(Machine) Learning Languages: Do Societies Strategically Reduce Linguistic Distance?

Arthur Blouin
University of Toronto

Julian G.A. Dyer
University of Toronto

March 6, 2020*

This paper empirically distinguishes between three mechanisms of language evolution: (1) random evolution; (2) natural absorption of language that might occur when groups interact; (3) purposeful reductions in language barriers. To distinguish between the latter two, we rely on the intuition that learning a language to reduce language barriers is less beneficial when more of the other group already speaks your language. This generates an asymmetry in language exchange that contrasts with the symmetric exchange that might result from two groups interacting, and absorbing each other's cultural traits. Our empirical tests require data on the extent and direction of language exchange, and gains from inter-group interaction. For the first, we use machine learning to predict, for (almost) all known words in all known languages, whether the word originated from a word in another language, and if so which one. We match this data to the output of an agricultural trade model, which provides us with exogenous trade incentives between language groups. Together, we use the data to document asymmetric patterns in language exchange that are systematically related to trade incentives. In our framework, this is only consistent with groups purposefully reducing linguistic distance to induce interaction.

*We are grateful to Gustavo Bobonis, Shari Eli, Benjamin Enke, James Fenske, Per Fredriksson, Rocco Macchiavello, Martina Miotto, Sharun Mukand, Naomi Nagy and Jordan Roper for helpful comments. Matthew Schwartzman and Frederick Gietz provided outstanding research assistance. We also thank seminar and workshop participants at University of Toronto, ASREC, CAGE, and University of Warwick. We gratefully acknowledge financial support from SSHRC and the Connaught fund.

1. Introduction

Linguistic diversity has been shown to be an important detriment to economic development (Easterly and Levine, 1997, Alesina et al., 2003), especially when linguistic distance is high (Desmet et al., 2012) or when there is inequality between language groups (Alesina et al., 2016).[1] Recently, attention has turned towards the forces that generate linguistic diversity and distance. This work has mostly adopted the view that cultures become more similar to each other whenever groups interact with each other more.[2] However, the theoretical trade literature postulates a different idea of language evolution. For instance, Kónya (2006) argues that groups make one-sided cultural investments to reduce language barriers with other groups that they would like to trade with. Empirical progress in distinguishing between these models has been limited by a dearth of data on linguistic exchange.

In this paper we construct a dataset to examine both the extent and direction of linguistic exchange for every language group in the world. Doing so allows us to verify the endogeneity of linguistic convergence. But it also opens the possibility to empirically distinguish between purposeful reductions in language barriers (which we will call *strategic convergence*), and the natural absorption of language that might occur when groups interact (which we will call *non-strategic convergence*). The patterns in the data suggest that strategic linguistic convergence is important, ruling out purely non-strategic convergence. To our knowledge, this is the first empirical evidence to identify strategic reductions in linguistic distance.

The first clear data requirement for this empirical exercise is a directed measure of linguistic exchange, as opposed to currently available measures of similarity, which are not able to identify asymmetries in linguistic exchange. For this reason we construct a dataset of loanwords - a word in one language that was horizontally transmitted from another[3] - covering a large number of language groups. We do this by using a machine learning algorithm to extend coverage from the few languages with loanwords verified by linguists to cover several thousand languages.

[1] Also see surveys by Alesina and Ferrara (2005) and Hale (2004).

[2] That is, cultural distance either falls due to cross-cultural interaction (Michalopoulos, 2012a, Dickens, 2019), or grows due to migration away from a core group (Ahlerup and Olsson, 2012, Ashraf and Galor, 2013).

[3] Vertical transmission indicates transmission from parent to child, while horizontal transmission indicates any other transfer.

One nice feature of the pairwise structure of our data is that it allows us to easily ignore the effect of colonialism on language borrowing, which - while interesting in its own right - is separate from our hypothesized mechanisms. To prevent our analysis from being driven by outliers in linguistic exchange resulting from colonialism, we restrict the pairs that include major colonial languages to include only neighbours that are adjacent to the colonial homeland. Specifically, this means that we restrict the neighbours for English, French, Portuguese and Spanish to include only their European neighbours.[4] The data reveals a surprisingly high degree of linguistic exchange, with nearly a quarter of each language comprising of loanwords, on average. This is encouraging for our study because it suggests that the data is not simply capturing words that originated from technological transfer.

Our aim is to use this dataset to empirically distinguish between strategic and non-strategic reduction linguistic distance. We formally develop three tests that allow us to identify strategic reductions in linguistic distance. All three tests follow from the intuition that under strategic convergence, interaction requires only one overlapping language.[5] So, even if gains to interaction are high, the benefit to learning the language of a group whose members already speak your language is low. On the other hand, because non-strategic convergence arises from absorbing some characteristics of the partner group, any interaction generates language exchange in both directions. In this case, we hypothesize that under strategic convergence (non-strategic prediction in brackets): borrowing and lending are negatively (positively) correlated; language exchange is heavily asymmetric (symmetric); and gains from interaction are positively correlated with borrowing but not lending (borrowing and lending).

Since our tests specify a mechanism for how language exchange relates to interaction, we need an exogenous measure of long-run economic incentives for two language groups to interact with each other. To do this, we rely heavily on data from FAO GAEZ. This provides us with the crop potential for most produceable crops around the globe. We combine this data with information on the nutrients

[4]Taking French as an example, we consider only pairwise relationships with languages such as Basque, Breton, Galician, etc. and exclude pairwise relationships in French West Africa or North America.

[5]We can think of investments to reduce language barriers as a public good with one group free-riding on the investments of the other.

needed for humans to survive. Nutritional requirements provide us with a utility function over crops, and the FAO GAEZ data provides us with nutritional endowments, via produceable crops. This allows us to structurally estimate the demand for each crop in each neighbourhood, and makes it straightforward to estimate trade incentives based only on the exogenously determined nutritional requirements and crop potential.[6] We measure incentive to trade at the societal level as the ratio of utility-under-trade to utility-under-autarky, while at the pairwise language level we construct the utility change from having (or not) each potential trading partner in a group's local trading network.

Using both the gains from trade and the language exchange data, we are able to test these hypotheses. In each case we find that the patterns are consistent with strategic convergence. We find (1) a strong negative (descriptive) correlation between linguistic lending and borrowing; (2) in a pairwise specification with pair fixed-effects we find that high incentive to trade is strongly positively correlated with linguistic borrowing.[7] The fixed effects framework that we employ means that this result is evidence of strong asymmetry in linguistic exchange, which is consistent only with strategic convergence in our framework. Finally, we find that (3) at the group level gains from trade matter for borrowing but not lending. We can empirically reject that the impact of trade incentives is the same for borrowing and lending, which constitutes a rejection of one prediction of the *pure* non-strategic convergence model. Each of our tests suggests that strategic reductions in linguistic distance are important.

Our analysis contributes directly to the literature on the relationship between economic development and diversity. While much of the literature views economic outcomes as a function of diversity, and explicitly assumes that diversity is not a function of economic outcomes (e.g. Alesina et al. (2003), Esteban et al. (2012)), our analysis puts this assumption into question. In particular, our analysis reveals that gains from trade, and hence economic outcomes are systematic drivers of

[6] And importantly, abstracting away from demand based on cultural preferences or taste - e.g. Atkin (2016) - which may be endogenous.

[7] Interestingly, this effect is non-linear. Being made much better off or much worse off by another group is positively correlated with language exchange. This is intuitive, as economic competitors may interact as frequently as trading partners, either to discuss coalitions, exchange technology (Michalopoulos, 2012a)

diversity in the first place. The fact that our estimates suggest that gains from trade can drastically reduce diversity suggests that the studies that place a causal interpretation on empirical models with economic development on the left hand side and diversity on the right may be negatively biased.

There is a set of the literature that is relatively unaffected by our analysis. This literature takes measures of diversity or distance that are based solely on vertical transmission, while we focus explicitly on horizontal transmission (Spolaore and Wacziarg, 2015, 2017, Desmet and Wacziarg, 2018). For instance, Desmet and Wacziarg (2018) measure linguistic distance by examining the distance in nodes on a language tree - which we should consider to be a function solely of vertical transmission, and more plausibly exogenous. Although, Ahlerup and Olsson (2012) conjecture that the process of group separation is itself endogenous.

In fact, the majority of the origins of diversity literature focuses on vertical transmission (Michalopoulos, 2012a, Ashraf and Galor, 2013, Dickens, 2019). Much of the literature tackles the problem from the viewpoint that diversity may be endogenous, but is driven by interaction-induced non-strategic convergence. For instance, Michalopoulos (2012a) considers migration and cross-cultural marriage. Dickens (2019) examines linguistic distance, as we do, but looks at Swadesh lists which are lists of words constructed by linguists to be most likely to be vertically transmitted.[8]

Finally, our results may be placed in the context of work examining the determinants of culture more broadly. The empirical literature has focused on culture as being determined by persistent historical shocks (Nunn and Wantchekon, 2011, Alesina et al., 2013) or co-evolution with institutions (Blouin, 2016, Lowes et al., 2016). There is a large literature on vertical versus horizontal transmission of culture, with notable theoretical contributions from Tabellini (2008) and Bisin and Verdier (2001), and empirical contributions from Algan and Cahuc (2010). Algan et al. (2013) focus on horizontal transmission of baby names, finding that the economic penalty associated with Arabic first names reduces horizontal transmission

[8]Changes in these words are therefore most likely due to subgroups breaking away from the main group, and experiencing linguistic drift. Our analysis instead explicitly discounts the Swadesh list words from the analysis to distinguish between cognate words (words that are vertically transmitted) from loan words (those that are horizontally transmitted) to suggest that strategic horizontal transmission aimed at inducing interaction are important.

in France. However, their study differs substantially, as their focus is on the value of cultural preferences, while ours is about the long run determinants of diversity.

2. Loanwords Background

We measure language exchange and investments by studying loanwords. A loanword is a word in one language whose sound and meaning enter the language's lexicon because it was copied from another language.[9] Loanwords are distinct from cognate words, which are vertically transmitted from a parent language. So, two language groups can have similar sounding words that mean the same thing either because they share a parent or because one borrowed the word from the other. 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.

Research on loanwords in the field of linguistics has traditionally been descriptive, and have typically been studied as part of a 'complete' understanding of a language and its influences. However, recently linguists have begun to go beyond a purely descriptive treatment of loanwords, and have begun asking questions such as 'why are words for body parts rarely borrowed but words for objects are?' This turns out not to be as straightforward as one might imagine. For instance, the English word *window* was borrowed from Old Norse even though English had previously used the word *eagpyrel* in precisely the same manner (Haspelmath and Tadmor, 2009a).

The field of linguistics has devoted significant effort to understanding the ancestry of languages and identify the age of branches in linguistic family trees (Vansina, 1990). This task requires *excluding* loanwords in order to focus on non-borrowed words that are indicative of parent languages and the timing of splits. Towards this end, linguists have identified lists of core meanings that are fundamental to human languages[10] that can be thought of as necessary and hence unlikely to be

[9] i.e. a loanword in language i was horizontally transmitted from language i.

[10] These are meanings such as 'man', 'woman', 'sun', 'night', 'eye', 'water', 'fire'. These meanings are essential and would almost certainly exist in any useable language, and are therefore less likely to be borrowed. Meanings outside these core concepts (such as for ideas and manufactured objects) are not necessarily an original part of all languages and are more likely to be borrowed from another language.

borrowed, since all languages likely had to develop or inherit their own word for these meanings. These Swadesh lists (named for Morris Swadesh) have become the foundation for many measures of linguistic distance used to measure ancestral distance among linguistic groups, like the Automated Similarity Judgment Program (ASJP) (Swadesh, 1950, Wichmann et al., 2016). For the inverse task of identifying loanwords, there is no such list of concepts that can be applied universally across languages since the lending and borrowing of words is so heavily influenced by power, economics and cultural openness (Haspelmath and Tadmor, 2009a). These factors – often an inconvenience to linguists with respect to understanding the evolution of languages – may be of importance to economists, and are the focus of this paper.

Historians have long used the existence of loanwords as evidence of exactly these factors. Furthermore, while we mostly avoid doing so in this paper, loanwords have also been heavily interpreted as indicators of cultural transfer/influence. This is often linked to economic and political power, as described in Frankopan (2016):

> *"Buddhism made sizeable inroads along the principal trading arteries to the west too [...] 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)

We are amenable to an interpretation of loanwords as a proxy for cultural exchange or influence more broadly, but we refrain from imposing this interpretation as it is both unnecessary for our hypothesis, and we are sympathetic to the conceptual issues that arise from conflating language and culture more broadly. Indeed, language as a standalone feature has been a prominent part of the economic development and diversity literature, and our focus is to contribute to that literature with a new dataset that allows us to speak specifically to the determinants of linguistic similarity between groups within a region.

3. Hypothesis, Assumptions and Predictions

We formally outline a hypothesis in order to be precise about our assumptions, and to demonstrate how any empirical results might identify strategic reductions

in linguistic distance under those assumptions.[11] We view this section as a guide to the empirical choices that we made throughout the project.

Consider two regions $r \in \{i, j\}$. Production is as follows:

$$Y_r = z_r \delta_r \cdot L_r^{\alpha} \cdot K_r^{(1-\alpha)} \tag{1}$$

The production function specifies the following variables: z_r is a productivity parameter that comes from spending time tending to the land. δ_r is the benefit from being able to trade with the other region that comes from learning their language. L is the size of the labour force. K is land, including quantity and quality.

Workers care about wages (y), which are equal to their MPL:

$$y_r = \alpha z_r \delta_r \cdot (K_r / L_r)^{(1-\alpha)} \tag{2}$$

We model this in an overlapping generations framework. So there are children and adults. Adults choose whether their children accumulate agricultural knowledge that enters as z, or learn a language, which enters as δ. Children become adults in the next period, and produce according to the choices their parents made, observe the state of the world, and based on that they make decisions for the knowledge of their own kids.

The choice is trivial, children spend time learning the other language if $z_r < \delta_r$. However, parents have to consider that the benefit of learning the language is higher if nobody in the other group speaks their language. People with only one overlapping language can trade. If everyone in the other group speaks their language already, then $\delta_r = 1$ If nobody does, then there is a huge set of people to profitably trade with if the other language is learned, so $\delta_r > 1$. z_r is exogenous and assumed to be greater than 1.

The share of individuals in a group who know the other groups language is:

$$\lambda_r = \frac{\sum_{l=1}^{L_r} \mathbb{1}(z_r^l < \delta_r^l)}{L_r} \tag{3}$$

[11] It is not really a model because there is no real concept of market clearing or equilibrium.

This implies that for region i the key decision is to learn the language if $z_i > \delta_i(\lambda_j)$ where $\delta_i'(\lambda_j) < 0$ & $\delta_i''(\lambda_j) > 0$. For instance, we could allow $\delta_i = c_i/\lambda_j$ and vice versa.

3.A. Timing

Parents observe the state of the world $\{\lambda_i, \lambda_j\}$ and decide whether their kids should engage in learning production (earning z_r next period) or in learning language (earning δ_r next period). There are always an equal number of kids and adults, so λ_r is always equally divided between the choice this and last period. So, in essence, we have $\delta_i^t = c_i/\lambda_j^{t-1}$, which combined with z_i^t determines λ_t which in turn is observed by the children in period t as they become adults at period $t+1$ and therefore determines δ_i^{t+1}.

In this case, unless both groups start off being identical (in which case we get multiple equilibria), eventually the more group i ends up learning the language of group j, the less group h learns of group i, as long as either has the incentive to learn the other.[12] This follows directly from the assumption that $\delta_i'(\lambda_j) < 0$.

3.B. Non-strategic convergence

However, this only accounts for strategic decisions. In other words, so far we are implicitly treating borrowing as a linear function of learning the language. It could also be that when group j learns the language of group i that people in group i borrow from group j. If these two factors were equally important, so that when group j learned the language of group i group i was equally likely to borrow from group j, then we would always have equal convergence. Define a borrowing function to be

$$\mathcal{L}_{ij} = f(\theta\lambda_i + (1-\theta)\lambda_j); \theta \in (1/2, 1) \tag{4}$$

If $\theta = 1/2$ then this implies that convergence is equal i.e. $\frac{\delta\mathcal{L}_{ij}}{\delta\mathcal{L}_{ji}} > 0$. In this case it would not matter which group learned the other group's language, we would just see cultural convergence. This is similar to how cultural change is modelled in

[12] Obviously, for example, if $c = 0$ for both groups, neither group ever learns the language of the other.

Michalopoulos (2012b). If $\theta = 1$ then we have the extreme situation outlined above where learning investments is all that matters, and interacting with someone who has learned your language has no impact on your language. In that case, because the more someone learns your language, the less worthwhile it becomes to learn theirs, we end up with $\frac{\delta \mathcal{L}_{ij}}{\delta \mathcal{L}_{ji}} < 0$. We therefore have:

Test 1: *If* $\theta = 1$ *then* $\frac{\delta \mathcal{L}_{ij}}{\delta \mathcal{L}_{ji}} < 0$. *If* $\theta = 1/2$ *then* $\frac{\delta \mathcal{L}_{ij}}{\delta \mathcal{L}_{ji}} > 0$

However, this is a purely descriptive test that ignores the endogeneity of interaction, making it difficult to say anything about mechanisms. To see how land endowments - which are exogenous - influence borrowing and lending, we can take our equation 4 and plug in our λ_i and λ_j.

$$\mathcal{L}_{ij} = f(\theta \frac{\sum_{l=1}^{L_i} \mathbb{1}(z_i^l < (c_i/\lambda_j)^l)}{L_i} + (1-\theta)\frac{\sum_{l=1}^{L_j} \mathbb{1}(z_j^l < (c_j/\lambda_i)^l)}{L_j}) \tag{5}$$

When $\theta = 1$ borrowing is high if $c_i > c_j$, since $z > 1$. For simplicity consider a case where everyone in the group has the same c and we start with $\lambda_i = \lambda_j$. Then we can write the first term as $\theta \frac{\sum_{l=1}^{L_i} \mathbb{1}(z_i^l < (c_i/c_j)^l)}{L_i}$ since λ_j is a function of c_j and $\lambda_i/\lambda_j = 1$. Then, strategic borrowing will only take place for the group with the higher c, and not at all for the group with lower c. Each period the benefit only gets higher as λ_i and λ_j adjust. The opposite is true for the second term, so the smaller is θ, the less strong is the pairwise correlation between c_i and $\mathcal{L}_{ij}$. Therefore:

Test 2: *If* $\theta = 1$, $c_i > c_j$ *implies* $\mathcal{L}_{ij} > \mathcal{L}_{ji}$ *and vice-versa. If* $\theta = 1/2$, $\mathcal{L}_{ij} = \mathcal{L}_{ji}$ *regardless of* $c_i \lessgtr c_j$.

If we start at $\lambda_i = \lambda_j$ then more broadly, $\delta c_i/\delta \mathcal{L}_{ij} > 0$. However, $\delta c_i/\mathcal{L}_{ji}$ is less clear. c_j enters negatively in the first term through λ_j (but for now we are holding λ fixed. It only explicitly appears in the numerator in the second term. If $\theta = 1$ the second term disappears, so we end up with no obvious correlation. If $\theta = 1/2$ then the correlation is positive. Our aim in the empirical analysis is simply to demonstrate that θ is sufficiently high that strategic considerations are important. While, $\delta c_i/\delta \mathcal{L}_{ij} > 0$ is generally consistent with high or low θ, $\delta c_i/\delta \mathcal{L}_{ji} < 0$ is only consistent with low θ.

Test 3: *c_i is correlated with $\mathcal{L}_{ij}$ regardless of θ. However, c_i is only correlated with $\mathcal{L}_{ji}$ for sufficiently low θ, otherwise the correlation is ambiguous.*

3.C. Summary of Empirical Tests

To summarize, if it is true that θ is high, so that strategic considerations are important, we expect the following patterns in the data (none of which would be expected if θ were low):

Test 1: $\frac{\delta \mathcal{L}_{ij}}{\delta \mathcal{L}_{ji}} < 0$

Test 2: *$c_i > c_j$ implies $\mathcal{L}_{ij} > \mathcal{L}_{ji}$ and vice-versa.*

Test 3: $\frac{\delta c_i}{\mathcal{L}_{ji}} = 0$

4. Data

4.A. From Theory to Empirics

To empirically test the predictions of the theory above, we need two pieces of data. First, we need data on linguistic transfer among groups, and further it is necessary that we observe pairwise directed transfer, i.e both $\mathcal{L}_{ij}$ and $\mathcal{L}_{ji}$. This is not possible with currently available data, and so we create a novel dataset of loanwords borrowed and lent among language group pairs to measure this pairwise directed linguistic transfer.[13] We accomplish this task by training a machine learning algorithm to extrapolate from a small number of languages with well-classified loanword status to a much larger dataset covering thousands of languages. We discuss the raw data for this task below in Section 4.B and the training and performance of our prediction algorithm in Section 5.A.

Second, we need to observe c_i and c_j, the amount that groups could potentially gain from interacting with each other. As discussed earlier, data on historic agricultural trade is scarcely available, and actual trade flows would be endogenous to other investments and confounding mechanisms. We therefore estimate a structural model to generate a dataset of potential gains from trade among neighbours.

[13]See Section 2 for a detailed discussion of our choice of loanwords for this task, and a brief review of how loanwords are used by historians as evidence of the phenomena we concern ourselves with in this paper.

This model is based on production capacity driven by geographic endowments, where groups make production and trade decisions maximize the population they are able to support, and where gains from trade are driven by nutritional complimentarity in crop production. We discuss the raw data used for this task in Section 4.C and the model and its estimation in Section 5.B.

4.B. Language Data

i) PanLex In order to construct data on loanwords and linguistic exchange, we need wordlists, or *lexicons*, from as many languages as possible. For this, we draw on the PanLex database. 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. The database takes thousands of translation dictionaries converted to a single common structure, and includes words from 5,700 languages, covering over 25,000,000 words.[14] The dataset is as close as we believe is possible to representing all known words in all known languages. The coverage of this dataset goes far beyond the coverage possible with sources based on textual and archival resources, which are restricted to languages with a significant body of written history. This breadth of coverage is a further advantage of the loanwords approach.

ii) World Loanword Database We combine PanLex lexicons with information on classified loanwords from the World Loanword Database (WoLD). WoLD is a scientific publication by the Max Planck Institute for Evolutionary Anthropology, and includes 41 recipient languages and 369 donor languages (See Figure 3 for a map of the spatial distribution of each type of language in WoLD). It is the first aggregated dataset of rigorously-identified loanwords under a consistent set of criteria, providing "...vocabularies (mini-dictionaries of about 1000-2000 entries) of 41 languages from around the world, with comprehensive information about the loanword status of each word." (Haspelmath and Tadmor, 2009b)[15] The data compiled into WoLD is the result of a long literature on loanwords by linguists.

[14] see https://panlex.org. The database is constantly being updated, in this paper we use the SQL database posted on September 1, 2019

[15] see: wold.clld.org

The WoLD data for a single language, Swahili, for example, is based on thirty-three academic publications by twenty-seven separate authors, published between 1861 and 2001.[16]

4.C. Potential Gains from Agricultural Trade

i) Potential Agricultural Production Our data on agricultural productivity comes from the Global Agro-Ecological Zones (GAEZ) dataset compiled by the Food and Agriculture Organization (FAO) (IIASA/FAO, 2012), which covers 49 crops at the 5 arc-minute grid-cell level for the entire world. This model combines agro-climatic potential yields with climatic and soil fertility to generate measures of potential production of crops under a variety of assumptions regarding the level of inputs and the method of water supply. To avoid concerns regarding endogenous investments in irrigation or in agricultural inputs, we use the potential yields assuming low-input and rain-fed agriculture that reflect long-run production. This is similar to the methodology used for generating the measures of crop productivity in Galor and Özak (2016). We combine this crop productivity data with digitized Ethnologue (Lewis, 2009) maps of ethnolinguistic groups to construct average measures of long-run potential production of crops.

ii) Nutritional Content and Requirements In addition to data on raw crop productivity, we include data on the nutritional content of crops and on dietary nutritional requirements in order to capture incentives for exchange. Data on nutritional content of crops comes from the FAO databases (FAO, 2017a,b)) which include content of twenty-three nutrients for forty-one of the forty-nine crops included in our agricultural productivity data.

To measure the required amounts of nutrients to sustain the average adult human, we use the Dietary Reference Intakes (DRI) tables produced by the Food and Nutrition Board of the Institute of Medicine, National Academy of Sciences (Institute of Medicine, 2006). We filter these recommended intake amounts by limiting to the sixteen nutrients in our crop content data that are also listed as essential nutrients in Chipponi et al. (1982), where "The dietary *essentiality* of an

[16] See Figure B1 for the full list.

organic compound signifies that it serves an indispensable physiological function, but cannot be synthesized endogenously".

iii) Neighbours and Location Data We use the Ethnologue map to define neighbours, and construct neighbourhoods that define the scope of possible historical agricultural trade. In specifications where we consider aggregate incentives to trade and interact, we consider trade between a group and its immediate neighbours. In pairwise specifications where we look at cultural influence between two neighbouring groups, we consider trade in a neighbourhood constructed as the union of the immediate neighbours of the two groups. See Figure 6 and Figure 7 for a graphical representation of the neighbourhoods of interest.

5. Data Processing

5.A. Machine Learning Model and Loanword Prediction Accuracy

As discussed in Section 4.B, WoLD is incomplete - while it is quite a large dataset, 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 where in the world it was borrowed from. To do this we train a machine learning prediction algorithm. Given the amount of labour required to identify loanwords by hand, a machine learning algorithm is the only feasible way to accomplish this - as mentioned, PanLex includes 25,000,000 words which results in $6.25 \cdot 10^{14}$ (625 trillion) word-pairs.[17][18]. From PanLex we created a word-pair level database, and from WoLD we had a good understanding - for a subset of those words - of both whether the word-pair constituted a loanword, and the direction of transfer. We used this

[17]Running a machine learning model for a dataset of this size requires considerable computational power. To implement this we relied heavily on SciNet, the largest supercomputer in Canada. The Niagara supercomputer at SciNet is owned by the University of Toronto, and includes a homogeneous cluster of 61,200 cores. Of this we were allocated 13.5 core-years, and our machine learning model ran for approximately 43,760 core-hours to apply this prediction algorithm to every candidate word-pair in PanLex. Running on Niagara, this took approximately one week using 300 cores. For a rough comparison, this would have taken approximately 1.25 years on a standard quad-core laptop.

[18]There were some important decisions to make in order to manage computational resources, even though we had access to the supercomputer. For details on the set-up and decisions relating to navigating our computational resources, please see Appendix 2.

subset of word-pairs as a training set, and estimated for all word-pairs whether one word originated from the other.

We first needed to generate the features of word-pairs from which our classifier could generate predictions.[19] For a potential word pair we generated features that fall into three categories. First, we generated measures that indicate the similarity of a target word to its own language, as a word that is an outlier relative to its own language is more likely to have been borrowed. Second, we generated the same own-language similarity measures for the potential source word, as a word that is an outlier in its own language is less likely to be the source of a transfer. Finally, we generated features to measure the phonetic and orthographic similarity of the word-pair, as more similar words are more likely to have been part of a transfer. Finally, we include a measure of the distance between the two languages in a language family tree, to allow our classifier to take this into account when setting thresholds.

Our training data is heavily unbalanced, with, the number of loanword pairs dwarfed by the number of non-loanwords and loanwords matched to non-source words. This is a potential problem because, by estimating that there has never, in any language, ever been a loanword, the algorithm could achieve very high accuracy, but this is clearly not what we want. We used a combination of two methods to deal with this issue. The simplest method is to under-sample the heavily-represented group. The second method we use is synthetic minority over-sampling, where 'synthetic' examples of the under-represented type of observation were resampled with replacement.[20] We used these methods to generate training sets for Random Forest classifiers, as well as an Extremely Randomized Tree,

[19]These features are listed and explained in detail in Appendix 2, including a description of how orthographic and phonetic measures were implemented. It is important to note here that family tree distance is the *only* feature included that is not based on orthographic and phonetic features of the word-pair. Our classifier does not observe variables that are directly indicative of the identities of the languages themselves (such as language family, lexicon size, population, or region, etc.). This means our algorithm is classifying on the characteristics of a word-pair, and not overfitting to simple (and problematic) rules such as 'Nilo-Saharan languages borrow a lot', or 'Smaller groups tend to borrow from bigger groups'.

[20]These synthetic examples were constructed as a convex combination of nearest neighbours of the same type within feature space. See Chawla et al. (2002) for a discussion of the theory of SMOTE over-sampling and see Lemaitre et al. (2017) for the details of the exact implementation used in this paper.

which is conceptually similar but further decreases overfitting.[21] From these three classifiers we built an ensemble Voting Classifier that is approximately 98% accurate on the test set. To show that our sample size is adequate, we bootstrap the training set at different sizes, and show that the accuracy of the classifier is no longer increasing as we reach the full training set (Figure 4). To ensure that the words we identify as loanwords are not false positives, we trained a second-stage classifier to further filter the pairs the Voting Classifier described above identifies as Loanword Pairs. This second-stage classifier is approximately 91% accurate, meaning that of the loanword pairs it identifies, 9% are false positives (Figure 5).

We then applied these classifiers to the full set of possible word-pairs in the PanLex lexical data, exactly as when we constructed the training set. We took these predicted word-pairs, and where two sources were identified for the same loanword, we kept the source word with the highest probability from the second stage classifier.

i) Construction of Dependant Variables of Interest We then collapsed these word-pair level results into a language-pair level dataset. This results in two sets of variables that we use throughout the analysis. First, at the pair level, we a define measures as follows:

$$\mathcal{L}_{ij} = \frac{\#(LoanWords_j \cap Words_i)}{\#(Words_i)} \tag{6}$$

Which is the pairwise borrowing by group i from group j. The notation specifies that in the numerator we have the cardinality of the intersection between all words in i and loanwords originating from j and in the denominator we specify the cardinality of all words in i. This is simply the share of words in language i borrowed from language j.

Similarly at the societal level, we have

$$\mathcal{L}_{i} = \frac{\#(LoanWords \cap Words_i)}{\#(Words_i)} \tag{7}$$

[21] As discussed in Mullainathan and Spiess (2017) and Varian (2014), these ensemble classifiers improve out-of-sample fit by ensuring that the learning algorithm does not over-fit to the training set.

We define the more general $\mathcal{L}_i$ to include the more general $LoanWords$ in the numerator, where $LoanWords$ is all loanwords regardless of source, so that $LoanWords \cap Words_i$ is the set of loanwords regardless of source that exist in language i.

5.B. Structural Estimation of Potential Gains from Agricultural Trade

i) Subsistence Utility Model We model groups as having an incentive to increase the population they can support, where each adult requires a subsistence bundle of calories and essential nutrients. This is similar to the approach of Galor and Ozak (2015) where caloric suitability is shown to dominate agricultural suitability. We differ, however, by considering the full range of nutritional requirements allowing for nutritional complimentarity in primary agricultural products, known to have been an important driver of pre-colonial trade (Gray and Birmingham, 1970). We define a nutritional utility function which takes the form of a Cobb-Douglas production function for a healthy population:

$$U(x_0, x_1, \cdots, x_{16}) = x_0^{\alpha_0} x_1^{\alpha_1} \cdots x_N^{\alpha_{16}} \tag{8}$$

where x_0 represents daily calories, and x_1 through x_{16} are the sixteen essential micronutrients. The weights for essential nutrients, α_i, are constructed as follows: $\alpha_i = \frac{\gamma_i}{\sum_j \gamma_j}$ where we use the Daily Reference Intake (DRI) amounts as γ_i, for $i \in \{1, 2, \cdots, 16\}$. For α_0, the weight for calories, we calibrate using observed population figures. 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.[22]

We model crop production linearly, where a group chooses an allocation of land ($\vec{l}$) to different crops, and output is land allocated to a crop multiplied by productivity, where the productivity vector ($\vec{q}$) is the average from the GAEZ dataset described in Section 4.C.

$$Y(\vec{q}, \vec{l}) = [y_0(q_0, l_0), \cdots, y_{41}(q_{41}, l_{41})] = [q_0 \cdot l_0, \cdots, q_{41} \cdot l_{41}] \tag{9}$$

[22] For a discussion of model validation, please se Section 3.D.

ii) Production & Utility under Autarky We first numerically solve for autarky production, the forty-one dimensional vector $\vec{l} = [l_0, l_1, \cdots, l_{41}]$ of land share l_c allocated to crop c that maximizes the nutritional utility function in Equation 8.[23]

$$U\left(N(\vec{l})\right) \text{ such that } \sum_{0}^{41} l_c = 1 \tag{10}$$

where the nutritional content from this land allocation, $N(\vec{l}) = [N(\vec{l})_0, N(\vec{l})_1, ..., N(\vec{l})_{16}]$ is the land share allocated to each crop, multiplied by average group productivity in that crop, multiplied by the nutritional content for that crop $N(\vec{l})_c = l_c \cdot q_c \cdot \vec{n}_c$, where q_c is average group productivity in crop c and $\vec{n}_c$ is the 1x16 vector of nutrient content per unit of crop c.[24]

iii) Production & Utility under Trade To move from autarky to the trading equilibrium, we repeat the process above for optimization under autarky, but where we instead maximize aggregate utility for all groups in a neighbourhood, subject to a constraint on each group's land allocation shares.

From the set of equilibrium prices together with land allocations under trade, we solve for the budget of each group in the neighbourhood. Using the properties of our utility function as above, we know that all groups consume in the same proportions so their individual consumption will be their share of the neighbourhood's total budget times the total crop output of the neighbourhood. Using this consumption bundle, we then compute utility under trade.

We similarly compute gains from trade without specific partners, without partner groups and without GE effects.[25]

iv) Mapping from the trade model to the empirical exercise: variable definition All of this comes together into a few variables as follows. At the pairwise level:

$$c_{ij} = \frac{U_i^{FT}}{U_i^{FT-j}} \tag{11}$$

[23] see Appendix B for details on data cleaning that were undertaken prior to estimation

[24] For the computational details on how our computer solves this optimization problem, see Appendix B.

[25] For details on each of these *comparative static* computations, please see Appendix B.

Which specifies the contribution of j to the trade utility of i. Note that $c_{ij} < 1$ if j is a direct competitor to i and $c_{ij} > 1$ if i and j make natural trade partners. We plot a histogram of c_{ij} in figure 8 on the left, and find that it is centred around one. At the societal level we compare free trade to autarky:

$$c_i = \frac{U_i^{FT}}{U_i^{NT}} \tag{12}$$

Which specifies the contribution of the entire trade network to the trade utility of i. Note that $c_i = 1$ if is indifferent towards trade, but is never less than one since i always has the option of choosing not to trade. $c_i > 1$ is typical, and represents the utility gains from trade for group i. This is plotted in figure 8 on the right.

6. Empirical Approach and Results

6.A. Test 1: Are linguistic lenders linguistic borrowers?

Our goal is to empirically identify strategic reductions in linguistic distance. The hypothesis that we outlined in section 3 generates three main predictions that would allow us to identify that strategic considerations are an important determinant of linguistic distance. We start with *Test 1* which outlines the descriptive fact that linguistic borrowing is positively correlated with linguistic lending if linguistic convergence occurs for non-strategic reasons, and negatively correlated if linguistic convergence occurs for strategic reasons. We are therefore interested in a conditional correlation between $\mathcal{L}_i$ and $\mathcal{L}_j$ (which should not be interpreted causally).

$$\mathcal{L}_i = \alpha + \lambda \mathcal{L}_N + \Lambda_i + \epsilon_i \tag{13}$$

Where $\mathcal{L}_i$ is linguistic borrowing by language i, as defined in section 5.A.1 and vice versa for $\mathcal{L}_N$. We specify both in logs. Λ is a vector of controls for characteristics of language i. We include population, the size of the group relative to their neighbourhood, and the average distance between them and the centroid of their neighbours.

We focus on λ and find that those that borrow most lend the least, and vice-

versa (table 3 columns 1-3). This is a puzzle for thinking about linguistic exchange as purely a function of non-strategic convergence. If it were true that those that interacted the most had the most convergence with their neighbours, we would expect that those that borrowed the most also shared the most - that is we would expect a positive correlation in table 3 columns 1-3. These correlations are not to be interpreted causally, but as a starting point we note that even this descriptive correlation is inconsistent with what we would expect from a conceptual model of language exchange based on solely on language convergence induced by economic activity.

6.B. Test 2: Introducing the trade incentives data

One interpretation of this negative correlation is that linguistic barriers represent a type of trade cost that may prevent interaction. These barriers can be reduced through investment in language learning, which evolves over time. Our hypothesis specifies in *Test 2* that the group that benefits more from interaction will invest more heavily in learning languages, and will converge more quickly. As this group learns more, the incentives are for the other group to invest less. This is because the returns to learning another language depend negatively on how much the other group already understands your own language.

In order to test this hypothesis we need a measure of trade utility. We get this from our model of agricultural trade (see section 5.B). What we expect is asymmetric linguistic adoption, with the group benefiting the most taking on the bulk of the language adoption. We investigate this by controlling for language pair fixed effects to control for any trade costs or other determinants of trade intensity, and conditional on these things we are interested in the correlation between $\mathcal{L}_{ij}$ and c_{ij}. In the pairwise setting, borrowing by i necessarily implies lending by j. To see if the larger c implies the heavier borrowing, we can therefore estimate the following:

$$\mathcal{L}_{ij} = \omega c_{ij} + \Omega_{ij} + \epsilon_{ij} \tag{14}$$

$\mathcal{L}_{ij}$ is borrowing of group i from group j (or alternatively, lending by group j to group i) and c_{ij} is a measure of gains from trade based on crop endowments of i

and j, as defined in equation 11.

We investigate this model in 4. In column (1) we estimate equation 16, and find a positive, moderate in magnitude, but not statistically significant effect of gains from trade on linguistic borrowing. We note that in figure 8 the gains from trade at the bilateral level can be positive or negative. Curious about whether there were some non-linearities stemming from these 'negative value trade partners' that were masking possible heterogeneity in the effect, we also examined a model similar to 16, but with a quadratic in $c_i j$. That model is estimated in column (2), where we estimate a strong U-shape relationship between gains from trade and linguistic borrowing. We see that the minimum of the U is a fairly large negative number (the absolute value is double the mean of c_{ij} - so seemingly for much of the distribution the effect of gains from trade is positive.

We take a closer look at non-linear gains from trade by splitting the variable into a number of binary variables at various parts of the gains from trade distribution. Indeed, we again find strong non-linearity. We find that linguistic borrowing is strongest ($\omega > 0$) for groups that are most influenced by the existence of the other in either direction, but in fact the point estimate for being in the bottom 25th percentile of the distribution is larger than that for being in the bottom 10th percentile. So, it does not appear true that the effect is simply driven by outliers. We find that the 'turning-point' of the U is near the point of indifference.

So, within the pair, the one that borrows less is always the more indifferent member of the pair. The fact that relatively higher gains from trade result in relatively higher linguistic borrowing is expected by *Test 2*. These societies have the largest incentive to learn the language of the other. The fact that economic rivals also exhibit high linguistic exchange is more surprising, however there are a number of mechanisms from the literature that might explain this.

First, there is evidence that more similar groups are more likely to engage in conflict (Spolaore and Wacziarg, 2016). In this case, the rivals are economic competitors in the sense that their geographic endowments make producing the same bundle of goods optimal. It is easy to see how either would benefit by conquering the other both because it would allow for less competitive pricing, but also because production on the new land would be most similar to production on the old land. The costs to conquering would be lowest, and the benefits highest.

Alternatively it could be a story similar to Michalopoulos (2012a) where those that produce similar goods interact more due to the gains from information exchange. In either case we find that both economic competitors and economic trade partners converge linguistically the most.

We can get away from the challenge with interpreting the 'negative trade incentives' in a societal-level analysis. This not only has the advantage of a more intuitive comparison of trade to autarky (where nobody is ever worse off by the option to trade) but it also allows for a separate analysis of borrowing and lending - which is not possible in the bilateral framework since one group's borrowing is necessarily the other group's lending. Our hypothesis however makes strong predictions on the societal-level differences between lending and borrowing, and we investigate that next.

6.C. Test 3: Trade incentives, linguistic borrowing and linguistic lending

The premise of the argument outlined in *Test 3* is similar to the stylized fact that we began with. Under non-strategic linguistic convergence we should expect that more trade benefit leads to both more borrowing *and* more lending, since one party having a high gain from trade means that on average the pair is more likely to interact, and interaction leads to both linguistic exchange in both directions (equally if $\theta = 1/2$ in our framework). Under pure strategic linguistic convergence (e.g. $\theta = 1$), a group with a high benefit of trading should invest in reducing linguistic distance with any potential partner such that trade becomes mutually beneficial. In this case the vast majority of the borrowing costs would be born by the group that was enthusiastic about the trade relationship, while the group that was relatively indifferent would lend but not borrow.

We look for the relationship between gains from trade and linguistic borrowing and lending at the societal level by estimating the following equations:

$$\mathcal{L}_i = \alpha + \gamma c_i + \Gamma_i + \epsilon_i \tag{15}$$

$$\mathcal{L}_N = \alpha + \mu c_i + M_i + \epsilon_i \tag{16}$$

Where c_i is as in 12, $\mathcal{L}_N$ is societal lending, as in equation **??**, $\mathcal{L}_i$ is as in 13 and Γ_i and M_i are both vectors of societal level controls.

Test 3 suggests that if θ is high, so that strategic considerations were important, then $\gamma > \mu$. However, if strategic considerations are not important, then $\gamma = \mu > 0$. Table 6 implements this test. In Column (1) we estimate γ, in column (2) we estimate μ and in column (3) we estimate $\gamma - \mu$. We find, consistent with the idea that θ is high and strategic considerations are important, that gains from trade influence borrowing but we fail to find evidence that gains from trade influence lending in a similar way.

Of course, failure to find a precise effect of gains from trade on linguistic lending could be either because there is truly not a significant correlation, or it could be because we simply estimate a very noisy estimate. Our hypothesis suggests that if $\theta = 1/2$ then $\gamma = \mu$, so using that as a guide we try to assess whether we are actually finding a small effect by testing whether we can rule out that μ is as large as γ. This estimate appears in column (3) - we find that the effect on lending is statistically significantly smaller than the effect on lending. If we take the model seriously, we can therefore rule out *pure* non-strategic convergence of language.

Interestingly, in these group-level regressions we find a precise positive effect of gains from trade on borrowing that we were unable to precisely estimate linearly in the pairwise regression. We mentioned that one reason we may fail to find a linear effect in the pairwise regression was the way we defined gains from trade allowed for negative values for groups that produced similar crops. At the group level, we define gains from trade as being the ratio of utility under trade to utility under autarky. In this case we do not find the same challenge as in the pairwise regression, since this definition rules out negative gains from trade.

To see this, consider again figure 8 for histograms of both variables, and notice that at the bilateral level the distribution to the left of 1 is non-trivial, indicating that the society would be better off in a trading network without that particular partner, while in the panel on the right none of the distribution is to the left of one, indicating that while some groups are indifferent between trade and autarky, trade does not make groups worse-off than under autarky.

We therefore look again for non-linear effects in the effect that we estimate in table 6 column (1). In table 7 we find a relatively monotonic and increasing

relationship between gains from trade and linguistic borrowing. This reinforces that the positive effects we had previously been finding at the pairwise level were driven by incentives to interact for reasons other than trade.[26]

7. Conclusion

In this paper we take a close look at the endogeneity of linguistic distance. We hypothesize two sources of endogeneity, one that has been proposed in the past based on directional linguistic drift, and another based on strategic investments. We find that controlling for interaction intensity, greater linguistic borrowing is correlated with greater linguistic exchange, which supports the idea of linguistic drift, as empirically identified in past studies (e.g. Michalopoulos (2012a), Dickens (2019), Ashraf and Galor (2013)). We also, however, find considerable evidence of asymmetric linguistic exchange, which is predicted by an exogenous measure of economic leverage.

This finding has important implications for the literature on the economic consequences of linguistic distance or ethnolinguistic diversity. Our findings suggest that while it remains possible that linguistic diversity is bad for development outcomes, it is likely true that economic context determines linguistic diversity. Furthermore, if the regions of the world where groups have the least incentives to engage with each other economically cause them also to be the ones where linguistic distance or fractionalization is highest, then the implications of ethnolinguistic diversity may be less drastic than the literature currently suggests.

Of course, not all of the literature suggests that diversity brings with it dire economic consequences. Notably Ashraf and Galor (2013) claim that a moderate amount of diversity is optimal, using migration distance as an instrument for diversity. It seems clear that physical distance is likely to be associated with more homogeneity simply because if one leaves a group they will be more alone. However, this may also lead to greater linguistic distance which means that the instrument is actually capturing the various dimensions of diversity itself in a non-linear way.

[26] For a sense of completeness, we estimate the same non-linear effects for the lending equation, but find, unsurprisingly, that there are no effects across the distribution.

Altogether, we argue that a re-analysis of the effects of diversity on economic outcomes would benefit from a more nuanced view of the origins of linguistic diversity, taking into consideration the fact that it is not determined independently of economic outcomes. However, we leave such analysis to future work.

References

Pelle Ahlerup and Ola Olsson. The roots of ethnic diversity. *Journal of Economic Growth*, 17(2):71–102, 2012.

Alberto Alesina and Eliana La Ferrara. Ethnic diversity and economic performance. *Journal of Economic Literature*, 43(3):762–800, 2005.

Alberto Alesina, Arnaud Devleeschauwer, William Easterly, Sergio Kurlat, and Romain Wacziarg. Fractionalization. *Journal of Economic Growth*, 8:155–94, 2003.

Alberto Alesina, Paola Giuliano, and Nathan Nunn. On the origins of gender roles: Women and the plough. *The Quarterly Journal of Economics*, 128(2):469–530, 2013.

Alberto Alesina, Stelios Michalopoulos, and Elias Papaioannou. Ethnic inequality. *Journal of Political Economy*, 124(2):428–488, 2016.

Yann Algan and Pierre Cahuc. Inherited trust and growth. *American Economic Review*, 100(5):2060–92, 2010.

Yann Algan, Thierry Mayer, Mathias Thoenig, et al. The economic incentives of cultural transmission: Spatial: Spatial evidence from naming patterns across france. Technical report, 2013.

Quamrul Ashraf and Oded Galor. Genetic diversity and the origins of cultural fragmentation. *American Economic Review*, 103(3):528–33, 2013.

David Atkin. The caloric costs of culture: Evidence from indian migrants. *American Economic Review*, 106(4):1144–81, April 2016.

Alberto Bisin and Thierry Verdier. The economics of cultural transmission and the dynamics of preferences. *Journal of Economic theory*, 97(2):298–319, 2001.

Arthur Blouin. Culture and contracts: The historical legacy of forced labour. *University of Toronto Working Paper*, 2016.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

J X Chipponi, J C Bleier, M T Santi, and D Rudman. Deficiencies of essential and conditionally essential nutrients. *The American Journal of Clinical Nutrition*, 35(5):1112–1116, 1982.

Klaus Desmet and Romain Wacziarg. The cultural divide. *National Bureau of Economic Research Working Paper Series*, No. 24630, 2018.

Klaus Desmet, Ignacio Ortuño-Ortín, and Romain Wacziarg. The political economy of linguistic cleavages. *Journal of Development Economics*, 97(2):322–338, 2012.

Andrew Dickens. The Historical Roots of Ethnic Differences: The Role of Geography and Trade. Working Papers 1901, Brock University, Department of Economics, June 2019. URL https://ideas.repec.org/p/brk/wpaper/1901.html.

William Easterly and Ross Levine. Africa's growth tragedy: policies and ethnic divisions. *The Quarterly Journal of Economics*, 112(4):1203–1250, 1997.

Joan Esteban, Laura Mayoral, and Debraj Ray. Ethnicity and conflict: An empirical study. *American Economic Review*, 102(4):1310–42, 2012.

FAO. *FAO/INFOODS Food Composition Database for Biodiversity Version 4.0, BioFoodComp4.0*. FAO. Rome, Italy, 2017a.

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

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

Oded Galor and Omer Ozak. Land productivity and economic development: Caloric suitability vs. agricultural suitability. *Working Papers - Brown University, Department of Economics*, (2015-5), 2015.

Oded Galor and Ömer Özak. The agricultural origins of time preference. *American Economic Review*, 106(10):3064–3103, 2016.

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

Henry E Hale. Explaining ethnicity. *Comparative Political Studies*, 37(4):458–485, 2004.

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

Martin Haspelmath and Uri Tadmor, editors. *WOLD*. Max Planck Institute for Evolutionary Anthropology, 2009b. URL https://wold.clld.org/.

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

Institute of Medicine. *Dietary Reference Intakes: The Essential Guide to Nutrient Requirements*. The National Academies Press, 2006.

Matthew A. Jaro. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84(406):414–420, 1989.

István Kónya. Modeling cultural barriers in international trade. *Review of International Economics*, 14(3):494–507, 2006.

M. Lalee, J. Nocedal, and T. Plantenga. On the implementation of an algorithm for large-scale equality constrained optimization. *Journal on Optimization*, 8 (3):682–706, 1998.

Guillaume Lemaitre, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017.

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

Sara Lowes, Nathan Nunn, James A. Robinson, and Jonathan Weigel. The evolution of culture and institutions: Evidence from the kuba kingdom. *Econometrica*, 2016.

Stelios Michalopoulos. The origins of ethnolinguistic diversity. *American Economic Review*, 102(4):1508–39, 2012a.

Stelios Michalopoulos. The origins of ethnolinguistic diversity: Theory and evidence. *Brown University Working Paper*, 2012b.

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori S. Levin. PanPhon: A resource for mapping IPA segments to articulatory feature vectors. In *COLING*, 2016.

Sendhil Mullainathan and Jann Spiess. Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106, 2017.

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

Nathan Nunn and Leonard Wantchekon. The slave trade and the origins of mistrust in africa. *American Economic Review*, 101(7):3221–3252, 2011.

Radim Rehurek and Petr Sojka. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50. ELRA, 2010.

Enrico Spolaore and Romain Wacziarg. War and relatedness. *The Review of Economics and Statistics*, 98(5):925–939, 2015.

Enrico Spolaore and Romain Wacziarg. Ancestry, language and culture. In Victor Ginsburgh and Shlomo Weber, editors, *The Palgrave Handbook of Economics and Language*. Palgrave Macmillan UK, 2016. doi: 10.1007/978-1-137-32505-1_7.

Enrico Spolaore and Romain Wacziarg. The political economy of heterogeneity and conflict. *National Bureau of Economic Research Working Paper Series*, No. 23278, 2017.

Morris Swadesh. Salish internal relationships. 16(4):157–167, 1950.

Guido Tabellini. The scope of cooperation: Values and incentives. *The Quarterly Journal of Economics*, 123(3):905–950, 2008.

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

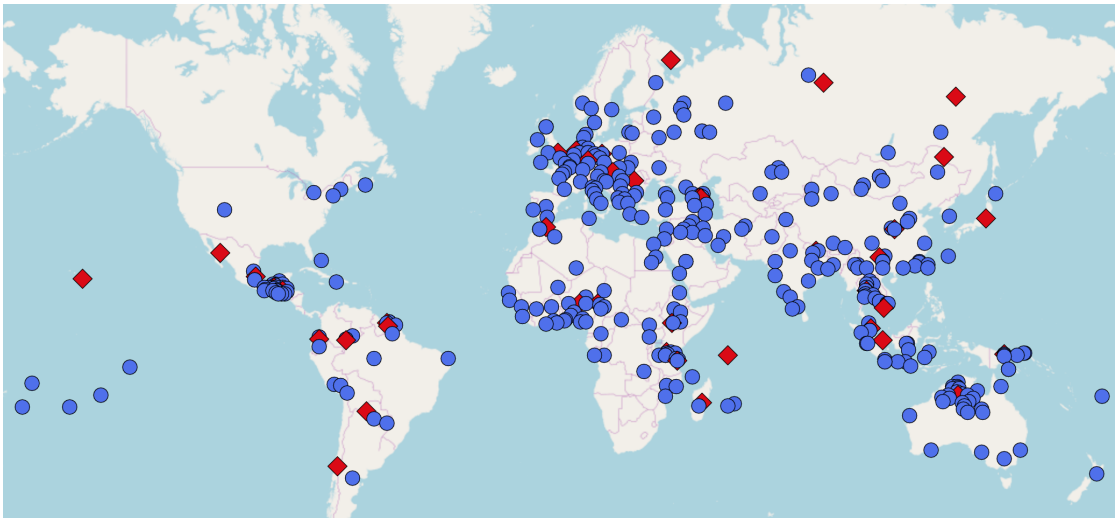
Hal R. Varian. Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2):3–28, 2014.

Soren Wichmann, Eric W. Holman, and Cecil H. Brown. The ASJP database (version 17). 2016.

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

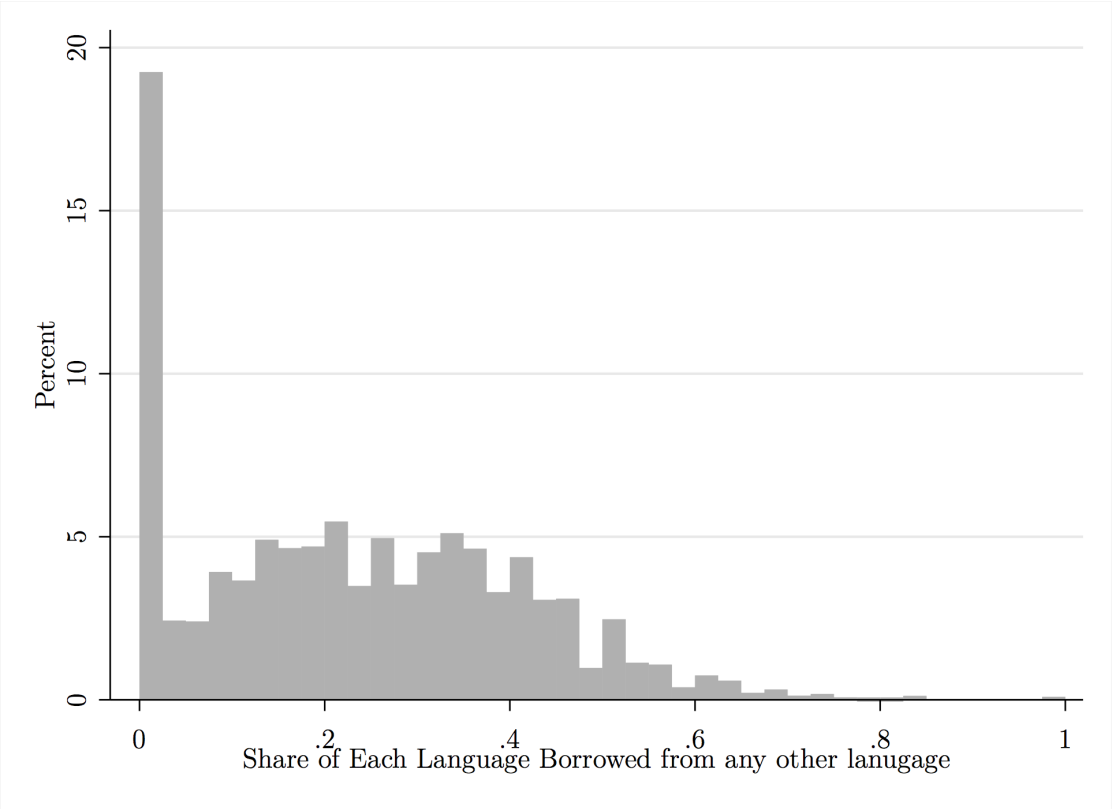
Main Figures

Figure 1: Map of WoLD language groups



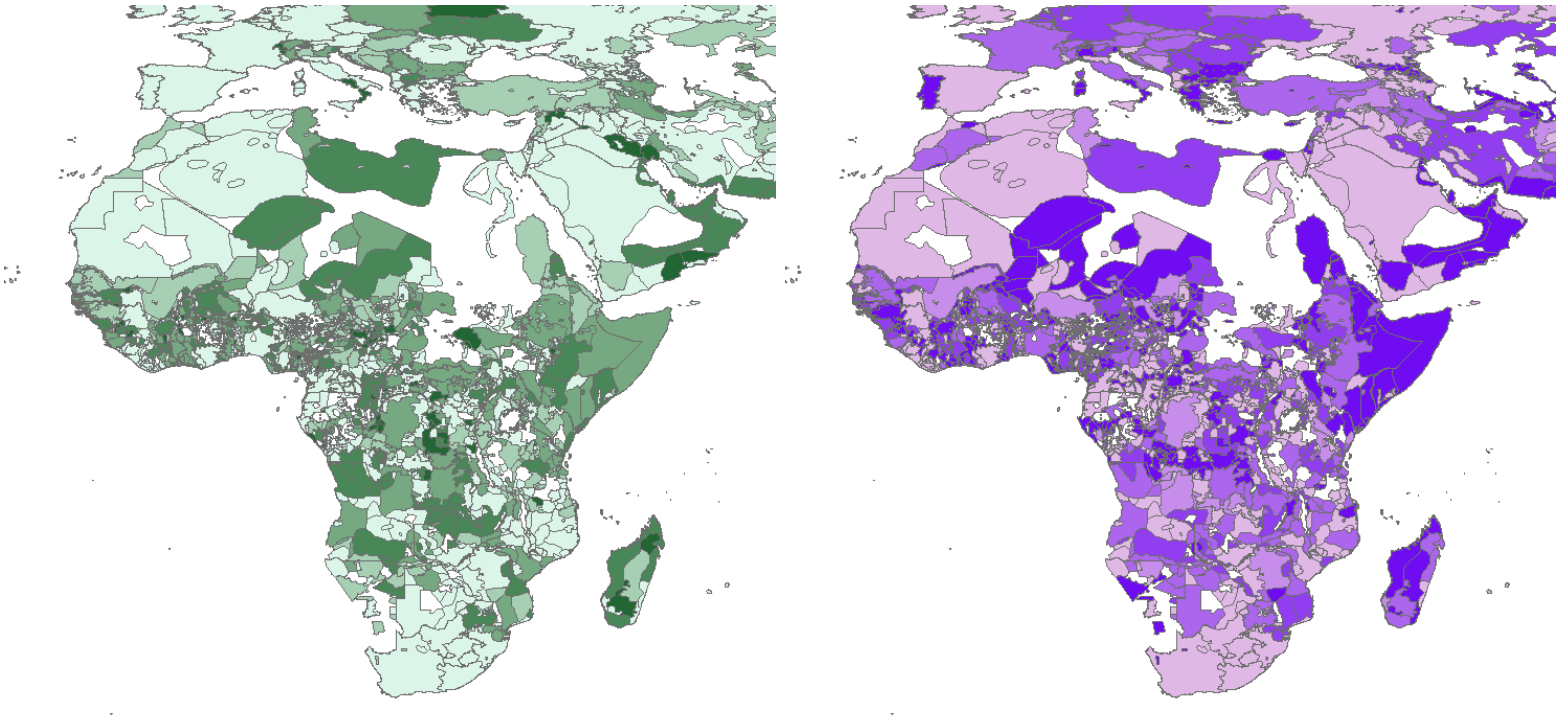
Note: This map shows each of the borrower and lender languages in the WoLD dataset. The blue dots represent lending languages while the red diamonds represent borrowing languages. In total there are 395 languages mapped, 41 of which are borrowers and 369 are lenders (this does not add to 395 because 15 languages are both lenders and borrowers).
Source: World Loanwords Database: https://wold.clld.org/language. Last Accessed October 10, 2019 5:00pm EST.

Figure 2: Histogram of Language Borrowing



Note: The figure shows the raw-data of the main dependent variable used throughout the paper, the share of any given language borrowed from one of their neighbours. 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.

Figure 3: Map of Linguistic Borrowing (left) and Gains From Trade (right)

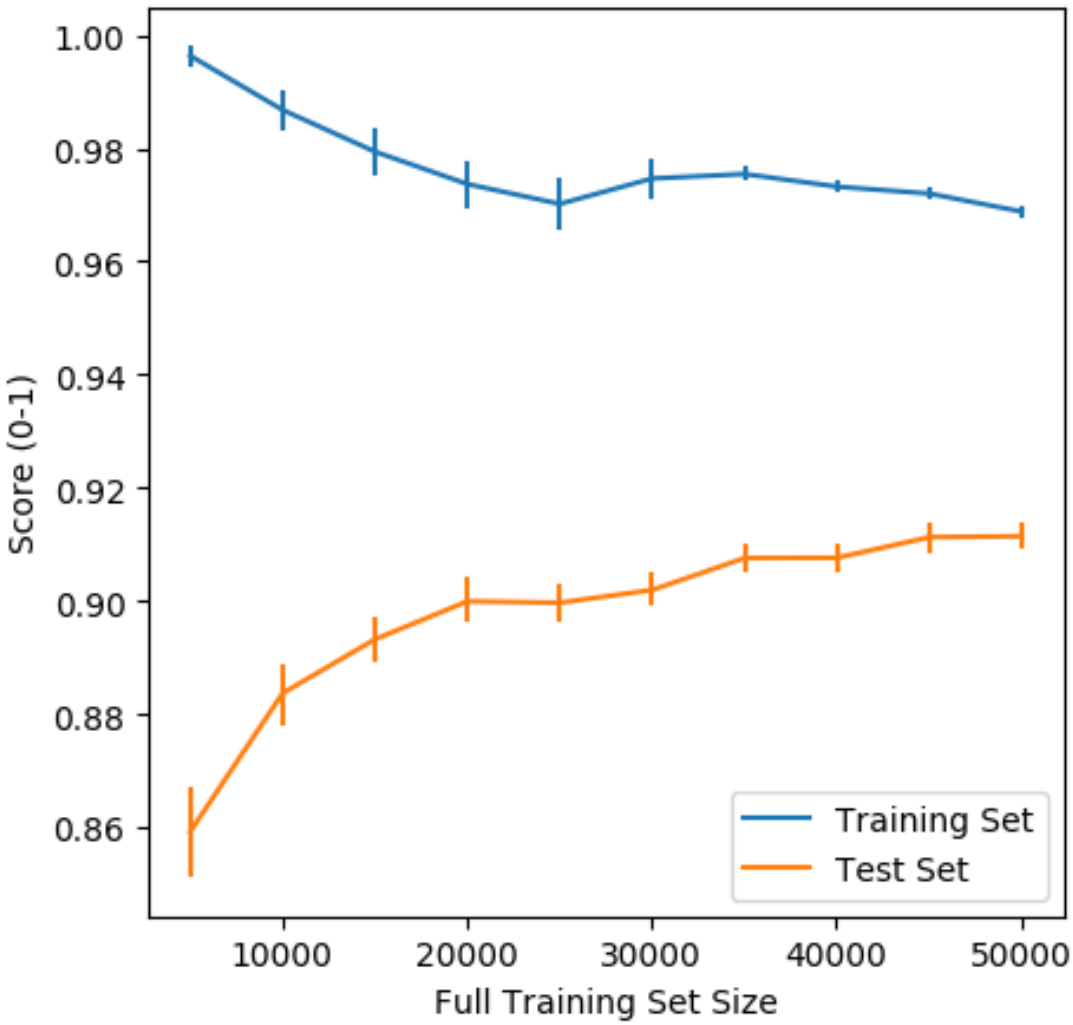


Note: These maps illustrated the main evidence presented in the text that linguistic distance is endogenous. On the left we map linguistic borrowing. Darker shades represent more borrowing in those regions. On the right we show gains from trade. Darker shades represent a larger gap in utility from trade relative to utility under autarky. The gains from trade are constructed exclusively from soil characteristics. We zoom in on Africa - where much of the diversity literature has focussed its attention - to show that there is a high correlation exogenously determined gains from trade and linguistic distance. This evidence is consistent with both linguistic drift as a byproduct of exchange and with strategic investments in language distance reduction.
Source: Author constructed. Data sources are described in the text.

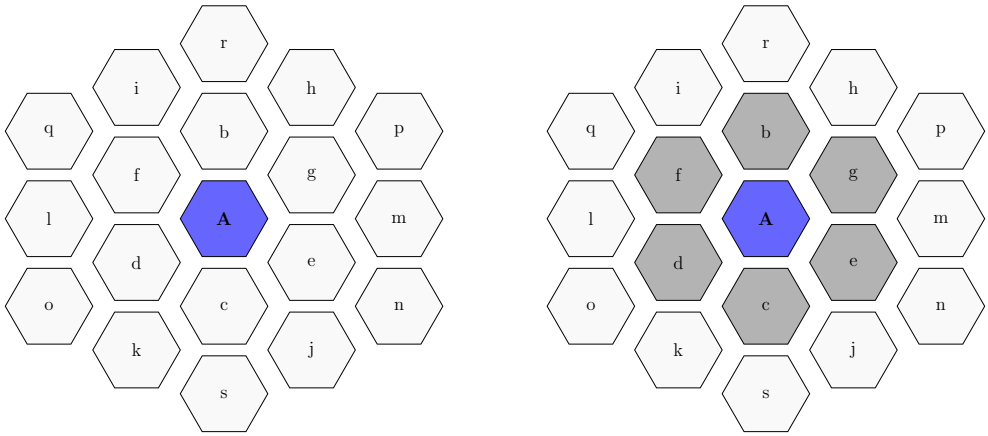
Figure 4: Accuracy of Voting Classifier

Note: The figure shows the accuracy of the machine learning algorithm by training set size. On the y-axis we show the share of words classified correctly by the algorithm. We contemplated adding observations to the training set, but the graphs suggest that adding additional words has not made a marginal improvement in accuracy for the past 10,000 words or so. Furthermore, we see that accuracy rates are quite high. In the test set we are classifying over 98% of words correctly for loanwords, getting the direction of borrowing right over 92% of the time, and are very rarely wrongly classifying a non-loanword as a loanword.
Source: Author constructed. Data sources are described in the text.

Figure 5: Accuracy of Phase Two Classifier



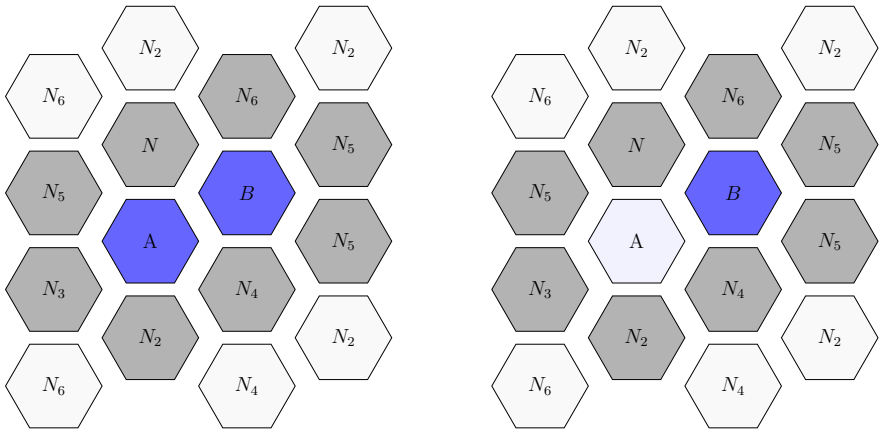
Note: This graph shows the results of the second-stage classifier described in the text. The second stage classifier classifies words correctly over 90% of the time, which includes all types of errors: correctly identifying a loanword but getting the direction wrong, correctly identifying a loanword but getting the source wrong, or incorrectly identifying a loanword.
Source: Author constructed. Data sources are described in the text.



(a) Autarky **(b)** Full Neighbourhood Trade

Figure 6: Neighbourhoods Used for Language-Level Trade Incentive

Note: This figure illustrates the counterfactual neighbourhoods used for our structural estimates of gains from trade at the language level. A dark shaded polygon indicates a group that is included in the given counterfactual neighbourhood. To generate aggregate predicted trade incentives for group A, we compare Autarky in panel a) (where Group A's consumption is its' own production) to a trading neighbourhood of A and immediate neighbours in panel b).

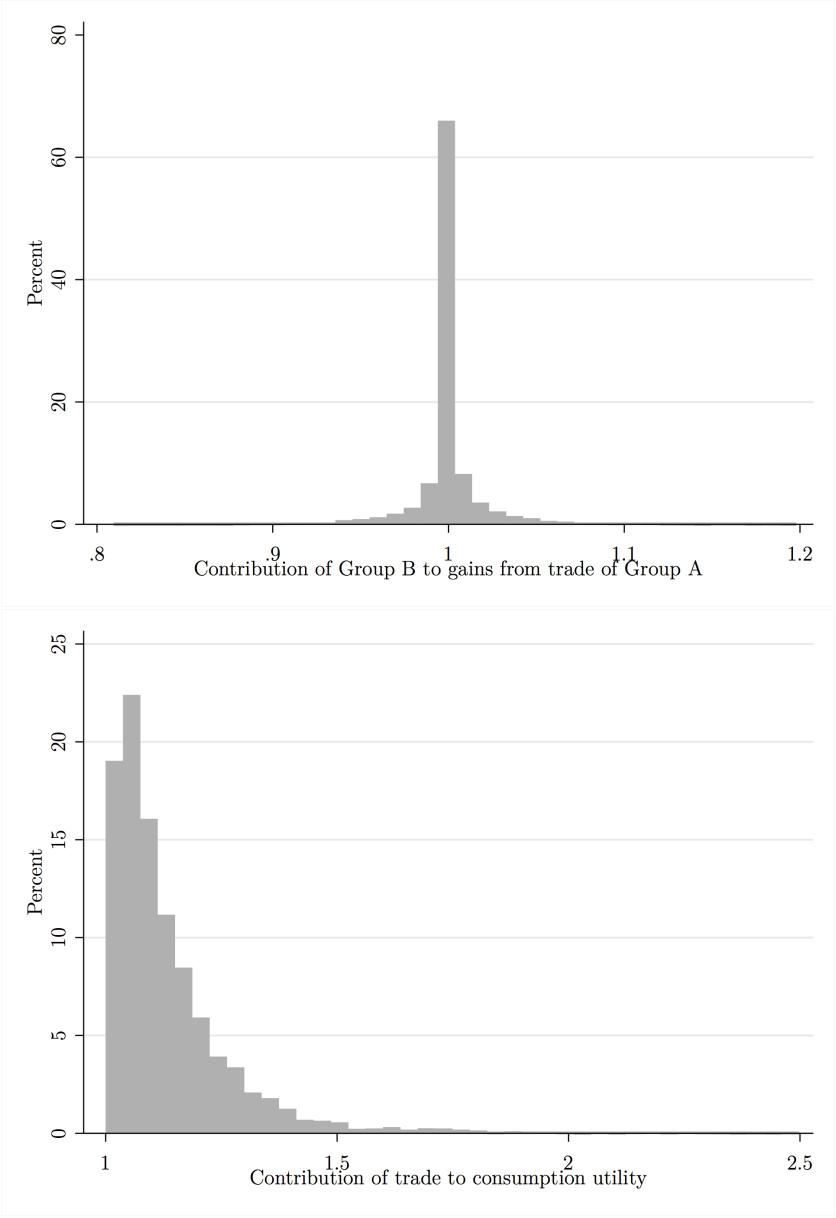


(a) Full Trade **(b)** Drop A

Figure 7: Minimal Neighbourhoods Used for Pairwise Specifications

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 group that is included in the given counterfactual neighbourhood. In panel a) we show the full minimal neighbourhood 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 i is dropped from the neighbourhood. By comparing j's utility under the two scenarios, we measure how much they would lose from i not being present, which we use as our measure of group i's leverage over group j.

Figure 8: Histogram of Gains From Trade



Note: The figure shows histograms of the output of the trade model. On the top we have the bilateral measure of gains from trade and on the bottom we have the societal level measure. Since on the top we compare full-trade with a partner to full trade without a partner, the measure can be greater or less than 1, and in fact it seems centred around 1 - indicating that most societies are indifferent towards the inclusion of most of their neighbours in their trading network. A value of less than one indicates that the society is worse-off due to the existence of their neighbour in their trading network - i.e. the societies are economic competitors. A value of greater than one indicates that we expect the societies have a profitable trading relationship. On the bottom we show our societal level measure which compares trade with the network as a whole to autarky. Here we only see values greater than one because societies always have the option of behaving as if there exists no trading network. By far most societies have values less than two. A value of two indicates that the society is twice as well off with the existence of their trading network relative to autarky.
Source: Author constructed. Data sources are described in the text.

Tables

Table 1: Descriptive Statistics

Variable	**Observations** (1)	**Mean** (2)	**SD** (3)	**Min** (4)	**Max** (5)
		Panel A: Language Data			
Share of Language Borrowed (overall)	11,926	23%	17%	0	100%
Share of Language Borrowed from a given other Language	11,926	0.28%	2%	0	8%
		Panel B: Linguistic Homeland Characteristics			
Population (1,000)	11,708	8,099	66,898	0	871,558
Arable Hectares (1,000)	11,708	17,439	156,356	0.2	2,154,896
Distance to Neighbour	11,708	225	461	0	6,841
		Panel C: Trade Data			
Utility Under Trade	11,708	2.64	1.72	0.003	15.27
Utility Under Autarky	11,708	2.39	1.57	0.00012	9.97
Utility Under Trade / Utility Under Autarky	11,708	1.18	1.89	0.06	123.34
Trade Utility without a neighbour	11,708	2.64	1.72	0.0008	15.29
Utility Under Trade / Trade Utility without a neighbour	11,708	1.006	0.33	0.072	36.48

Note: The table shows descriptive statistics for the main variables used throughout the empirical analysis. We have word-level data for 11,926 society-pairs, 11,708 of which can be matched to the Ethnologue data. Notably linguistic sharing is substantial, with the average society having borrowed about 23% or their language from their neighbours. The population data comes directly from the Ethnologue, while the Arable Hectares is constructed through a location match of the Ethnologue and the FAO GAEZ data. Distance to neighbour is author constructed based on the Ethnologue centroids. The Utility data all comes from the trade model, which is described in section 5.B. Utility under trade and utility under autarky have meaningless units, but the share of these variables suggests that on average societies are 18% better off due to trade, and on average almost 1% better off due to the existence of any given neighbour.

Table 2: Structural Model Diagnostics

	Dep. Var: Actual Population		
Model Generated Maximum Supportable Population under:	(1)	(2)	(3)
Autarky 1200kcal	0.04***		0.96***
	(0.001)		(0.008)
Trade 1200kcal		0.03***	0.96***
		(0.001)	(0.01)
$\sqrt{Autarky \cdot Trade}$			-1.9***
			(0.02)
R^2	0.14	0.07	0.63
N	11,780	11,780	11,780

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 even collect population) and all of the population data is from the Ethnologue (who do not even collect nutritional data). We conclude from column three that our trade model is capturing the variation that we would like reasonably well.

Table 3: Test 1: Are Borrowers also Lenders?

	Dep. Var.: log(Ling. Borrowing)		
	(1)	(2)	(3)
log(Linguistic Lending)	-0.056***	-0.056***	-0.053***
	(0.013)	(0.013)	(0.013)
Population	No	No	No
Area Share	No	Yes	Yes
Distance Polynomial	No	No	Yes
R^2	0.004	0.005	0.02
N	2,995	2,995	2,995

Note: *, **, *** denote significance at 10%, 5% and 1% respectively. The unit of observation is a language. Robust standard errors are reported. We observe that societies that lend a lot borrow very little. This could only emerge under strategic language investment.

Table 4: Test 2a: Loanwords and Trade at the Relationship Level

Dependant Variable:	Linguistic Borrowing	
	(1)	(2)
Crop-based potential gain from trade (c_{ij})	1.17	-107.196***
	(0.745)	(28.05)
(Crop-based potential gain from trade)^2 (c_{ij}^2)		52.19***
		(13.55)
Value where $\delta\mathcal{L}_{ij}/\delta c_{ij} = 0$		-2.8
p-value for H_0 : argmin $c_{ij} = 0$		0.02
Mean of crop-based potential gain from trade	1.006	
Standard error of crop-based potential gain from trade	0.0004	
Relationship Fixed Effects	Yes	Yes
R^2	0.517	0.517
N	11,708	11,708

Note: *, **, *** denote significance at 10%, 5% and 1% respectively. The unit of observation is a language-pair. Standard errors are two-way clustered at the lending and borrowing language.
This table examines language exchange and economic incentives at the pairwise level. For a particular trading pair linguistic exchange is lowest when partners are economically indifferent towards each other. Linguistic exchange is highest when partners are either competitors or good trading partners. The positive correlation with potential trading partners was what we expected - as more interaction is likely to mean more linguistic exchange. It may be surprising to see that economic competitors also feature a high degree of linguistic exchange. This could be because there are incentives to discuss coalitions or technology, or because there are incentives to go to war. We interpret this as also being related to increased interaction from non-trade economic activities.

Table 5: Prediction 2b: Loanwords and Trade at the Relationship Level (distribution effects)

	Dependant Variable: Linguistic Exchange				
	(1)	(2)	(3)	(4)	(5)
Potential gains from trade < 10th pctl (trade competitors)	0.219**				
	(0.094)				
Potential gains from trade < 25th pctl (trade competitors)		0.279**			
		(0.098)			
25th pctl < Potential gains from trade < 75th pctl (indifferent)			-0.367***		
			(0.120)		
Potential gains from trade > 75th pctl (trade partners)				0.32**	
				(0.132)	
Potential gains from trade > 90th pctl (trade partners)					0.222**
					(0.109)
Relationship Fixed Effects	Yes	Yes	Yes	Yes	Yes
R^2	0.515	0.516	0.518	0.517	0.516
N	11,708	11,708	11,708	11,708	11,708

Note: *, **, *** denote significance at 10%, 5% and 1% respectively. The unit of observation is a language-pair. Standard errors are two-way clustered at the lending and borrowing language.
This table examines language exchange and economic incentives at the bilateral level. For a particular trading pair linguistic exchange is lowest when partners are economically indifferent towards each other. Linguistic exchange is highest when partners are either competitors or good trading partners. The positive correlation with potential trading partners was what we expected - as more interaction is likely to mean more linguistic exchange. It may be surprising to see that economic competitors also feature a high degree of linguistic exchange. This could be because there are incentives to discuss coalitions or technology, or because there are incentives to go to war. We interpret this as also being related to increased interaction from non-trade economic activities.

Table 6: Test 3: Asymmetry in the effect of gains from trade on borrowing vs. lending

Dependant Variable:	log(Borrowing) (1)	log(Lending) (2)	(3)
Crop-based potential gain from trade ($log(c_i)$)	0.042***	-0.012	
	(0.015)	(0.017)	
H_0 : borrowing effect - lending effect			0.054**
χ^2 for H_0			5.28
Population	Yes	Yes	Yes
Mean population of neighbours	Yes	Yes	Yes
Area	Yes	Yes	Yes
Mean area of neighbours	Yes	Yes	Yes
Land Quality	Yes	Yes	Yes
R^2	0.0249	0.193	.
N	2,995	2,995	2,995

Note: *, **, *** denote significance at 10%, 5% and 1% respectively. The unit of observation is a language group. Standard errors are two-way clustered at the lending and borrowing language. This table examines language exchange and economic incentives at the group level. For a particular group linguistic borrowing is lowest when partners are economically indifferent towards each other, but lending is unaffected by economic incentives to trade.

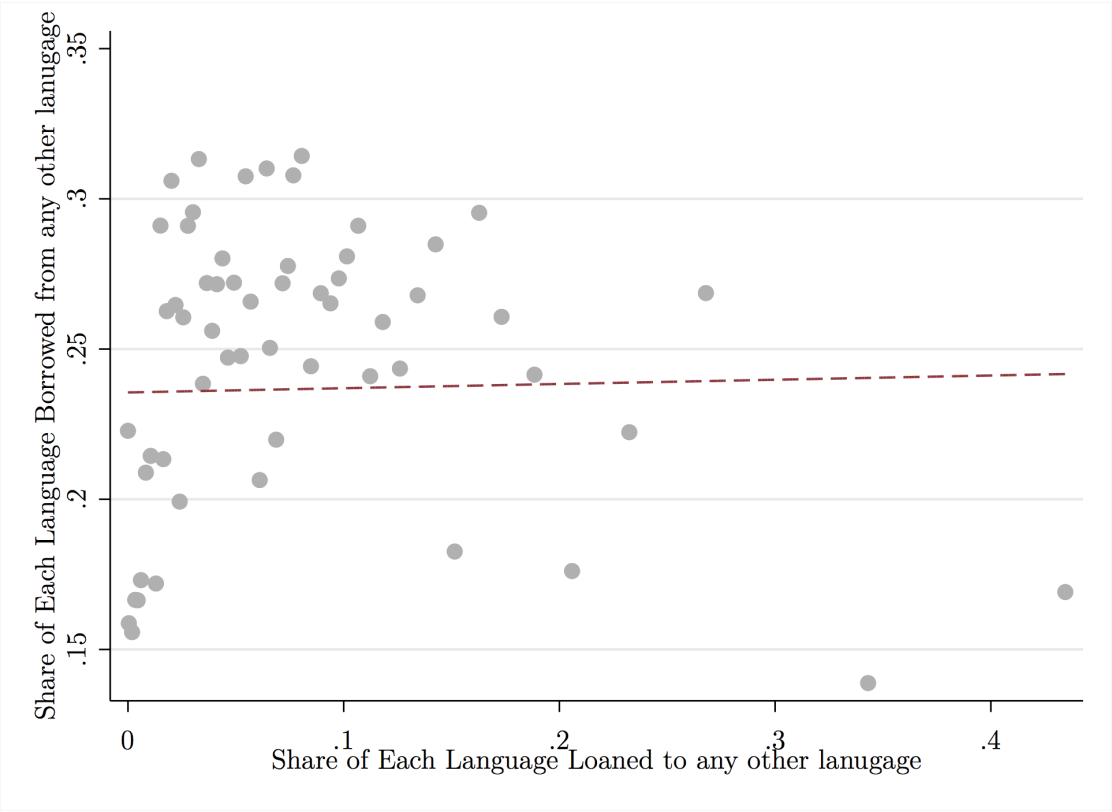
Table 7: Do gains from trade matter for borrowing non-linearly?

	Panel A Dependant Variable: Linguistic Borrowing				
	(1)	(2)	(3)	(4)	(5)
Gains from trade < 10th pctl	-0.020*				
	(0.012)				
Gains from trade < 25th pctl		-0.017**			
		(0.008)			
75th pctl > Gains from trade > 25th pctl			0.0001		
			(0.006)		
Gains from trade > 75th pctl				0.012*	
				(0.007)	
Gains from trade > 90th pctl					0.025***
					(0.009)
Population	Yes	Yes	Yes	Yes	Yes
Mean population of neighbours	Yes	Yes	Yes	Yes	Yes
Area	Yes	Yes	Yes	Yes	Yes
Mean area of neighbours	Yes	Yes	Yes	Yes	Yes
Land Quality	Yes	Yes	Yes	Yes	Yes
R^2	0.023	0.024	0.022	0.023	0.025
N	2,995	2,995	2,995	2,995	2,995

Note: *, **, *** denote significance at 10%, 5% and 1% respectively. The unit of observation is a language group. Standard errors are two-way clustered at the lending and borrowing language. This table examines societal level borrowing. We find, not surprisingly, that more economic incentives to trade mean more linguistic borrowing. This is consistent with both language convergence due to trade, and with strategic language investments to induce trade. It is however, inconsistent with exogenous linguistic distance. We do not find the same non-linearities as in the pairwise empirical model.

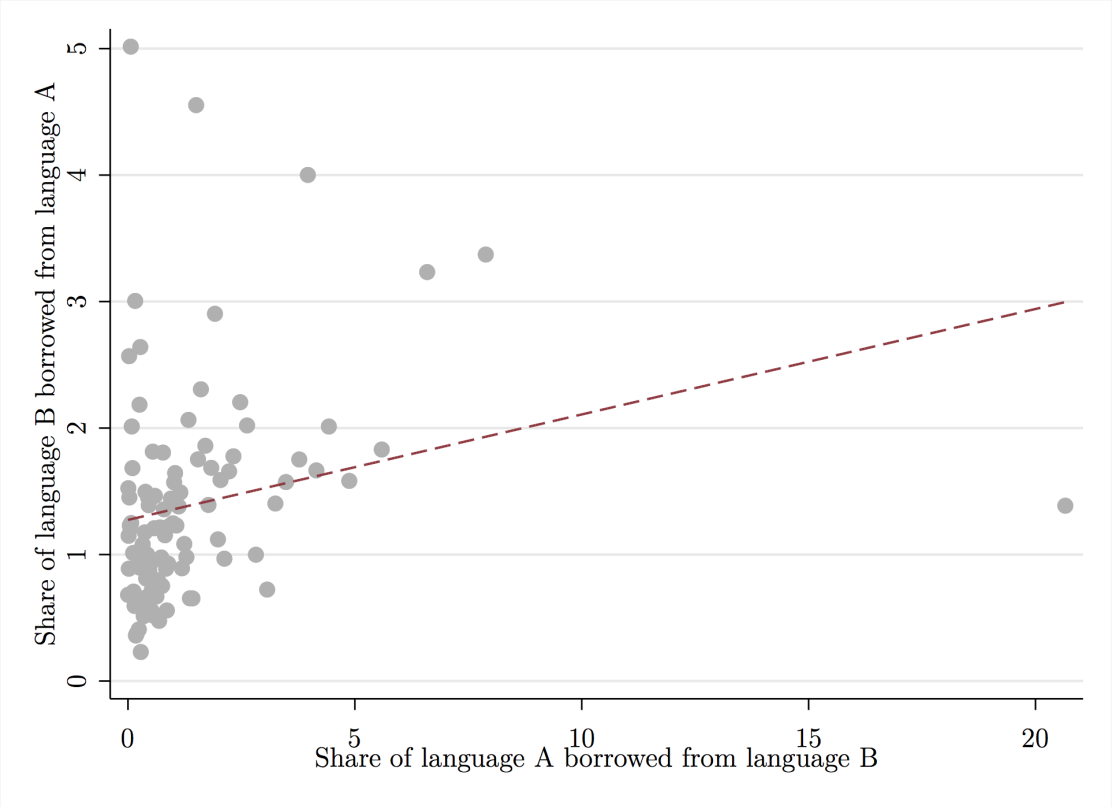
Appendix A. Additional Evidence (TO BE PUBLISHED ONLINE)

Figure A1: Relationship between aggregate borrowing and aggregate lending



Note: This figure shows an unconditional bin-scatter plot of lending and borrowing between languages at the societal level. There appears to be a mild positive relationship, but certainly not as strong as one might expect based purely on linguistic convergence.
Source: author constructed, see text for data sources.

Figure A2: Relationship between pairwise borrowing and pairwise lending



Note: This figure shows an unconditional bin-scatter plot of lending and borrowing between languages at the bilateral level. There appears to be a positive relationship. This would not be expected if only strategic lending drove language exchange, and it is consistent with linguistic convergence.
Source: author constructed, see text for data sources

Table A1: Do gains from trade matter for lending non-linearly?

	Dependant Variable: Linguistic Lending				
	(1)	(2)	(3)	(4)	(5)
Gains from trade < 10th pctl	0.005				
	(0.013)				
Gains from trade < 25th pctl		0.00025			
		(0.009)			
75th pctl > Gains from trade > 25th pctl			0.007		
			(0.007)		
Gains from trade > 75th pctl				-0.009	
				(0.008)	
Gains from trade > 90th pctl					-0.008
					(0.01)
Population	Yes	Yes	Yes	Yes	Yes
Mean population of neighbours	Yes	Yes	Yes	Yes	Yes
Area	Yes	Yes	Yes	Yes	Yes
Mean area of neighbours	Yes	Yes	Yes	Yes	Yes
Land Quality	Yes	Yes	Yes	Yes	Yes
R^2	0.019	0.019	0.020	0.020	0.019
N	2,995	2,995	2,995	2,995	2,995

Note: *, **, *** denote significance at 10%, 5% and 1% respectively. The unit of observation is a language group. Standard errors are two-way clustered at the lending and borrowing language. This table examines non-linearities in societal level lending. We find no relationship between more economic incentives to trade mean more linguistic borrowing. This is inconsistent with language convergence due to trade. We do not find the same non-linearities as in the pairwise empirical model.

Appendix B. Lexical Data & Loanword Prediction (TO BE PUBLISHED ONLINE)

B.1. Overview

The preparation of the dataset follows the following rough order:

1. Data Extraction & Epitranscription
2. Self-Dissimilarity Measures
3. Contextual Similarity
4. Word-Pair Construction and Pairwise Distance

Table B1: Sources Compiled for WoLD-Swahili Data

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.

5. Classifier Training

6. Classifier Implementation: Predicted Loanwords

B.2. Data Extraction and Phonetic Transcription

The first task in creating this dataset was extracting data on expressions from the PanLex dataset.

The next stage was preparing the necessary features for each expression, and transcribing orthographic representations into phonetic representations. We use the `Epitran` package[27] to convert orthographic text into International Phonetic Alphabet (IPA). this package, however, relies on mappings between orthographic and phonetic units and, for languages with complex orthographies, pre- and post-processors that are applied before or after the mapping. Coverage for these mappings is not complete for all languages in our PanLex sample. The `Epitran` module includes 64 language-script pairs (e.g 'eng-Latn', for English in Latin script, and 'tir-Ethi' for Tigrinya in Ethiopic script), but some language families were not represented. We therefore coded orthographic-phonetic mappings using orthography tables from OmniGlot 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 that is included in our augmented list of Epitran language-scripts.

For each Epitran language-script, we build a dataframe including all expressions from associated languages, and extract the following information for each expression:

- Unique Expression ID
- Raw Text
- Degraded text (no accents, etc.)
- Language code
- Meaning ID
- Epitranscribed raw text

Meaning identifiers in the PanLex dataset 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. Additionally,

[27]Developed by David Mortensen (Carnegie Mellon University, http://www.davidmortensen.org). GitHub: https://github.com/dmort27/epitran

meaning identifiers may be associated with one or more definitions in the dataset, where PanLex definitions are a short string of words that describe the concept in a particular language.

B.3. Own-Dissimilarity Measures

A core requirement for identifying loanwords is that we can determine which words appear to be outliers in their language, that are likely to have been borrowed, and which ones are unlikely to have been introduced from another language. We therefore generate the following measures of own-language dissimilarity:

- Average raw Jaro Winkler Distance
- Average squared Jaro Winkler Distance
- Average Expected Phonetic 2-gram occurrence
- Average Expected Phonetic 3-gram occurrence
- Average Expected Swadesh Phonetic 2-gram occurrence
- Average Expected Swadesh Phonetic 3-gram occurrence

i) Jaro-Winkler Metric The Jaro-Winkler 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 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 is constructed as the average Jaro-Winkler distance from every other word in the same language, as well as the average squared Jaro-Winkler distance, which gives greater weight to closer matches.

ii) Phonetic n-gram We construct measures of whether the combinations of phonetic units, or *phonemes*, that make up a word are typical for the language. 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 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.

iii) Swadesh Phonetic n-gram 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. To improve this measure, and compare words to the 'core' words in a language that are highly unlikely themselves to be loanwords, we construct a similar expected occurrence score for 2- and 3-grams based on observed occurrence in words that are part of the Swadesh list for that language. The Swadesh lists are a classic compilation of concepts that are basic, universal concepts Swadesh (1950) and are therefore likely to be inherited from a proto-language, rather than borrowed horizontally from another language.

We therefore construct measures of whether the combinations of phonetic units, or *phonemes*, that make up a word are typical for words from the Swadesh list for a language. Our source of Swadesh words is the 40 word lists compiled as part of the Automatic Similarity Judgement Program.Wichmann et al. (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.

B.4. Contextual Distance

To identify loanword pairs and 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 dataset. This model has a vocabulary of roughly three million expressions, and can be used to generate measures of contextual similarity for English words. This contextual similarity is implemented by the `Gensim` package Rehurek and Sojka (2010). For all meanings in the PanLex dataset with an English denotation, or a definition in English, 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.

This word vector measure of contextual similarity of expressions is less conservative than restricting only to expressions that are translations, and broadens the space of potential loanword pairs while preventing nonsensical matches between expressions denoting entirely unrelated concepts.

B.5. 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 additional pairwise distance measures.

We restrict our analysis to expressions who are above a threshold of contextual

similarity, set arbitrarily at 0.7. 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 dataset, we create pairwise matches with all words in all other languages. As 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 those word pairs where words are contextually similar, as described above. We then calculate a number of pairwise distance measures between the two words, as follows.

i) Articulatory Feature-Edit Distance Metrics The first set of pairwise distance metrics we create is exploits detailed information on the phonemes that make up the phonetic representations of words. Using the PanPhon package developed in Mortensen et al. (2016), 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. 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.

ii) Jaro-Winkler Distance As with the own-language dissimilarity measures, we compute the pairwise Jaro-Winkler orthographic distance for the candidate word-pair.

iii) Language Family Cladistic Distance For the candidate wordpair, 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.

iv) 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.

B.6. Train Machine Learning Classifier

In the steps above, for each word pair, we calculate the following features which are inputs to the machine learning algorithm:

1. raw Jaro-Winkler Own-Language Dissimilarity – Word 1
2. squared Jaro-Winkler Own-Language Dissimilarity – Word 1
3. Phoneme Bigram Own-Language Dissimilarity – Word 1
4. Phoneme Trigram Own-Language Dissimilarity – Word 1
5. Phoneme Bigram Swadesh Dissimilarity – Word 1
6. Phoneme Trigram Swadesh Dissimilarity – Word 1
7. raw Jaro-Winkler Own-Language Dissimilarity – Word 2
8. squared Jaro-Winkler Own-Language Dissimilarity – Word 2
9. Phoneme Bigram Own-Language Dissimilarity – Word 2
10. Phoneme Trigram Own-Language Dissimilarity – Word 2
11. Phoneme Bigram Swadesh Dissimilarity – Word 2
12. Phoneme Trigram Swadesh Dissimilarity – Word 2
13. Pairwise Contextual Similarity Score
14. Pairwise Language Family Cladistic Distance
15. Pairwise Hamming Articulatory Feature-Edit Distance
16. Pairwise Weighted Hamming Articulatory Feature-Edit Distance
17. Pairwise Jaro-Winkler Distance
18. Difference in raw Jaro-Winkler Own-Language Dissimilarity
19. Difference in squared Jaro-Winkler Own-Language Dissimilarity

20. Difference in Phoneme Bigram Own-Language Dissimilarity
21. Difference in Phoneme Trigram Own-Language Dissimilarity
22. Difference in Phoneme Bigram Swadesh Dissimilarity
23. Difference in Phoneme Trigram Swadesh Dissimilarity

i) Classified Training Sets Having created these features, we match datasets of words with manually classified origins to the dataset of PanLex words. The three major sources of words with confirmed etymologies are the World Loanwords Database (WoLD), the Oxford English Dictionary and Wiktionary.

Appendix C. Trade Model Appendix (TO BE PUBLISHED ONLINE)

C.1. Cleaning neighbourhoods & aggregate small groups

Before optimizing production and solving the nutritional trade model, we first clean up 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.

C.2. Computational solution details for utility under autarky

We solve this optimization problem using a SciPy[28] implementation of the Byrd-Omojokun Trust-Region SQP method (see Lalee et al. (1998), Nocedal and Wright (2006)). This method smooths out 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. Solving over a forty-one dimensional 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

[28] https://docs.scipy.org/doc/scipy/reference/

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 autarky landuse, consumption and utility.

C.3. Computational solution details for utility under trade

Using autarky land allocations as the starting points for the optimization routine over each group's allocation to each crop, the algorithm also solves this using the same two-step procedure where we optimize over all possible crops for all groups in the neighbourhood at coarse tolerance, then restrict to those with non-trivial ($\geq$1% land share) then optimize again at finer tolerance.

We then solve for the system of prices which can support such an allocation of land. Here we assume that trade is frictionless and that all groups in a neighbourhood face the same prices for all goods. Using the social planner's solution for production and consumption under trade obtained as above, we derive supply-side and demand-side side constraints on prices.

i) Comparative Statics: Trade Excluding Partner Group To estimate the language-specific aggregate gains from trade, we use the method described above to estimate production patterns and prices under trade and autarky. To estimate aggregate gains from trade with neighbours relative to autarky, we simply compare the utility from the optimal autarky bundle to the utility of the bundle under full trade with immediate neighbours.

ii) Comparative Statics: Trade Excluding Partner Group To estimate the pair-specific economic leverage between neighbours, we use the method described above to estimate utility under full trade for the minimal neighbourhood. For the pairwise specification (group A, group B), the neighbourhood is not just the immediate neighbours of group A, and is instead the union of A's neighbours and B's. We then estimate utility for group A where group B was dropped from the neighbourhood. To do this, we use the exact same process as described above, but where the pair is excluded. We repeat this for each of the two neighbourhoods in the pair.

We compute this measure of utility gain for each of the pair, and take the difference between them to measure the difference in economic benefit from the relationship. Intuitively, this means we are measuring economic leverage as how much group A gains from including group B in its trading network relative to the gain B receives from including group A. If group A gains much more from group B being present than vice versa, we interpret this as group B having greater economic leverage over group A.

iii) Comparative Statics: Gains from trade without General Equilibrium Effects
We also restrict to utility gains that only come through crops that are directly traded between groups to eliminate general equilibrium effects arising from changes made by other groups in the neighbourhood. To do this, we first look at the full trade equilibrium described above and use production and consumption to compute net imports for each crop by group. For the pair of language groups under consideration, we identify which crops involve direct interaction between groups A and B, where one is a net importer and the other is a net exporter.

To compute this direct utility effect of group B on group A, we fix all prices for non direct interaction crops at the equilibrium prices in the scenario where group B is not included in the neighbourhood and compute the utility of group A. We then generate a new off-equilibrium price vector where we change the prices of direct interaction crops for A & B to be the prices under full trade when group B is included. This comparative static therefore only includes utility changes for A coming from price changes in the crops where it is directly interacting with group B.

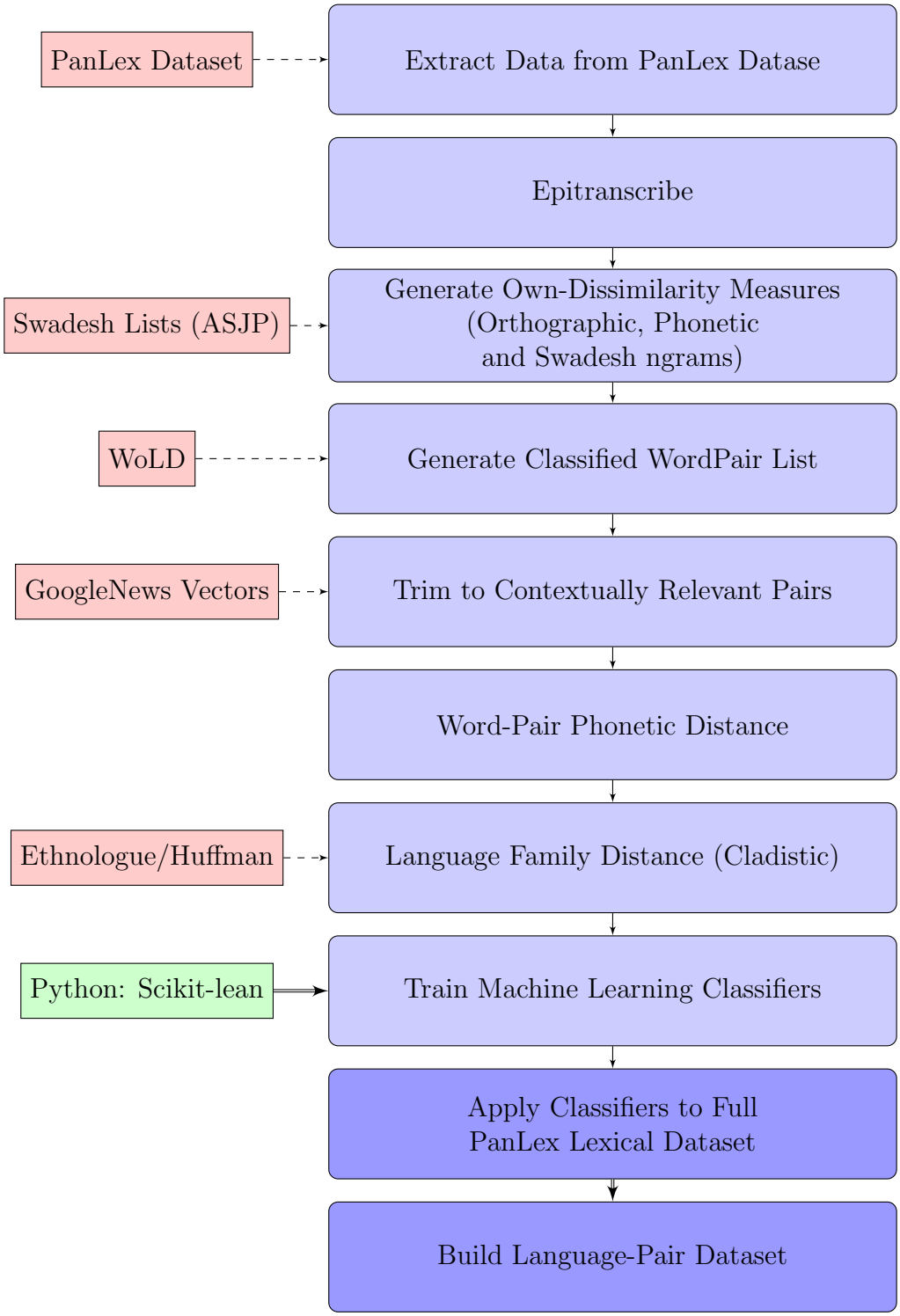
C.4. Model Validation

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 2. 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 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.

Figure B1: Code & Data Flowchart



Note: