that two words did not originate from the same source. Descriptions are taken from David R. Mortensen et al. 2016 unless otherwise specified.

B.2. Train Machine Learning Classifier

For each word pair, we calculate the features described below, which are the inputs to the machine learning algorithm. These are the features the classification algorithm uses to decide if a given word-pair is likely an adopted loanword.

i) Own-Dissimilarity Measures A core requirement for linguists while identifying loanwords is to determine which words appear to be outliers in their language, as these are more likely to have been adopted. Conversely, more typical sounding words are less to have been introduced from another language. As an example of how linguists approach this:

> In general, a word can only be recognized with certainty as a loanword if both a plausible source word and a donor language can be identified. [...] This is the case, in particular, if the word is phonologically aberrant in a way that would be explicable by a borrowing history of the word.(Haspelmath and Tadmor 2009, p. 44)

We approximate and automate this step by generating a few measures of own-language dissimilarity. First, the Jaro-Winkler similarity metric computes the minimum edit distance between two words, accounting for transpositions, where greater weight is given to characters near the beginning of the word (Jaro 1989; Winkler 1990). As loanwords are often likely to be adapted with added suffixes, this metric is suitable for measuring likelihood of a word being introduced from another language. This measure is between 0 and 1, with 1 being identical spellings. Our measure computes the Jaro-Winkler distance (i.e. one minus the Jaro-Winkler similarity) with other words in the same language, and compute the deciles of this distribution to use as features.

To give a concrete example of how this approximates the methods used by linguists, consider the following discussion of loanwords in Hausa, where the the unusualness of the first syllable *'ta-'* is used to identify it as a loanword, and potential sources where this first syllable is not so dissimilar to the rest of the language:

> Hausa tafasa ('to boil') could link to several good reflexes of a root p/f-s- in other Chadic and Afroasiatic languages [...] but we lack an explanation for the origin of the first syllable ta- from within Hausa (Awagana, Wolff, and Löhr 2009, p.153).

We also construct the Jaro-Winkler measure restricting only to the spellings of words in the language with similar meanings. To do this we use the same threshold for contextual similarity as when we generate word-pairs, described in more detail below, and compute quintiles of this distribution as features. For these contextually-similar words we also compute the phonetic difference and also use quintiles of this distribution as features.

In addition to looking at expectedness of spellings of how words are written, we also look at the phonetics similarity to other words in a language. Again, to link this to how a linguist would approach phonetic similarity, we return to a linguists loanwords discussion, this time focusing on Swahili. The presence of the unusual sound *'h'* was in this case a sign of borrowing:

> The consonant h occurs natively only in grammatical morphemes (i.e., proximal and referential demonstratives, and the habitual and the negative pre-initial markers hu- and ha-); its appearance in lexical items is a sign of borrowing, e.g. huru 'free' < Arabic 'urr), muhanga 'aardvark' (< Sambaa mhanga) Schadeberg 2009, p. 98.

We construct measures of whether the combinations of phonetic units, or *phonemes*, that make up a word are typical for the language. Once again, the presence of outliers in terms of phonetic composition is often used by linguists to identify borrowing and the likely source:

> This is the case, in particular, if the word is phonologically aberrant in a way that would be explicable by a borrowing history of the word. For example, Thurgood ... notes that many loanwords from Mon-Khmer languages into Chamic languages (of the Austronesian family) can be recognized by their loan phonemes, sounds which occur only in borrowed words (e.g. implosives; thus, Chamic *ia+ 'little' seems to have a Mon-Khmer origin (Thurgood 1999, p. 313).

Using the phonetic transcriptions of PanLex expressions, we build a list of all 2- and 3-grams of phonemes contained in a language and compute the expected number of occurrences of this n-gram, and the position of this n-gram in words from that language. For each word, we then take the average of this score for all contiguous sequences of two or three phonemes making up a word. This captures whether a combination of sounds , and whether a given combination of sounds in a given part of an expression, is unusual.

In the basic phonetic n-gram measure we create above, we create an expected occurrence score for 2- and 3-grams of a word based on observed occurrence in all words in the language. We generate an additional measure of similarity to ancestral languages, by comparing words to the 'core' words in a language that are likely inherited from an ancestor language. The possibility of mistaking cognates for loanwords is a crucial factor, always taken into account by linguists, e.g.:

> Most importantly, of course, we need to exclude the possibility of descent from a common ancestor, which is a very common reason for word similarities across languages (Haspelmath and Tadmor 2009, p. 44).

To construct a measure of similarity to ancestors we therefore construct measures of whether a word's phonemes are typical, given words from the Swadesh list for a language. The Swadesh lists are a list of words that are almost certainly core to a language, and are considered unlikely to have been borrowed. Our source for Swadesh words is the word lists compiled as part of the Automatic Similarity Judgement Program (Wichmann, Holman, and Brown 2016). Using the phonetic transcriptions of these Swadesh words, we build a list of all 2- and 3-grams of phonemes contained in a language and compute the expected number of occurrences of this n-gram in Swadesh words, then take the average of this score for all contiguous sequences of two or three phonemes making up a word.

ii) Semantic Similarity Routine To restrict the space of candidate word pairs we consider, we generate a measure of the contextual distance between concepts. To do so, we use a pre-trained model of word vectors trained from the Google News data-set and two hundred and ninety-four multiple language versions of Wikipedia compiled by Bojanowski et al. (2017). This contextual similarity is implemented by the `Gensim` package (Rehurek and Sojka 2010). The intuition of this procedure is to represent or 'embed' words as vector values in semantic space. Here, semantic space is a 300-dimensional vector space where each of these dimensions captures some feature or characteristic of words that captures the similarity between their meanings. Each of these dimensions is intuitively related to a characteristic that captures the relationship between two words.[83] For all meanings in the PanLex data-set in one of the covered languages we can assign a contextual similarity score, between 0 and 1. For all expressions with the same meaning identifier, we assign a similarity score of 1. For all other word-pairs, we consider them semantically similar if they are above a threshold of 0.8. This word vector measure of contextual similarity of expressions is less restrictive than considering only expressions that are strict translations, and broadens the space of potential loanword pairs while still preventing nonsensical matches between expressions denoting entirely unrelated concepts. This semantic similarity routine also considers word associations from hundreds of languages across the world, so it is not overly biased towards word associations in a single language.

iii) Word-Pair Construction and Pairwise Distance Having created own language dissimilarity measures for expressions and mapped them into space of contextual similarity, we create word pairs that are candidates for being loanwords, and generate pairwise

[83] Mikolov, Yih, and Zweig (2013) give a commonly used example of how these vectors follow intuitive logical arithmetic: *we examine the vector-space word representations that are implicitly learned by the input-layer weights. We find that these representations are surprisingly good at capturing syntactic and semantic regularities in language, and that each relationship is characterized by a relation-specific vector offset. This allows vector-oriented reasoning based on the offsets between words. For example, the male/female relationship is automatically learned, and with the induced vector representations, "King - Man + Woman" results in a vector very close to "Queen."*

distance measures. This is typically the first step linguists consider when identifying loanwords. e.g.

> Linguists identify words as loanwords if they have a shape and meaning that is very similar to the shape and meaning of a word from another language (Haspelmath and Tadmor 2009, pp. 43–44).

We restrict our potential source words to those that are above a threshold of contextual similarity. For each expression, we identify the top one thousand most similar expressions, and within those restrict to those above a threshold of 0.8 (from a range of 0 to 1) as expressions representing meanings that are similar enough to be plausible source words. We merge all associated meaning identifiers onto the expressions to identify potential source meanings for the given word. This threshold is low enough that we consider a broad range of related meanings, but is high enough to be practical and reduce the number of comparisons made to a level that can be carried out with a reasonable amount of computing time.

For each word in our data-set, we create pairwise matches with all words in all other languages. Since each *expression* may be mapped to multiple *meanings*, we create pairwise matches at the word-meaning level, and restrict to the most similar meaning pair for each word-pair where words have multiple meanings. We then restrict to pairs of words that have meaning-pairs identified as contextually similar, as above. We then calculate a number of pairwise distance measures between the two words, as follows.

iv) Articulatory Feature-Edit Distance Metrics The first set of pairwise distance metrics we create exploits detailed information on the phonemes that make up the phonetic representations of words. We map each phoneme to a vector of twenty-one articulatory features describing the way a spoken sound is actually produced, such as tongue position, open or closed mouth, etc.[84] This level of detail means that phoneme differences can be weighted by how similar the two phonemes sound.

Using these articulatory vector representations, we construct two pairwise minimum edit distances. The Hamming Feature-Edit Distance computes the minimum distance between two words, allowing for insertion and deletion of phonemes and accounting for the difference in phonemes weighted by difference in articulatory features.

The Weighted Hamming Feature-Edit distance is similar to the unweighted Hamming distance, but where the cost of articulatory feature edits are differently weighted depending on their class and subjective variability. This takes into account how likely it is that differences in sound result from a borrowed word naturally changing, or not, which could make the potential source word more or less plausible. See, for example, the discussion

[84] This is done using the PanPhon package developed in David R. Mortensen et al. 2016

of loanword integration in Kanuri from Löhr, Wolff, and Ari Awagana 2009 where loanwords are adjusted to fit the borrowing language. The Weighted Hamming Feature-Edit therefore accounts for the type of likely natural adjustments made to the source word as part of the integration processes.

v) Jaro-Winkler Distance As with the own-language dissimilarity measures, we compute the Jaro-Winkler orthographic distance for the candidate wordpair.

vi) Language Family Cladistic Distance For the candidate word-pair, we also compute the pairwise cladistic distance between the two languages. This data is based on the Ethnologue language family trees (Lewis 2009), where the measure of linguistic family distance is equal to the share of nodes in the first language's tree that are also in the second language's family classification. As discussed above, considering the possibility of inheritance from a common ancestor is a crucial step in how linguists identify loanwords.

vii) Pairwise Difference in Own-Language Dissimilarity In addition to these measures of pairwise difference between words, we also calculate the *difference* in all of the own-language dissimilarity measures generated above. By including the differences in these measures as features in the machine learning algorithm, we allow the classifier to explicitly decide whether one word in a pair appears more likely to be an outlier than the other. A similar technique is often used by linguists to gain information about the direction of borrowing:

> However, there are a number of criteria available that often give us a clear indication of the borrowing direction. First, if the word is morphologically analyzable in one language but unanalyzable in another one, then it must come from the first language. For instance, German Grenze *'border'* must have been borrowed from Polish granica *'border'* rather than the other way round, because *-ica* is a well recognized suffix in Polish, and the stem *gran-* occurs elsewhere, whereas German Grenze is not analyzable in this way (Haspelmath and Tadmor 2009, p. 45).

Therefore, when linguists identify possible loanword pairs and want to learn about the direction, they compare the relative dissimilarity of the two words to their own languages to identify cases where one word is clearly an outlier and the other word is not. To replicate this, we compute the difference in the own-language dissimilarity for the two words in the candidate word-pair.

B.3. Training Random Forest Classifiers

Having generated our observations and features from PanLex, we matched by spelling to the World Loanwords Database (WoLD). As discussed in section 3, WoLD is a data-set of loanwords, with origins, manually classified by linguistic experts.

We first built a training set including a number of word-pairs of important categories, including word-pairs that are known loanwords target-source pairs, as well as known loanwords target-source pairs with the direction *inverted.*[85] This way the classifier could learn more about identifying the direction of loanwords borrowing. We also included in the training set known non-loanwords matched to potential source words with similar meanings as well as known loanwords matched to incorrect source words with similar meanings.

As discussed in the body of the paper, loanword pairs are a very small share of all possible word-pairs, so we use a combination of under-sampling (where we select a random sample of the non-loanwords and loanwords matched to incorrect source words) as well as oversampling the under-represented categories (loanword pairs and inverted loanword pairs).

After training a Random Forest and an Extremely Random Forest, we combine these into a Voting Classifier. We then expand the random subsample of non-loanwords and loanwords matched to the wrong source in the training set and then restrict to only those classified as loanword pairs by the first-stage classifier to build a training set only of reasonably plausible loanword pairs. We then train another Random Forest and Extremely Random Forest on this new training set which we use to make our final classification. The first-stage classifier can be thought of as a coarse filter to plausible pairs, and the second-stage classifier is a more focused classifier that chooses from among these plausible pairs.

B.4. Generating Measures for Analysis

From the second stage classifier, there remains one consideration to identify loanwords. What happens when a word is attached to multiple sources? Where there are multiple source words identified by the second-stage classifier, we sort in descending order by second-stage classifier's predicted probability of being a loanword and keep the first candidate. We then exclude any loanword pairs where the source word also appears in the list of target words. This allows us to remove any ambiguous cases where the direction of borrowing is unclear, cases where a word is indirectly borrowed via an intermediary, and to deal with *'areal roots'* or *'wanderwörter'*, which are words that occur commonly

[85] We used all known loanword pairs and their inversions, including those whose meanings were not included in our list of semantically similar meanings, that could be matched by spelling. This does not impact performance as semantic similarity is not used as a classifier feature.

and often appear as outliers.[86] (Awagana, Wolff, and Löhr 2009, p. 149)

In the construction of our loanwords data, we exclude chains of loanword borrowing, and would not consider a word that was borrowed from language A ⟶ language B ⟶ language C. This choice was made to reduce ambiguity, as well as avoid cycles of borrowing among similar languages. This would underestimate the degree of borrowing of group B words by group C if this influence is primarily captured by words group B borrowed from group A. For this to be a concern for the empirical analysis, it would have to be correlated with gains from trade. This would potentially be a concern with endogenous gains from trade influenced by B's interaction with A, but not in our empirical exercise using pair-wise nutritional complementarity.

Appendix C. Validation of Language Data

C.1. Cross-validation and algorithm performance

We took a few approaches to examine the validity of the loanwords data. One such approach was to consult with a linguist with expertise in a language that was not in the training set. We sent the expert - who is a linguistics professor with a specialization in east African loanwords - a small sub-sample of representative words from a language of their expertise. Half of the word-pairs were ones that our classifier listed as probable loanwords, and half were not. The words were randomly ordered, and we did not reveal the predictions of the classifier.[87] We asked the expert for a ranking from most to least likely to be a loanword. Overall, the correlation between the raw expert ranking and the classifier's was 0.73. While this is fairly promising - it is only one case study of a small sub-sample of word-pairs.

i) Overall accuracy: To get a more comprehensive picture of accuracy, we employ standard cross-validation techniques, and report the results in table 1. Panel A reports the overall accuracy, which is a combination of first and second stage accuracy. Panel B reports the results from the first-stage classification only. Panel C reports second stage accuracy on the observations that were suspected to be possible loanwords in the first stage. The table suggests that the classifier is approximately 98% accurate overall (table 1, panel A, column 1 as well as figure C2). However, this high accuracy score could be misleading if it comes at a cost of very low accuracy among some types of classifications. For example, if the classifier achieved 0% accuracy on actual loanwords and 100% accuracy on non-loanwords - i.e. it categorized everything as a non-loanword - it would

[86] These words are often of *onomatopoeic* origin, meaning the word mimics a real sound, and so might mistakenly be classed as loanwords. By removing these 'cycles' of borrowing, we reduce the likelihood of including these.

[87] We did share some contextual information that was needed to make an assessment, like the candidate source and donor languages.

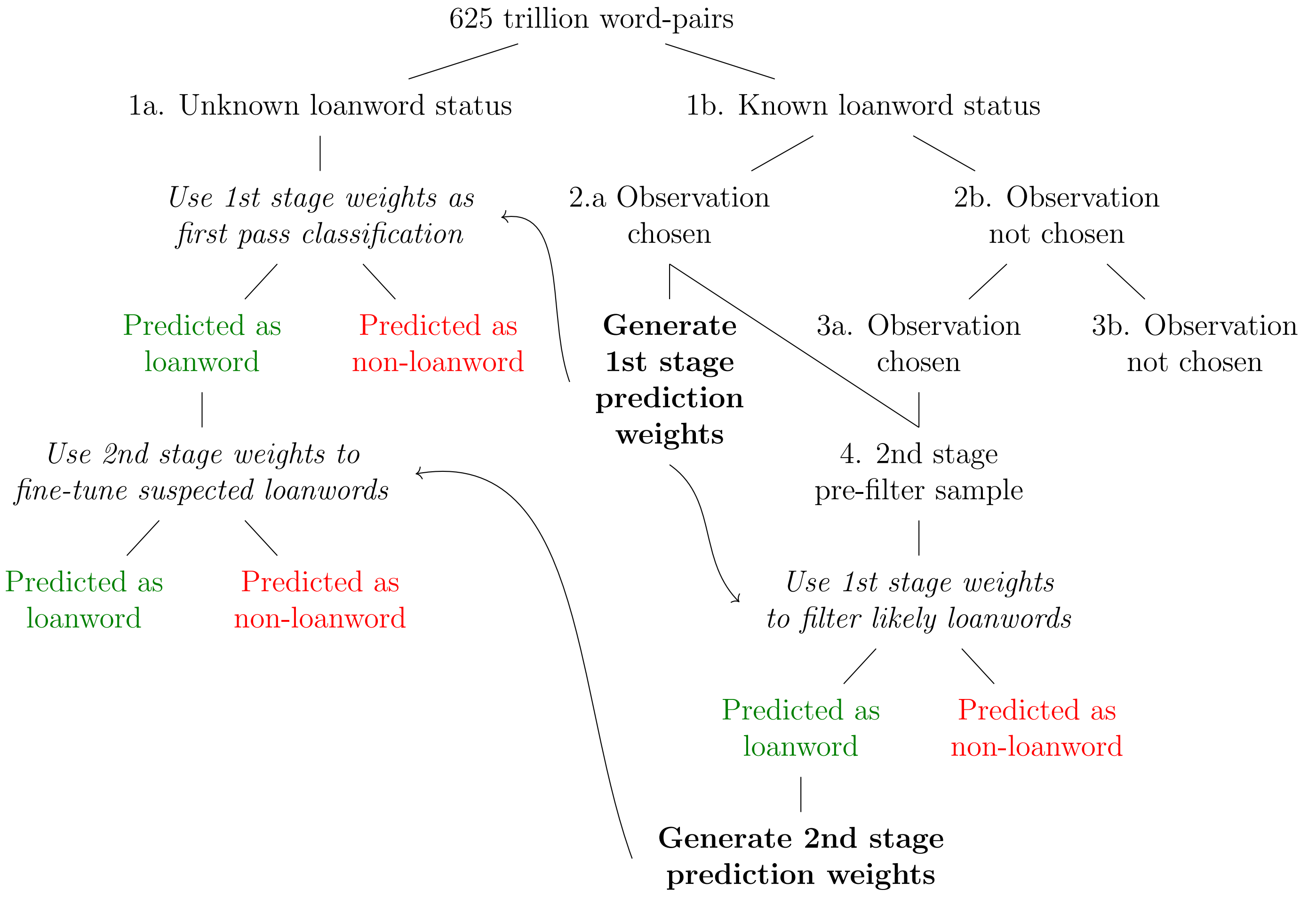

625 trillion word-pairs
1a. Unknown loanword status
1b. Known loanword status
Use 1st stage weights as first pass classification
2.a Observation chosen
2b. Observation not chosen
Predicted as loanword
Predicted as non-loanword
Generate 1st stage prediction weights
3a. Observation chosen
3b. Observation not chosen
Use 2nd stage weights to fine-tune suspected loanwords
4. 2nd stage pre-filter sample
Predicted as loanword
Predicted as non-loanword
Use 1st stage weights to filter likely loanwords
Predicted as loanword
Predicted as non-loanword
Generate 2nd stage prediction weights

Figure B1: Machine Learning algorithm flow-chart

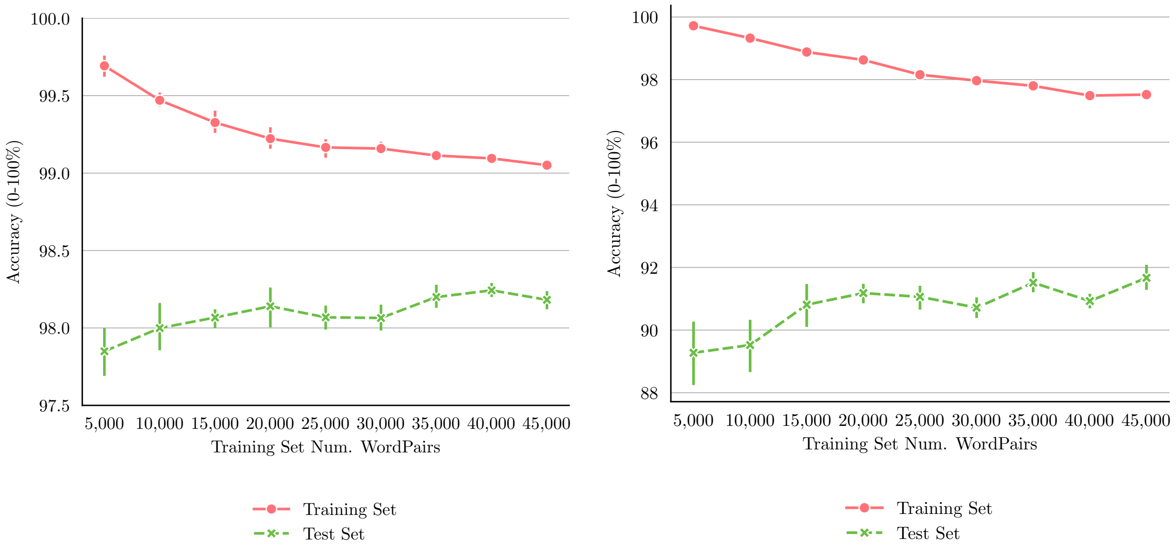
Table B3: Example: Sources Compiled by WoLD for Swahili

Author	**Year**	**Publication Title**
Beaujard, Philippe	1998	Dictionnaire malgache (dialectal)-français: dialecte tañala, sud-est de Madagascar.
Besha, Ruth M	1993	A classified vocabulary of the Shambala language with outline grammar
Brauner, Siegmund	1986	Chinesische Lehnwörter im Swahili Zeitschrift für Phonetik
Dozy, Reinhart P. A.	1881	Supplément aux dictionnaires arabes.
Grosset-Grange, Henri & Alain Rouaud	1993	Glossaire nautique arabe ancien et moderne de l'Océan Indien
Höftmann, Hildegard, and Irmtraud Herms	1979	Wörterbuch Swahili-Deutsch.
Höftmann, Hildegard	1963	Suaheli-Deutsches Wörterbuch.
Holes, Clive	2001	Dialect, culture, and society in Eastern Arabia
Johnson, Frederick	1939	A standard Swahili-English dictionary
Kazimirski, A. de Biberstein	1860	Dictionnaire arabe-français, contenant toutes les racines de la langue arabe
Kirkeby, Willy A	2000	English-Swahili dictionary
Kisbey, W. A	1906	Zigula-English dictionary
Knappert, Jan	1970	Contribution from the study of loanwords to the cultural history of Africa
Knappert, Jan	1972 - 1973	The study of loan words in African languages
Knappert, Jan	1983	Persian and Turkish loanwords in Swahili
Krumm, Bernhard	1940	Words of oriental origin in Swahili
Lane, Edward William	1863 - 1893	An Arabic-English lexicon
Lang Heinrich, F	1921	Schambala-Wörterbuch
Lodhi, Abdulaziz Y	2000	Oriental influences in Swahili: a study in language and culture contact
Maganga, Clement, and Thilo C. Schadeberg	1992	Kinyamwezi: grammar, texts, vocabulary
Nurse, Derek, and Thomas J. Hinnebusch	1993	Swahili and Sabaki: a linguistic history
Platts, John T	1884	A dictionary of Urdu, classical Hindi and English
Sacleux, Ch.	1939	Dictionnaire swahili-français
Steingass, Franz	1892	A comprehensive Persian-English dictionary including the Arabic words and phrases to be met with in Persian literature.
Taasisi ya Uchunguzi wa Kiswahili (TUKI)	1981	Kamusi ya Kiswahili Sanifu
Taasisi ya Uchunguzi wa Kiswahili (TUKI)	1996	English-Swahili dictionary / Kamusi ya Kiingereza-Kiswahili
Taasisi ya Uchunguzi wa Kiswahili (TUKI)	2001	Kamusi ya Kiswahili-Kiingereza / Swahili-English dictionary
Velten, Carl	1910	Suaheli-Wörterbuch – 1
Velten, Carl	1933	Suaheli-Wörterbuch –2
Wagenaar, H. W., S. S. Parikh, D. F. Plukker, and R. F. Veldhuyzen van Zanten	1993	Allied Chambers transliterated Hindi-Hindi-English dictionary
Wilkinson, R. J	1901 - 1902	A Malay-English dictionary
Wilkinson, R. J	1932	A Malay-English dictionary (romanised)
Worms, A	1898	Wörterverzeichniss der Sprache von Uzaramo

Note: This table shows all the sources required in order to compile the classified WoLD data for a single language, Swahili. This demonstrates the enormity of the task of classifying loanwords, and motivates our use of a big-data approach and significant high-performance computing resources.

achieve greater than 99.9% accuracy overall. However, a classifier that did this would obviously be useless.

For this reason we also report a number of different diagnostic measures that balance different types of errors. For instance, *precision* is the share of predicted positives that are true positives, while *recall* is the share of actual positives that are predicted. Building on



(a) First Pass **(b)** Second Pass

Figure C1: Accuracy of Voting Classifiers by Sample Size

Note: The figure shows the accuracy of the machine learning algorithm by training set size. Accuracy is defined as the share of word-pairs that are correctly classified. On the y-axis we show the share of word-pairs classified correctly by the algorithm. Vertical lines represent 95% confidence intervals. In figure C2 we further break this down and show accuracy on the different classes of word-pair in the sample.

our previous example, suppose the algorithm predicted only one loanword, but that it was correct - the predicted word-pair was a true loanword pair. In this case the recall score would be approximately 0 since it only picked up a tiny share of the actual loanwords, but the precision score would be 1, since 100% of the loanwords that it identified were actually loanwords.[88]

The classifier still does well on these additional checks, with an 85% recall score (panel A column 5), and a precision score of 83% (panel A column 4).[89] The F1-score and balanced accuracy scores account for both of these different error types, but in different ways, and so they are useful summary statistics.[90] In this case, since the precision and recall scores are similar to each other it is not surprising that the F1-scores and balanced accuracy scores are also similar.[91]

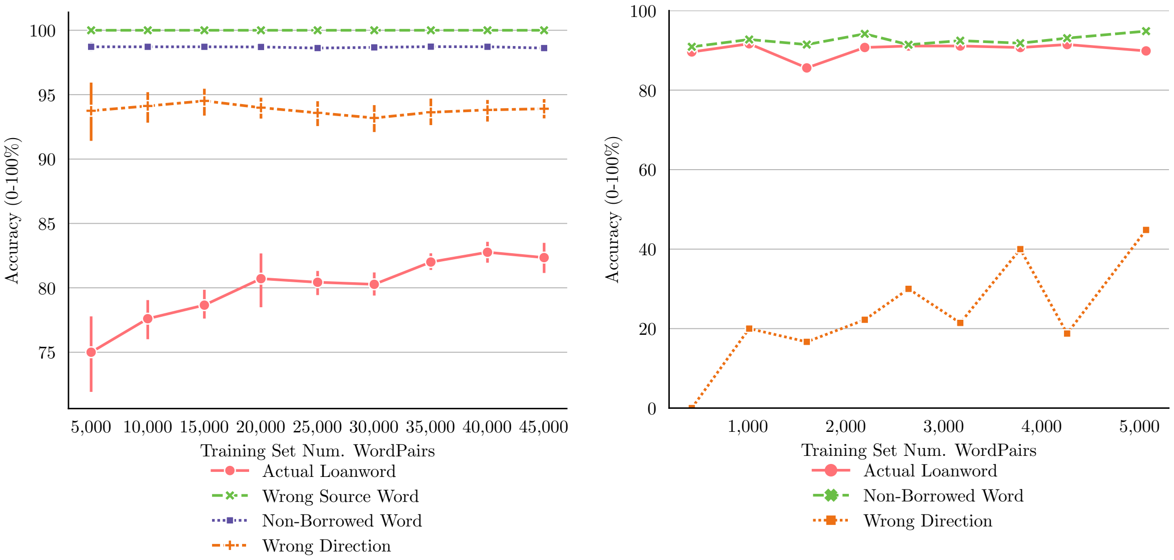
Figure C2 goes a bit further, by showing exactly where the accuracy is coming from. The pink line represents accuracy on word-pairs that are actually loanword-pairs. This is

[88] The accuracy score, meanwhile, would be approximately equal to the share of words that were not loanwords. So it too would be high, as previously discussed.

[89] Incidentally, this lines up with the expert consultation, where the recall score is 80% if we treat the expert's top-half as 'true-positives' (note that these also come with some uncertainty).

[90] Balanced accuracy is the mean of the true positive rate and the true negative rate. The F1-score is the harmonic mean of precision and recall.

[91] We also checked for heterogeneity in accuracy by society observable characteristics. We look at accuracy among the sub-sample of colonisers and colonists; above / below the median coverage in PanLex; large and small societies; high and low land quality; and by continent. Across the board, we find very little difference in either overall accuracy, or balanced accuracy, which gives equal weighting to the different types of errors (table C1).



(a) First Pass

(b) Second Pass

Figure C2: Accuracy of Voting Classifiers by Word-pair Type

Note: The figure shows the accuracy of the machine learning algorithm by training set size for different types of word-pairs. Accuracy is defined as the share of word-pairs that are correctly classified. On the y-axis we show the share of word-pairs classified correctly by the algorithm. When deciding on when to stop adding observations to the training set, we use the point where adding observations no longer made meaningful improvements in accuracy.

essentially the recall score. For actual loanwords, the classifier correctly identifies them as loanwords about 89% of the time. Next, going bottom to top (on both the figure and legend), the figure displays the category of word-pairs that are actually a loanword pair, but the direction of transfer is flipped. In these cases the classifier correctly identifies the pair as not being a loanword-pair almost 95% of the time. It correctly categorizes words that were not borrowed at all as non-loanwords over 98% of the time. Finally, it essentially never mis-categorizes actual loanwords as being borrowed from the wrong potential source word in another language. We also present balanced accuracy scores by additional sources of heterogeneity. Table C1 shows heterogeneity by observable characteristic.

C.2. Loanwords and Societal Similarity

While our machine learning classifier is accurate at identifying loanwords, we still need to verify that loanwords can be interpreted as a proxy for cross-societal transmission of characteristics. If loanwords do capture horizontal transmission, the share of loanwords in one language originating from another should positively co-vary with the share of common traits between those two societies. While this is simple to compute, a challenge arises from the fact that this is clearly not the only way to achieve societal similarity.

Groups could be similar for any of three reasons: (1) they adopted traits from each other; (2) their common environment produced similar features; (3) they share an an-

Table C1: Accuracy Heterogeneity by Language Category

	Lender		Borrower	
	Accuracy (1)	Balanced Accuracy (2)	Accuracy (3)	Balanced Accuracy (4)
Colonist	0.9997	0.9998	0.9996	0.8435
Colonised Group	0.9993	0.9232	0.9996	0.8431
Above Median Wordcount	0.9996	0.9224	0.9996	0.9224
Below Median Wordcount	0.9999	.	.	.
Population Above Median	0.9995	0.9248	0.9996	0.9446
Population Below Median	0.9999	0.7999	0.9998	0.8803
Land Quality Above Median	0.9995	0.9190	0.9996	0.9299
Land Quality Below Median	0.9997	0.9249	0.9996	0.9211
Africa	0.9996	0.9613	0.9994	0.8664
Americas	0.9994	0.9222	0.9996	0.9405
Asia	0.9998	0.7306	0.9997	0.8471
Europe	0.9990	0.9393	0.9995	0.9336
Pacific	0.9999	0.8999	0.9997	0.9421

Note: This table gives the accuracy and balanced accuracy for a random sample of word-pairs in PanLex whose loanword status is known from WoLD. These are disaggregated by language category. In columns (1) and (2) the score is for potential loanword pairs where languages from the given category are the potential loanword *lenders*, and in columns (3) and (4) we consider word-pairs where languages of the given category as the potential *borrowers*. The random draw for the below median word count category did not include any actual loanwords. So the balanced accuracy (which takes a mean of accuracy on loanwords and non-loanwords) could not be computed. We cannot compute below median word count for the borrower because none of the WoLD borrowing languages are below median word count in PanLex.

cestry, and inherited characteristics that have not changed.[92] Accordingly, we would like to decompose cross-societal similarity into these three components, to assess whether horizontal transmission is an important source of reduction in cross-societal differences relative to the other two mechanisms.

Fortunately, there already exist very good measures of environmental and ancestral similarity. For example, to measure common ancestry, Blouin 2021 examines distance in language tree nodes, a strategy that suits this context particularly well given the focus on language. Environmental similarity is even easier. For instance, geographic distance

[92] A further challenge arises from the possibility that societies that were already similar to each other may have interacted more, and may therefore have experienced more horizontal transmission.

between two groups will also account for much of the differences in geography, institutions, and history. We can assess the relative strength of explanatory variables effects with and without country (c) fixed effects to account for any geographic or institutional features that might have generated similar societal traits between the two groups (groups i and j). Likewise, we can include language family (l) fixed effects to account for the possibility that some language groups are more amenable to linguistic adoption than others.

With all of this included, we get the following regression equation:

$$\begin{aligned}(14) \qquad [\frac{100 \cdot \text{\# common societal traits}}{\text{\# societal traits observed}}]_{ijcl} &= \alpha_{ci} + \alpha_{cj} + \alpha_{li} + \alpha_{lj} \\ &+ \beta_1[\frac{100 \cdot \text{\# loanwords}}{\text{\# words in lexicon}}]_{ijcl} \\ &+ \beta_2 AncestralSimilarity_{ijcl} \\ &+ \beta_3 GeographicDistance_{ijcl} + \epsilon_{icl}\end{aligned}$$

We measure linguistic borrowing as the share of words (# words in lexicon) that one society borrowed from another (# loanwords). We multiply by 100 so that this ratio ranges from 0-100. Likewise, $AncestralSimilarity_{ijcl}$ is the share of language tree nodes that the two societies have in common, and also ranges from 0-100. $GeographicDistance_{ijcl}$ measures distance between Ethnologue centroids and proxies for the similarity between the environmental conditions faced by each group. Each of α_{ci}, α_{cj}, α_{li}, α_{lj}, represents fixed effects by country (c) or language family (l) for societies i and j. To measure societal similarity between group i and group j we measure the share of identically measured traits in the Ethnographic Atlas, relying on data-set provided by Giuliano and Nunn 2018. We also estimate a version of this regression using a similar societal similarity measure, but from the Folklore data (Michalopoulos and Xue 2021). Of course the advantage of the word-level approach is that we can investigate adoption directly, while in this exercise we can capture similarity but not who adopted from whom.

We are interested in the estimate of β_1, which may help to determine whether loanwords are capturing horizontal transmission. Since we are interested in getting a sense of the importance of horizontal transmission, we also report the standardized-β coefficient for loanwords relative to $AncestralSimilarity$. That is, we report the standardized-β_1 divided by the standardized-β_2. We interpret this as the importance of horizontal transmission relative to vertical transmission.

One consideration that warrants some discussion is measurement error. Any classical measurement error in observed loanwords would generate an underestimate of β_1, while error in the measures of vertical or environmental transmission would likely generate an overestimate. The latter is because any residual variation in - for example - ancestral similarity, is expected to positively co-vary with loanwords, as more similar groups are expected to interact more, and therefore feature more horizontal transmission. For the same reason, measurement error in loanwords is likely to generate an overestimate of β_2. The most likely scenario is that loanwords is measured with more error than ancestral distance since it is an estimate rather than a direct measure. If so, the estimate of the relative importance of horizontal transmission should be treated as a lower bound.

Table C2: Loanwords and societal similarity

Dependent Variable	Share of common cultural traits in the:				
	Ethnographic Atlas				Folklore
	(1)	(2)	(3)	(4)	(5)
Share of borrower's language borrowed from lender (%)	1.155***	1.128***	1.118***	1.092***	0.0897***
	(0.203)	(0.200)	(0.174)	(0.179)	(0.0142)
Similarity in family tree between borrower and lender	40.23***	40.21***	40.11***	39.94***	2.763***
	(0.571)	(0.570)	(0.569)	(0.568)	(0.0550)
Distance between lender and borrower centroids (1,000km)	-0.503***	-0.497***	-0.528***	-0.551***	-0.0357***
	(0.00924)	(0.00916)	(0.00872)	(0.00835)	(0.000732)
$\frac{Standardised - \beta^{loanwords}}{Standardised - \beta^{LanguageFamily}}$	20.8%	19.4%	19.4%	19.4%	17.0%
$\frac{Standardised - \beta^{loanwords}}{Standardised - \beta^{LanguageFamily}}$ (measurement-error corrected)	157.5%	158.4%	73.3%	64.7%	162.2%
Lexicon size		✓	✓	✓	✓
Distance to Capital		✓	✓	✓	✓
Language Family FE			✓	✓	✓
Country FE				✓	✓
Obs.	15,439,425	15,380,469	15,325,613	15,325,613	7,542,687
R sq.	0.064	0.072	0.211	0.252	0.699
Dependent Variable Mean	43.95	43.95	43.95	43.95	50.92

Note: The regression is run at the society-pair level. Standard errors are two-way clustered by the two societies in a society-pair. *, **, *** denotes 10%, 5%. 1% significance respectively. Language Ancestry is defined as the number of overlapping nodes in the society's language tree. Lexicon size accounts for the PanLex coverage of the borrowing group. Language Family FE are FE for the 4-level from root language family based on language trees for both the borrower and the lender. Country fixed effects are included for both the borrower and the lender.

i) Results for overall intensity of adoption: The estimates are presented in table C2. The estimates for the three possible determinants of group traits, without any additional controls are displayed in column 1. The loanwords estimate is positive, it is approximately equal to one, and is very precise, with a t-statistic of 5.6. Furthermore, the estimates for vertical transmission (t-statistic$>$ 70) and environmental similarity (proxied using geographic distance, t-statistic< -54) are also extremely precise, which mitigates the concerns that there is substantial measurement error in either variable.

To further capture remaining residual variation, we add controls and fixed effects in columns 2-4. Column 2 presents the same estimates, but with the inclusion of controls for the lexicon size and distance to the capital. The reason for the first is that a group with more lexicographic coverage has more opportunities that one of the words is a loanword; and for the second, as we will see later on, distance to the capital is important for colonial adoption.[93] The estimate, however, remains essentially unchanged with the inclusion of these controls. With the inclusion of the country fixed effects (column 3), the estimate of β_1 again remains essentially unchanged at 1.118. The same is true in column 4 which adds language family fixed effects.

In column 5 we see if the same relationship holds using the folklore data, and we again estimate a positive and statistically significant estimate. The magnitude of the estimate, however, is much smaller, which is attributable to the differences in the two data sources. In particular, the dependent variable variance in the folklore data is much lower. We feel comfortable attributing the differences in magnitude to this as the standardized-β from the Folklore sample is very similar to the standardized-β that is estimated using the Ethnographic Atlas.

The standardized-βs also reveal that a standard deviation change in horizontal transmission is associated with cross-societal similarity by about 20% as much as a standard deviation change in ancestral similarity. This is a relatively conservative estimate since, as we discussed above, we expect that (a) there is more measurement error in loanwords than linguistic similarity; and (b) more similar groups are likely to exhibit more horizontal transmission.

To confirm this, we also try the classic Wald 1940 measurement error correction. The idea is to instrument for each variable with a randomly determined threshold in that variable. If the measurement error is classical, the error above and below the random threshold is the same, so using the threshold as an instrument for the variable with error, in theory, generates a clean estimate.[94] As expected, accounting for the measurement error in loanwords and ancestral similarity generates a larger estimate of the relative importance of horizontal transmission in every column. At a minimum the importance of loanwords relative to ancestral similarity with the correction is about 3-times larger than the same ratio without the correction. In fact, with the measurement error correction, the estimates suggest that that vertical and horizontal transmission are about equally important, though these estimates vary quite a lot from column to column (Table C2).

ii) Results for topic-specific intensity We have shown above that greater linguistic adoption is associated with greater overall cultural similarity. We can now disaggregate the data by topic-level and identify whether greater language adoption related to a specific topic is associated with greater cultural similarity, holding borrowing of other topics constant. We present the results of this analysis in table C3, and show that greater borrowing

[93] While not a focus of this paper, this also accounts for the influence of other state languages, such as the imposition of French over other languages within the borders of France, e.g Breton, Basque, etc.

[94] The random threshold is exogenous since it is random; it is excludable if the measurement error is classical; it is relevant by construction; and it is difficult to see how the monotonicity assumption could be violated. The downside of this method is lower precision, but that is less of a concern in our case, given that we have sample sizes in the millions of observations.

Table C3: Validity of loanwords as a proxy for influence at the topic level

Dependent Variable	Share of common cultural traits in the Folklore Data					
	Construction (1)	Geography (2)	Conflict (3)	Science (4)	Institutions (5)	Psychology / Philosophy (6)
Share of borrower's language borrowed from lender (%)	0.0785*** (0.0124)	0.0649*** (0.0102)	0.0521*** (0.00810)	0.0370*** (0.00718)	0.0659*** (0.00995)	0.0157*** (0.00272)
Distance between lender and borrower centroids (1,000 km)	-0.0155*** (0.000516)	-0.0150*** (0.000503)	-0.0157*** (0.000528)	-0.0163*** (0.000626)	-0.0220*** (0.000779)	-0.0149*** (0.000485)
Similarity in family tree between borrower and lender	2.151*** (0.0912)	2.153*** (0.0963)	2.148*** (0.0947)	2.341*** (0.120)	2.699*** (0.137)	2.134*** (0.0962)
Other Borrowing	✓	✓	✓	✓	✓	✓
Lexicon size	✓	✓	✓	✓	✓	✓
Distance to Capital	✓	✓	✓	✓	✓	✓
Country FE	✓	✓	✓	✓	✓	✓
Language Family FE	✓	✓	✓	✓	✓	✓
Obs.	7,793,190	7,793,190	7,793,190	7,793,190	7,793,190	7,793,190
R sq.	0.198	0.197	0.196	0.194	0.184	0.195
Dependent Variable Mean	0.352	0.349	0.337	0.336	0.462	0.341

Note: The regression is run at the society-pair level. Standard errors are two-way clustered by the two societies in a society-pair. *, **, *** denotes 10%, 5%. 1% significance respectively. Lexicon size accounts for the PanLex coverage of the borrowing group. Language Family FE are FE for the 4-level from root language family based on language trees for both the borrower and the lender. Country fixed effects are included for both the borrower and the lender. Other borrowing is constructed based on the residualized share of language borrowed after differencing out the topics considered in the table. This is done to avoid cases where the words associated with the topic makes up a large share of overall borrowing.

in a given category is associated with greater similarity in culture, as measured in the Folklore data (Michalopoulos and Xue 2021).

iii) Language adoption and the diffusion of development Before analysing the relationship between trade incentives and linguistic convergence, we first conduct validation exercises. We show in tables C2 that more language borrowing is associated with greater cultural similarity between groups. We also show in table 7 that groups are more likely to have borrowed language from those with whom they had more contact, focusing in particular on nearby groups and distant groups who were colonisers. We next establish that linguistic borrowing is negatively correlated with barriers to the diffusion of development. This is important support for the argument in this paper that language borrowing responds to incentives for groups to reduce cross-group barriers. We do this by including our measure of language borrowing in the country-pair analysis in Spolaore and Wacziarg 2009. They regress the absolute difference in log income in 1995 on the genetic relatedness of the populations of two countries. We do the same, adding in loanwords adoption as well as the same controls that they do in their paper. The results are in table C4.

Table C4: Loanwords and the horizontal transmission of development

Dependent Variables	Absolute difference in log income (WB), 1995						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Weighted Fst Genetic Distance	1.878*** (0.541)	2.270*** (0.656)	2.189*** (0.686)	2.661*** (0.684)	2.697*** (0.680)	2.692*** (0.680)	0.424 (1.490)
Share of borrower's language borrowed from lender (%)		-0.126 (0.0923)	-0.180** (0.0902)	-0.181** (0.0892)	-0.180** (0.0887)	-0.180** (0.0886)	-0.420*** (0.146)
Lexico-statistical % cognate (indo-Euro only)							-0.624*** (0.207)
$\frac{Standardised-\beta^{loanwords}}{Standardised-\beta^{LanguageFamily}}$		10.7%	15.7%	13.3%	13.1%	13.2%	271.4%
$\frac{Standardised-\beta^{loanwords}}{Standardised-\beta^{LanguageFamily}}$ (measurement-error corrected)		6.9%	27.7%	49.1%	48.4%	47.6%	12.4%
Absolute difference in latitudes and longitude			✓	✓	✓	✓	✓
Geographic distance (most populated cities)				✓	✓	✓	✓
Either country is an island					✓	✓	✓
Freight rate (surface transport)						✓	✓
Obs.	13,203	7,875	7,875	7,750	7,750	7,750	1,431
R sq.	0.022	0.033	0.071	0.073	0.075	0.075	0.111
Dependent Variable Mean	1.273	1.327	1.327	1.320	1.320	1.320	1.056

Note: The unit of observation is a country-pair. Loanword intensity is aggregated from the society-pair level using the mean borrowing by the societies in one country from the societies in the other. Standard errors are two-way clustered by each country. ***, **, * represents significance at the 1%, 5%, 10% levels respectively. Each variable other than loanwords comes from the replication materials in Spolaore and Wacziarg 2009, accessed from: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/TNCU8K on April 1, 2022.

[95] We find that loanwords are a robust, consistent predictor of development diffusion, just as genetic distance is. This holds even when we control for cognate share (i.e. the share of words in a language that come from a common parent language). While this cognate data is only available for Indo-European languages, our results are quite robust to the inclusion of cognates as a control, further reinforcing the idea that our loanwords measure is able to accurately distinguish between actual loanwords and cognates.

We interpret the negative and significant loanwords estimates as evidence that loanwords are associated with the diffusion of technology, institutions, or aspects of culture that may matter for economic development. There are (at least) two possible reasons for this association. The first is that loanwords directly proxy for the diffusion of the specific items that are important for spurring development. Alternatively, it could be that groups make investments in cross-societal relationships more generally. In this case loanwords may not be specifically capturing the diffusion of the specific items that are important for development, but may instead simply capture a strong relationship between countries. Transmission may stem from strong relationships, regardless of whether the words for the most important features for development are adopted with greater or lesser frequency than other types of words. We once again report the relative importance of horizontal transmission to genetic relatedness using the ratio of standardized-β

[95] A version of the table with country fixed effects - which are not included in the original paper - appears in table C5. However, the inclusion of these fixed effects does not meaningfully impact the results for either genetic relatedness or loanwords.

Table C5: Horizontal transmission of development with Country Fixed Effects

Dependent Variables	Absolute difference in log income (WB), 1995						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Weighted Fst Genetic Distance	2.919*** (0.635)	4.016*** (0.827)	3.545*** (0.823)	3.489*** (0.841)	3.492*** (0.841)	3.483*** (0.838)	17.70*** (4.025)
Share of borrower's language borrowed from lender (%)		-0.221* (0.127)	-0.208* (0.120)	-0.174* (0.101)	-0.173* (0.101)	-0.171* (0.100)	-0.195 (0.133)
Lexico-statistical % cognate (indo-Euro only)							-0.407** (0.197)
Country Fixed Effects (both)	✓	✓	✓	✓	✓	✓	✓
Absolute difference in latitudes and longitude			✓	✓	✓	✓	✓
Geographic distance (most populated cities)				✓	✓	✓	✓
Either country is an island					✓	✓	✓
Freight rate (surface transport)						✓	✓
Obs.	13,201	7,873	7,873	7,748	7,748	7,748	1,429
R sq.	0.308	0.320	0.332	0.336	0.336	0.338	0.532
Dependent Variable Mean	1.273	1.327	1.327	1.320	1.320	1.320	1.057

Note: The unit of observation is a country-pair. Loanword intensity is aggregated from the society-pair level using the mean borrowing of societies by the societies in one country from the societies in the other. Standard errors are two-way clustered by each country. ***, **, * represents significance at the 1%, 5%, 10% levels respectively. Each variable other than loanwords comes from the replication materials in Spolaore and Wacziarg (2009), accessed from: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/TNCU8K on April 1, 2022.

coefficients. We do this both for the raw estimates, as well as for the estimates with the Wald-correction for measurement error. Based on the Wald-corrected estimates, loanwords explain as much as 50% of the variation of genetic distance. So it remains quite important, but not as important as two heavily related countries.

Appendix D. Construction of Topic-Level Outcomes

For each of the topics we consider (*technology*, *geography*, *conflict*, *science*, *politics*, and *psychology/philosophy*), we capture representative words using a seed word approach. The idea is to use a set of words that represent a concept, and then to algorithmically find all of the words that are either directly synonymous, or heavily related. From this expanded set of words we can then construct the share that were borrowed.

The original seed words were generated based on the Library of Congress (LoC) categorization of various topics. The LoC classifies all of their holdings into one of 21 different classifications, some examples are: politics; psychology, philosophy and religion; science; warfare. Each classification includes a large number of sub-classifications, each with a description of the sub-class. We used web-scraping techniques to download and organize the entire classification schema.

Using this text-data, we generated seed-words by topic, in a manner that was as hands-off as possible. The seed-words that we relied upon were the 15 words (or terms) that occurred with the highest excess-frequency within a classification category. To be more

Table D1: Seed Words Used to Generate Concept Level Measures

Technology	Geography	Conflict	Science	Politics	Psychology / Philosophy
engineering	games	military	astronomy	political	church
construction	physical	artillery	chemistry	international	religious
machinery	human	cavalry	physics	state	churches
electric	relative	drill	anatomy	government	bible
industrial	sports	regulations	geographical	municipal	christian
building	customs	tactics	mechanics	administration	buddhism
applied	environmental	troops	plant	institutions	testament
environmental	athletic	arms	distribution	executive	ethics
water	hemispheres	defense	magnetism	legislative	literature
mining	ice	warfare	biochemistry	public	islam
power	recreation	services	particle	relations	languages
industry	water	air	radiation	emigration	christ
energy	amusements	field	radioactivity	states	religions
mechanical	cultural	armor	theoretical	nations	worship
mechanics	folk	armored	animal	islamic	catholic

Note: The table shows the seed words used to generate concept-level variation in both borrowing and societal similarity (using the folklore data).
Note on Exclusions: We excluded words that appear in archival contexts that would be misleading to include among the keywords, or that indicate specific geographic contexts or specific historical periods. The list of excluded words is as follows: *general, journal, journals, works, printed, periodical, periodicals, proceedings, branch, special, life, biography, annals, chapter, chapters, including, covering, new, small, description, map, maps, atlas, atlases, century, thirteenth, fourteenth, fifteenth, sixteenth, seventeenth, eighteenth, nineteenth, zone, regional, region, country, countries, general, africa, europe, india, asia, pacific, atlantic, islands, ocean, oceans, coast, americas, america, american, british, french, african, european, asian, antarctica, soviet*

specific, we excluded standard English stopwords,[96] as well as common archival terms that do not represent the topics of interest.[97] We then computed term frequency–inverse document frequency (tf-idf) scores. We define td-idf as the product $\text{tf-idf}(t,d) = \text{tf}(t,d) \cdot \text{idf}(t)$, where $\text{tf}(t,d)$ is the frequency of term t in document d (i.e. $\text{tf}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$), and $\text{idf}(t)$ is the inverse document-frequency of term t (i.e. $\text{idf}(t) = \log \frac{1+n}{1+df(t)} + 1$). Here $f_{t,d}$ is the count of term t in document d, n is the count of documents, and $df(t)$ is the share of documents including term t.[98] The full list of topics, and the respective seed-words generated using this approach is in table D1.

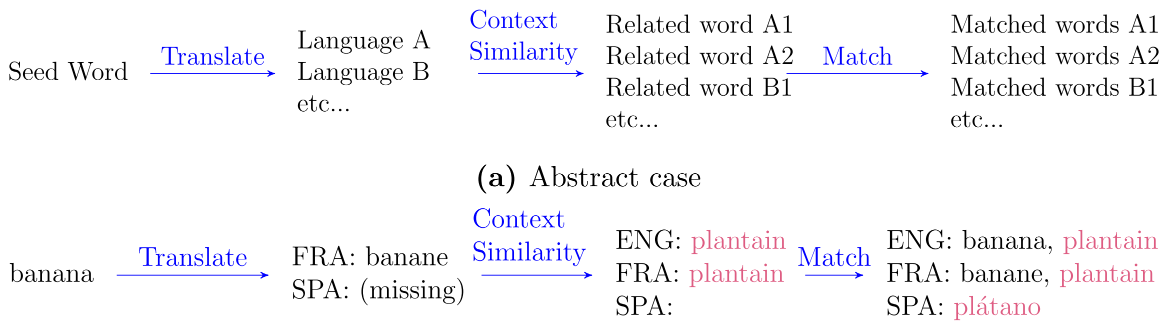
One drawback of this approach is that the LoC categorizations are pretty western-centric. To address this issue we implemented a semantic analysis routine (based on Bojanowski et al. (2017)), after generating the seed-words from the LoC. This routine first finds direct translations of the seed-words in as many languages as possible from the

[96] A stopword is a term from the field of Natural Language Processing to refer to words like 'the' or 'of' that are typically ignored by text analysis algorithms.

[97] These words are specific to the language used in the LoC classification schema, for example, words like: 'annals,' 'proceedings,' 'collection,' etc.

[98] For an intuitive explanation, the words 'of' or 'and' might be the most common words overall in each group, but the word 'political' is among most frequently used words within the Politics category, relative to its frequency in other categories.

PanLex meaning IDs. The intuition is to then take these words and their translations, and search through models trained on Wikipedia to find words that are used in semantically similar ways. This results in a much larger list of related concepts in nearly three hundred languages. We take these 'similar expressions' 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 our various topics, in a way that goes beyond English conceptual associations. An example is illustrated in figure D1.



(a) Abstract case

(b) Concrete example

Figure D1: Semantic analysis illustration: capturing semantically similar words

There are a few important advantages to doing this aside from the main goal of obtaining seed-words that are less heavily western-biased. The first one is that we think 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 horizontal transmission, and the validation exercises are all in this spirit. We have purposely avoided getting into the process of defending the loanword status of specific word pairs - which is the focus of linguists[99] - and perhaps not of direct interest to most economists. Our philosophy is to acknowledge that any automated approach will come with some error, and we should accordingly manage that error to the best of our ability. One way of doing this is by exploring averages of larger sub-samples, whenever possible. Second, expanding the set of seed-words 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.

With the concepts of interest matched to a large set of relevant words, we can reconstruct the topic-specific share of language adopted in exactly the same way as for the overall linguistic adoption. Only in this case the result is the share of words adopted by each group from each other group, within a specific topic.

[99] 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.

Appendix E. Estimated Trade Incentives Appendix

E.1. Calibration of Nutritional Utility Model

For the sixteen essential nutrients we consider, we are able to use the Daily Reference Intake amounts from the NIH as a source for the weights they are assigned in the nutritional utility function. The weight assigned to calories, however, is not as easily drawn from reference sources, as this may reflect differences in living standards and using a contemporary caloric intake may be unreasonable. The weight assigned to calories therefore reflects the tradeoff between meeting the ideal nutrient requirements and simply meeting the minimal caloric requirement for an adult to survive.

In order to calibrate this and estimate the weight we assign to calories, we estimate the trade and production model (using the DRI weights for the essential nutrients) at various weights for calories and evaluate which produces the most reasonable estimates of population levels. Here, a higher weight to calories therefore implies more focus on providing minimal caloric intake, and less focus on meeting the ideal nutritional profile. To evaluate whether population figures are reasonable, we estimate the upper- and lower- bound populations that could possible be supported under a given nutritional consumption bundle. We use 3000 kcal as the absolute maximum calories required for an adult, and 1,500 as the minimum necessary for an adult to survive. We therefore solve the optimization algorithm at a number of potential weights for calories, and choose the weight that leads to the largest share of these lower-bound, upper-bound ranges of model-estimated population density including the observed population density. We find that a target of 2,700 kcal relative to an individual's DRI is the optimal weight. Given that the NHS suggests a healthy caloric intake of 2,000 for women and 2,500 for men, this suggests that our calibration exercise places a higher weight on calories than a contemporary Western diet, consistent with a priority to sustain a larger population while somewhat sacrificing the ideal nutritional profile (National Health Service 2019).

E.2. Solving for prices

Since the production function is linear, prices must be such that all crops a group chooses to grow yield at least as much income per unit of land as any crop that the society chooses not to grow.[100] This yields the following constraints on prices:

$$\begin{aligned} p_c \cdot \Omega_{ic} \geq & p_{c'} \cdot \Omega_{ic'} \\ & \forall i \\ & \forall c \in \{\text{Crops Grown by } i\} \\ & \forall c' \in \{\text{Crops Not Grown by } i\} \end{aligned} \tag{15}$$

[100] This is necessary to support the production decisions in the social planner's solution.

where p_c is the price of a crop and Ω_{ic} is an element of the matrix Ω, defined above. This also means that the income per unit of land for all crops a group chooses to grow must be equal:

$$p_c \cdot \Omega_{ic} = p_{c''} \cdot \Omega_{ic''} \qquad \forall i \qquad \forall c, c'' \in \{\text{Crops Grown by } i\} \tag{16}$$

Taken together, equations 15 and 16 define the supply side constraints on prices.

The way that the utility function is specified also generates a set of demand-side constraints on prices. Given that Cobb-Douglas utility features identical, homothetic preferences, the ratio of consumption of different goods will be the same at all income levels (Feenstra 2004).[101] Therefore, each society will consume in the same proportions as the aggregated regional land shares, since they all face the same prices. Accordingly, the following conditions must hold for the optimal I by C consumption matrix, κ^*:

$$\left(p_c \middle/ \frac{\partial U(\kappa^*)}{\partial c} \right) - \left(p_{c'''} \middle/ \frac{\partial U(\kappa^*)}{\partial c'''} \right) = 0 \qquad \forall c, c''' \in \{\text{Crops grown in Region}\} \tag{17}$$

We then numerically solve this system by finding the price vector $\vec{p} = [p_0, p_1, ..., p_C]$ that minimizes Equation 17 subject to Equations 15 and 16. Between Λ - which was solved for in equation 5 - and this price vector, we can get income (and therefore income shares) for each society.

Agricultural income is of some independent interest so we control for the model estimated agricultural incomes of each society throughout the analysis. However, more importantly, income shares determine the level of consumption of each society, which in turn determines the crop-specific consumption. This, once again, makes use of the fact that homothetic preferences means that all societies consume in equal proportions. With crop consumption in hand, equation 3 generates each society's utility under trade, which was the goal in the first place.

E.3. Accounting for indirect effects of trade partners.

Our goal here is to compute the trade benefit from a trade partner in a minimal neighbourhood without any general equilibrium (GE) effects that do not potentially result from direct interaction between group i and group j.

[101] Since nutrients are an identical linear function of crops, it is straightforward to show that the nutritional utility function can also be written as a homothetic function of crops themselves. This simplifies matters because it means that we do not have to distinguish between consumption of crops and consumption of nutrients.

The concern is that some groups may have indirect effects on each other when they enter the trading neighbourhood. So, when group j enters the neighbourhood, they might change the prices for goods that they don't produce. This might hurt group i where group j's presence causes the price of i's imports to go up or the price of i's exports decreases. However, this impact doesn't directly lead to interaction between i and j.

So, to compute the direct component of i's change in utility from j's presence, we shut down the channel of prices for crops not produced by exactly one of i or j in the full-trade equilibrium. This means i will only be impacted by change in price for crops grown by i and consumed by j in equilibrium or vice versa. This will isolate changes that come from direct interaction, i.e importing or exporting between i & j as much as is feasible in this model. Given the setup of the model, if a crop is grown by j and by some other group k in equilibrium, the price will not be fixed, as we cannot specifically identify whether i is purchasing from j or k.

We therefore start from the full-trade equilibrium, then identify the crops to have prices shut down. These are crops not produced by i or j. We identify the prices for these crops in the full-trade equilibrium, and set these as fixed in the optimization routine. We then run the optimization routine again with group j dropped from the trading neighbourhood. Prices for crops that had been produced by j are set to be equal to those from the full-trade equilibrium.

Appendix F. Validating using historical trade data

F.1. *Validity of the gains from trade measure:*

The basic strategy of using observed relative land productivity to model agricultural trade is valid at the country level (Costinot and Donaldson 2012). However, since in our case we both model subnational trade, and impose more structure on the demand side, we would ideally like to validate our measure of gains from trade against actual local trade flows in primary agricultural products among subnational language groups. These data, however, do not exist to our knowledge for a large set of language groups. We take a number of alternative approaches to validating our measure.

The main strategy is to test whether the model predicts actual production of regionally traded crops, controlling for the FAO productivity of all crops. The crop production data come from Monfreda, Ramankutty, and Foley 2008, who report the share of land allocated to each crop within a 5 arc-minute cell for the whole world.[102]

For each crop, we regress the actual production (denoted $\tilde{Y}$) on Y and Ω. This results in a number of regressions at the society-level, producing a vector of estimates β that

[102] We focus on the society mean of that variable for each crop with more than 0.0001% mean land share for both actual and estimated mean land allocation, that exist in both the FAO and the Monfreda, Ramankutty, and Foley 2008 data-set.

denotes, for each crop, how well the model predicts actual production. This is represented by the following system of regressions:

$$\tilde{Y} = \beta Y + \Gamma\Omega + \epsilon \tag{18}$$

There are some crops that are typically globally traded, like tobacco, wheat and maize, where contemporary production is clearly not related to the local trade dynamics captured by the model. Accordingly, we define any crop as being predominantly determined by global trade dynamics (and therefore not relevant for our regional trade model) if it had more than five billion USD in global trade in 2008 according to the FAO.

This delineation was based on a a natural gap in export dollars that exists in the data (figure F1), and results in 14 crops in the global trade group and 8 crops in the group we expect to be relevant for our model.

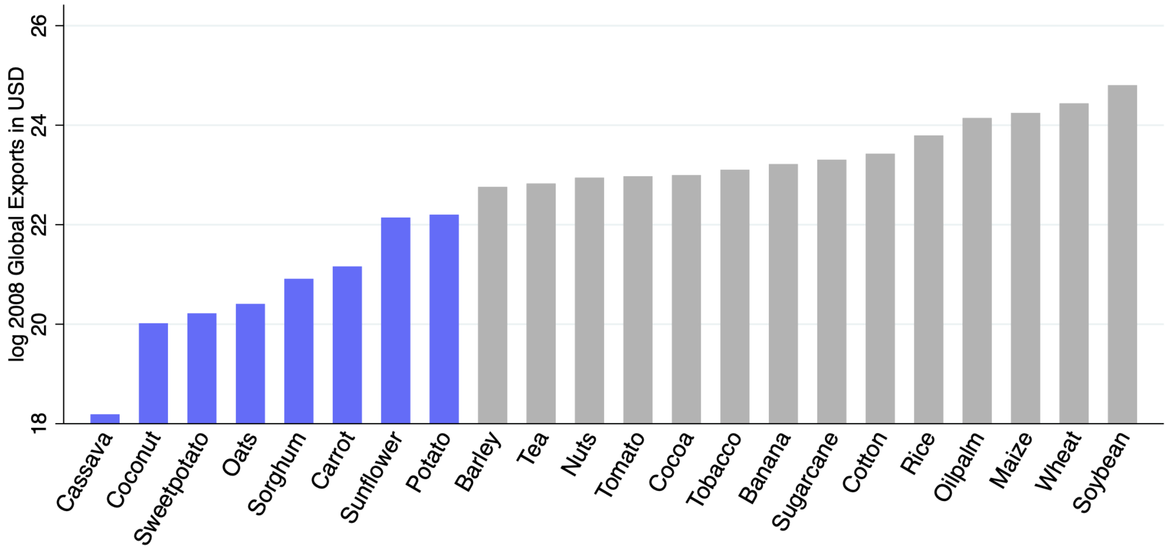


Figure F1: Delineation of global crops used for trade model validation

Note: The figure shows the log exports of 2008 trade as per the FAO. We use these trade numbers to determine which crops are globally traded and therefore not relevant to our regional trade model. All crops coloured grey are determined to be global crops, and we chose this cutoff based on the small break in the data on either side of the cutoff.
Source: Author constructed based on data from FAO accessed August 12, 2020. http://www.fao.org/faostat/en/#data/TP

The bulk of our focus is on the crops that we expect to be relevant, and we report crop-by-crop estimates (i.e. the elements of β) for each. We report an aggregate for the crops that are predominantly traded globally, which serves as a placebo estimate.

Table F1 reports estimates for crops that are relevant for our trade model in columns 1-8. Each one has (precisely estimated) predicted production that is positively associated with actual production. For the globally traded crops the model is not predictive, as anticipated (column 9).

Our second approach is to look at historical market prices for crops across a number of cities, sourced from David S Jacks 2004; David S. Jacks 2005. We matched these cities to

Table F1: Validating the trade measure against actual crop production

	Dependent variable: Actual Land Allocation								
	Sweet potato (1)	Carrot (2)	Sunflower (3)	Sorghum (4)	Coconut (5)	Cassava (6)	Oats (7)	Potato (8)	Global (9)
Model allocation	0.0874***	0.0959***	0.0831***	0.0887**	0.170***	0.233***	0.540***	0.524***	-0.0117
	(0.0238)	(0.0305)	(0.0299)	(0.0407)	(0.0331)	(0.0560)	(0.163)	(0.126)	(0.0993)
Crop suitability	✓	✓	✓	✓	✓	✓	✓	✓	✓
N	2,606	2,606	2,606	2,606	2,606	2,606	2,606	2,606	2,606
R^2	0.248	0.203	0.366	0.381	0.377	0.347	0.378	0.359	0.190
Dep. Var. Mean	0.002	0.0001	0.001	0.007	0.006	0.006	.0002	0.001	0.054

Note: The unit of observation is a society. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Model allocation is the predicted production of a particular crop, based on the trade model described in the text. Crop suitability includes the potential production per hectare of all FAO crops. Global crops are those with more than $USD5 billion in yearly trade. Actual production is the share of land allocated to each crop within a 5 arc-minute cell.

our language groups to test whether our model-predicted gains from trade between a pair of neighbouring groups is correlated with market price integration.[103] This gives us only sixty pairs that we can match to our data, but on this small sample we show in table F3 that model-predicted gains from trade are associated with greater price integration. This suggests greater trade volume among these pairs.

F.2. Validating using population

We next conduct a third validation exercise based on population that further reinforces the validity of our trade model. While we are not able to directly validate our main model-based measures since units are in utils, we can run essentially the same model to estimate the maximum population that the model believes the society can support under some assumptions of caloric intake per person, to see how much of the variation in actual population this explains. We do this with the caveat that since we use the potential production without negative shocks or inefficiencies so the estimated population will be larger than the actual populations. If *systematic* variation in these inefficiencies is relatively small compared to the importance of trade and the land itself for productivity, we might still find a significant correlation between the model-estimates of population and the actual population numbers.

We investigate this relationship in table F2. In columns 1 and 2 we examine our model-estimated supportable population under autarky and trade respectively. As expected, the maximum population under fully efficient production is much larger than the actual population, so an additional estimated supportable person is only associated with a 0.03-0.04 additional actual persons, although both estimates are extremely precisely estimated. These univariate regressions explain 14% and 7% of the variation in actual population for autarky and trade respectively. Although there may be unobserved variation in the

[103] See appendix F for details on this data and the matching procedure.

Table F2: Structural Model Diagnostics

	Dep. Var: Actual Population		
Model Generated Maximum Supportable Population under:	(1)	(2)	(3)
Autarky 1200kcal	0.0384***		-0.600***
	(0.00111)		(0.0178)
Trade 1200kcal		0.0287***	-4.107***
		(0.00159)	(0.0551)
$\sqrt{Autarky \cdot Trade}$			4.270***
			(0.0680)
R^2	0.111	0.033	0.794
N	9,564	9,564	9,564

Note: This table demonstrates that our trade model is producing data that is highly correlated with actual hand collected data, but also that it is able to explain a surprisingly high degree of variation in that data. We see extremely precise, though small estimates in columns one and two. The low estimate is due to the fact that our model assumes societies produce at 100% efficiency on all dimensions, so the output is the maximum sustainable population - and the actual population should be a relatively small fraction of that. The third column accounts for the fact that in columns one and two, autarky production explains population much better than trade. When we flexibly allow autarky and trade to predict population we are able to explain over 60% of the variation in population even though all of the trade data is from the FAO GAEZ (who do not collect population) and all of the population data is from the Ethnologue (who do not collect nutrition). We conclude from column three that our trade model is capturing the variation that we would like reasonably well.

degree of isolation, and autarky may explain some populations better, while trade may explain others - the population under trade and autarky are highly correlated so they may both be picking up the same actual population variation. When we include both variables, in a fully-saturated model, the two variables jump to explaining over 60% of the variation in actual population (column 3). We therefore feel relatively confident that our trade model is producing fairly reliable estimates.

F.3. Validating using commodity prices

The data sources for wheat prices is (David S Jacks 2004; David S. Jacks 2005). The data files include commodity prices for 515 cities across 33 countries as defined by the borders of 19th century for the total time period of 1700-1940.

We extracted Jacks' price data for wheat and rice from each city for the relevant time period and transferred them to Stata. When annual data were available for the time periods that fall within the set brackets, they were used. Otherwise, we averaged weekly, bi-weekly or monthly data for each year. Units of measurement were also harmonized. Archaic units of measurement were converted into kilograms using either Jacks' conversion rates where possible or looked up on the internet in historical sources to ensure the conversions were valid for the time period and city in question, as units of weight measurement had not been standardized yet.

Jacks' price data includes different currencies depending on the original source of the

Table F3: Correlates of gains from trade

Dependent variable:	Regional price variance (1)
Gains from trade with neighbours	-0.381** (0.169)
Observations	64
R-squared	0.999
Dependent Variable Mean	2.627

Note: *** p<0.01, ** p<0.05, * p<0.1. The unit of observation is a society pair, we do not aggregate to the society level due to very few observations. Standard errors in parentheses are two-way clustered by each society. Gains from trade with neighbours is the percentile rank in the distribution (range [0,1]) of c_i as defined in equation 7.

data. The exchange rates specified in the data were used to convert to more common currencies (mostly European) whenever available. After that, historical exchange rates for 19th century found on a projects website of University of Exeter were used to convert currencies as well as gold and silver coins to US dollars. (Davies 2019).

The price data were merged by country and city with a data-set of world cities coordinates (latitude and longitude) obtained from the University of Toronto library. We manually added coordinates for cities that were not included in the world cities data-set or did not merge properly because of changes to country or city names or borders. The country names were preserved as in Jacks' data files. For instance, city Riga is listed under country Russia, although it is presently in Latvia. Lastly, all price data were converted to 2018 dollars using historical inflation rates from U.S. Official Inflation Data by taking a product of cumulative inflation rates from the year of interest to contemporary prices (U.S. Official Inflation Data, Alioth Finance 2019).

To match to our societal data, we do a fuzzy geographic match, taking a one decimal-degree radius around the city in the Jacks data as well as the group centroid in the Ethnologue data. We match prices to a society if the radii intersect. For each group in a pair we match the prices for each, and take our integration measure to be the absolute value of the difference in prices. We interpret price similarity as a sign of economic integration and trade, so that a lower value on our measure implies higher trade.

Results using this data are presented in table F3. Consistent with our other evidence, we find that our gains from trade data are negatively correlated with price variation, implying more economic integration in regions with more gains from trade. While the sample is small, this evidence reinforces our main approach to validating the gains from trade measure, as well as our next exercise which validates the same measure by comparing the predicted populations implied by the model and the actual populations.

Appendix G. Supporting Evidence, Tables, and Robustness Exercises

G.1. Figures

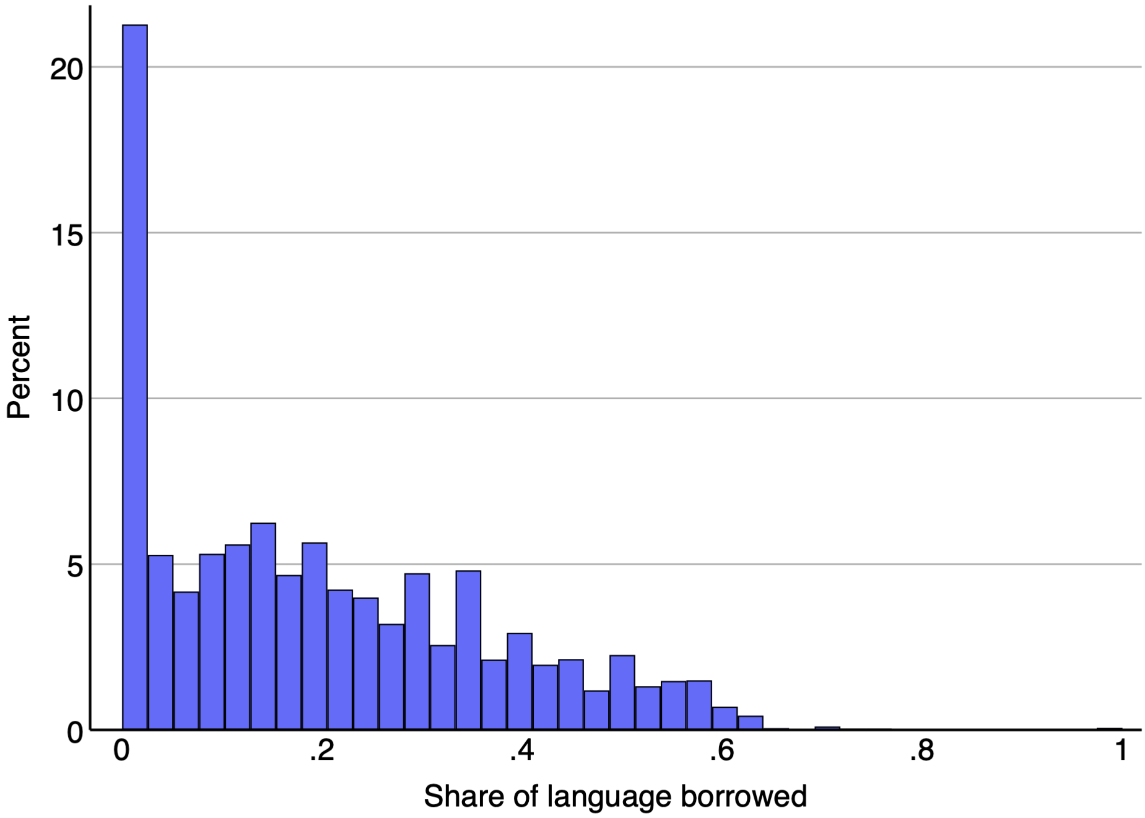
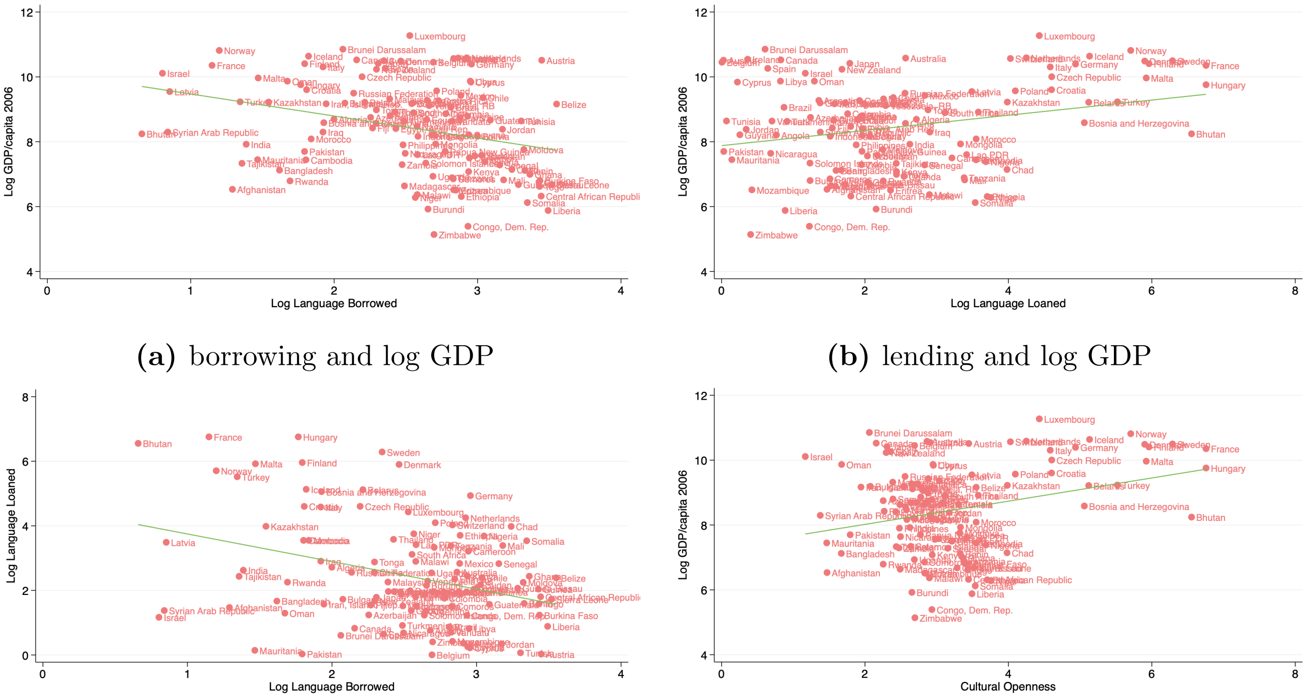


Figure G1: Histogram of Language Borrowing

Note: The figure shows the raw-data of the loanword share, which is an aggregated society-level version of $[\frac{100 \cdot \#\text{ loanwords}}{\#\text{ words in lexicon}}]_{ij}$, the main dependent variable used throughout the paper (equations 1, 12, 13, 11). Notably, while about 20% of societies do not borrow at all, a non-trivial share of societies borrowed between 20% and 60% of their language. This justifies a focus on loanwords, and illustrates that it is a non-trivial source of variation in linguistic distance.
Source: Author constructed. Data sources are described in the text.



(a) borrowing and log GDP

(b) lending and log GDP

(c) lending and borrowing

(d) Cultural openness and log GDP

Figure G2: Country-level correlates of lending and borrowing

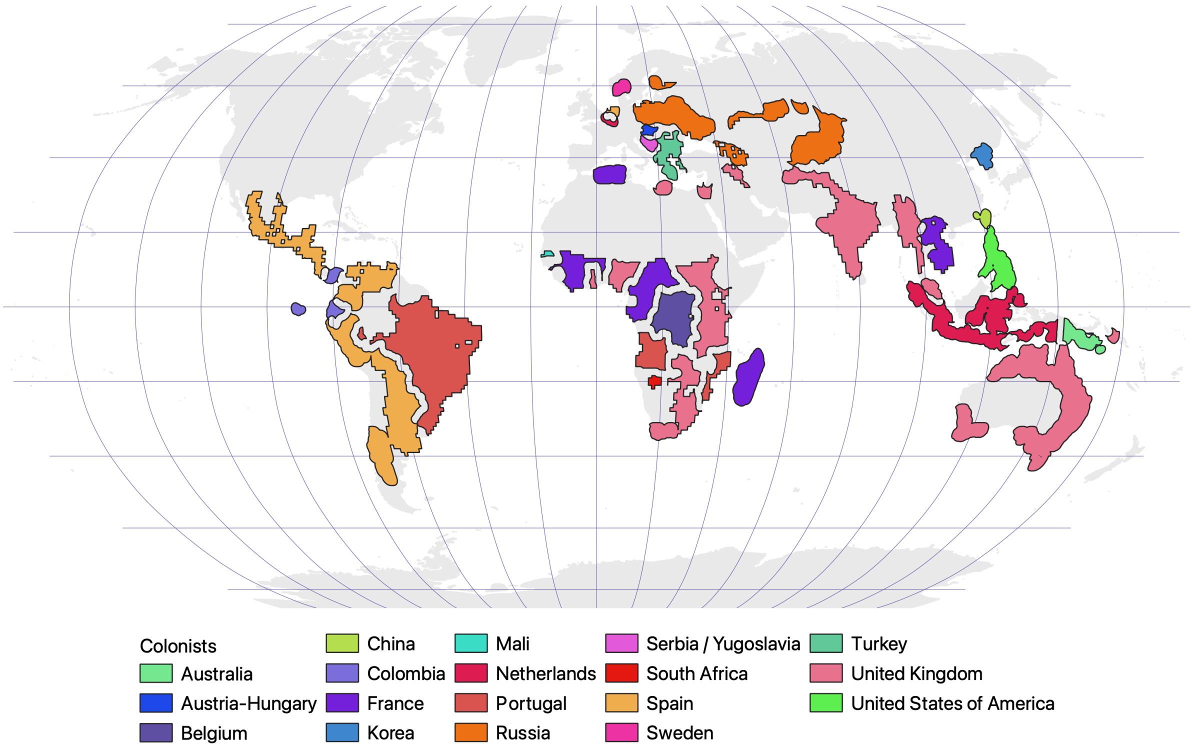


Figure G3: Colonialism map used to generate the centroids of contiguous colonial holdings

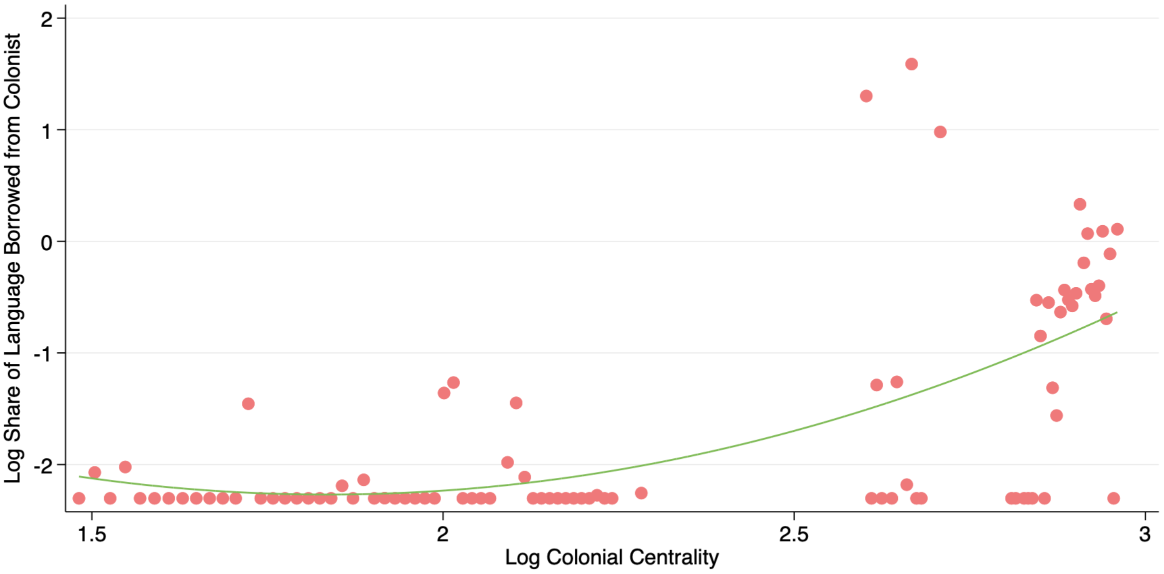
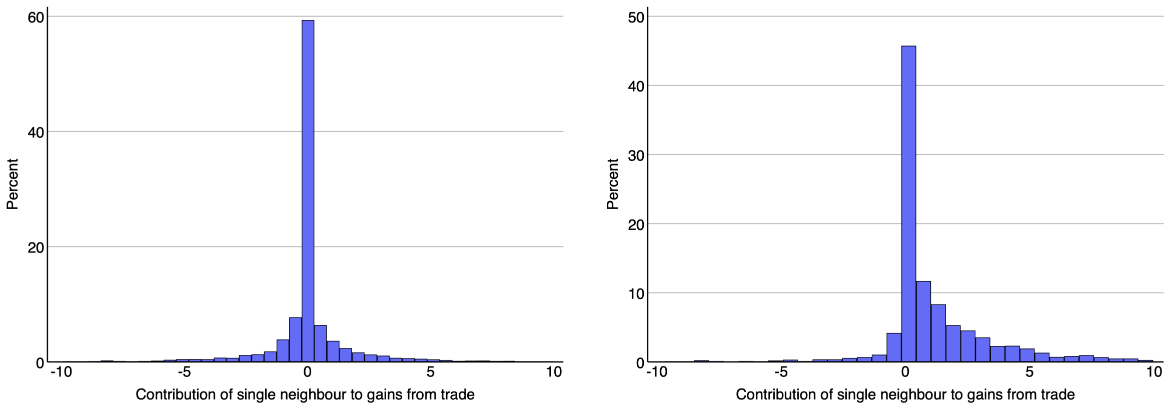


Figure G4: Relationship between colonial intensity and colonial borrowing

Note: The figure shows the correlation between language borrowing from the colonist and proximity to the centroid of the contiguous colonial holdings. The scatterplot groups observations into equidistant x-axis bins of size 0.1, and reports the mean borrowing within each bin. The fit line is quadratic and the graph excludes outliers with values less than 1, who are extremely far from the centroid.



(a) Relationship-level

(b) Society-level

Figure G5: Histogram of Gains From Trade

Note: The figure shows histograms of the output of the trade model. in panel (a) we have the bilateral measure of gains from trade (equation 6) and in panel (b) we have the societal level measure (equation 7). Source: Author constructed. Data sources are described in the text.

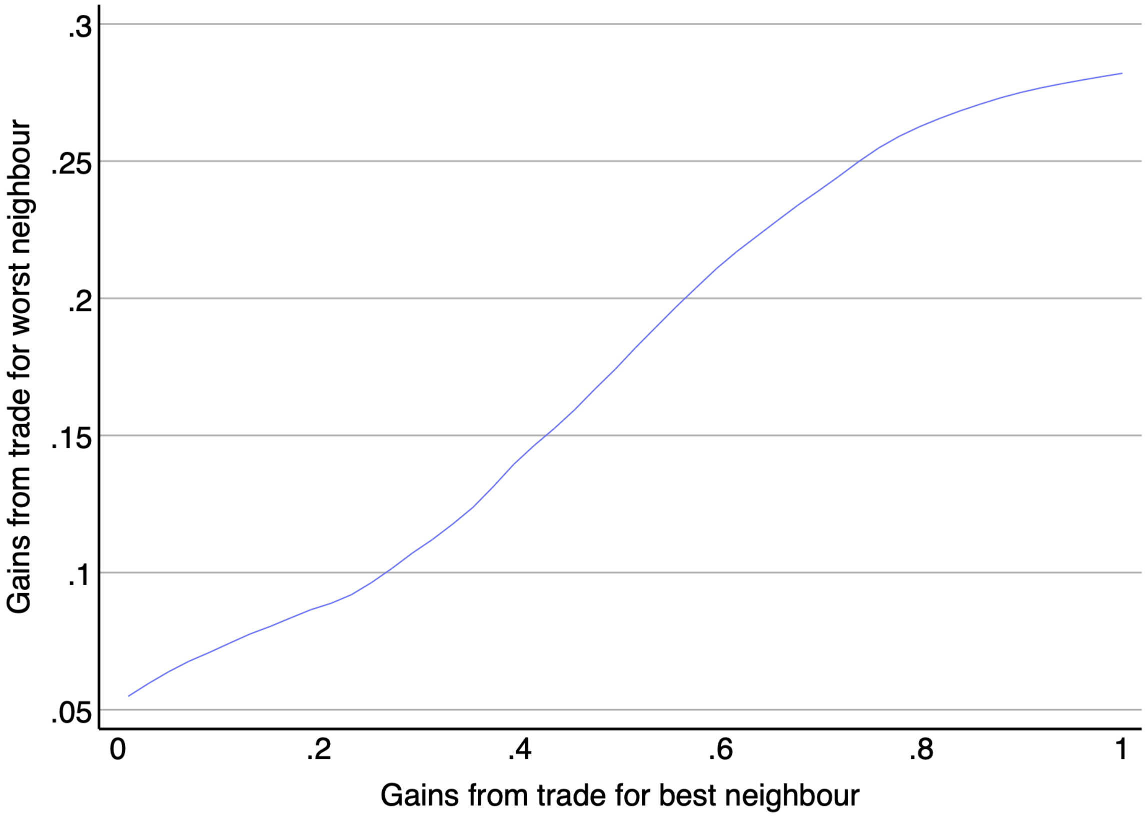


Figure G6: Correlation of best and worst neighbours

Note: This figure shows that good neighbourhoods are typically good on multiple dimensions. So, the best neighbour being good implies the worst is likely good too. Given this, it may be surprising that trade with a best neighbour leads to exchange but trade with worst does not, despite that this is a clear prediction of our hypothesis. The fit line in each graph is based on a biweight kernal of degree 1, with a bandwidth of 0.025.

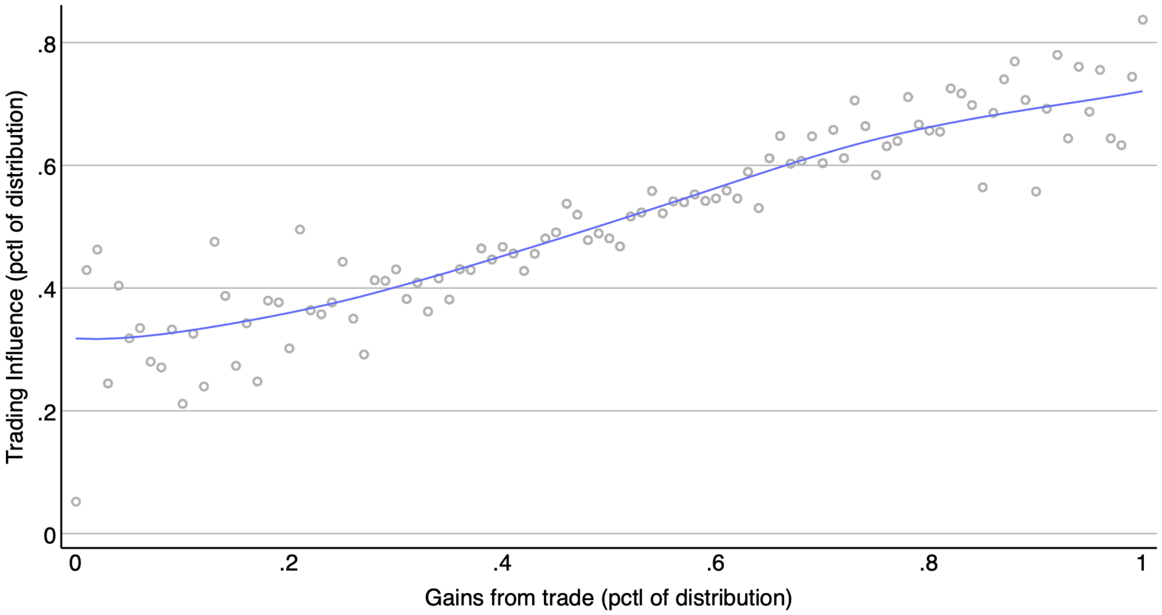


Figure G7: Gains from trade and trade influence

Note: The figure shows the correlation between gains from trade (c_i, as defined in equation 7) and trade influence (ι_i as defined in equation 9). The graph uses a society as the unit of observation, and takes the mean across neighbours for both gains and influence. The scatterplot groups observations into 0.01 gains from trade bins. The fit line in each graph is based on a biweight kernal of degree 1, with a bandwidth of 0.025.

G.2. Tables

Table G1: Summary of Observations in Analysis

Panel A: Main Language Data

	Observations		Intersection with restriction above	
	Societies	Pairs	Societies	Pairs
	(1)	(2)	(3)	(4)
Ethnologue	9,239	85,349,882	9,239	85,349,882
PanLex	4,258	18,126,306	4,206	17,686,230
Language Family Data	4,197	17,610,612	4,197	17,610,612
Matched Ethnographic Atlas	3,982	15,852,342	3,920	15,383,684

Panel B: Neighbourhood Economic Incentives Data

	Observations		Intersection with restriction above	
	Societies	Pairs	Societies	Pairs
	(1)	(2)	(3)	(4)
Language Data	3,920	16921	3,920	16921
Removed No Neighbours	.	.	3,049	12,096
Removed no-arable and colonial neighbours	.	.	2,646	9,596
Model Failed to Solve	.	.	2,606	9,436

Note: The table shows the number of observations in each data source used in the analysis and the overlap with the other data sources. For each we report the number of societies and the number of society-pairs that we observe.

Table G2: Intensity of Coverage in the PanLex Data by Country

	Number of Words in PanLex (100,000)												
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
Log Language Borrowed	-342.3												
	(371.0)												
Dummy: East Asia and Pacific		906.0											
		(3,710)											
Dummy: Eastern Europe and Central Asia		1,653											
		(3,696)											
Dummy: Middle-East and North Africa		-655.3											
		(3,722)											
Dummy: central Asia		2,182											
		(3,861)											
Dummy: Western Europe		3,032											
		(3,716)											
Dummy: Sub-Saharan Africa		-499.4											
		(3,658)											
Dummy: Latin America and Caribbean		-329.9											
		(3,705)											
population density (persons/km2)			1.510										
			(1.933)										
Real GDP p.c. (2006)				0.0412*									
				(0.0226)									
luminosity per capita					4,428								
					(3,495)								
Colonial origin indicator: Spanish						148.6							
						(1,417)							
Colonial origin indicator: British						686.0							
						(1,293)							
Colonial origin indicator: French						-462.0							
						(1,386)							
Colonial origin indicator: Portuguese						184.9							
						(1,978)							
Ruggedness (Terrain Ruggedness Index, 100 m.)							-203.7						
							(281.0)						
% Fertile soil								9.641					
								(13.18)					
Average distance to nearest ice-free coast (1000 km.)									552.6				
									(974.7)				
% Within 100 km. of ice-free coast									-5.561				
									(11.68)				
Ethnic Fragmentation Index										-2,090			
										(1,341)			
altitude above min altitude (m)											-0.327		
											(0.597)		
avg. annual precipitation (mm)												-0.494	
												(0.382)	
suitability for agriculture													2,010
													(1,289)
Obs.	137	137	137	137	137	88	137	137	137	125	137	137	135
R sq.	0.006	0.121	0.005	0.024	0.012	0.022	0.004	0.004	0.012	0.019	0.002	0.012	0.018
Dependent Variable Mean	2180	2180	2180	2180	2180	1426	2180	2180	2180	2341	2180	2180	2205

Note: The table regresses the intensity (measured by number of words included) of coverage in the PanLex data against country-characteristics to get a sense of any geographic bias in PanLex inclusion. An observation is a country, and the coverage in a country is the sum of the words in each language group covered withn each country. *, **, *** denotes 10%, 5%. 1% significance respectively.

Table G3: Intensity of Coverage in the WoLD Data by Country

	Inclusion in WoLD												
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
Log Language Borrowed	0.0148												
	(0.0496)												
Dummy: East Asia / Pacific		-0.278											
		(0.505)											
Dummy: Eastern Europe / Central Asia		-0.571											
		(0.503)											
Dummy: Middle-East / North Africa		-0.688											
		(0.506)											
Dummy: central Asia		-0.571											
		(0.525)											
Dummy: Western Europe		-0.706											
		(0.506)											
Dummy: Sub-Saharan Africa		-0.605											
		(0.498)											
Dummy: Latin America / Caribbean		-0.421											
		(0.504)											
Population density (persons/km2)			-0.00006										
			(0.000258)										
Real GDP p.c. (2006)				-0.000004									
				(0.000003)									
luminosity per capita (2006)					-0.713								
					(0.464)								
Colonial origin: Spanish						0.328							
						(0.226)							
Colonial origin: British						0.382*							
						(0.206)							
Colonial origin: French						0.307							
						(0.221)							
Colonial origin: Portuguese						0.357							
						(0.315)							
Ruggedness (Terrain Ruggedness Index, 100 m.)							-0.0417						
							(0.0373)						
% Fertile soil								0.00234					
								(0.00175)					
Average distance to nearest ice-free coast (1000 km.)									-0.113				
									(0.130)				
% Within 100 km. of ice-free coast									-0.00150				
									(0.00156)				
Ethnic Fragmentation Index										-0.0519			
										(0.175)			
Altitude above min altitude (m)											-0.00007		
											(7.94e-05)		
avg. annual precipitation (mm)												0.00007	
												(5.08e-05)	
suitability for agriculture													0.418**
													(0.168)
bs.	137	137	137	137	137	88	137	137	137	125	137	137	135
R sq.	0.001	0.082	0.000	0.010	0.017	0.040	0.009	0.013	0.008	0.001	0.005	0.016	0.044
Dependent Variable Mean	0.453	0.453	0.453	0.453	0.453	0.466	0.453	0.453	0.453	0.464	0.453	0.453	0.452

Note: The table regresses a dummy for the inclusion in the WoLD data against country-characteristics to get a sense of any geographic bias in WoLD. An observation is a country, and a country is counted as included in WoLD if any langauge group from that country is included. *, **, *** denotes 10%, 5%. 1% significance respectively.

Table G4: Distance to centre of contiguous colonial holdings and distance to capital

	Distance to Capital			
	(1)	(2)	(3)	(4)
Colonial centrality	-20.24***	-20.23***	-33.77***	-46.84***
	(6.835)	(6.821)	(8.282)	(17.06)
Population		✓	✓	✓
Language Family FE	✓	✓	✓	✓
Country FE				✓
Obs.	2,573	2,573	2,495	2,461
R sq.	0.171	0.171	0.671	0.817
Dependent Variable Mean	440.1	440.1	433.9	436

Note: An observation in the table is a language group. *** denotes statistical significance at the 1% level. Both distance to the capital and distance to the colonial centroid are measured in 1,000km.

Table G5: Robustness: Alternate distance thresholds and definitions of Colonial Relationship

	Share of borrower's language borrowed from lender (%)			
	(1)	(2)	(3)	(4)
Within 25km		0.0596*** (0.0203)		
Within 50km	0.0444*** (0.0108)	0.0260*** (0.00796)		
Within 100km	0.0294*** (0.00389)	0.0384*** (0.00398)	0.0437*** (0.00538)	0.0437*** (0.00538)
Within 500km	0.00980*** (0.00103)		0.00608*** (0.000926)	0.00608*** (0.000926)
Within 1,000km			0.00400*** (0.000387)	0.00400*** (0.000387)
Similarity in family tree between borrower and lender	0.444*** (0.0500)	0.442*** (0.0506)	0.456*** (0.0491)	0.456*** (0.0491)
Colonial Relationship			0.00372 (0.0347)	
European lender			-0.000684 (0.00522)	
Colonial Relationship x European lender			0.180** (0.0897)	
Colonial Relationship	0.172** (0.0747)	0.172** (0.0748)		
Colonial Relationship (EU-only)				0.184** (0.0826)
Lexicon size	✓	✓	✓	✓
Distance to capital	✓	✓	✓	✓
Country FE	✓	✓	✓	✓
Language Family FE	✓	✓	✓	✓
Obs.	16,670,484	16,670,484	16,670,484	16,670,484
R sq.	0.016	0.016	0.016	0.016
Dependent Variable Mean	0.00350	0.00350	0.00350	0.00350

Note: The regression is run at the society-pair level. Standard errors are two-way clustered by the two societies in a society-pair. *, **, *** denotes 10%, 5%, 1% significance respectively. Colonial relationship codes the pair as 1 if the adopting society was colonized by the lending society, but not vice-versa. Lexicon size controls for PanLex coverage of the adopting language. Language Family FE are FE for the 4-level from root language family based on language trees for both the borrower and the lender. Country fixed effects are included for both the borrower and the lender.

Table G6: Colonial Centrality: alternate measure of colonial intensity

	(1)	(2)	(3)	(4)	(5)
Dependent Variables	Share of borrower's language borrowed from lender (%)				
Within 100km	0.0451***	0.0451***	0.0451***	0.0452***	0.0479***
	(0.00655)	(0.00657)	(0.00656)	(0.00657)	(0.00654)
Within 500km	0.00534***	0.00533***	0.00573***	0.00585***	0.00521***
	(0.00109)	(0.00110)	(0.00114)	(0.00114)	(0.00154)
Within 1,000km	0.00372***	0.00372***	0.00347***	0.00360***	-0.00102
	(0.000572)	(0.000572)	(0.000438)	(0.000438)	(0.000899)
Distance to centroid of contiguous colonial holdings	-0.000138***	-0.000126***	-0.000236***	-0.000281*	0.000105
	(1.88e-05)	(1.99e-05)	(5.61e-05)	(0.000168)	(0.000145)
Colonial Relationship x Distance to colonial centre	-0.0395***	-0.0397***	-0.0380***	-0.0294***	-0.0405***
	(0.00605)	(0.00612)	(0.00620)	(0.00664)	(0.00419)
Similarity in family tree between borrower and lender	0.421***	0.420***	0.420***	0.419***	0.405***
	(0.0494)	(0.0496)	(0.0495)	(0.0494)	(0.0500)
Lexicon Size		✓	✓	✓	
Distance to Capital		✓	✓	✓	
Language Family FE			✓	✓	
Country FE				✓	
Country-pair FE					✓
Obs.	10,978,803	10,953,261	10,790,487	10,790,487	10,817,968
R sq.	0.006	0.006	0.013	0.016	0.024
Dependent Variable Mean	0.00374	0.00374	0.00376	0.00376	0.00375

Note: The regression is run at the society-pair level. Standard errors are two-way clustered by the two societies in a society-pair. *, **, *** denotes 10%, 5%, 1% significance respectively. Colonial relationship codes the pair as 1 if the adopting society was colonized by the lending society, but not vice-versa. Lexicon size accounts for the PanLex coverage of the borrowing group. Language Family FE are FE for the 4-level from root language family based on language trees for both the borrower and the lender. Country fixed effects are included for both the borrower and the lender.

Table G7: Robustness: loanword categorization thresholds (society-level)

Dependent Variable:	Language Borrowed		Language Loaned	
Threshold Used	60%	70%	60%	70%
	(1)	(2)	(3)	(4)
Gains from trade with neighbours	0.00412**	0.00185**		
	(0.00182)	(0.000904)		
Influence on trade with neighbours			0.0104***	0.00277**
			(0.00394)	(0.00134)
Trade wealth (structurally estimated)	✓	✓	✓	✓
Population	✓	✓	✓	✓
Land Share	✓	✓	✓	✓
Land diversity	✓	✓	✓	✓
Distance to Neighbour(s)	✓	✓	✓	✓
Linguistic Distance	✓	✓	✓	✓
Language Family FE	✓	✓	✓	✓
Colonizer FE	✓	✓	✓	✓
Continent FE	✓	✓	✓	✓
N	2,606	2,606	2,606	2,606
R^2	0.110	0.174	0.129	0.261
Dependent Variable Mean	0.00436	0.00150	0.00436	0.00150

Note: The unit of observation is a society. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Gains from trade with neighbours is c_i as defined in equation 7, and analogously, influence on trade with neighbours is ι_i (equation 9). Language Borrowed (range [0,100]) is the share of a language made up of borrowed loanwords, while Language Loaned (range [0,100]) is the lending analogue. Distance to neighbours is a mean distance to neighbours, and in this case captures the density of the neighbourhood. Thresholds are based on the probability provided by the algorithm that a word-pair is a loanword. We typically use a threshold of 50%, but larger thresholds produce similar results.

Table G8: Agricultural trade and linguistic exchange outside of the horserace specification

Dependent Variable:	Language Borrowed		Language Loaned		Bilingualism	
Utility measure	percent change	percentile rank	percent change	percentile rank	percent change	percentile rank
	(1)	(2)	(3)	(4)	(5)	(6)
Gains from trade with neighbours	0.0914***	0.958***			0.0171***	0.114***
	(0.0340)	(0.214)			(0.00597)	(0.0337)
Influence on trade with neighbours			0.264**	1.740***		
			(0.107)	(0.458)		
Trade wealth (structurally estimated)	✓	✓	✓	✓	✓	✓
Population	✓	✓	✓	✓	✓	✓
Land Share	✓	✓	✓	✓	✓	✓
Land diversity	✓	✓	✓	✓	✓	✓
Distance to Neighbour(s)	✓	✓	✓	✓	✓	✓
Linguistic distance	✓	✓	✓	✓	✓	✓
Language Family FE	✓	✓	✓	✓	✓	✓
Colonizer FE	✓	✓	✓	✓	✓	✓
Continent FE	✓	✓	✓	✓	✓	✓
N	2,606	2,606	2,606	2,606	2,606	2,606
R^2	0.117	0.120	0.151	0.152	0.195	0.196
Dependent Variable Mean	0.951	0.951	0.951	0.951	0.331	0.331

Note: The unit of observation is a society. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Gains from trade with neighbours is c_i as defined in equation 7, and analogously, influence on trade with neighbours is ι_i (equation 9). Language Borrowed (range [0,100]) is the share of a language made up of borrowed loanwords, while Language Loaned (range [0,100]) is the lending analogue. Distance to neighbours is a mean distance to neighbours, and in this case captures the density of the neighbourhood.

Table G9: Trade and language exchange - Old World only

	Language Borrowed		Language Loaned		Bilingual	
	(1)	(2)	(3)	(4)	(5)	(6)
Gains from trade with neighbours	0.115***	0.979***	0.0923	0.312	0.0157**	0.0841*
	(0.0391)	(0.360)	(0.0842)	(0.337)	(0.00683)	(0.0439)
Influence on trade with neighbours	-0.0335	0.153	0.287**	1.730***	0.00931	0.0759
	(0.0440)	(0.402)	(0.141)	(0.594)	(0.00767)	(0.0479)
Trade wealth (structurally estimated)	✓	✓	✓	✓	✓	✓
Population	✓	✓	✓	✓	✓	✓
Land Share	✓	✓	✓	✓	✓	✓
Land diversity	✓	✓	✓	✓	✓	✓
Distance to Neighbour(s)	✓	✓	✓	✓	✓	✓
Linguistic Distance	✓	✓	✓	✓	✓	✓
Language Family FE	✓	✓	✓	✓	✓	✓
Colonizer FE	✓	✓	✓	✓	✓	✓
Continent FE	✓	✓	✓	✓	✓	✓
N	2,272	2,272	2,272	2,272	2,272	2,272
R^2	0.113	0.116	0.147	0.148	0.178	0.179
Dependent Variable Mean	0.987	0.987	0.956	0.956	0.346	0.346

Note: The unit of observation is a society. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Gains from trade with neighbours is the c_i measure defined in equation 7, and analogously, influence on trade with neighbours is ι_i (equation 9). In each case, in order to aggregate to the societal level we take the maximum value from the society's neighbours. Distance to neighbours is a mean distance to neighbours, and in this case captures the density of the neighbourhood. Language Borrowed (range [0,100]) is the share of a language made up of borrowed loanwords, while Language Loaned (range [0,100]) is the lending analogue. Bilingualism is a binary variable denoting whether the society is heavily bilingual in any of its neighbours' languages. The number of observations differs from 4 because we restrict the sample to societies in one of Europe, Africa, Asia and Oceana, and exclude societies in North America and South America.

Table G10: Gains from trade & linguistic exchange with other neighbours

	Exchange with best		Average exchange		Exchange with worst	
	Borrowed	Loaned	Borrowed	Loaned	Borrowed	Loaned
	(1)	(2)	(3)	(4)	(5)	(6)
Gains from trade with neighbours	0.0738**		0.0143		-0.00560	
	(0.0305)		(0.0138)		(0.0114)	
Influence on trade with neighbours		0.140**		0.0142		-0.0119***
		(0.0573)		(0.0110)		(0.00444)
Trade wealth (structurally estimated)	✓	✓	✓	✓	✓	✓
Population	✓	✓	✓	✓	✓	✓
Land Share	✓	✓	✓	✓	✓	✓
Land diversity	✓	✓	✓	✓	✓	✓
Distance to Neighbour(s)	✓	✓	✓	✓	✓	✓
Linguistic distance	✓	✓	✓	✓	✓	✓
Language Family FE	✓	✓	✓	✓	✓	✓
Colonizer FE	✓	✓	✓	✓	✓	✓
Continent FE	✓	✓	✓	✓	✓	✓
N	2,606	2,606	2,606	2,606	2,606	2,606
R^2	0.107	0.146	0.111	0.131	0.084	0.120
Dependent Variable Mean	0.799	0.634	0.291	0.209	0.0790	0.0493

Note: The unit of observation is a society. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Gains from trade with neighbours is c_i as defined in equation 7, and analogously, influence on trade with neighbours is ι_i (equation 9). Language Borrowed (range [0,100]) is the share of a language made up of borrowed loanwords, while Language Loaned (range [0,100]) is the lending analogue. Distance to neighbours is a mean distance to neighbours, and in this case captures the density of the neighbourhood.

Table G11: Heterogeneity by population size

	Language Borrowed			
	Society-level		Society-pair-level	
	percent change	percentile rank	percent change	percentile rank
	(1)	(2)	(3)	(4)
Gains from trade with neighbours	0.101***	0.941***	0.0732***	0.725***
	(0.0354)	(0.319)	(0.0272)	(0.264)
Influence on trade with neighbours	-0.0326	0.0792		
	(0.0390)	(0.358)		
Gains from trade with neighbours x Population difference	-0.000000351	-0.00000496	0.0000000140	0.0000000243
	(0.000000386)	(0.00000307)	(0.0000000116)	(0.0000000336)
Influence on trade with neighbours x Population difference	-0.000000461	-0.00000446*		
	(0.000000381)	(0.00000251)		
Trade wealth (structurally estimated)	✓	✓		
Population	✓	✓		
Land Share	✓	✓		
Land diversity	✓	✓		
Distance to Neighbour(s)	✓	✓		
Linguistic Distance	✓	✓		
Language Family FE	✓	✓		
Colonizer FE	✓	✓		
Continent FE	✓	✓		
Language pair FE			✓	✓
N	2,606	2,606	5,693	5,693
R^2	0.118	0.121	0.518	0.519
Dependent Variable Mean	0.951	0.951	0.278	0.278

Note: The unit of observation is a society. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Gains from trade with neighbours is c_i as defined in equation 7, and analogously, influence on trade with neighbours is ι_i (equation 9). Language Borrowed (range [0,100]) is the share of a language made up of borrowed loanwords, while Language Loaned (range [0,100]) is the lending analogue. Distance to neighbours is a mean distance to neighbours, and in this case captures the density of the neighbourhood. Unviable relationships are ones without any positive gains from trade.

Table G12: Heterogeneity by linguistic distance

	Language Borrowed			
	Society-level		Society-pair-level	
	percent change	percentile rank	percent change	percentile rank
	(1)	(2)	(3)	(4)
Gains from trade with neighbours	0.00765	-0.143	0.0286	-0.0894
	(0.0881)	(0.715)	(0.0303)	(0.209)
Influence on trade with neighbours	-0.0505	0.974		
	(0.0951)	(0.876)		
Gains from trade with neighbours x Linguistic Distance	0.147	1.752*	0.0746	1.447*
	(0.124)	(1.041)	(0.0731)	(0.743)
Influence on trade with neighbours x Linguistic Distance	0.0240	-1.462		
	(0.129)	(1.158)		
Trade wealth (structurally estimated)	✓	✓		
Population	✓	✓		
Land Share	✓	✓		
Land diversity	✓	✓		
Distance to Neighbour(s)	✓	✓		
Linguistic Distance	✓	✓		
Language Family FE	✓	✓		
Colonizer FE	✓	✓		
Continent FE	✓	✓		
Language pair FE			✓	✓
N	2,606	2,606	5,693	5,693
R^2	0.118	0.121	0.518	0.521
Dependent Variable Mean	0.951	0.951	0.278	0.278

Note: The unit of observation is a society. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Gains from trade with neighbours is c_i as defined in equation 7, and analogously, influence on trade with neighbours is ι_i (equation 9). Language Borrowed (range [0,100]) is the share of a language made up of borrowed loanwords, while Language Loaned (range [0,100]) is the lending analogue. Distance to neighbours is a mean distance to neighbours, and in this case captures the density of the neighbourhood. Unviable relationships are ones without any positive gains from trade.

Table G13: Robustness: loanword categorization thresholds (relationship-level)

Dependent Variable:	Language Borrowed			
	60% threshold		70% threshold	
	(1)	(2)	(3)	(4)
Gains from trade with neighbours	0.00656**	0.00349**	0.00293***	0.00104
	(0.00294)	(0.00175)	(0.00108)	(0.000788)
Influence on trade with neighbours		0.00395		0.000635
		(0.00271)		(0.000974)
Relationship Fixed Effects	✓		✓	
Society Fixed Effects (both)		✓		✓
N	5,693	5,693	5,693	5,693
R^2	0.502	0.589	0.504	0.717
Dependent Variable Mean	0.00129	0.00129	0.000459	0.000459

Note: The unit of observation is a society-pair. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Gains from trade with neighbours is the c_{ij} measure defined in equation 7, and analogously, influence on trade with neighbours is ι_{ij} (equation 9). In each case we aggregate to the society level by taking the maximum value from the society's neighbours. Thresholds are based on the probability provided by the algorithm that a word-pair is a loanword. We typically use a threshold of 50%, but larger thresholds produce similar results.

Table G14: Unviable Trading Relationships (relationship-level)

Dependent Variable:	Language Borrowed		Bilingualism	
Utility measure:	percent change	percentile rank	percent change	percentile rank
	(1)	(2)	(3)	(4)
Gains from trade with neighbours	0.0232	0.306	-0.0118	-0.0955
	(0.0565)	(0.304)	(0.0147)	(0.0660)
Society Fixed Effects (both)	✓	✓	✓	✓
Baseline controls	✓	✓	✓	✓
N	3,871	3,871	3,871	3,871
R^2	0.727	0.727	0.897	0.897
Dependent Variable Mean	0.232	0.232	0.108	0.108

Note: The unit of observation is a society-pair. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Gains from trade with neighbours is the c_{ij} measure defined in equation 7, and analogously, influence on trade with neighbours is ι_{ij} (equation 9). In each case we aggregate to the society level by taking the maximum value from the society's neighbours. Viable trading relationships are any relationships where at least one of the two parties can gain from trade. Language Borrowed (range [0,100]) is the share of a language made up of borrowed loanwords, while Language Loaned (range [0,100]) is the lending analogue. Controls are as follows: trade wealth (estimated); population; land share; land diversity.

Table G15: Loanwords by word-type and trade incentives

	Crop names	Non-crop words	Economic transaction words	All but crop/transaction
	(1)	(2)	(3)	(4)
Gains from trade with neighbours	0.167*	0.702***	0.0945***	0.680***
	(0.0881)	(0.192)	(0.0366)	(0.194)
Trade wealth (structurally estimated)	✓	✓	✓	✓
Population	✓	✓	✓	✓
Land Share	✓	✓	✓	✓
Land diversity	✓	✓	✓	✓
Distance to Neighbour(s)	✓	✓	✓	✓
Linguistic Distance	✓	✓	✓	✓
Language Family FE	✓	✓	✓	✓
Colonizer FE	✓	✓	✓	✓
Continent FE	✓	✓	✓	✓
N	2,606	2,606	2,534	2,606
R^2	0.097	0.113	0.089	0.114
Dependent Variable Mean	0.247	0.634	0.106	0.604

Note: The unit of observation is a society. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Gains from trade with neighbours is c_i as defined in equation 7, and analogously, influence on trade with neighbours is ι_i (equation 9). Language Borrowed (range [0,100]) is the share of a language made up of borrowed loanwords, while Language Loaned (range [0,100]) is the lending analogue. In each case we consider only words within a given category. Distance to neighbours is a mean distance to neighbours, and in this case captures the density of the neighbourhood. Unviable relationships are ones without any positive gains from trade.

Appendix H. Supplementary Empirical Exercises

H.1. Migration

Migration is a tricky issue in this context, as the theoretical predictions are ambiguous. On the one hand, reductions in cultural distance should be expected to come along with more migration, as the cultural cost to living in another society is similarly reduced as the transaction costs of inter-cultural trade are reduced. On the other hand, we might expect more migration between geographically homogenous regions because production would be easier in the new location (Michalopoulos 2012). Since trade partners are unlikely to produce the same things, we might expect land complementarity to be negatively associated with migration.

Thus, the issue is an empirical question. To address it we use the World Migration Matrix 1.1 from Putterman and Weil 2010. The matrix provides " for each of 165 countries, an estimate of the proportion of the ancestors in 1500 of that country's population today that were living within what are now the borders of that and each of the other countries" Putterman and Weil 2010. This is, to the best of our knowledge, the closest we can get to long-run migration data that maps into our language data. The migration data come at the country-pair, a higher unit of aggregation than the language data, but aside from that seems ideal for assessing the role of migration in driving the estimates that we observe.

To match the country-pair level of aggregation in the migration data, we aggregate the language data to the country-pair level, and match it to the migration matrix. To do this we follow with the convention in the rest of the paper, and consider the trade incentives of the most suitable trading partners when we aggregate. We also show a specification with the mean of the welfare gain, and mean of the rank, since when there are many groups in a country the maximum of those groups may not have a large effect on migration. For example, there could be a very small group in a country who is the only group with a viable trade relationship in the other county. This may generate a high maximum gains from trade between the country-pair, but low levels of migration since the group driving that high value is itself very small.

The regression that we analyze tests for a relationship between our main independent variable - gains from agricultural trade, and migration since 1500.

$$Migration_{ab} = \alpha_a + \alpha_b + \beta_1 c_a + \beta_1 \iota_a + \epsilon_{ab} \tag{19}$$

In this regression the unit of observation is a country-pair, for countries a and b. $Migration_{ab}$ is the variable from Putterman and Weil 2010, and c_a and ι_a are the aggregated gains from trade and trade influence variables. We estimate the equation with country fixed effects for each country (α_a and α_b). The results are in table H1.

Table H1: Migration

	Migration since 1500			
	Max GFT		Mean GFT	
	(1)	(2)	(3)	(4)
Gains from trade with neighbours	-0.00452 (0.00384)	0.00262 (0.0168)	-0.00476 (0.00519)	0.0192 (0.0196)
Influence on trade with neighbours	-0.00133 (0.00296)	0.00689 (0.0191)	-0.00420 (0.00286)	-0.00924 (0.0190)
Country Fixed Effects (both)	✓	✓	✓	✓
Observations	420	420	420	420
R-squared	0.767	0.765	0.766	0.765
Dependent Variable Mean	0.0180	0.0180	0.0180	0.0180

Note: The unit of observation is a country-pair. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Gains from trade with neighbours is the percentile rank in the distribution (range [0,1]) of c_i as defined in equation 7. In each case, to aggregate to the country-pair level we take the mean or maximum value from the society's neighbours. Distance to neighbours is a mean distance to neighbours, and in this case captures the mean distance between groups in each country. The dependent variable is taken directly from Putterman and Weil 2010.

In columns 1 and 2 we present estimates where the aggregation takes the maximum gains from trade and trade influence, as in the rest of the paper. In column 1 we show results using the raw welfare gains, and in column 2 we show the rank. In both cases, we see no clear relationship between agricultural gains from trade and cross-country migration. In columns 3 and 4 we show the estimates from same regression specification, but with the gains from trade data aggregated to the country-pair level using the mean instead of the maximum values. However, using this variable construction does not change the interpretation of the estimates much - once again we find small and insignificant results.

Even though migration generally does not appear to be driving the estimates, one concern might be that the results are driven primarily by New World groups as a result of colonial conquest. Accordingly, we re-estimate the main result from table 4 but removing North and South America. Those results can be seen above in table G9.

We can see in the table that the results are nearly identical to the main table 4. In each column we still estimate a positive and significant estimate for the appropriate parameter in the horserace, and not for the other. We lose a little bit of significance on the bilingualism estimate, which is expected since that variable is noisily measured in the first place, and in any case this should be expected, simply for power reasons, when we drop two continents from the analysis. Between the estimates in this table and the table based on the Putterman and Weil 2010 data, we are confident that migration is not a first-order driver of our main findings.

H.2. Loanword timing

Of interest is the timing of loanwords, especially given that our trade model is based on local agricultural trade incentives that are likely still relevant for most language groups in the world, but certainly made up a greater share of trade historically.

The first thing to note is that the gains from trade measure, which is computed based on contemporary agricultural trade incentives, is actually very highly correlated with pre-1500 trade incentives. To compute pre-1500 trade incentives, we recompute gains from trade based on the crops that each society would have had access to prior to the Colombian Exchange.[104] The strong correlation between the two measures indicates that our main results may be capturing either persistent historical changes or more recent ones.

When we attempt to empirically distinguish between historical and more recent borrowing, it becomes clear that both pre-1500 and post-1500 agricultural incentives matter, and statistically, it is not possible to say which plays a larger role. To show this, we revisit the language-group level regression from equation 10, but using the pre-1500 trade incentives. Those results can be seen in columns 1-3 of table H2. We see that for each of lending, borrowing and bilingualism, the estimates correspond closely with the main results in table 4.

In columns 4-6 of the same table, we also show results that attempt to distinguish between pre-1500 trade and post-1500 trade. This is slightly more difficult since we do not have a good measure of purely post-1500 trade, since the main measure includes those crops available prior to the Colombian Exchange, as well as after. Measuring pre-1500 trade is more straightforward because we can restrict the gains from trade measure to the crops available at that period, but we do not have an analogous natural experiment where there were crops historically available that no longer are. Accordingly, the best we can do to isolate contemporary trade incentives is to regress the main measure on historical trade incentives, and examine the residual.

That is:

$$\begin{aligned} GFT_i &= \beta_0 + \beta_1 Pre1500GFT_i + \epsilon_i \\ &\approx \beta_0 + \beta_1 Pre1500GFT_i + Post1500GFT_i \end{aligned} \tag{20}$$

Where GFT_i is the main gains from trade measure for society i using all crops, $Pre1500GFT_i$ is the analogous measure using only the crops available prior to the Colombian Exchange, and the assumption is that $\epsilon_i \approx Post1500GFT_i$. This assumes additive separability which

[104] Several recent works have used the Colombian Exchange strategy. For example, see Nunn and Qian 2011; Blouin 2021; Dickens et al. 2022

Table H2: Timing of Language Exchange

	Borrowed (1)	Loaned (2)	Bilingual (3)	Borrowed (4)	Loaned (5)	Bilingual (6)
Pre-1500 Gains from trade	0.641***	0.473**	0.142***	0.694***	0.385	0.137***
	(0.247)	(0.234)	(0.0395)	(0.250)	(0.239)	(0.0401)
Pre-1500 Trade influence	0.204	1.547***	-0.000525	0.121	1.632***	0.00284
	(0.256)	(0.520)	(0.0432)	(0.291)	(0.542)	(0.0438)
Post-1500 Gains from trade				0.802*	-0.143	0.0221
				(0.422)	(0.389)	(0.0453)
Post-1500 Trade influence				-0.0311	1.016**	0.0805
				(0.337)	(0.423)	(0.0511)
Trade wealth (structurally estimated)	✓	✓	✓	✓	✓	✓
Population	✓	✓	✓	✓	✓	✓
Land Share	✓	✓	✓	✓	✓	✓
Land diversity	✓	✓	✓	✓	✓	✓
Distance to Neighbour(s)	✓	✓	✓	✓	✓	✓
Linguistic Distance	✓	✓	✓	✓	✓	✓
Language Family FE	✓	✓	✓	✓	✓	✓
Colonizer FE	✓	✓	✓	✓	✓	✓
Continent FE	✓	✓	✓	✓	✓	✓
N	2,606	2,606	2,606	2,606	2,606	2,606
R^2	0.119	0.153	0.198	0.120	0.154	0.199
Dependent Variable Mean	0.951	0.951	0.331	0.951	0.951	0.331

Note: The unit of observation is a society. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Gains from trade with neighbours is the percentile rank in the distribution (range [0,1]) of c_i as defined in equation 7. In each case, to aggregate to the societal level we take the maximum value from the society's neighbours. Distance to neighbours is a mean distance to neighbours, and in this case captures the density of the neighbourhood. Language Borrowed (range [0,100]) is the share of a language made up of borrowed loanwords. All word-type borrowing outcomes are winsorized at the 0.1% level to deal with outliers.

may be a strong assumption, so this measure should be treated only as suggestive.

References

Ager, Simon (2019). *Omniglot*.

Awagana, Ari, H. Ekkehard Wolff, and Doris Löhr (2009). "Loanwords in Hausa, a Chadic language in West Africa". In: *Loanwords in the World's Languages: A Comparative Handbook*. Ed. by M. Haspelmath and U. Tadmor. De Gruyter Mouton.

Blouin, Arthur (2021). "Axis-orientation and knowledge transmission: Evidence from the Bantu expansion". In: *The Journal of Economic Growth* 26.4, pp. 359–384.

Bojanowski, Piotr et al. (2017). "Enriching Word Vectors with Subword Information". In: *Transactions of the Association for Computational Linguistics* 5, pp. 135–146.

Costinot, Arnaud and Dave Donaldson (2012). "Ricardo's theory of comparative advantage: old idea, new evidence". In: *American Economic Review* 102.3, pp. 453–58.

Davies, Roy (2019). *Current Value of Old Money*. Website. URL: http://projects.exeter.ac.uk/RDavies/arian/current/howmuch.html (visited on 08/17/2019).

Dickens, Andrew et al. (2022). "Understanding ethnolinguistic differences: The roles of geography and trade". In: *The Economic Journal* Forthcoming.

Feenstra, Robert C. (2004). *Advanced international trade: theory and evidence*. Princeton, N.J: Princeton University Press.

Giuliano, Paola and Nathan Nunn (2018). "Ancestral characteristics of modern populations". In: *Economic History of Developing Regions* 33.1, pp. 1–17.

Haspelmath, M. and U. Tadmor (2009). *Loanwords in the World's Languages: A Comparative Handbook*. De Gruyter Mouton.

Hayes, Bruce (2009). *Introductory phonology*. Blackwell textbooks in linguistics 23. OCLC: 212893710. Malden, MA ; Oxford: Wiley-Blackwell.

Jacks, David S (2004). "Market Integration in the North and Baltic Seas, 1500-1800". In: *The Journal of European Economic History* 33.2.

— (2005). "Intra- and international commodity market integration in the Atlantic economy, 1800–1913". In: *Explorations in Economic History* 42.3. Publisher: Elsevier, pp. 381–413.

Jaro, Matthew A. (1989). "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida". In: *Journal of the American Statistical Association* 84.406, pp. 414–420.

Lewis, Paul M. (2009). *Ethnologue : languages of the world*. Texas: SIL International.

Löhr, Doris, H. Ekkehard Wolff, and Ari Awagana (2009). "Loanwords in Kanuri, a Saharan language". In: *Loanwords in the World's Languages: A Comparative Handbook*. Ed. by M. Haspelmath and U. Tadmor. De Gruyter Mouton.

Michalopoulos, Stelios (2012). "The origins of ethnolinguistic diversity". In: *American Economic Review* 102.4, pp. 1508–39.

Michalopoulos, Stelios and Melanie Meng Xue (2021). “Folklore”. In: *The Quarterly Journal of Economics* 1, p. 54.

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig (June 2013). “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*. Atlanta, Georgia: Association for Computational Linguistics, pp. 746–751.

Monfreda, Chad, Navin Ramankutty, and Jonathan A Foley (2008). “Farming the planet: 2. Geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000”. In: *Global biogeochemical cycles* 22.1.

Mortensen, David R. et al. (2016). “PanPhon: A Resource for Mapping IPA Segments to Articulatory Feature Vectors”. In: *COLING*.

Mortensen, David R, Siddharth Dalmia, and Patrick Littell (2018). “Epitran: Precision G2P for many languages”. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

National Health Service (2019). *What should my daily intake of calories be?* Website. URL: https://www.nhs.uk/common-health-questions/food-and-diet/what-should-my-daily-intake-of-calories-be/ (visited on 02/18/2022).

Nunn, Nathan and Nancy Qian (2011). “The Potato’s Contribution to Population and Urbanization: Evidence from an Historical Experiment”. In: *The Quarterly Journal of Economics* 126.2, pp. 593–650.

Putterman, Louis and David N Weil (2010). “Post-1500 population flows and the long-run determinants of economic growth and inequality”. In: *The Quarterly journal of economics* 125.4, pp. 1627–1682.

Rehurek, Radim and Petr Sojka (May 2010). “Software Framework for Topic Modelling with Large Corpora”. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, pp. 45–50.

Schadeberg, Thilo C. (2009). “Loanwords in Swahili”. In: *Loanwords in the World’s Languages: A Comparative Handbook*. Ed. by M. Haspelmath and U. Tadmor. De Gruyter Mouton, pp. 77–102.

Spolaore, Enrico and Romain Wacziarg (2009). “The diffusion of development”. In: *The Quarterly journal of economics* 124.2, pp. 469–529.

Thurgood, Graham (1999). “From Ancient Cham to Modern Dialects: Two Thousand Years of Language Contact and Change: With an Appendix of Chamic Reconstructions and Loanwords”. In: *Oceanic Linguistics Special Publications* 28. Publisher: University of Hawai’i Press, pp. i–407.

U.S. Official Inflation Data, Alioth Finance (2019). *Inflation Calculator*. URL: https://www.officialdata.org (visited on 08/17/2019).

Wald, Abraham (1940). “The fitting of straight lines if both variables are subject to error”. In: *The annals of mathematical statistics* 11.3, pp. 284–300.

Wichmann, Soren, Eric W. Holman, and Cecil H. Brown (2016). “The ASJP Database (version 17)”. In:

Winkler, William (Jan. 1, 1990). *String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage.*