How Cultures Converge:
An Empirical Examination of Language Exchange

Columbia Departmental Seminar

Arthur Blouin and Julian Dyer

University of Toronto and University of Exeter

September 10, 2020

Cultural Convergence?

"On one thing the whole world seems to agree: Globalization is homogenizing cultures."

–Tyler Cowen, 2002

- Governments regularly operate as if cross-societal trade erodes local culture, manifesting in a variety of protectionist policies
  - Empirical evidence supporting the homogenizing effect of cross-cultural trade is extremely sparse
  - In what way do cultures homogenize? What is the mechanism?
- We introduce global data on the *direction* of cultural convergence to answer these questions.
- Ask whether & how trade incentives impact cultural convergence between societies.

This Paper 1: Culture data

- To focus the question, we investigate the adoption of one (fundamental) element of culture: language
  - We study loanwords: a word that is used in one language, that was adopted (but not inherited) from another
  - Accurately and consistently measured for all groups globally

- Linguists have identified 50,000 loanwords to date, we use ML to identify loanword status of the remaining 625-trillion (known) word pairs.

- Aggregate the word level data to the society-pair level to measure the intensity of (directional) transfer

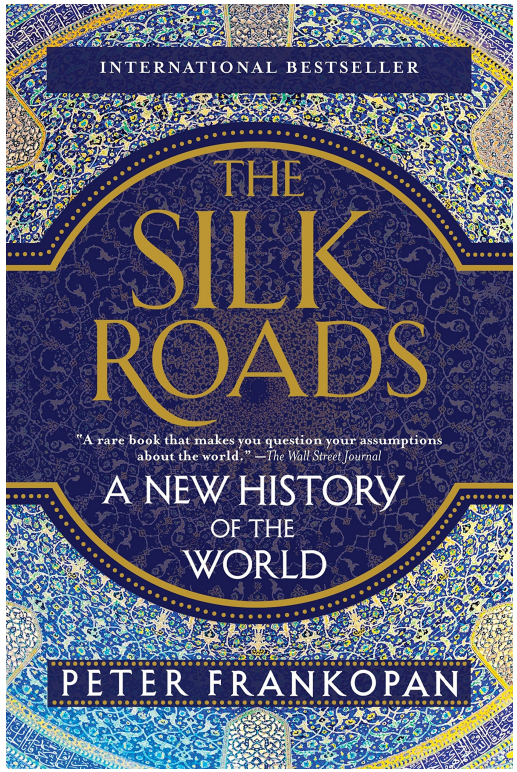
This Paper 2: Gains from trade

- We show that loanword adoption is associated with incentives for agricultural trade.
- We focus on agricultural trade because it makes us more comfortable with causal claims.
  - structurally compute gains from trade based only on the soil characteristics of the land inhabited by a particular society
  - identifying assumption: language exchange does not generate soil characteristic complementarity
- For each society, we estimate welfare both under full trade, and under the counterfactual where they can trade with all but one neighbour (for each neighbour)

Background to loanwords

- Measuring cultural influence is near-impossible with currently available data
  - Need to observe direction of cultural transfer to identify who is influential and who is being influenced
  - Languages are typically made-up of 20%-50% loanwords, and typically include loanwords from 3-15 neighbouring languages.

"Buddhism made sizeable inroads along the principal trading arteries to the west [. . .] The rash of Buddhist **loan words in Parthian also bears witness to the intensification of the exchange of ideas** *in this period" (Frankopan, 2016)*



Loanword adoption

- Loanword adoption based on "the nature and extent of cultural contacts" (Scotton and Okeju, 1973)
- Loanwords documented as part of a complete understanding of a literature, and typically based on language case-studies
- Dominant theory of adoption among socio-linguists is the **contact hypothesis** (Gumperz and Wilson, 1971):
  - Language exchange is the unintended byproduct of contact (economic or otherwise)
  - Puzzle: many cases involving significant contact but no exchange, and cases with little contact but a lot of exchange.

Contact hypothesis or strategic adoption?

- The contact-hypothesis implies **symmetric** language borrowing (on average) when societies interact
- An alternative to the contact-hypothesis implies **asymmetric** language borrowing:
  1. individuals have foreign-speaking potential trading partners
  2. allow for a strategic decision: decide to invest in learning a foreign language or not (*adoption*):
     - one-sided cost (F) to facilitate trade
     - individual earns π' if they adopt, otherwise π
     - WIP: embedding into a war of attrition with uncertainty framework (e.g. Abreu and Pearce (2002)) to get that the one with the larger gains from trade is the one to adopt.
  3. foreign language is then *diffused* throughout own-language by adopters

Strategic adoption toy-model: 2 society case

- each individual (n) in society with population N faces the decision to adopt, subject to π_n, π'_n, F.
- define L as the share of people in the society that decided to become bilingual.
  - $L = \frac{\sum_n^N 1[(\gamma_n - F) \geq 0]}{N}$, where $\gamma_n = \pi'_n - \pi_n$ represents the gains from trade
- For a word from another language to be conversationally useful and potentially diffuse, it must be both known by the using party and understood by the receiving party.
  - probability that a conversation between 2 people in a society could effectively make use of a loanword is L^2
  - loanwords in a society: $\mathcal{L} = \rho L^2$ where ρ is rate that a conversation that could use a loanword results in diffusion.

Strategic adoption toy-model: multiple neighbours

- Intuition: if everyone invests in a different second language (of J options) you could get 100% bilingualism, but 0% diffusion.
- Consider the introduction of a new viable trading partner L_v (writing shares as lower-case, so $\ell_j = \frac{L_j}{L}$):

$$\frac{d\mathcal{L}}{dL_V} = \underbrace{2\rho L\left(\sum \ell_j^2\right)\frac{dL}{dL_V}}_{\text{Change in bilingualism}} + \underbrace{2\rho L^2\left(\sum \ell_j \frac{d\ell_j}{dL_V}\right)}_{\text{Change in composition of bilingualism}}$$

- This derivative is only unambiguously positive for the most popularly adopted language.
  - In fact, adding worse - but still marginally beneficial - trading partners with few (but nonzero) adopters decreases diffusion by fractionalizing billingualism

Predictions of strategic adoption

This generates three predictions that we can take to the data:

P1 An increase in partner quality only **unambiguously** increases the total share of loanwords borrowed for a society's best neighbour

P2 In contrast to the contact hypothesis, adoption is **asymmetric**. *WIP: All else equal, the society that gains the most from trade will be the society that bares the cost of acquiring a common language.*

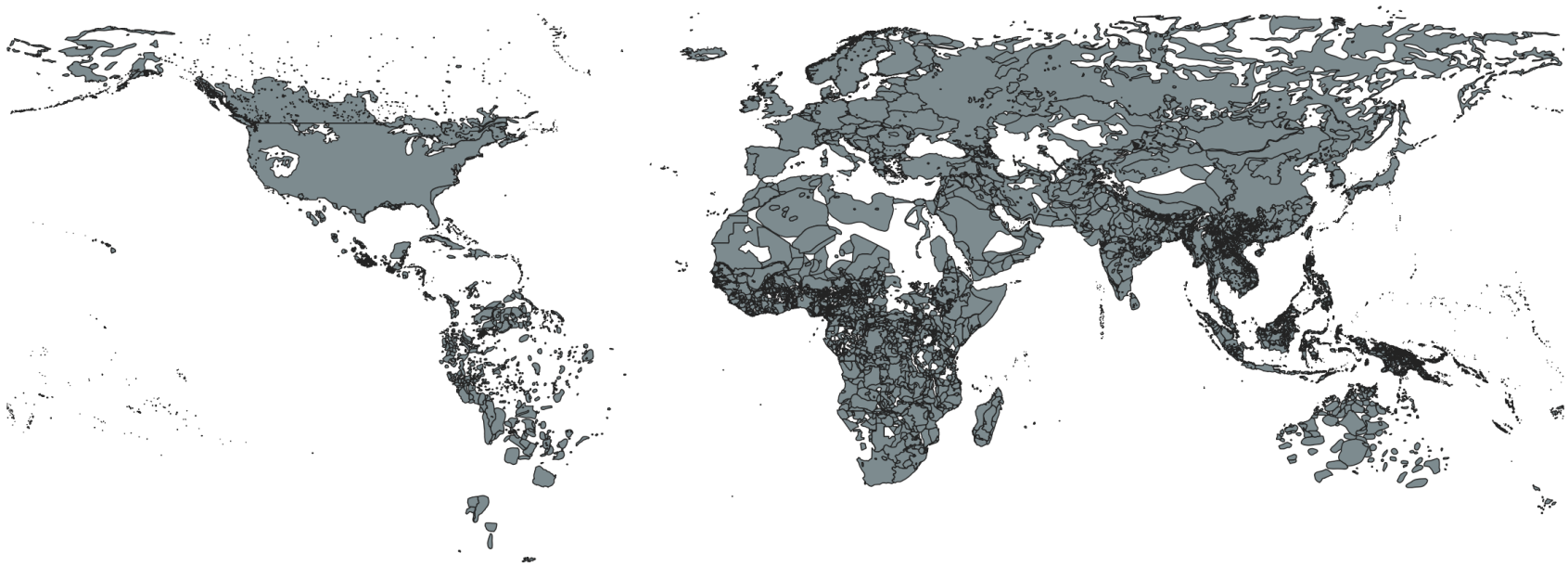
P3 Total borrowing is **inverse-U shaped** in the number of potential viable trading partners, while total lending is non-decreasing.

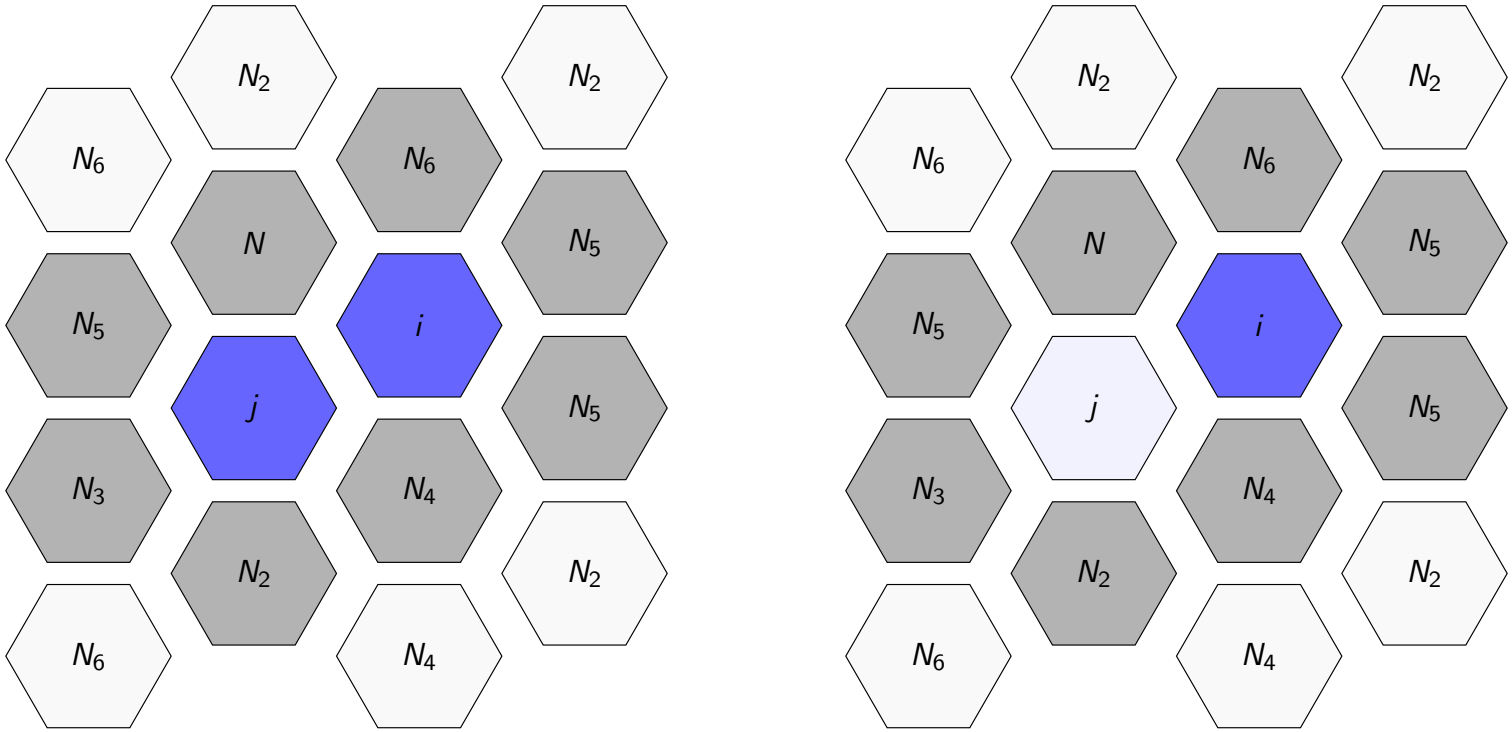
Incentive to Trade – Data

1. *Productivity*: Data from Ethnologue on language group areas matched to FAO GAEZ data on potential crop production.
   - Tells us about land productivity of 41 crops (kg/hectare) for each language group, as well as total area.
   - Also indicates pairs of neighbouring languages.

2. *Demand*: We use National Institutes of Health data to capture recommended amounts of the essential nutrients necessary for survival

Ethnologue Boundaries Map





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

Figure 1: 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.

Gains from Trade – Production

- Focus on local trade in agricultural production - model crop production similarly to Costinot and Donaldson (2012)
- each society chooses an allocation of land ($\vec{l}$) to different crops, and output is land allocated to a crop multiplied by productivity,
- the productivity vector ($\vec{q}$) is the average from the GAEZ land suitability dataset.

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

Gains from Trade – Demand

- Costinot and Donaldson estimate using observed prices
  - we do not observe historical prices, so we need to impose some (more) structure.
- We focus on component of welfare based on health / survival
  - as with our focus on agricultural trade, this helps us to avoid endogenous tastes / preferences.
- We specify a very simple agricultural trade model:
  - utility based on consumption of the nutritional elements required to survive
  - estimate potential gains from trade due to comparative advantage in producing these nutrients

Gains from Trade – Utility

- Define a nutritional utility function which takes the form of a Cobb-Douglas production function over nutrients x_i for a healthy population:

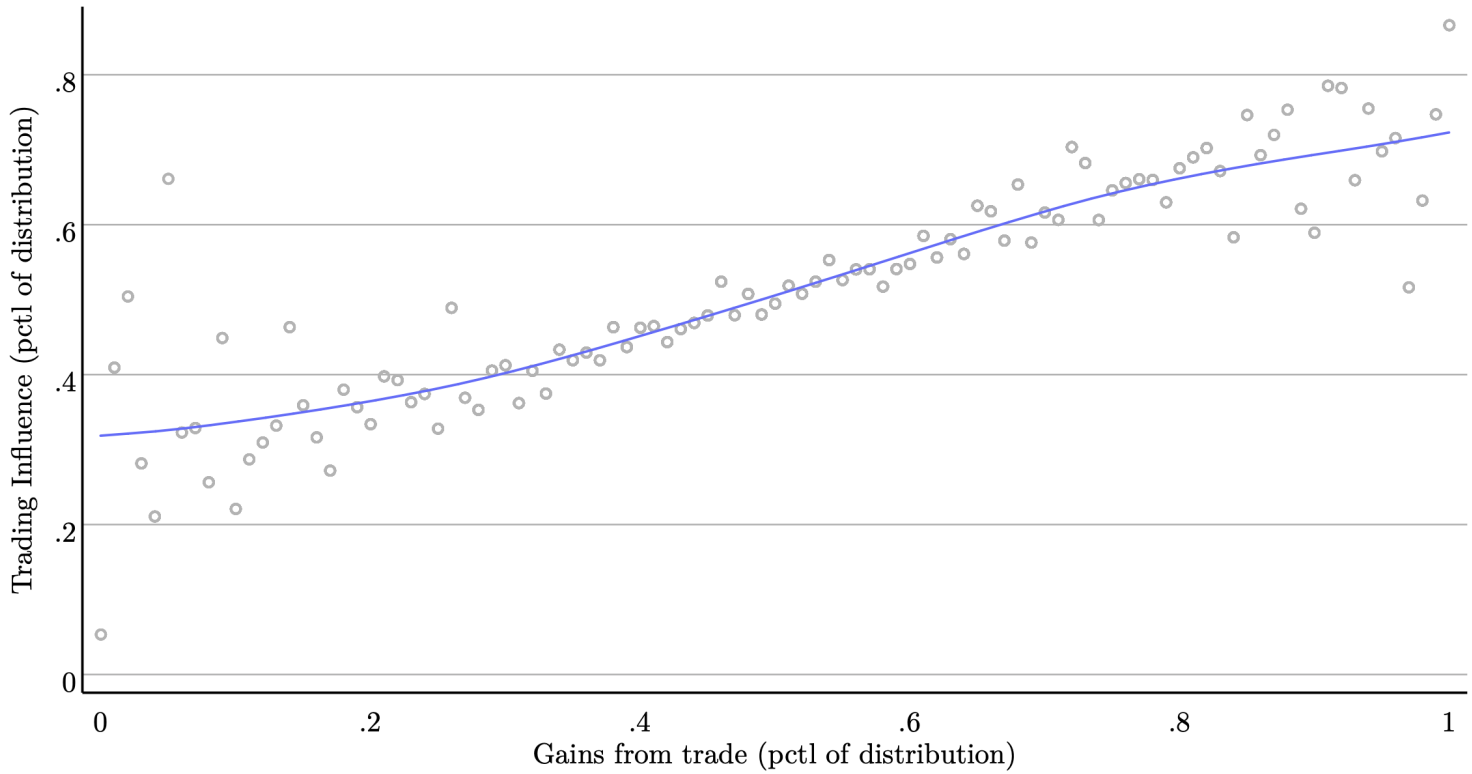
$$U(x_0, x_1, \cdots, x_{16}) = x_0^{\alpha_0} x_1^{\alpha_1} \cdots x_N^{\alpha_{16}} \tag{1}$$

- 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.

Gains From Trade – Solving Numerically

- We first match up production and utility to solve for a price-vector.
- From the set of equilibrium prices we compute land allocations under trade (societies are simple maximizers).
  - Cobb-Douglas means everyone consumes in the same proportion, regardless of income
    - so we only need to solve for the neighbourhood aggregate land allocations
  - Using this consumption bundle, compute utility under trade.
- We similarly compute welfare under full-trade and under counterfactuals without any given trade partner

Gains from trade and regional economic influence



Gains From Trade – Validity

- The basic strategy of using observed relative land productivity to model agricultural trade is valid at the country level (Costinot and Donaldson, 2012)
- We impose more structure due to:
  a. do not observe prices
  b. need to estimate welfare
  c. need sub-national trade
- We therefore undertake a similar exercise to Costinot and Donaldson (2012)
  - Is predicted production correlated with actual production, controlling for land productivity?

Gains From Trade – Validity

- We get actual crop shares for crop *i* from Monfreda et al. (2008).
  - share of land to each crop across globe (5 arc-minute cell)
  - use any crop with both actual & estimated mean land share $>$ 0.0001%, that overlap in FAO and Monfreda et al.

$$Production_i = \beta_0 + \beta_1 \hat{Production}_i + \Gamma FAOSuitability_i + \epsilon_i$$

- We are interested in approximating *local* trade. Many crops are really only global (e.g. tobacco, coffee).
  - separate crops into local v. global
  - report local crops individually; global crops as a group

Table 1: Validating the trade measure against actual crop production

	Dependent variable: Actual Land Allocation								
	Sweetpotato (1)	Carrot (2)	Sunflower (3)	Sorghum (4)	Coconut (5)	Cassava (6)	Oats (7)	Potato (8)	Global (9)
Model allocation	0.0977***	0.0711**	0.0919***	0.0615*	0.146***	0.262***	0.530***	0.523***	-0.00412
	(0.0209)	(0.0278)	(0.0274)	(0.0348)	(0.0298)	(0.0501)	(0.158)	(0.119)	(0.0883)
Crop suitability	✓	✓	✓	✓	✓	✓	✓	✓	✓
N	2,995	2,995	2,995	2,995	2,995	2,995	2,995	2,995	2,995
R^2	0.249	0.183	0.361	0.396	0.394	0.350	0.362	0.353	0.187
Dep. Var. Mean	0.002	0.0001	0.001	0.007	0.006	0.006	.0002	0.001	0.054

Note: The unit of observation is a society. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Constructing the main regressor

- Pairwise level: U_i^{FT} is society i's utility under full trade; U_i^{FT-j} is i's trade utility without society j

$$c_{ij} = \frac{U_i^{FT}}{U_i^{FT-j}} \tag{2}$$

- We interpret as the contribution of j to the trade utility of i.
- Societal level (guided by P1) we focus on 'best' neighbour:

$$c_i = max\{\frac{U_i^{FT}}{U_i^{FT-j}}\} \tag{3}$$

Constructing the main regressor

- Similarly for trade influence:

$$\iota_{ij} = \frac{U_j^{FT}}{U_j^{FT-i}} \tag{4}$$

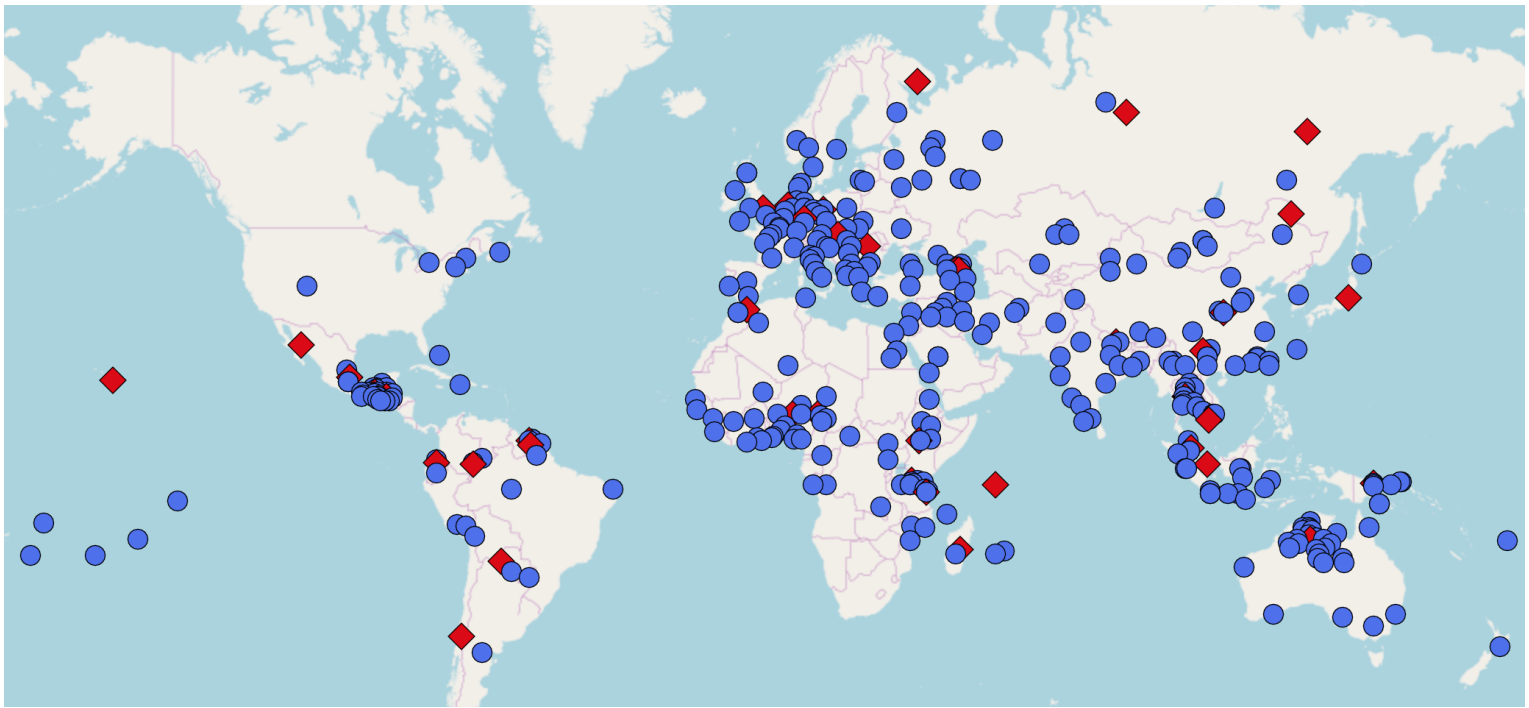
$$\iota_i = \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} \frac{U_j^{FT}}{U_j^{FT-i}} \tag{5}$$

Loanwords: Language Data

- All words in all languages
  - PanLex: lexicon with meanings for every language
  - We are interested in knowing the extent of language sharing between all language pairs
- Machine Learning: estimate the number of words in each language borrowed (loaned) from (to) other languages
- Construct a training set based on (almost) all words for which the field of linguistics has identified a word origin.
  - Main source: World Loanwords Database (WoLD)
  - Linguistics has identified loanword status with for about 50,000 words in 41 languages

WoLD

Figure 2: Map of the known loanwords (i.e. our training set)



Note: Red diamonds indicate the forty-one borrowing languages, blue circles indicate source languages.

Computation

Panlex contains 25,000,000 words $\Longrightarrow$ $6.25x10^{14}$ (625 trillion) word-pairs.

- Linguistics has determined the loanword status for 0.2% of all (known) words, which forms our training-set. We estimate the loanword status of all remaining (known) words.
- Running prediction algorithm on the U of Toronto supercomputer took 43,760 core-hours
- If we ran this on Julian's quad-core laptop, it would have taken 1.25 years of continuous computation

Prediction algorithm

We detect loanword pairs using a ML algorithm that looks at **only** spelling & pronunciation. Our approach (for comp. efficiency):

1. How phonetically dissimilar is each word from its own language?
   - use data on spelling, phonetics, Swadesh lists.
2. How similar is a pair of words across languages?
   - use contextual, phonetic & orthographic similarity and language family distance.
3. Take all word pairs, coarsely filter to those in same semantic space and algorithm decides if it is a loan-source word-pair.
   - if multiple source words identified, choose the most likely one.

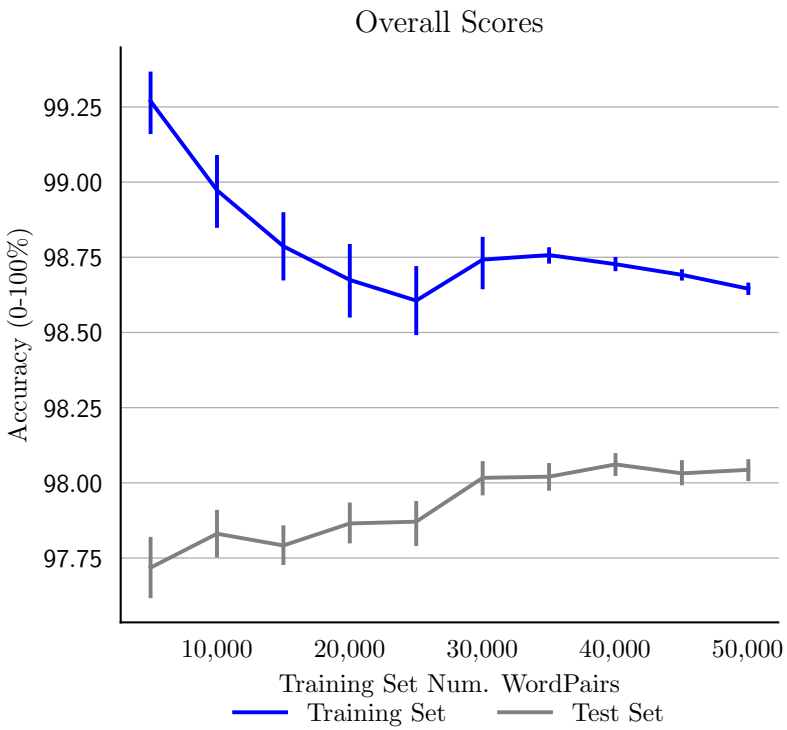
Random Forest Classifier

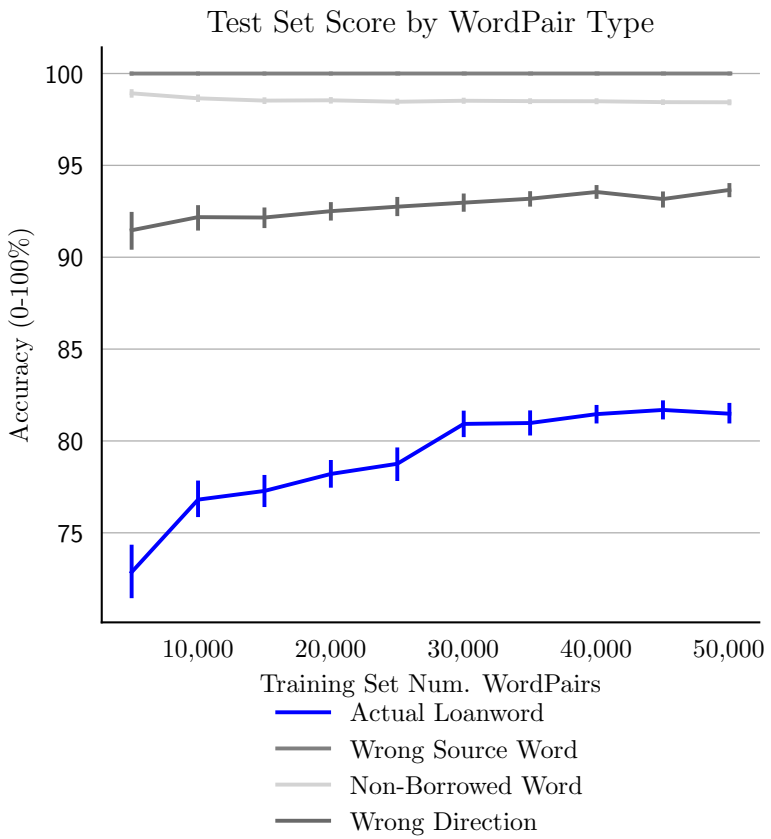
- We use a relatively straightforward Random Forest algorithm, in addition to:
  - under-sampling and synthetic minority oversampling (SMOTE) to deal with imbalance
  - 'Extremely Randomized Forest' to reduce overfitting
  - The three are included in an emsemble Voting Classifier.

Accuracy

- Overall test set accuracy: 98.1 %
- If we just aim for overall accuracy, we can get over 99.9%, but we end up with LOTS of known loanwords classified as own-language words.
- For the baseline results we accept 0.35% of non-loanwords to be classified as loanwords
- This results in us classifying over 90% of known loanwords correctly (of the word-pairs we identify as loanwords, at most 9% of these are incorrect)

Classifier Performance





▸ Category Label Details

Prediction Accuracy

Table 2: Machine learning model accuracy by accuracy metric

	Output score	F1 score	Balanced accuracy score	Precision score	Recall score
	(1)	(2)	(3)	(4)	(5)
Random Forest - undersample	0.948	0.892	0.919	0.925	0.862
Random Forest	0.984	0.826	0.865	0.949	0.731
Random Forest - SMOTE	0.980	0.809	0.911	0.785	0.834
Extra Trees - SMOTE	0.974	0.737	0.858	0.745	0.729
Voting Classifier	0.981	0.812	0.908	0.798	0.826
Random Forest Phase 2	0.902	0.866	0.892	0.880	0.852

Note: The table shows prediction accuracy statistics for the main predictive models used in our machine learning routine.

Constructing the main outcomes

We collapse the word-pair level results to the society-pair level:

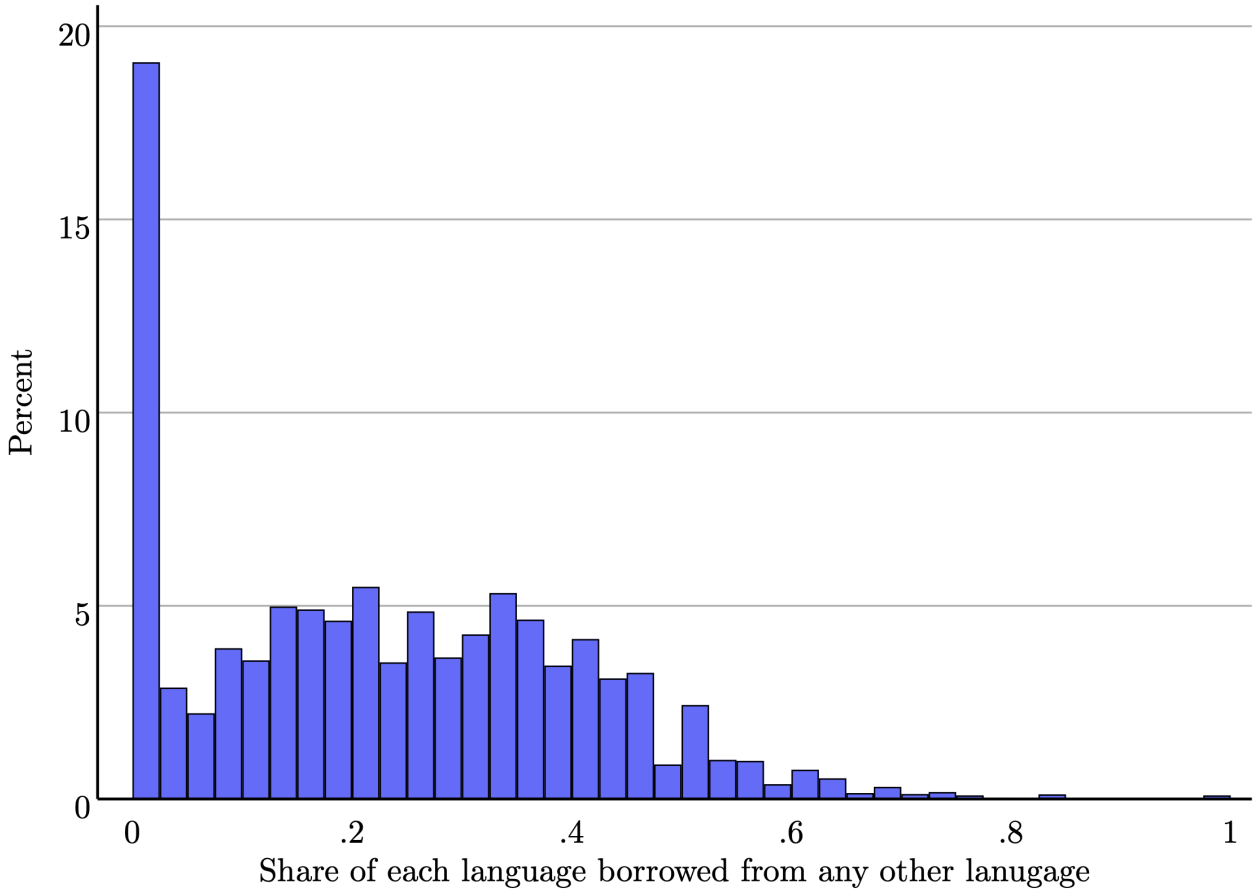
$$\mathcal{L}_{ij} = \frac{\sum 1(Word_i = LoanWord_{ij})}{\sum 1(Word_i)} \tag{6}$$

- pairwise borrowing by society i from society j.
- Similarly at the societal level, we have:

$$\mathcal{L}_i = \sum_{j=1}^{J} \mathcal{L}_{ij} \tag{7}$$

- The more general $\mathcal{L}_i$ is the sum of loanwords from each of the various neighbours j.

Aggregating: Share of language borrowed



Contact hypothesis? Borrowing and lending negatively correlated.

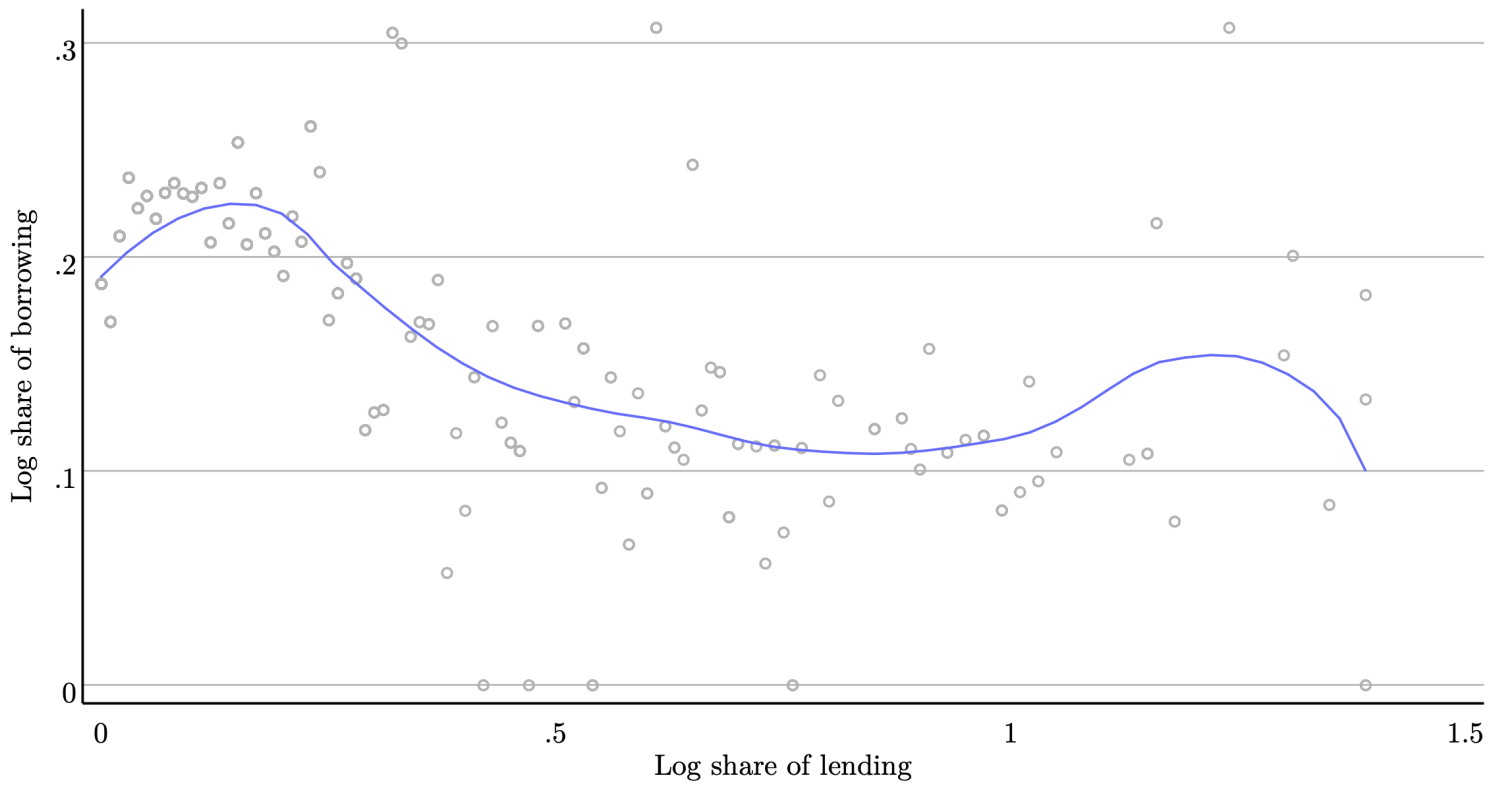


Table 3: Descriptive Statistics

Variable	Observations	Mean	SD	Min	Max
	(1)	(2)	(3)	(4)	(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 **??**. 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

Results

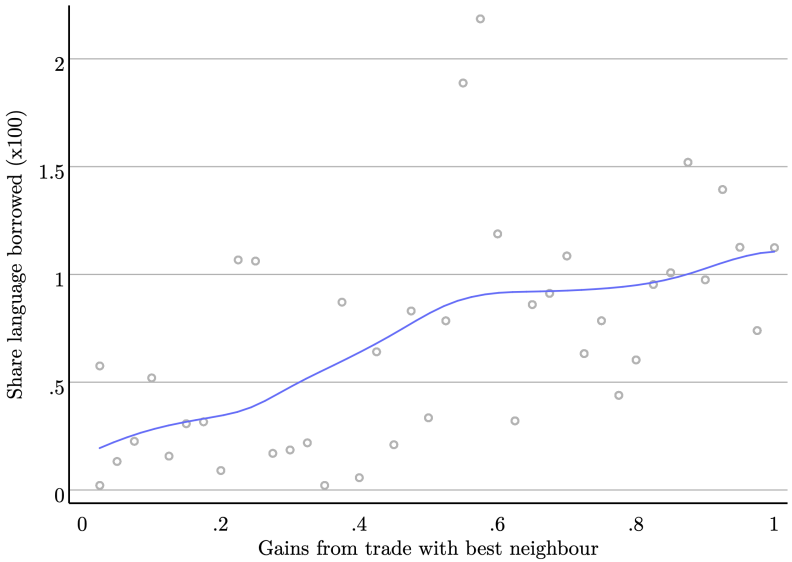
Bringing trade and language data together. The idea of the paper is simple:

1. Is an exogenous measure of incentive to trade correlated with cultural adoption?

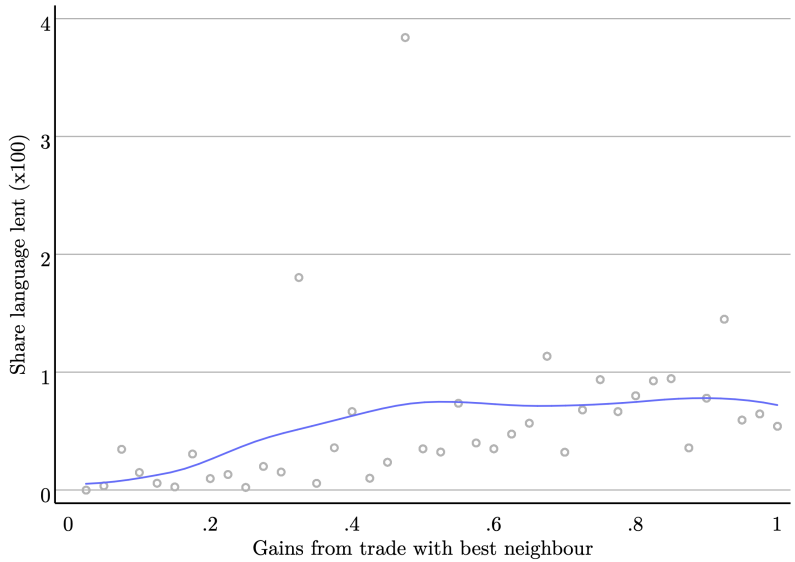
2. If so, what is the mechanism? Are the patterns in the data more consistent with the contact hypothesis (symmetric exchange) or strategic adoption (asymmetric exchange)?

3. What role does diffusion play?

Gains from trade and borrowing and lending

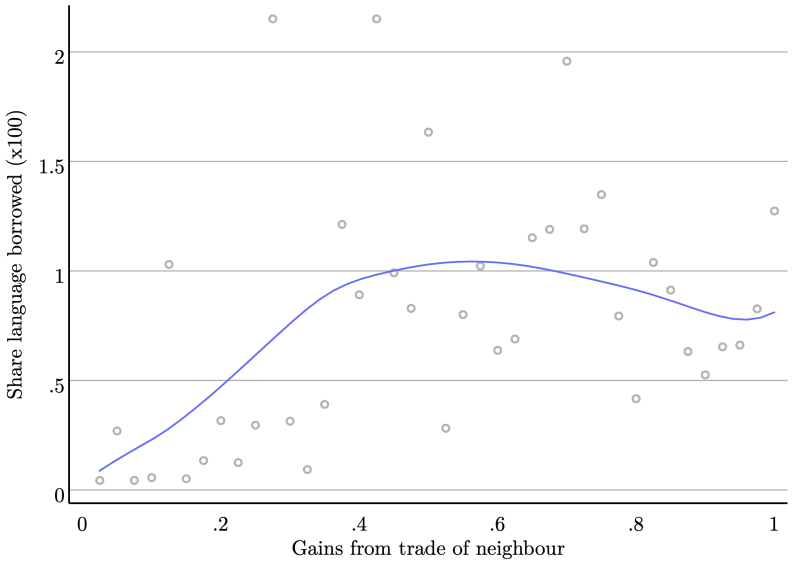


(a) Gains from trade and language **borrowing**

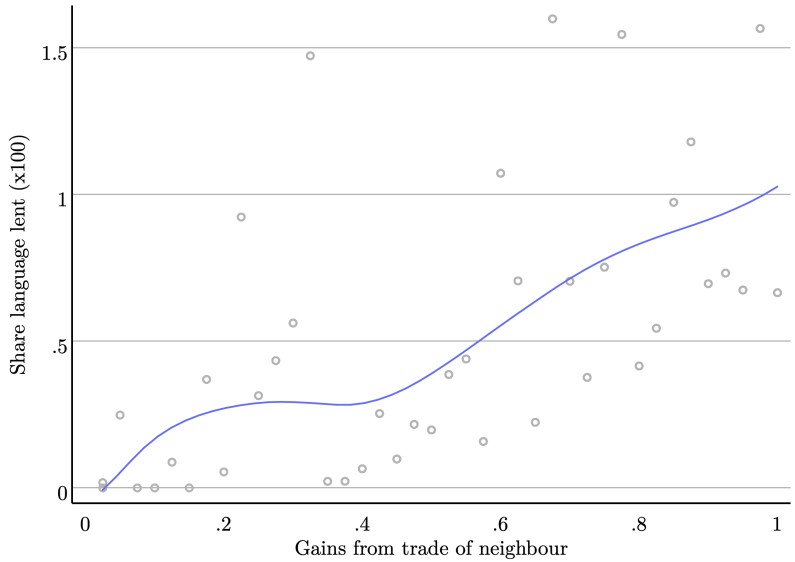


(b) Gains from trade and language **lending**

Language exchange and value as a trading partner



(a) Economic influence and language **borrowing**



(b) Economic influence and language **lending**

Taking stock so far

- Clear correlation between trade gains and language exchange.
- Mechanism is less clear:
  1. Consistent with contact hypothesis (and not strategic adoption):
     - Similar convergence in both borrowing & lending
  2. Consistent with strategic adoption (and not contact hypothesis):
     - Group-level borrowing and lending negatively correlated
     - Could be confounded as gains from trade are very positively correlated with trade influence
- We need a horse-race specification to tease this out

Empirical Specifications

Society level regression:

$$\mathcal{L}_i = \alpha_{colonizer} + \alpha_{continent} + \beta_1 c_i + \beta_2 \iota_i + X_i' \Gamma + \epsilon_i \quad (8)$$

Society-pair level regressions:

$$\mathcal{L}_{ij} = \alpha_{ij} + \beta_1 c_{ij} + \epsilon_{ij} \quad (9)$$

$$\mathcal{L}_{ij} = \alpha_i + \alpha_j + \beta_1 c_{ij} + \beta_2 c_{ji} + \epsilon_{ij} \quad (10)$$

Table 4: Gains/influence from agricultural trade and language borrowing/lending

Dependent Variable:	Language Borrowed		Language Loaned	
	(1)	(2)	(3)	(4)
Gains from trade with neighbours	0.773***	0.671**		0.0269
	(0.206)	(0.281)		(0.177)
Influence on trade with neighbours		0.219	0.629***	0.613***
		(0.293)	(0.201)	(0.222)
Trade wealth (structurally estimated)	✓	✓	✓	✓
Population	✓	✓	✓	✓
Land Share	✓	✓	✓	✓
Land diversity	✓	✓	✓	✓
Distance to Neighbour(s) (quintic polynomial)	✓	✓	✓	✓
Colonizer FE	✓	✓	✓	✓
Continent FE	✓	✓	✓	✓
Observations	2,808	2,808	2,808	2,808
R-squared	0.117	0.117	0.260	0.260
Dependent Variable Mean	0.842	0.842	0.610	0.610

Note: The unit of observation is a society. Robust standard errors in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$. Language Borrowed (range [0,100]) is percentage of language borrowed from neighbours, while Language Loaned (range [0,100]) is the lending analogue. Gains from trade with neighbours is the percentile rank in the distribution (range [0,1]). Influence on trade with neighbours is the percentile rank in the distribution (range [0,1]). In each case, to aggregate to the societal level we take the maximum value from the society's neighbours.

Table 5: Loanwords and trade incentives at the relationship level

Dependent Variable:	Language Borrowed			
Sample:	All	Viable Only		
	(1)	(2)	(3)	(4)
Gains from trade with neighbours	0.327*	0.979***	0.318**	0.324**
	(0.178)	(0.369)	(0.154)	(0.154)
Influence on trade with neighbours				0.153
				(0.236)
Relationship Fixed Effects	✓	✓		
Society Fixed Effects (both)			✓	✓
Observations	11,690	6,882	6,882	6,882
R-squared	0.516	0.523	0.793	0.793
Dependent Variable Mean	0.275	0.284	0.284	0.284

Note: The unit of observation is a society-pair. Standard errors two-way clustered by each society within a society-pair are in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$. Language Borrowed (range [0,100]) is percentage of language borrowed from neighbours, while Language Loaned (range [0,100]) is the lending analogue. Gains from trade with neighbours is the percentile rank in the distribution (range [0,1]). Influence on trade with neighbours is the percentile rank in the distribution (range [0,1]). Viable trading relationships are those where at least one of the two parties can gain from trade. 'Society Fixed Effects (both)' means we separately include a society fixed effect for each society in the relationship; 'Relationship Fixed Effects' means we include a fixed effect for a specific society-pair.

Word-Type Disaggregation

- Nature of linguistic adoption:
  - Purely functional for trade or relating to cultural distance?
- Our data is at word-level, allowing disaggregation:
  - Crop names, economic transaction keywords
  - 'cultural' words not in above
- Start with English wordlists, identify semantically similar words in other languages
  - Use names of the forty-two crops in our dataset
  - Ten-word lists reflecting other categories
- ▸ Wordlists ▸ Semantic Tagging Routine Details

Table 6: Loanwords by word-type and trade incentives

	Crop words	Non-crop words	Economic transaction words	All but crop/transaction
	(1)	(2)	(3)	(4)
Gains from trade with neighbours	0.169*	0.615***	0.0864**	0.614***
	(0.0911)	(0.190)	(0.0396)	(0.191)
Trade wealth (structurally estimated)	✓	✓	✓	✓
Population	✓	✓	✓	✓
Land Share	✓	✓	✓	✓
Land diversity	✓	✓	✓	✓
Distance to Neighbour(s) (quintic polynomial)	✓	✓	✓	✓
Colonizer FE	✓	✓	✓	✓
Continent FE	✓	✓	✓	✓
Observations	2,718	2,718	2,718	2,718
R-squared	0.022	0.140	0.020	0.140
Dependent Variable Mean	0.296	0.617	0.128	0.576

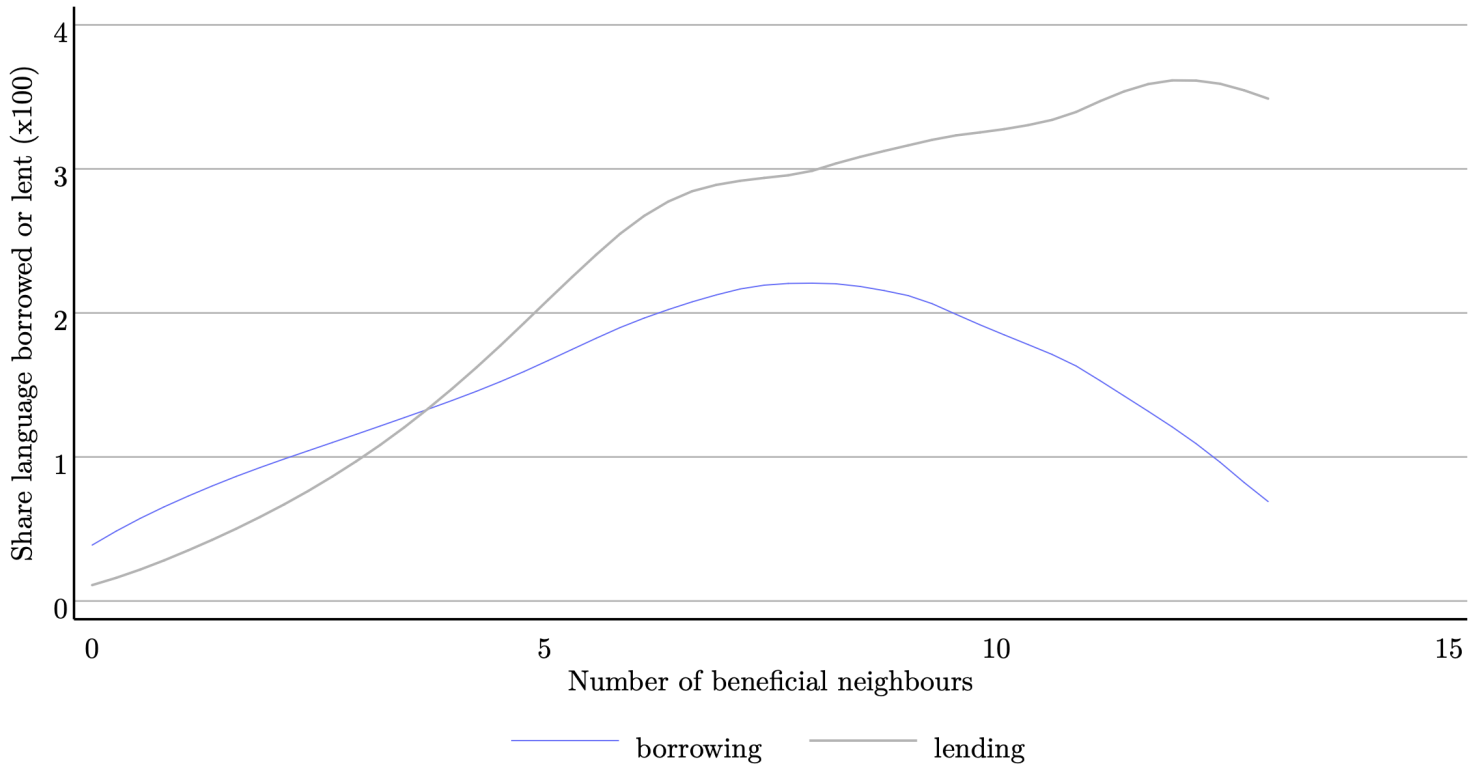
Note: *** p<0.01, ** p<0.05, * p<0.1. We lose 90 observations relative to the sample in table 4 because there are some languages where we find no english equivalents for one category of word-type. All word-type borrowing outcomes are winsorized at the 0.1% level to deal with outliers.

▸ Wordlists ▸ Semantic Tagging Routine Details

Viable Neighbours and Language Exchange

- So far, we focused on Prediction 1: strategic adoption
  - Asymmetric exchange responds to strategic economic incentives
- Now, Prediction 2 concerns diffusion of linguistic adoption
  - Predicts that borrowing follows inverse-U in number of viable trading partners:
    - As the number of viable partners increases, more individuals adopt other languages, **increasing** diffusion
    - but at a certain point adoption of other languages becomes fragmented enough to **decrease** diffusion
  - Predicts that total lending is nondecreasing in number of viable partners

Total borrowing and the supply of viable trade partners



Discussion

- We find evidence that linguistic convergence reacts to strategic economic incentives
  - Strong evidence of asymmetry, with less influential partners converging towards to more economically influential
- Thinking about language adoption as a (one-sided) individual cost required to initiate trade + diffusion explains most patterns in the data.

Classifier Performance

If word **A** was borrowed from source word **B** in another language, we look at classifier performance on different types of word-pairs.

- We want to detect the following type of word-pair:
  - **B** $\rightarrow$ **A**: Loanword pair in correct direction
- and distinguish from other types of word-pair:
  - **A** $\rightarrow$ **B**: Loanword pair in incorrect direction of transfer
  - (Other Source Words) $\rightarrow$ **A**: Loanword but incorrect source
  - (Potential Source Words) $\rightarrow$ **C**, where word **C** not borrowed

◂ Classifier Performance

Implications 1: Re-evaluation of Colonialism Literature

- Colonialism literature has been shown to have big implications on ethnic divisions e.g. **?**; **?** Michalopoulos and Papaioannou (2016); Blouin (2020)
  - Almost entirely based on case studies without much insight into heterogeneity of the effect (Michalopolous and Papaiannou (2018))
- Our results suggest a framework to think about heterogeneity of colonial shock:
  - If colonization creates powerful centres, then it will create incentives for everyone to culturally align with one group.
  - This incentive will be especially powerful if colonists happen to interact with already influential groups

Colonial centrality – methodology

- One possibility: a colonizer will strategically focus their presence closer to the centre of contiguous, populated blocks under their control
  - Identification strategy in the spirit of **?**

- We test this by identifying centroids for each colonial cluster:
  - Start with standard map of colonies
  - Restrict to areas with potential caloric yield above threshold of 1000
    - (**?**)
  - Focus on functionally contiguous clusters
    - split clusters connected by narrow 'bridges' using buffer zones
  - Identify cluster centroids

Colonial centrality

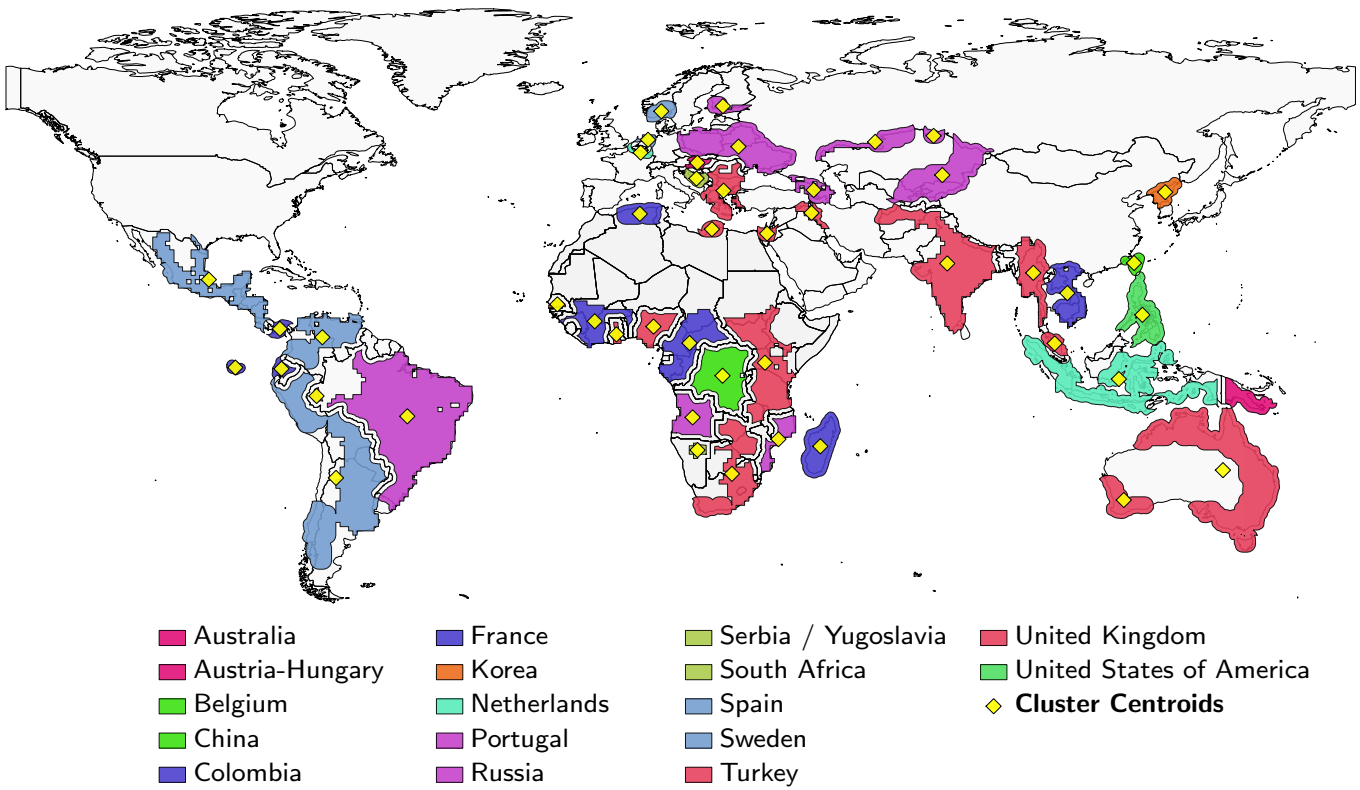


Figure 5: Colonial clusters & centroids

Note: This map shows the contiguous, functional clusters of populated colonies with their centroids.

Colonial centrality – Africa

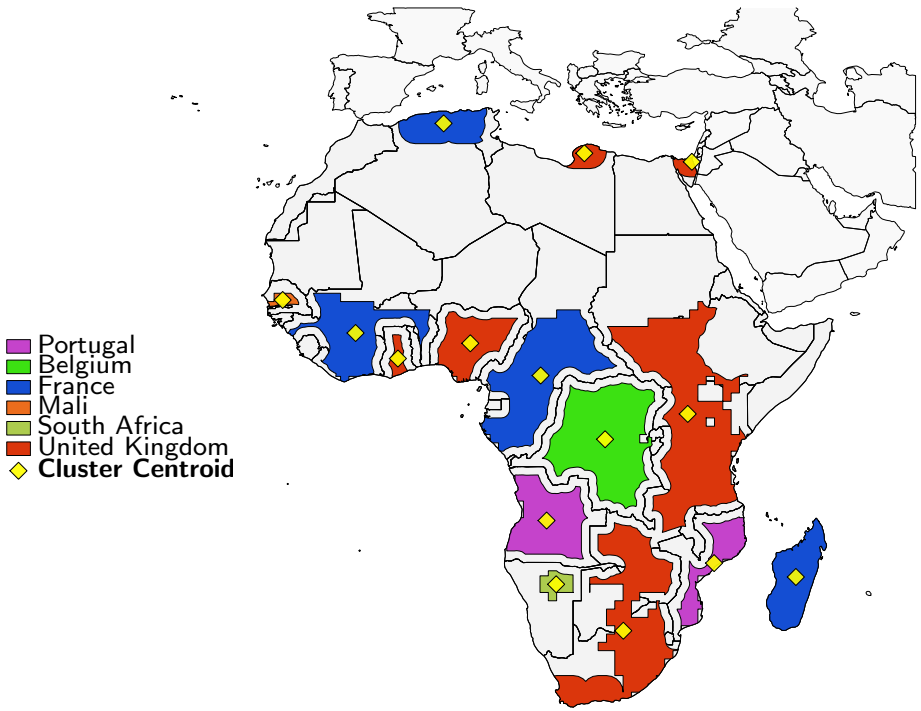
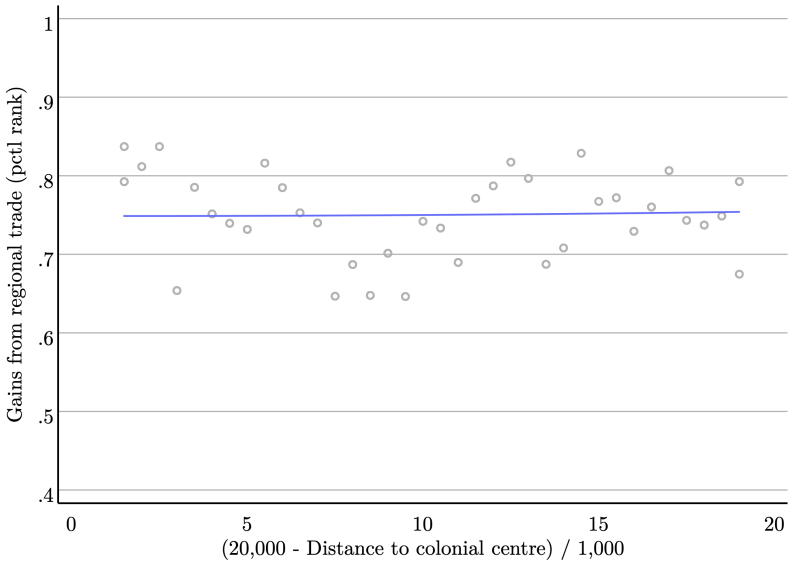


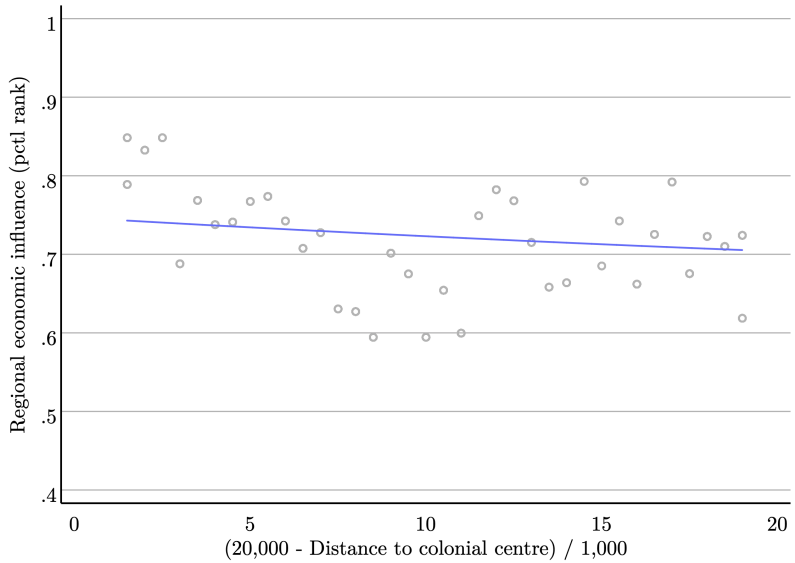
Figure 6: Colonial clusters & centroids in Africa

Note: This map shows the contiguous, functional clusters of populated colonies with their centroids, focusing only on colonies in Africa to give greater detail.

Colonial group selection

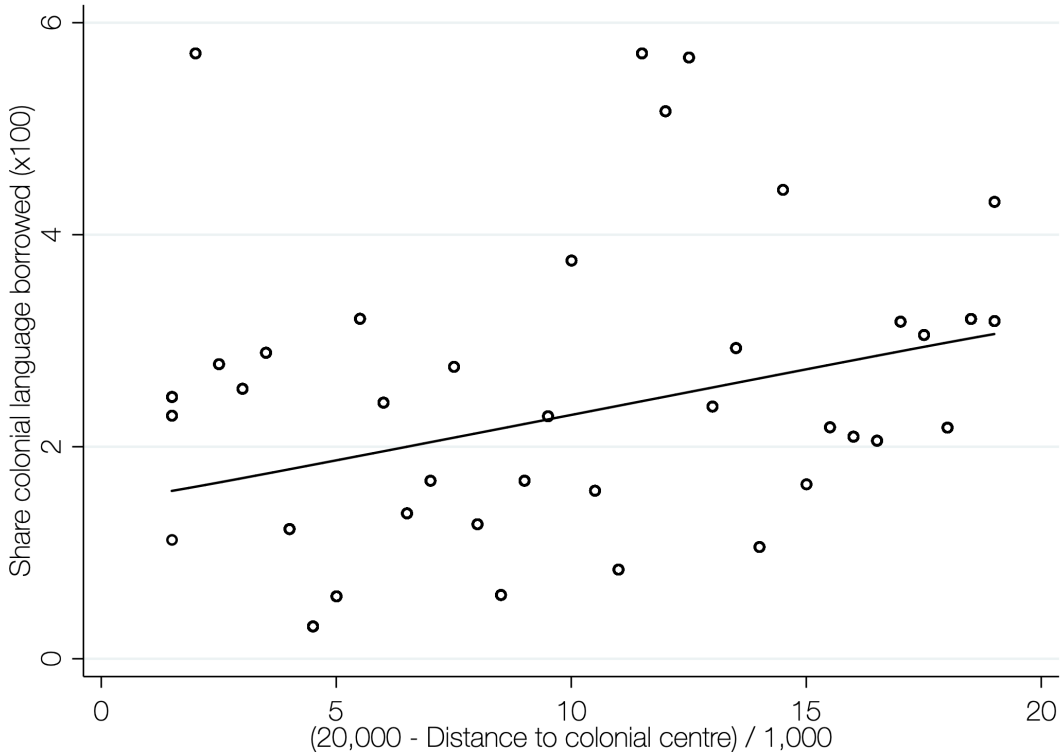


(a) Gains from agricultural trade



(b) Economic influence

Colonial centrality and colonial language adoption



Colonial centrality and colonial borrowing

Dependent Variable	share of colonial language borrowed			
	(1)	(2)	(3)	(4)
Colonial centrality	0.0666***	0.117***	0.0664***	0.117***
	(0.0198)	(0.0379)	(0.0198)	(0.0379)
Reliance on trade	0.363	0.352		
	(0.425)	(0.425)		
Influence on trade			-0.0689	-0.0377
			(0.444)	(0.443)
Trade wealth (struct. estimation)	0.102	0.363**	0.103	0.363**
	(0.0796)	(0.164)	(0.0799)	(0.164)
Trade wealth x colonial centrality		-0.0205		-0.0205
		(0.0125)		(0.0125)
Distance polynomial	Yes	Yes	Yes	Yes
Population	Yes	Yes	Yes	Yes
Land size	Yes	Yes	Yes	Yes
N	2,579	2,579	2,579	2,579
R^2	0.045	0.046	0.045	0.046

Hypotheses: regional dynamics

Not obvious how/if colonial centrality influences **local** processes of linguistic exchange. Possibilities:

1. **No Effect**
   - Only impacts colonizer-colonized interaction, no reshaping of local dynamics
2. **Decreased Interaction**
   - Increased oppression & extraction reduce gains to interaction: (**?**)
3. **Increased Interaction**
   - Colonial intensity increases strategic gains to to interaction through increased economic/political power
   - Cultural power & colonial domination (**?**)

Colonial centrality and regional borrowing

Dependent Variable	local languages					
	borrowed (1)	loaned (2)	borrowed (3)	loaned (4)	borrowed (5)	loaned (6)
Colonial centrality	0.0291**	0.0964***	0.0292**	0.0964***	0.00492	0.0440
	(0.0143)	(0.0244)	(0.0144)	(0.0245)	(0.0303)	(0.0333)
p-value H_0: difference = 0	0.023		0.023		0.40	
Reliance on trade	0.743***		0.642**	0.0903	0.407	
	(0.185)		(0.269)	(0.293)	(0.289)	
Influence on trade		0.973***	0.209	0.919**		0.216
		(0.302)	(0.316)	(0.370)		(0.376)
Centrality x trade					0.0314	0.0716*
					(0.0329)	(0.0391)
p-value H_0: difference = 0					0.42	
Distance polynomial	Yes	Yes	Yes	Yes	Yes	Yes
Trade wealth	Yes	Yes	Yes	Yes	Yes	Yes
Population	Yes	Yes	Yes	Yes	Yes	Yes
Land size	Yes	Yes	Yes	Yes	Yes	Yes
N	2,579	2,579	2,579	2,579	2,579	2,579
R^2	0.072	0.433	0.073	0.433	0.073	0.433

Implications for colonialism and diversity

- Large literature showing colonialism drove wedge between groups, with lasting consequences
- Our results suggest a framework to think about heterogeneity
  - A strong centre of power generates homogenizing forces; especially strong when colonists happened to interact most with already strong groups.
  - Incentives for all to culturally align with newly central group.
  - Increased cultural distance for those not (now) central
- Socioeconomic shocks may impact diversity if they shift regional economic / political / social power dynamics

◂ Return to Introduction ◂ Return to Results

Word-Type Analysis

- We begin with a list of words in English
- We match these words by spelling to expressions in the PanLex dataset for English and recover the list of associated Meaning-ID's
- We first identify direct translations in 293 languages covered by contextual similarity models by identifying all expressions that share exactly the meaning IDs in English
- take the combined list of meaning identifiers described above, and for each of the languages, we extract a list of expressions corresponding to these meanings, and run these through the contextual similarity model for each language to identify words that are contextually similar at different thresholds.

◂ Word Type Disaggregation ◂ Word Type Results

Wordlists - Crop Names

- alfalfa, banana, plantain, barley, buckwheat, cacao, canary grass, carrot, cassava, manioc, chickpea, lemon, lime, orange, coconut, cotton, cowpea, dry pea, flax, foxtail millet, millet, green grams, groundnut, peanut, maize, corn, miscanthus, silvergrass, oat, palm tree, oil palm, palm, olive, onion, phaseolus, bean, pigeon pea, pea, rye, sorghum, soybean, soya, beet, sugarbeet, sugarcane, sugar, sunflower, sweet potato, sweetpotato, switchgrass, bunchgrass, tea, tobacco, tomato, rice, wheat, potato, yam

◂ Word Type Disaggregation ◂ Word Type Results

Wordlists - Other Topics I

- Transactions:
  - money, exprensive, price, trade, exchange, loan, delivery, buy, product, contract
- Technology:
  - plough, book, boat, harvest, irrigation, medicine, map, machine, planting, fishing, husbandry
- Religion
  - God, priest, afterlife, spirit, pray, worship, sacred, church, temple, mosque, astrology

◂ Word Type Disaggregation ◂ Word Type Results

Wordlists - Other Topics II

- Politics
  - leader, ruler, capital, government, policy, law, council, jurisdiction, justice, authority
- Gender & Broader Human Rights
  - queen, housewife, equality, slave, freedom, agency, empowerment, chivalry, child labour, effeminate
- Social Organization
  - polygamy, polygyny, marriage, husband, wife, adoption, cousin, inheritance, ancestor, ancestry, kinship

◂ Word Type Disaggregation ◂ Word Type Results

References I