Geographical Origins and Economic Consequences of Language Structures*

Oded Galor†, Ömer Özak‡, and Assaf Sarid§

October 28, 2016

Abstract

This research explores the economic causes and consequences of language structures. It advances the hypothesis and establishes empirically that variations in pre-industrial geographical characteristics that were conducive to higher return to agricultural investment, larger gender gap in agricultural productivity, and more hierarchical society, are at the root of existing cross-language variations in the presence of the future tense, grammatical gender, and politeness distinctions. Moreover, the research suggests that while language structures have largely reflected the coding of past human experience and in particular the range of ancestral cultural traits in society, they independently affected human behavior and economic outcomes.

Keywords: Comparative Development, Cultural Evolution, Language Structure, Future Tense, Politeness Distinctions, Grammatical Gender, Human Capital, Education

JEL Classification: I25, J24, O1, O10, O11, O12, O40, O43, O44, Z10

*The authors are grateful to Delia Furtado, Paola Giuliano, Luigi Guiso, Stelios Michalopoulos, and David Weil, as well as conference participants at the "Annual Meetings of the American Economic Association", 2016; "Deep Rooted Factors in Comparative Development", Brown, 2016, and seminar participants at Brown, and Clark Universities, for useful comments and discussions.

†Department of Economics, Brown University; NBER, CEPR, IZA, CES-Ifo. E-mail: Oded_Galor@brown.edu

‡Department of Economics, Southern Methodist University. E-mail: ozak@smu.edu

§Department of Economics, University of Haifa. Email: asarid@econ.haifa.ac.il

1 Introduction

The origins of the vast inequality in the wealth of nations have recently been attributed to the persistent effect of an uneven distribution of pre-industrial geographical, cultural, institutional and human characteristics across the globe.[1] In particular, evidence suggests that regional variations in the geographical environment in the distant past have contributed to the differential formation of cultural traits and their lasting effect on comparative economic development across countries, regions and ethnic groups.[2] In light of the coevolution of cultural and linguistic characteristics in the course of human history,[3] the evolution of language has conceivably reinforced the persistent effect of cultural factors on the process of development. Nevertheless, the significance of these evolutionary processes and their potential common geographical roots for the understanding of the process of development and the unequal distribution of wealth across nations has remained obscured.

This research explores some of the most fundamental and intriguing mysteries about the origins of the coevolution of linguistic and cultural traits and their impact on the development process:[4] Has the coevolution of linguistic and cultural traits contributed to the stability and persistence of cultural characteristics and their lasting effect on economic prosperity? Has the evolution of languages reflected economic incentives, promoting an efficient economic exchange? Have language structures merely encoded existing cultural traits or have they influenced human behavior and values and contributed directly to the development process? What are the geographical roots of the coevolution of linguistic and cultural traits? Are the geographical characteristics that triggered the coevolution of culture and language critical for the understanding of the contribution of cultural and linguistic characteristics for the wealth of nations?

The analysis uncovers geographical origins of the coevolution of cultural and linguistic characteristics in the course of human history and their significance for the understanding of the unequal distribution of wealth across the globe. It advances the hypothesis and establishes empirically that pre-industrial geographical characteristics that were conducive for the emergence and progression of complementary cultural traits triggered an evolutionary process in language structures that has fostered the transmission of these cultural traits and has magnified their impact on the process of development.

In particular, the research establishes that regional variations in pre-industrial geographical characteristics that were conducive to higher return to agricultural investment, and thus to the emergence of long-term orientation, are at the root of existing cross-language variations in the presence of the future tense. Moreover, the study uncovers the effect of geographical characteristics on two additional language structures: grammatical gender and politeness distinctions. First, it suggests that

[1] E.g., Gallup et al. (1999), Guiso et al. (2004, 2006), Tabellini (2010), Acemoglu et al. (2001), Glaeser et al. (2004), and Ashraf and Galor (2013b).

[2] Specifically, Alesina et al. (2013) and Galor and Özak (2016).

[3] E.g., Cavalli-Sforza et al. (1994), Cavalli-Sforza (2000), and Richerson et al. (2010).

[4] Existing economic research predominantly views languages as an identifier of cultural and ethnic groups. Linguistic fractionalization as well as linguistic distance have been extensively used as a proxy for ethnic fractionalization and cultural distance in the exploration of the effect ethnic diversity on economic growth and the impact of cultural distance on the diffusion of development (Alesina et al., 2003; Alesina and Ferrara, 2005; Desmet et al., 2012; Easterly and Levine, 1997; Fearon, 2003; Harutyunyan and Özak, 2016; Miguel et al., 2004).

differences in the suitability of land for the adoption of agricultural technology, and its differential effect on agricultural productivity across genders, have contributed to the presence of cross-language variations in grammatical gender. Second, it indicates that regional variations in the diversity of soil quality that have contributed to specialization and trade, and hence to the emergence of the state and hierarchical structures, are at the origin of existing variations in the presence of politeness distinctions across languages.

The research further suggests that while language structures have been largely a reflection of past human experience, and in particular ancestral cultural traits, they have played a pivotal role in the persistent effect of cultural characteristics on the comparative economic development. Moreover, the evidence cannot refute the presence of a direct and independent effect of language structures on human behavior and contemporary economic outcomes.

The hypothesized coevolution of culture and language structures generates various predictions and novel insights. First, in light of the communication function of language, emerging language structures in the process of development have conceivably facilitated efficient communication across individuals, while enhancing the transmission of cultural values. Hence, natural selection across language structures gave an evolutionary advantage to the ones that reflected the dominating cultural traits. The theory therefore suggests that the geographical environment and the corresponding economic incentives, which have governed the evolution of cultural traits in the course of human history, have also triggered the emergence and evolution of complementary language structures. In particular, regional differences in geographical characteristics that have contributed to the emergence of variations in cultural traits should have also contributed to the cross-language variations of complementary language structures.

Second, in view of the pivotal role of language in the transmission of knowledge and values, language structures have plausibly affected the diffusion of cultural values and therefore the derived human behavior across members of society, reinforcing existing cultural traits and their intergenerational transmission. Moreover, it is not inconceivable that language structures per se may have directly influenced individuals' mindsets and thus human behavior, beyond the cultural transmission channel.[5]

Third, considering the pivotal role of language as a coordination device across members of society, the evolution of language structures necessitated and reflected the adoption of linguistic mutations by society as a whole. Unlike the feasibility of a unilateral deviation by individuals from existing cultural norms, the diffusion of unilateral linguistic innovations is rather limited and language structures therefore tend to be more persistent than cultural traits. Thus, inevitably, cultural traits encoded in language structures would be expected to be more persistent across time and space.

The proposed hypothesis about the interaction between the geographical environment and the coevolution of cultural and linguistic traits is exemplified in three distinct settings. Consider a hierarchical society characterized by obedience, conformity, and power distance. Conceivably, language structures that reinforced the existing hierarchical structure and cultural norms were likely to emerge and persist in this unequal society. In particular, politeness distinctions in pronouns (e.g., "tu" and "usted" in Spanish, "Du" and "Sie" in German, and "tu" and "vous" in French) were likely to ap-

[5]The Oxford English Dictionary defines mindset as "[a]n established set of attitudes, esp. regarded as typical of a particular group's social or cultural values; the outlook, philosophy, or values of a person;"... "an incident of a person's *Weltanschauung* or philosophy of life".

pear and endure in this hierarchical society.[6] Thus, geographical characteristics that were conducive to the development of hierarchical societies, (e.g., ecological diversity (Depetris-Chauvin and Özak, 2016; Fenske, 2014; Litina, 2014)) would be expected to be associated with the emergence of politeness distinctions as well.

Similarly, in a society characterized by distinct gender roles and consequently by the existence of gender bias, grammatical gender that could have fortified the existing social structure and cultural norms may have emerged and persisted over time. Moreover, agricultural characteristics that were conducive to a gender gap in agricultural productivity (e.g., crops and soil characteristics that were complementary for the usage of the plow (Alesina et al., 2013; Pryor, 1985)), and thus to distinct gender roles in society, may have fostered the emergence and the prevalence of grammatical gender. Finally, in societies characterized by long-term orientation, the use of the future tense could have affected the efficiency of communication. Moreover, pre-industrial agro-climatic characteristics that were conducive to higher return to agricultural investment and therefore to the prevalence of long-term orientation (Galor and Özak, 2016) may have affected the use of the future tense.

The hypothesized coevolution of culture and language structures partly reflects the demand for communication technology that complements the manifestation of existing cultural traits as well as the effects of two opposing linguistic forces that govern the emergence and evolution of language structures (Deutscher, 2010). First, the emergence of complementary language structures reflects the principle of *expressiveness*, permitting individuals to communicate more coherently and effectively.[7] Second, the subsequent evolution of language structures is guided by the principle of *efficiency*, which erodes frequently used language structures over time.[8] Importantly, the prevalence of complementary cultural traits and speaker population size affect the strength with which the opposing forces of expressiveness and efficiency operate. In particular, higher prevalence of complementary cultural traits bolsters the effect of expressiveness, while a larger population intensifies the effect of efficiency.[9]

The proposed hypothesis is tested in two stages. In the initial stage, the empirical analysis explores the origins of language structures, focusing on the geographical roots of the future tense, sex-based grammatical gender systems and politeness distinctions in pronouns. The analysis confirms the proposed hypothesis establishing the geographical origins of language structures. In particular, the empirical analysis examines whether agro-climatic characteristics, that have governed the return to agricultural investment and are thus associated with long-term orientation (Galor and Özak, 2016), have influenced the structure of the future tense. In a language-level analysis, the research establishes a negative robust association between pre-1500CE potential crop return and the existence of a future tense in a language. The estimated association is statistically and economically significant suggesting that, accounting for regional fixed-effects and confounding geographical characteristics of the language homeland, a one standard deviation increase in crop return is associated with a 12-23 percentage points

[6]Politeness distinctions could have emerged in order to mitigate the coordination cost in the interaction between individuals from various social strata (Brown and Levinson, 1987; Brown and Gilman, 1989; Helmbrecht, 2003, 2005).

[7]For instance, the number of words for ice or snow in eskimo languages is significantly higher than in languages which are spoken in tropical regions of the globe.

[8]All daughter languages of Latin, for instance, lost the *case structure* of nouns, which was prevalent in Latin.

[9]Lupyan and Dale (2010) find a negative correlation between the size of the speaker population and morphological complexity as captured by different language structures.

decrease in the probability of presence of a future tense in a language. Reassuringly, crop return is not associated with other language structures (e.g., the presence of a past or perfect tense, grammatical gender, possessive), suggesting that the future tense is indeed encoding long-term orientation.

The estimated effect of potential crop return (associated with agro-climatic conditions that are orthogonal to human intervention) on the structure of the future tense overcomes potential concerns about reverse causality, reflecting the effect of long-term orientation on cultivation methods, the choice of technologies, and actual crop returns. Moreover, accounting for a large set of confounding geographical characteristics and regional fixed-effects, mitigates concerns about the role of omitted geographical, institutional, cultural, or human characteristics that might have determined long-term orientation and are correlated with potential crop return. Furthermore, the results are robust to spatial auto-correlation, and to potential biases due to omitted factors.

The empirical analysis further explores the impact of potential crop return within the ancestral homeland of contemporary language families on the presence of the future tense in individual daughter languages. In light of the observation that contemporary languages within a language family descended from a common proto-language (Bouckaert et al., 2012; Pagel et al., 2013), the hypothesis further suggests that crop return in the ancestral homeland of the proto-language would have had a persistent effect on the presence of a future tense in its daughter languages. Consistent with this prediction, the analysis establishes that the share of daughter languages within a language family in which the future tense is present is negatively associated with crop return in the in the ancestral homeland of the proto-language.

The empirical methodology addresses potential concerns regarding omitted variables and sorting. By focusing on languages located outside the ancestral homeland of their proto-language and accounting for regional fixed effects, the analysis mirrors the epidemiological approach to cultural diffusion, thus addressing potential concerns regarding omitted variables at the host-region level and providing support to the view that the future tense was formed mostly in the proto-language. Moreover, the study exploits the descent of contemporary languages from proto-languages to establish the persistent effect of geographical characteristics in the proto-language homeland, rather than sorting in the course of the demic diffusion, on the evolution of the future tense. In particular, neither the change in crop return during the demic diffusion, nor the crop return in the contemporary homeland are associated with the existence of a future tense, alleviating concerns about sorting in the observed association.

Furthermore, the analysis explores the potential economic mechanisms through which crop return, and thus long-term orientation, might have affected the presence of the future tense. In particular, it establishes that the pattern of subsistence among the speakers of a language (i.e., agricultural intensity), and its effect on the scale and the complexity of society, had a significant negative effect of the presence of a future tense, reflecting the hypothesized reinforcement of the principle of efficiency over expressiveness in larger societies.

Similarly, the study uncovers the effect of geographical characteristics on two additional language structures: grammatical gender and politeness distinctions. The analysis establishes that variations in agricultural productivity across genders are associated with the emergence and prevalence of grammatical gender, while ecological diversity contributes to the emergence and prevalence of politeness

distinctions, reflecting the domination of the principle of expressiveness over efficiency. Interestingly, while the geographical characteristics of the ancestral homeland of their proto-language have a persistent effect on both the existence of the future tense and of sex-based grammatical gender systems, it is the geographical characteristics of the daughter languages' homeland that affect the existence of politeness distinctions, consistent with evidence about the greater adaptability of the latter language structure to environmental changes.

In its second stage, the empirical analysis examines the effects of language structures on contemporary economic outcomes, conceivably via their potential impact on the persistence of ancestral cultural traits as well as on individual behavior. In particular, it explores whether speaking a language with a future tense affects contemporary college attendance.[10] The analysis establishes a direct and indirect negative effect of speaking a language with a future tense (and thus having low long-term orientation) on college attendance.

In line with the proposed hypothesis, the empirical analysis establishes that second-generation migrants in the US, who speak languages characterized by the presence of the future tense, face a lower probability of attending college. Following the epidemiological approach (Fernández, 2012; Galor and Özak, 2016; Giuliano, 2007), the analysis exploits variations across second-generation migrants to identify the effect of the presence of the future tense on college attendance, accounting for ethnic and labor market conditions at the county level, individual level characteristics, and some some ancestral characteristics. The finding suggests that speaking a language with a future tense lowers the probability of attending college by about 20 percentage points.

Nevertheless, this conventional use of the epidemiological approach permits the identification of the persistent effect of cultural traits, but does not account for all other ancestral characteristics in the parental country of origin. Thus, the association between the presence of the future tense and college attainment may capture the persistent effect of other ancestral cultural characteristics rather than the direct effect of the language structure per se.

Fortunately, however, the focus on language structures permits this study to be the first within the epidemiological approach to account for the host county as well as the parental country of origin fixed-effects, allowing for a finer isolation of the direct effect of language from those of ancestral cultural characteristics. In particular, variations in the languages spoken by second-generation migrants with the same parental country of origin can be exploited to distinguish between the effect of a language structure per se and the effect of ancestral cultural traits. Moreover, and in contrast to existing studies that have followed the epidemiological approach, the cultural traits of second-generation migrants are uniquely mapped to the parental country of origin, the analysis overcomes the potential biases that could be generated by omitted ancestral characteristics.[11]

Interestingly, once ancestral cultural characteristics as well as the parental characteristics (i.e., education and the level of proficiency in the local language) are accounted for, the effect of the

[10] Given space limitations and the data requirements for the identification of the effects of language (explained below), the analysis focuses only on the effects of the future tense.

[11] This strategy can be employed in other settings, as long as languages spoken by second-generation migrants individuals are reported. Furthermore, linguistic homelands can be used as an alternative method of adjusting cultural, institutional, and bio-geographical factors for the ancestral composition of a population in a manner similar to Putterman and Weil (2010).

future tense falls by about 75%, but remains statistically and economically significant. In particular, the estimated effect suggests that speaking a language with a future tense lowers the probability of attending college by 4 percentage points.[12] Thus, one cannot refute the presence of a direct effect of language on human behavior.

This research is the first attempt in the economic literature to explore the geographical origins of language structures and to advance the hypothesis that regional variations in geographical characteristics have contributed to cross-language variations in language structures such as the presence of the future tense, grammatical gender, and politeness distinctions.[13] Furthermore, in order to overcome limitations of the existing studies about the association between language structures and economic outcomes (Chen, 2013; Roberts et al., 2015), the empirical methodology advanced in the course of this research augments the epidemiological approach and permits the study of the persistent effect of linguistic and other cultural factors, while accounting for ancestral characteristics. In particular, it suggests that variations in the languages spoken by second-generation migrants originated from the same ancestral region can be exploited to account for country of origin fixed-effects and thus to isolate cultural from linguistic persistence.

Moreover, it sheds additional light on the geographical and bio-cultural origins of comparative development (e.g., Ashraf and Galor, 2013b; Diamond, 1997), the interaction between the evolution of human traits and the process of development (Galor and Moav, 2002; Spolaore and Wacziarg, 2013), and the persistence of cultural characteristics (e.g., Alesina et al., 2013; Bisin and Verdier, 2000; Fernández, 2012; Nunn and Wantchekon, 2011).

2 Data

This section presents the data used in the empirical analysis of the origins of language structures. In particular, it introduces the data on the future tense, sex-based grammatical gender and politeness distinctions in nouns, as well as measures of their hypothesized geographical determinants.

2.1 Main Variables of Interest: Language Structures

This subsection introduces the main variables of interest in the analysis, namely the existence of an *inflectional future tense*, *sex-based grammatical gender systems* and *politeness distinctions in pronouns* across languages, based on The World Atlas of Language Structures - WALS - (Dryer, 2013), which is the most comprehensive database of language structures gathered from descriptive materials by 55 authors.

The analysis encodes the existence of a language structure S in language ℓ, S_ℓ, as follows:

$$S_\ell = \begin{cases} 1 & \text{if the structure exists in language } \ell, \\ 0 & \text{if the structure does not exist in language } \ell. \end{cases}$$

[12] The qualitative results are robust to the exclusion of the main languages spoken in the US.

[13] In contrast, Michalopoulos (2012) and Ashraf and Galor (2013a) explore the geographical attributes (diversity of soil quality and migratory distance from Africa) that contributed to variation in the number of languages within a geographical region.

In order to link linguistic characteristics of a language to the history and geography of the people that speak that language, the analysis creates a correspondence table that links the languages in WALS to other datasets. In particular, the analysis merges the linguistic data from WALS with the *Ethnologue* (Lewis et al., 2009) in order to identify the geographical regions where languages are spoken today. This allows the assignment of geographical characteristics of the linguistic homeland to each language, where the linguistic homeland of a language is the indigenous region where the language is spoken today.[14] Additionally, the research links these two datasets to the Ethnographic Atlas (Murdock, 1967) and the Standard Cross Cultural Sample (Murdock and White, 1969) in order to link each language to the ethnographic data of its aboriginal speakers.

2.1.1 Inflectional Future Tense

According to linguists, languages differ in the structure of when and how they mark future events (Dahl and Velupillai, 2013). In particular, in some languages, a change in the structure of the verb when making reference to the future is required, while in others it is not. A language is classified as having an inflectional future tense if verbs display morphological variation, i.e., a change in the form of the verb (typically the ending), to express the future tense (Dahl, 1985, 2000; Dahl and Velupillai, 2013). Importantly, inflectional markings are obligatory and have a wide range of uses (Dryer, 2013).[15] For example, French is a language with inflectional future tense, since speakers are required to change the form of the verb when referring to the future:

Present:	Future:
"Il *fait* froid aujour d'hui."	"Il *fera* froid demain."
(It is cold today)	(It will be cold tomorrow)

Hence, the conjugation of the verb *faire* changes when used to speak about the future. On the other hand, in Finnish, the present tense is used in reference to both the present and the future:

Present:	Future:
"Tänään *on* kylmää."	"Huomenna *on* kylmää."
(Today is cold)	(Tomorrow *is* cold)

Thus, in this case the conjugation of the verb *olla* does not change when referencing the future.

Dahl and Velupillai (2013) provides data on the existence of future tense in 222 languages. The analysis expands this set of languages using data from Dahl (1985) and Dahl (2000). The expanded dataset on the existence of a future tense includes a total of 275 contemporary languages, from 76 different language families. Importantly, 90% of languages included in Ethnologue belong to these

[14] Thus, the linguistic homeland of Spanish is Spain rather than the regions that speak Spanish outside of Spain.

[15] In the literature of typological linguistics, linguists often separate between *strong* future-time reference vs. *weak* future-time reference. This dichotomy is very similar to the one used in this analysis. In particular, weak future-time reference is equivalent to the lack of an inflectional future tense (see discussion in Chen, 2013).

families, which constitute about 1/3 of all language families in Ethnologue.[16]

Table D.1 and Figure 1 describe the distribution of the future tense in the dataset.[17] They demonstrate wide regional variations in the existence of future tense. In particular, in most regions about 50% of the languages in the sample do not have a future tense.

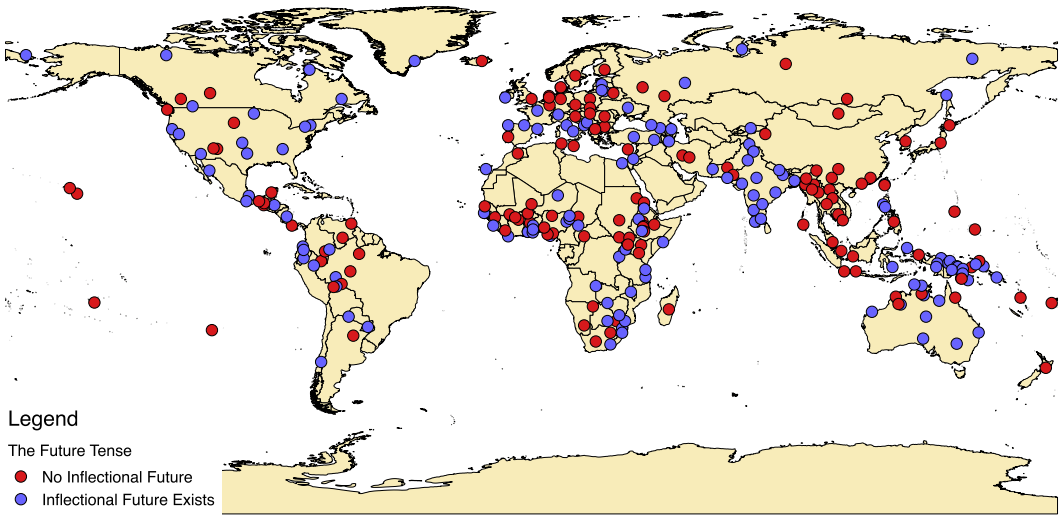


Figure 1: Global Distribution of Inflectional Future Tense

2.1.2 Sex-Based Grammatical Gender Systems

Sex-based grammatical gender systems vary across languages. Corbett (2013) provides data on sex-based grammatical gender systems for 227 languages across 76 language families. Table D.2 and Figure 2 describe the distribution of this measure of sex-based grammatical gender systems in the dataset. They demonstrate wide regional variations in the existence of future tense. In particular, about 37% of the languages in the sample do have a sex-based grammatical gender systems.

2.1.3 Politeness Distinctions in Pronouns

Politeness distinctions in pronouns (e.g., "tu" and "usted" in Spanish, "Du" and "Sie" in German, and "tu" and "vous" in French) vary across languages. Helmbrecht (2013) provides data on second-person politeness distinctions for 207 languages across 69 language families. Table D.3 and Figure 3 describe the distribution of this measure of politeness distinctions in the dataset. They demonstrate wide regional variations in the existence of politeness distinctions. In particular, about 34% of the languages in the sample do have a politeness distinctions.

[16]Many linguistic genera or families only have one language in either WALS or Ethnologue, which mostly precludes the analysis from using within-genus or within-family variation.

[17]Any reference to a future tense in the text is implicitly referring to an inflectional future tense.

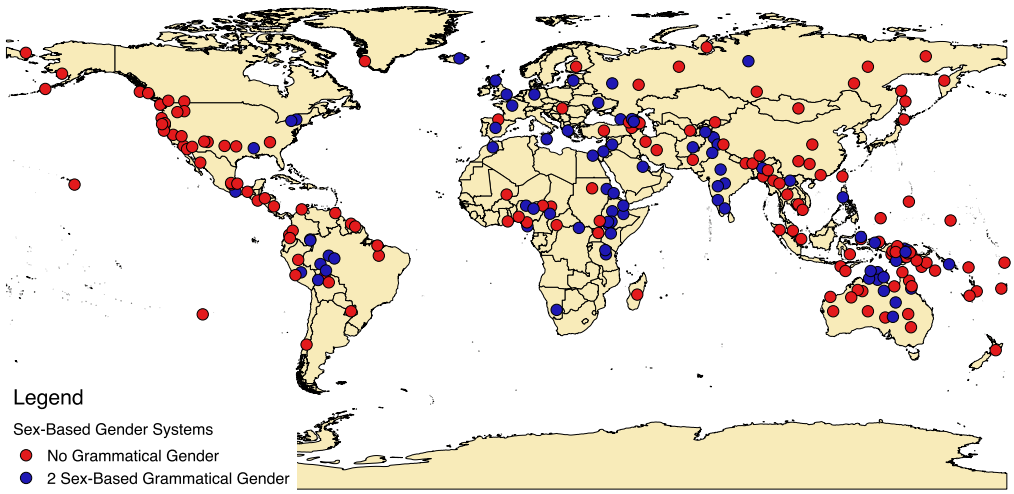


Figure 2: Global Distribution of Sex-Based Grammatical Gender System

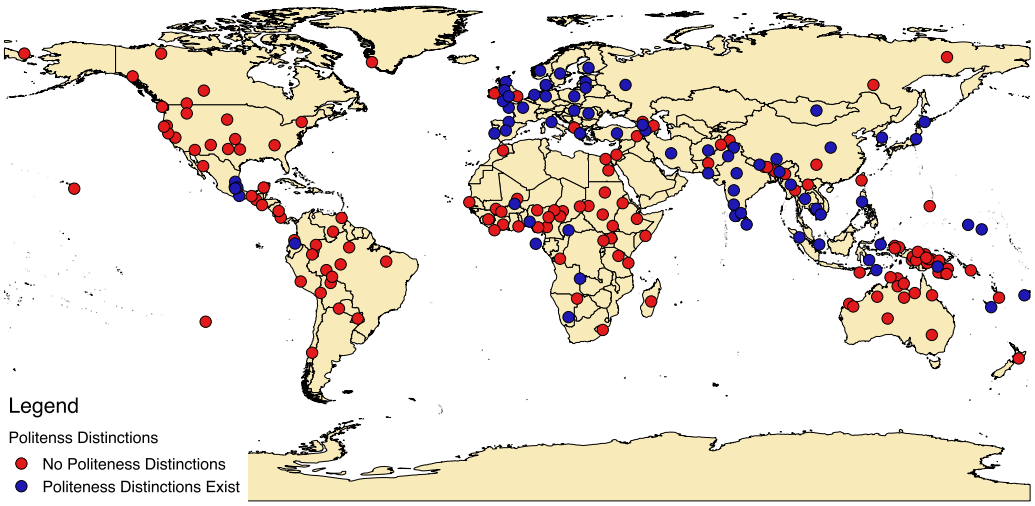


Figure 3: Global Distribution of Politeness Distinctions

2.2 Main Independent Variables

2.2.1 Pre-1500 CE Crop Return

This subsection introduces the historical potential crop return in a linguistic homeland, which is the hypothesized geographical origin of the future tense. In particular, the historical potential crop return in a location measures the potential daily calories from cultivating the crop with maximal caloric yield during the pre-1500CE era in that location. These measures are based on the Caloric Suitability Index - CSI - (Galor and Özak, 2015, 2016), which provide measures of historical (pre-1500CE) potential crop yield and growth cycles for each grid across the globe.

The measures of historical crop yield and growth cycles constructed by Galor and Özak (2015, 2016) are based on data from the Global Agro-Ecological Zones (GAEZ) project of the Food and Agriculture Organization (FAO), which supplies global estimates of crop yield and crop growth cycle

for 48 crops in grids with cells size of $5' \times 5'$ (i.e., approximately 100 km^2). Moreover, the CSI measures are based on the agro-climatic estimates under low level of inputs and rain-fed agriculture.[18] These restrictions remove the potential concern that the level of agricultural inputs, the irrigation method, and soil quality, reflect endogenous choices that could be potentially correlated with time preference (or other cultural traits), and thus, with the existence of future tense.

In order to capture the nutritional differences across crops, and thus, to ensure comparability of yields across crops, Galor and Özak (2015, 2016) convert each crop's yield in the GAEZ data (measured in tons, per hectare, per year), into caloric yield (measured in millions of kilo calories, per hectare, per year) using the caloric content of crops provided by the United States Department of Agriculture Nutrient Database for Standard Reference. Given the caloric yield of each crop in a cell, Galor and Özak (2015, 2016) assign to each cell the yield and growth cycle of the crop that maximizes the yield in that cell.

The analysis employs the potential pre-1500 caloric yield and crop growth cycle to construct a potential pre-1500CE caloric return index for each linguistic homeland. In particular, the analysis assigns to each linguistic homeland the average pre-1500CE daily caloric return per hectare (Galor and Özak, 2016). More specifically, the potential pre-1500CE caloric return per hectare per day in the homeland of language ℓ, R_ℓ, is given by

$$R_\ell = \frac{1}{|C_\ell|} \sum_{c \in C_\ell} \left(\frac{y_c}{g_c} \right), \tag{1}$$

where C_ℓ is the set of cells in the homeland of language ℓ, $|C_\ell|$ is the cardinality of this set, y_c and g_c are the potential pre-1500CE crop yield and growth cycle in cell c of the crop that maximizes caloric output in that cell. Figure 4 depicts the global distribution of the potential pre-1500CE crop return at the cell level.

2.2.2 Ecological Diversity & Plow Suitability

This subsection introduces the measures of ecological diversity and plow suitability in a linguistic homeland, which are the hypothesized geographical origins of politeness distinctions in pronouns and sex-based grammatical gender systems.

Following Fenske (2014), ecological diversity within a linguistic homeland is a Herfindahl index of the share of each territory that is occupied by different ecological zones. In particular, the ecological

[18] For each crop, GAEZ provides estimates for crop yield based on three alternative levels of inputs – high, medium, and low - and two possible sources of water supply – rain-fed and irrigation. Additionally, for each input-water source category, it provides two separate estimates for crop yield, based on agro-climatic conditions, that are arguably unaffected by human intervention, and agro-ecological constraints, that could potentially reflect human intervention. The FAO dataset provides for each cell in the agro-climatic grid the potential yield for each crop (measured in tons, per hectare, per year). These estimates account for the effect of temperature and moisture on the growth of the crop, the impact of pests, diseases and weeds on the yield, as well as climatic related "workability constraints". In addition, each cell provides estimates for the growth cycle for each crop, capturing the days elapsed from the planting to full maturity.

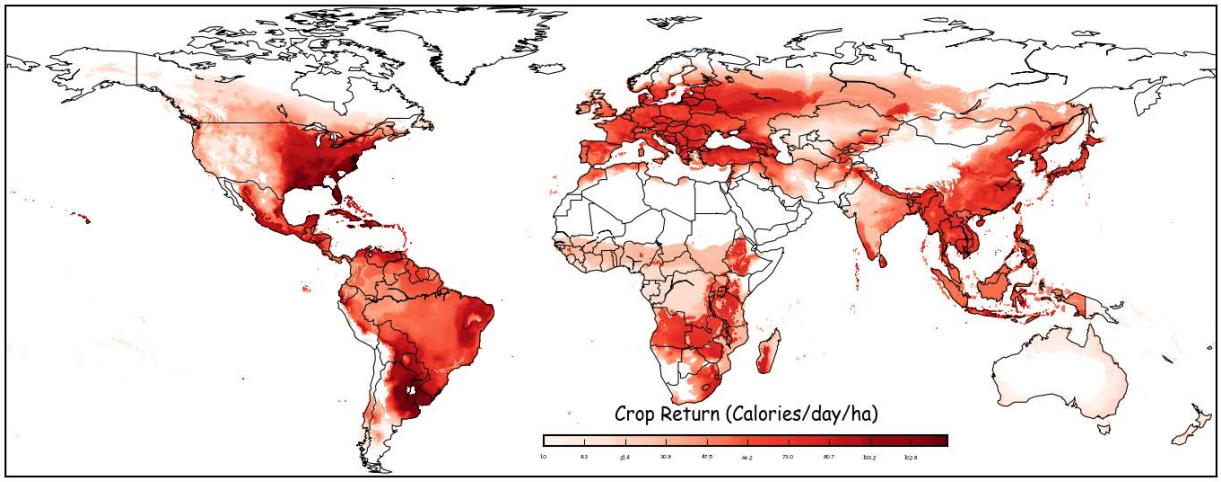


Figure 4: Crop Return (Pre-1500CE)

diversity E in the homeland of language ℓ is

$$E_\ell = 1 - \sum_{j=1}^{16} \left(\theta_{\ell j}\right)^2 \tag{2}$$

where $\theta_{\ell j}$ is the share of the homeland of language ℓ in ecological zone j, $j = 1, \ldots, 16$.[19]

Furthermore, following Galor and Özak (2015, 2016), the analysis constructs measures of average caloric suitability across *plow positive* and *plow negative* crop as defined by Pryor (1985).[20] Figure 5 depicts the global distribution of these two measures.

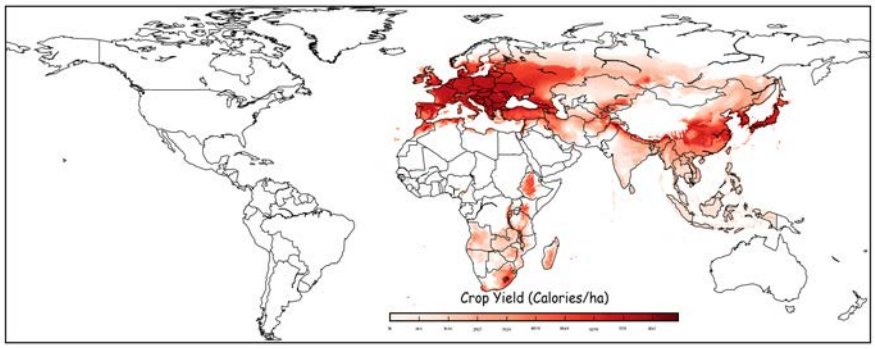
2.3 Additional Controls

The measure of crop return is potentially correlated with other geographical characteristics that may have affected the evolution of the future tense. Hence, the analysis accounts for the potential confounding effects of a wide range of geographical factors of the linguistic homeland such as absolute latitude, average elevation, terrain ruggedness, coast length, climatic conditions (average, standard deviation, volatility and spatial correlation) such as temperature and precipitation.[21] Additionally, the analysis accounts for the length of the unproductive period, which measures the potential number of days between the last harvest in one year and the first harvest of the next, in order to account for additional effects that agriculture and its temporal structure might have on the existence of the

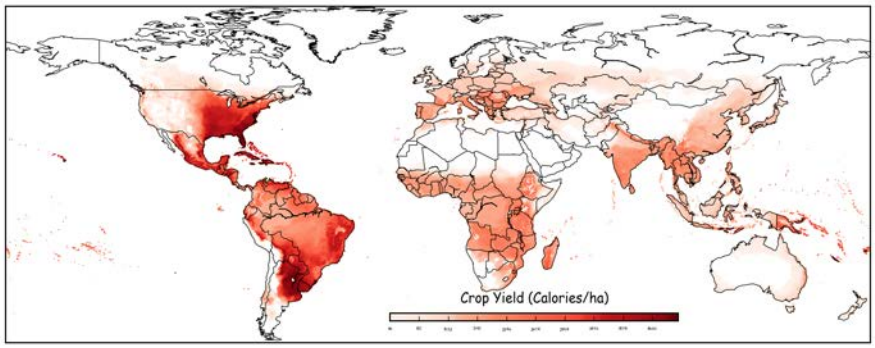
[19]Olson et al. (2001) provide a global dataset of biomes with 16 ecological zones: Boreal Forests/Taiga; Deserts and Xeric Shrublands; Flooded Grasslands and Savannas; Inland Water; Mangroves; Mediterranean Forests, Woodlands and Scrub; Montane Grasslands and Shrublands; Rock and Ice; Temperate Broadleaf and Mixed Forests; Temperate Conifer Forests; Temperate Grasslands, Savannas and Shrublands; Tropical and Subtropical Coniferous Forests; Tropical and Subtropical Dry Broadleaf Forests; Tropical and Subtropical Grasslands, Savannas and Shrublands; Tropical and Subtropical Moist Broadleaf Forests; Tundra.

[20]Plow positive crops include wheat, barley, rye, buckwheat, teff, and wet rice. Plow negative crops include all root crops, tree crops, and grains such as millet, sorghum, dry rice, and maize. Non-caloric based measures were computed by Alesina et al. (2013).

[21]The summary statistics and description of all variables used in the analysis is provided in Table D.4.



(a) Plow Positive



(b) Plow Negative

Figure 5: Average Caloric Suitability Index (Pre-1500CE) for Crops that are Complementary (Plow Positive) and not Complementary (Plow Negative) to the use of the Plow

future.[22]

Furthermore, the analysis accounts for regional fixed-effects, capturing unobserved region-specific geographical and historical characteristics that may have codetermined the global distribution of the future tense. Moreover, for each language the analysis employs additional data on its language structures, taken from WALS (Dryer, 2013), in order to overcome the potential concern that the results are driven by a general characteristic of a language. Finally, for each language, the analysis employs ethnographic data on its speakers from the Ethnographic Atlas and the Standard Cross-Cultural Sample (Murdock, 1967; Murdock and White, 1969) in order to analyze the mechanisms suggested by the theory.

3 The Origins of Future Tense

This section explores the empirical relation between the historical return to investment in agriculture and the existence of a future tense across languages as well as language families.

[22]This captures the effect of winter in temperate regions, and similar effects in other parts of the world.

3.1 Empirical Strategy

In order to explore this relation, the following empirical specification is estimated via a probit model:

$$P\left(S_\ell \mid R_\ell, \{X_{\ell j}\}, \{\delta_c\}\right) = \Phi\left(\beta_0 + \beta_1 R_\ell + \sum_j \gamma_{0j} X_{\ell j} + \sum_c \gamma_c \delta_{\ell c}\right), \tag{3}$$

where S_ℓ denotes whether language ℓ has a future tense, R_ℓ denotes the pre-1500CE crop return in the homeland of language ℓ; $X_{\ell j}$ is geographical characteristic j of the homeland of language ℓ; $\{\delta_{\ell c}\}$ is a complete set of regional fixed-effects. For robustness and also in order to develop additional analyses, the research also estimates a Linear Probability Model by means of Ordinary Least Squares (OLS).

The identification of the effect of pre-1500CE crop return on future tense is subject to various potential concerns. First, if as proposed by the theory, future tense encodes time preference and long-term orientation, the estimated effect may reflect the consequences of variations in the latter on the choice of technologies and therefore actual crop returns. Hence, to overcome this concern about reverse causality, this research exploits variations in potential (rather than actual) crop returns associated with agro-climatic conditions that are orthogonal to human intervention.

Second, the results may be biased by omitted geographical, institutional, cultural, or human characteristics that might have determined time preference and long-term orientation and are correlated with potential crop return. Thus, several strategies are employed to mitigate this concern: (i) The analysis accounts for a large set of confounding geographical characteristics (e.g., absolute latitude, average elevation, terrain ruggedness, coast length, and climatic conditions measured by the average, standard deviation, volatility and spatial correlation of temperature and precipitation). (ii) It accounts for regional fixed-effects, capturing unobserved time-invariant heterogeneity at the regional level. (iii) It explores the size and sign of the potential bias generated by omitted factors.

3.2 Crop Return and Future Tense

This section analyzes the relation between crop return and the emergence of the future tense using contemporary languages as the unit of analysis. In particular, Table 1 explores the effect of the pre-1500CE crop return on the existence of future tense in a language for the full sample of languages.[23] Column (1) shows the unconditional correlation between the pre-1500CE crop return and the existence of future tense. The estimated coefficient is negative and statistically significant at the 5%, and suggests that a one standard deviation increase in crop return reduces the probability of having a future tense in a language by 6%.

Column (2) accounts for regional fixed-effects and, therefore, for any unobserved time-invariant heterogeneity at the regional level. Reassuringly, the coefficient on pre-1500CE crop return becomes more negative and increases its statistical significance, suggesting that unobserved time-invariant factors at the regional level may have biased the coefficient towards zero. The estimated coefficient suggests that a one standard deviation increase in crop return is associated with a reduction of 8% in the probability of having a future tense.

[23] Table A.1 shows that the results are similar when using the linear probability model (OLS).

Table 1: Crop Return and Future Tense (Probit)

	Existence of Future Tense								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Crop Return (pre-1500CE)	-0.06**	-0.08**	-0.08**	-0.08**	-0.08**	-0.08**	-0.09**	-0.09***	-0.12***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Absolute Latitude			-0.10*	-0.10*	-0.10*	-0.08	-0.07	-0.09	-0.13
			(0.05)	(0.05)	(0.05)	(0.05)	(0.06)	(0.10)	(0.10)
Elevation				-0.00	-0.02	-0.03	-0.01	-0.04	-0.03
				(0.03)	(0.04)	(0.04)	(0.04)	(0.05)	(0.05)
Ruggedness					0.04	0.04	0.02	0.02	0.02
					(0.04)	(0.04)	(0.04)	(0.05)	(0.04)
Coast Length						-0.10***	-0.08***	-0.07**	-0.08**
						(0.03)	(0.03)	(0.03)	(0.04)
Precipitation							0.00	0.01	-0.00
							(0.08)	(0.08)	(0.08)
Precipitation (std)							-0.09***	-0.05	-0.05
							(0.04)	(0.06)	(0.05)
Precipitation Volatility							0.05	0.03	0.04
							(0.08)	(0.08)	(0.08)
Precipitation Spatial Correlation							-0.02	-1.05***	-0.97***
							(0.04)	(0.31)	(0.31)
Temperature								-0.06	-0.06
								(0.08)	(0.08)
Temperature (std)								-0.05	-0.05
								(0.05)	(0.05)
Temperature Volatility								0.04	0.08
								(0.09)	(0.09)
Temperature Spatial Correlation								1.04***	0.96***
								(0.31)	(0.31)
Unproductive Period (pre-1500CE)									-0.10***
									(0.03)
Regional FE	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Pseudo-R^2	0.01	0.04	0.05	0.05	0.05	0.07	0.09	0.12	0.14
Observations	275	275	275	275	275	275	275	275	275

Notes: This table establishes the negative, statistically, and economically significant effect of a region's pre-1500CE potential crop return on the existence of future tense in the language spoken in this region, accounting for regional fixed-effects and other geographical characteristics. Geographical controls include absolute latitude, mean elevation, terrain ruggedness, and coast length, as well as other agriculture-related controls as precipitation and temperature means and standard deviations. All independent variables have been normalized by subtracting their mean and dividing by their standard deviation. Thus, all coefficients can be compared and show the effect of a one standard deviation in the independent variable on the probability of having a future tense in the language. Heteroskedasticity robust standard error estimates are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.

In columns (3)-(6) the analysis accounts also for other potential confounding geographical characteristics of the linguistic homelands. In particular, the analysis accounts for the homeland's absolute latitude, mean elevation above sea level, terrain ruggedness, and the length of its sea coast. Reassuringly, accounting for the effects of these geographical factors and unobserved regional heterogeneity does not alter the results.

Columns (7) and (8) additionally account for the potential confounding effects of climate as captured by measures of temperature and precipitation. In particular, given that the pre-1500CE crop return is based on climatic factors, it might be capturing any potential direct effects of (average) climate on the existence of future. Moreover, variability of climate may affect the riskiness of agricultural investment, thus reducing any potential effects of crop return. Nevertheless, accounting for the effect of average temperature and precipitation, their standard deviations and volatility, and the potential for spatial diversification of risk due to precipitation and temperature does not alter the qualitative results. Accounting for all of these climatic characteristics does not change the magnitude of the coefficient of the crop return, but increases its statistical significant to the 1% level. Furthermore, most of the controls have no significant effect on the emergence of future tense.

Finally, given that the future tense might be associated with planning in general, and in agriculture in particular, the relation between existence of future and historical crop return might spuriously be capturing, e.g., the effect of the length of the time between crop harvests. In order to overcome this potential concern, column (9) accounts for the number of days between the last harvest in one year and the first potentially feasible harvest in the following year. Hence, during this period a region is not agriculturally productive and people are forced to plan for survival during this period. Reassuringly, accounting for this unproductive period does not alter the main result. In fact, the coefficient becomes even more economically and statistically significant. In particular, its magnitude (in absolute value) is the largest obtained in the analysis and suggests that a one standard deviation increase in pre-1500CE crop return is associated with a 12% reduction in the probability of having a future tense. Moreover, the effect of the unproductive period is also negative and economically and statistically significant, suggesting that planning for the future has a similar effect on the existence of a future tense as crop return.[24]

These results lend credence to the idea that crop return, through its effect on time preference and long-term oriented behavior, decreases the probability of existence of a future tense in a language. Moreover, additional sources of variation in the requirements for planning, as captured by the yearly agricultural unproductive period, also are associated with a decrease in the probability of existence of future tense. Still, the results might be biased due to omitted variables, precluding a causal interpretation of the estimated coefficients.

[24]Nevertheless, the analysis reveals that unlike the crop return, the relation between the unproductive period and the existence of a future tense is not robust to the other specifications analyzed in Table 1, and its semi-partial R^2 is very low.

3.2.1 Robustness to Omitted Variables, Clustering and Spatial-Autocorrelation

This section explores the robustness of the previous results to omitted variables, clustering and spatial auto-correlation. In particular, as mentioned above, omitted variables can potentially bias the estimated effect of pre-1500CE crop return on the probability of existence of future tense in a language. Moreover, the existence of a future tense might not be independent across languages that belong to the same language genus or that are spatially closely located.

Table 2: Crop Return and Future Tense (OLS)
Robustness to Spatial-Autocorrelation, Clustering and Omitted Variables

	Existence of Future Tense								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Crop Return (pre-1500CE)	-0.06**	-0.08**	-0.08**	-0.08**	-0.09**	-0.08**	-0.09**	-0.09**	-0.12***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)
	([0.04])	([0.04])	([0.04])	([0.04])	([0.04])	([0.04])	([0.04])	([0.04])	([0.03])
	[0.04]	[0.04]	[0.04]	[0.04]	[0.04]	[0.04]	[0.04]	[0.04]	[0.03]
	{0.03}	{0.03}	{0.03}	{0.03}	{0.03}	{0.03}	{0.03}	{0.03}	{0.03}
Regional FE	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Altonji et al		-4.14	-3.86	-3.86	-3.45	-3.55	-3.28	-3.16	-2.09
δ		-0.32	-0.40	-0.40	-0.37	-0.55	-0.67	-1.05	-0.75
β-Oster		-0.28	-0.25	-0.25	-0.27	-0.21	-0.20	-0.16	-0.23
R^2	0.01	0.06	0.07	0.07	0.07	0.09	0.11	0.15	0.17
Adjusted-R^2	0.01	0.03	0.04	0.04	0.04	0.05	0.06	0.09	0.11
Observations	275	275	275	275	275	275	275	275	275

Notes: This table shows the robustness of the results to selection by unobservables. It presents the Altonji et al. (2005) AET ratio as extended by Bellows and Miguel (2009). Additionally, it presents the δ and $\beta^*(1,1)$ statistics suggested by Oster (2014). All statistics suggest that the results are not driven by unobservables. Heteroskedasticity robust standard error estimates are reported in parentheses, clustered at the language genus in parenthesis and squared brackets, spatial auto-correlation corrected standard errors (Conley, 1999) in squared brackets and Cliff-Ord ML in curly brackets; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.

In order to analyze these issues further, Table 2 replicates the analysis of Table 1 using a Linear Probability Model. Interestingly, the estimated average marginal effects of the probit model in Table 1 are identical to the estimated effects using OLS. Additionally, Table 2 establishes the robustness of the results to clustering at the language genus level and to spatial auto-correlation. In particular, in all columns the statistical significance of the results is not affected by the method used to overcome the concerns due to the potential violation of the independence assumption.

Furthermore, Table 2 explores the size and sign of the potential bias generated by omitted variables. In particular, using statistics on the selection on observables and unobservables (Altonji et al., 2005; Bellows and Miguel, 2009; Oster, 2014), it establishes that the degree of omitted variable bias is low and is unlikely to explain the size of the estimated effect of crop return. More specifically, the research analyzes the change in the estimated coefficient once observables are controlled for. The results suggest that omitted factors would need to be 1-4 times more strongly negatively correlated with crop return than all the controls accounted for in order to explain the estimated effect of the crop return on the

emergence of the future tense. Thus, the estimated coefficient should be considered a lower bound of the true effect.

Thus, the analysis suggests that the true effect of historical returns to agricultural investment on the probability of existence of future in a language is economically and statistically significant. Indeed, in all specifications, the bias-adjusted estimated effect of pre-1500CE crop return is strictly negative and at least twice as large than the OLS estimate (Oster, 2014). In particular, the bias corrected estimate in column 9, which assumes the unobservables are as strongly correlated with pre-1500CE crop return as the set of observables accounted for, implies that a one standard deviation increase in crop return decreases the probability of existence of a future tense by 25%.

3.2.2 Crop Return and Other Language Structures

A potential concern with the previous results, is that the pre-1500CE crop return in a linguistic homeland might be capturing some general aspect about a language and the culture of the people who speak it. This concern is partially mitigated by the results of Galor and Özak (2016) who established that pre-1500CE crop return only affects time preference and does not have a significant effect on the other cultural traits studied by them. Still, it is possible that time preference and long-term orientation are encoded in other aspects or structures of a language or that crop return is associated with cultural traits not previously studied.

In order to address this potential concern, Table 3 explores the relation between pre-1500CE crop return and other temporal and non-temporal structures in a language. Given that not all the outcomes are binary, the analysis in Table 3 uses OLS in order to estimate the various relations. Column (1) replicates the analysis of column (9) in Tables 1 and A.1 for comparison. As previously established, pre-1500CE crop return is economically and statistically significantly negatively associated with the existence of a future tense. Columns (2) and (3) examine whether crop return is associated with other tenses and aspects in the verbal system, specifically with the existence of a past tense and a perfect tense. The results suggest that crop return is not statistically significantly associated with these additional temporal structures of language.

Similarly, columns (4)-(9) explore the relation between pre-1500CE crop return and other non-temporal characteristics of a language such as (i) the number of gender distinctions it has, (ii) whether its has possessive classifications, (iii) whether it has coding for evidentiality, (iv) the number of consonants, (v) the ratio of consonants to vowels, and (vi) the number of colors in the language (Dryer, 2013). Again, crop return does not have a statistically significant association with any of these structures.[25]

These results suggest that crop return is not significantly associated with other temporal and non-temporal structures of languages. On the contrary, pre-1500CE crop return is statistically and economically significantly associated only with the existence of a future tense. Thus, in consonance with the evidence of Galor and Özak (2016), these results support the hypothesis that pre-1500CE crop return only affects time preference and long-term orientation, which are encoded only in the

[25]The results are even stronger if one estimates a Probit model for the language structures that are coded as binary response variables. In particular, the possessive structure is not statistically significantly associated with crop return even at the 15% significance.

Table 3: Crop Return and Language Structures

	Language Structure								
	Temporal Structures			Non-Temporal Structures					
	Future	Past	Perfect	Gender	Possessive	Evidentiality	Consonants	C/V Ratio	Colors
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Crop Return (pre-1500CE)	-0.12***	-0.06	0.05	0.03	-0.07*	0.00	0.08	-0.08	0.06
	(0.03)	(0.04)	(0.04)	(0.03)	(0.04)	(0.03)	(0.06)	(0.05)	(0.34)
All Geographic Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Regional FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted-R^2	0.11	0.08	0.14	0.20	0.15	0.20	0.31	0.19	-0.03
Observations	275	218	218	244	224	386	540	541	117

Notes: This table establishes the negative, statistically, and economically significant effect of pre-1500CE potential crop return on the existence of future tense in a language, and not with any other language structure. The analysis accounts for regional fixed-effects and other geographical characteristics as in previous tables. Other language structures include the existence a past tense, a perfect tense, the number of genders, the existence of obligatory possessive inflections, semantic distinctions of evidentiality, the number of consonants, the ratio of consonants to vowels and the number of colors. Heteroskedasticity robust standard error estimates are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.

future tense.

3.3 Persistent Effect of Crop Return in Languages' Ancestral Homeland

In light of the view that contemporary languages within a language family descend from a common proto-language, if the future tense started forming during these pre-historic times, the theory suggests that the crop return in the ancestral homeland of the proto-language (i.e., the Urheimat of a language family) would be expected to have a persistent effect on the presence of a future tense in its daughter languages. Similarly, crop return in the Urheimat should have a persistent effect on the share of languages with future tense within a language family.

Table 4 explores the association between pre-1500CE crop return in a language family's Urheimat and the probability that a daughter language has a future tense.[26] Column (1) shows the negative and significant unconditional association between the crop return in a language family's Urheimat and the existence of a future tense in a daughter language. The estimated coefficient is twice as large as the one established in column (1) in Table 1, suggesting a stronger association between the existence of a future tense and crop return in the Urheimat compared to the contemporary homeland of the daughter language. Moreover, the explanatory power of the Urheimat's crop return is larger than the one of the contemporary homeland. In fact, the pseudo-R^2 and semi-partial R^2 of the Urheimat's crop return are 4 and 20 times larger, respectively, than the ones corresponding to the return in the contemporary homeland.

Column (2) establishes that once additional geographical characteristics of the Urheimat as well as

[26]Given the lack of data on the location of a Urheimat for the Khosian family, the analysis in this section excludes this family.

time invariant regional unobserved heterogeneity are accounted for, the absolute value of the estimated negative coefficient on crop return increases by 61%. Thus, the estimated coefficient on crop return in a language's Urheimat is twice the size of the estimated effect of the crop return in its contemporary homeland (column (9) in Table 1). Moreover, the estimated coefficient is similar to the biased adjusted estimated effect in column (9) in Table 2.

Table 4: Urheimat's Crop Return and Future Tense

	Existence of Future Tense							
	All Languages				Languages In/Near Urheimat			
					All	$\Delta R <$ 0.5SD	$\Delta R <$ 0.25SD	$\Delta R <$ 0.01SD
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Urheimat Crop Return	-0.13***	-0.21***	-0.12***	-0.20***	-0.17***	-0.15***	-0.20**	-0.24**
	(0.04)	(0.06)	(0.04)	(0.06)	(0.06)	(0.04)	(0.08)	(0.07)
Change in Crop Return (ΔR)			0.03	0.05	0.08	0.04	0.14	6.62
			(0.05)	(0.04)	(0.05)	(0.17)	(0.41)	(31.72)
Regional FE	No	Yes	No	Yes	Yes	Yes	Yes	No
Urheimat Geographical Charac.	No	Yes	No	Yes	Yes	Yes	Yes	No
Change in Geographical Charac.	No	No	No	Yes	Yes	Yes	Yes	No
Pseudo-R^2	0.04	0.22	0.04	0.30	0.28	0.35	0.43	0.19
Observations	273	273	273	273	233	166	120	20
Language Families	75	75	75	75	74	70	56	20

Notes: This table explores the association between pre-1500CE crop return in a language family's Urheimat and the probability that a daughter language has a future tense. Heteroskedasticity robust standard error estimates clustered at the language family level are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.

Columns (3) and (4) establish that the persistent negative effect of crop return in a language family's Urheimat on the existence of a future tense in a daughter language remains economically and statistically significant, even after accounting for the change in crop return, as well as in other geographical characteristics, generated by the migration out of the Urheimat and into the language's homeland.

The results in columns (1)-(4) in Table 4 suggest that the origins of the future tense in contemporary languages are found in cultural processes that took place during the formation of proto-languages. In particular, since contemporary languages are descendants of proto-languages, the analysis in columns (1)-(4) in Table 4 captures the spirit of the analyses that focus on descendants (e.g., second-generation migrants) to identify the effect of culture (Galor and Özak, 2016; Giuliano, 2007).

Finally, the analysis further explores the relative contributions of pre-1500CE crop return in the homeland vs. the Urheimat to the presence of future tense in a daughter language. In particular, Table 5 establishes that the existence of a future tense among daughter languages located outside the Urheimat of their proto-language is only significantly negatively associated with crop return in the Urheimat. Thus, the results further suggest the deep-historical origins of the future tense and its

association with crop return and long-term orientation. In particular, a one standard deviation increase in crop return in the Urheimat is associated with 52 percentage points decrease in the probability of existence of a future tense in a daughter language. Importantly, by focusing on languages located outside the Urheimat of their proto-language and accounting for regional fixed effects, the analysis mirrors the epidemiological approach to cultural diffusion, thus addressing potential concerns regarding omitted variables at the host-region level and providing support to the view that the future tense was formed mostly in the proto-language.

Table 5: Persistent Effect of Urheimat Characteristics on Future Tense: Languages Outside Urheimat

	Existence of Future Tense			
	Homeland		Urheimat	
	(1)	(2)	(3)	(4)
Crop Return (Pre-1500CE)	-0.01	-0.03	-0.14*	-0.52***
	(0.06)	(0.04)	(0.07)	(0.07)
Regional FE	No	Yes	No	Yes
Homeland Geographical Characteristics	No	Yes	No	No
Urheimat Geographical Characteristics	No	No	No	Yes
Adjusted-R^2	-0.01	0.12	0.04	0.17
Observations	163	163	163	163
Language Families	19	19	19	19

Notes: This Table explores the relative contributions of pre-1500CE crop return in the homeland vs. the Urheimat to the presence of future tense in a daughter language. Heteroskedasticity robust standard error estimates clustered at the language family level are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests; All regressions include a constant.

3.4 Sorting

This section exploits the descent of contemporary languages from proto-languages to explore the relative contributions of the persistent effect of geographical characteristics in the proto-language's homeland and sorting in the course of the demic diffusion of languages on the evolution of the future tense.

A potential concern with the results in section 3.3 is that individuals who spoke languages characterized by the presence (or absence) of the future tense could have sorted into regions with high return to agricultural investment. Although, this sorting would not affect the nature of the association (i.e., variations in the return to agricultural investment across Urheimats of languages would still be the origin of variations in the existence of a future tense across languages), it may alter the cultural interpretation.

The analysis in columns (5)-(8) in Table 4 addresses this potential concern by replicating the analysis for languages that remained in the same region as their Urheimat and had small changes in their crop return. More specifically, by constraining the set of languages to those which remained in

the same region as the proto-language, the analysis excludes languages that were subjected to longer migratory processes. Additionally, by constraining the differences in return between the homeland and the Urheimat, the analysis constrains the potential incentives that might have caused people to sort themselves across locations. Reassuringly, the qualitative results remain unchanged, mitigating concerns about the effect of sorting in the course of the demic diffusion of languages on the evolution of the future tense. Moreover, neither the change in crop return (columns (3)-(8)) nor crop return in the contemporary homeland (Table 5) are associated with the existence of a future tense, further alleviating concerns about sorting in the observed association.

3.5 Robustness to Sample Selection Bias and Measurement Error

This section explores the robustness of the analysis to potential sample selection bias and measurement error. In particular, a potential concern with the previous results is that the sample of languages for which data on the existence of a future tense is available is not representative of the universe of languages. Thus, some genera or families might be over or under-represented and drive the results. Moreover, if language structures originated in the proto-languages that generated the different families, and are, thus, shared within language families, then languages within a language family might not contribute real independent information. Reassuringly though, this last concern has been, at least partially, addressed by (i) clustering at the family level to account for the lack of independence across observations (Table 2) and (ii) accounting for other geographical characteristics of the Urheimat (Table 4).

Table 6: Persistent Effect of Urheimat Characteristics:
Share of Daughter Languages with Future Tense

	Share of Daughter Languages with Future Tense					
	(1)	(2)	(3)	(4)	(5)	(6)
Crop Return (pre-1500CE)	-0.19**	-0.25***	-0.25***	-0.24***	-0.20***	-0.23***
	(0.06)	(0.04)	(0.05)	(0.05)	(0.05)	(0.06)
Regional FE	No	Yes	Yes	Yes	Yes	Yes
Main Geographical Controls	No	No	Yes	Yes	Yes	Yes
Precipitation Controls	No	No	No	Yes	Yes	Yes
Temperature Controls	No	No	No	No	Yes	Yes
Unproductive Period	No	No	No	No	No	Yes
Observations	74	74	74	74	74	74

Notes: This table establishes the negative statistically and economically significant effect of crop return in a language family's Urheimat on its share of daughter languages with a future tense. Coefficients are average marginal effects of a zero-inflated fractional regression, in which observations are weighted to account for missing languages and future tense data. Heteroskedasticity robust standard error estimates are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.

The analysis further addresses these potential concerns in two ways. First, it explores the relation between the crop return in a language family's Urheimat and the share of daughter languages that have a future tense. In particular, the theory suggests that if language structures started forming in

the proto-language, crop return in the Urheimat should have a persistent effect on the share of languages with future tense within a language family. Table 6 establishes the robust negative statistically and economically significant relation between the Urheimat's crop return and the share of daughter languages that have a future tense. The coefficients in the table are the average marginal effects of increasing crop return in the Urheimat in a zero-inflated fractional regression where observations are weighted to account for missing language data within the family.[27] The estimates imply that a one standard deviation increase in the Urheimat's crop return is associated with a decrease of 23 percentage points in the share of daughter languages that have a future tense. Figure 6 depicts the association in an OLS regression that accounts for the same controls as column (6).

Second, in order to account for a potential mismeasurement in the data, due to either mismeasurement of the existence of a future tense in a daughter language or the location of the Urheimat, the research additionally replicates the analysis using various strategies. In particular, if language structures had developed in the proto-languages and never changed after that, all daughter languages ought to share the same language structures. Thus, any within-family variation would be generated by mismeasurement of the existence of the future tense. In order to address this potential concern, Tables A.3 and A.4 replicate the analysis assuming that the proto-language of each family had a future tense if the majority of contemporary languages in the family have a future tense. These tables establish that the probability of a proto-language having a future tense is decreasing in the crop return in its Urheimat.

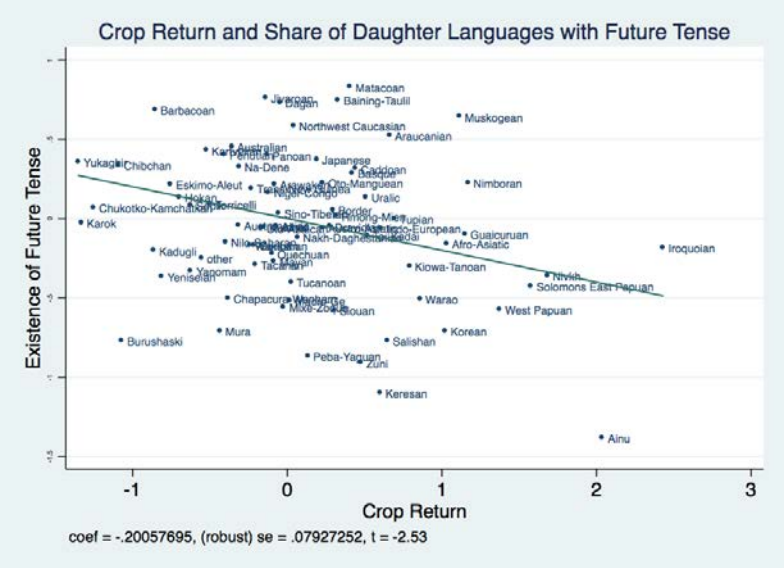


Figure 6: Persistent Effect of Urheimat's Crop Return
Share of Daughter Languages with Future Tense

As an additional strategy to address these potential concerns, the analysis explores the relation between crop return and future tense at the language family and genus levels. In particular, Tables A.5 and A.6 analyze the relation between pre-1500CE crop return and future at the language family level, where the family level values are given by the mean and median value, respectively, within

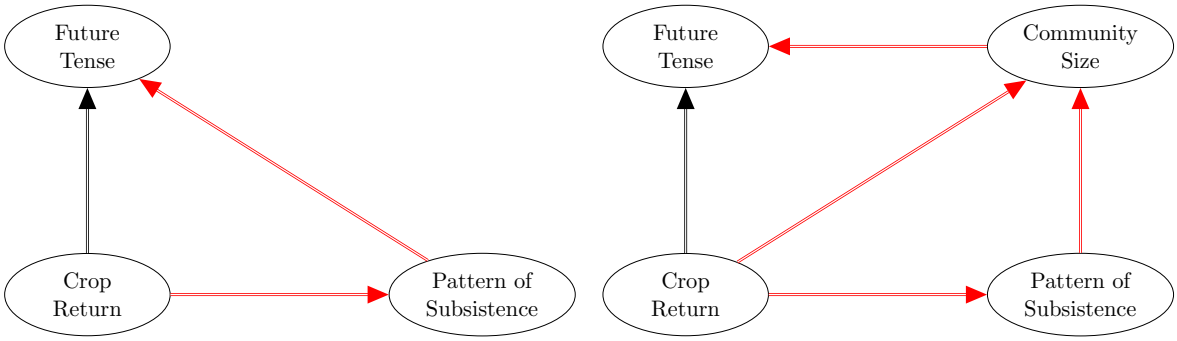
[27] Table A.2 establishes that similar results are obtained if observations are not weighted.

each language family. Reassuringly, the results remain qualitatively unchanged and imply that a one standard deviation increase in crop return is associated with a 44 percentage points decrease in the share of daughter languages with a future tense. Alternatively, the results imply a 43 percentage points decrease in the probability a proto-language has a future tense. Moreover, similar results are obtained at the language genus level (Tables A.7 and A.8).

Although reassuring, these results may still be biased if the proto-language is not well approximated by the mean or median of the languages in the family. Thus, as an additional approach, the analysis uses individual languages in the family to approximate the proto-language. In particular, Table A.11 shows the average estimated association between crop return and future tense obtained in 5000 simulations, where in each simulation the sample of language families is generated by selecting randomly one language within each family. Again the results suggest an economically and statistically significant negative association between pre-1500CE crop return and the existence of a future tense. Moreover, similar results are obtained if only geography is sampled, while existence of a future tense is assumed to be given by the language family's median (Table A.12).

3.6 Mechanisms

This section explores potential mechanisms through which cultural traits are encoded in language structures. In particular, the theory suggests that long-term orientation is encoded in the existence of a future tense, which following Galor and Özak (2016) suggests that crop return affects the existence of a future tense through the pattern of subsistence of its speakers (Figure 7(a)). Additionally, the theory suggests a negative association between long-term orientation and the existence of a future tense, if the linguistic expressiveness mechanism is dominated by the linguistic efficiency mechanism.[28] This suggests in particular that larger societies, where the social cost of having complex language structures increases, should be less likely to speak a language with a future tense (Figure 7(b)).



(a) Crop Return & Pattern of Subsistence (b) Crop Return & Population Scale

Figure 7: Mechanisms

Tables 7 and 8 present supporting evidence for the proposed mechanism in Figure 7(a). First, Table 7 establishes the robust positive association between crop return and the pattern of subsistence

[28] Indeed, Lupyan and Dale (2010) find that contemporary languages with larger societies tend to have simpler morphological structures. Similarly, Perkins (1992) argues that simpler morphological structures appear among more complex societies.

in a pre-modern society. In particular, using data from the Ethnographic Atlas (Murdock and White, 1969) on the importance of patterns of subsistence – hunting, gathering, fishing, animal husbandry and crop cultivation – the analysis explores the effect of crop return on a pre-modern society's agricultural intensity, i.e., the level of dependence on agriculture. Columns (1) and (2) explore the relation for all societies in the Ethnographic Atlas, while columns (3) and (4) constrain the analysis to the set of societies that speak languages for which data on the existence of the future tense is available. In line with the proposed theory, the results suggest that societies inhabiting regions with higher crop return have higher levels of agricultural intensity. In particular, the results imply that a one-standard deviation increase in crop return is associated with a 0.3 standard deviations increase in agricultural intensity.

Table 7: Agricultural Intensity and Crop Return

	Agricultural Intensity			
	Full Sample		Future Sample	
	(1)	(2)	(3)	(4)
Crop Return (pre-1500CE)	0.19***	0.22***	0.27***	0.30***
	(0.03)	(0.02)	(0.07)	(0.06)
Regional FE	No	Yes	No	Yes
All Geographical Controls	No	Yes	No	Yes
Adjusted-R^2	0.04	0.64	0.07	0.61
Observations	1306	1306	264	264

Notes: This table establishes the positive statistically and economically significant effect of a language homeland's crop return on the level of agricultural intensity of a pre-modern society that speaks that language. Standardized coefficients. Heteroskedasticity robust standard error estimates are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.

Second, Table 8 establishes the robust negative association between the level of agricultural intensity in a pre-modern society and the probability of existence of a future tense in the language it speaks. The results suggest that a one standard deviation increase in agricultural intensity is associated with a 10 percentage point decrease in the probability of existence of a future tense in the society's language. Although these results cannot be given a causal interpretation, they are in line with the proposed mechanism in Figure 7(a). Moreover, if the depicted causal graph were satisfied, i.e., crop return did not affect the existence of a future tense through any other channels, then it would be a valid instrument for agricultural intensity. In fact, Table A.13 replicates the analysis of Table 8 using OLS and also instrumenting agricultural suitability with crop return and the length of the unproductive period, both of which affect a society's agricultural intensity. Reassuringly, the OLS estimates are similar to the Probit ones, while the IV estimates are 3.4 times larger, suggesting that a one standard deviation increase in agricultural suitability would decrease the probability of the existence of a future tense by 0.34 percentage points. While this hints that the estimates in Table 8 might be biased towards zero, the overidentification test in Table A.13 suggests that the IV does not satisfy the exclusion restriction. In part, this could be explained by the proposed mechanism in Figure 7(b).

Table 8: Agricultural Intensity and Existence of Future Tense

	Existence of Future Tense				
	(1)	(2)	(3)	(4)	(5)
Agricultural Intensity	-0.07**	-0.10***	-0.10**	-0.09**	-0.10**
	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)
Continental FE	No	Yes	Yes	Yes	Yes
Main Geographic Controls	No	No	Yes	Yes	Yes
Main Precipitation Controls	No	No	No	Yes	Yes
Main Temperature Controls	No	No	No	No	Yes
Pseudo-R^2	0.02	0.07	0.09	0.10	0.16
Observations	264	264	264	264	264

Notes: This table establishes the negative statistically and economically significant effect of the level of agricultural intensity of a pre-modern society on the existence of a future tense in the language spoken by the society. Standardized coefficients. Heteroskedasticity robust standard error estimates are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.

The mechanism depicted in Figure 7(b) is explored in Tables 9 and 10. In particular, Table 9 establishes the positive and significant association between the size of local communities and the level of agricultural intensity for the sample of all ethnicities in the Ethnographic Atlas (columns (1) and (2)) and for the sample of ethnicities for which data on the future tense is available (columns (3) and (4)). Additionally, it establishes that in both samples crop return is positively associated with the size of local communities, even after accounting for the effect of agricultural intensity (columns (5)-(8)). The analysis suggests that a one standard deviation increase in agricultural intensity is associated with 0.56 standard deviations increase in the size of local communities. Additionally, a one standard deviation increase in crop return is associated with 0.24 standard deviations increase in the size of local communities, above and beyond its indirect effect through agricultural intensity.

Finally, Table 10 establishes the negative association between the size of local communities and the existence of a future tense. As before, these results cannot be given a causal interpretation, but they support the proposed mechanism in Figure 7(b). Moreover, Table A.14 replicates the analysis using OLS and additionally instrumenting the size of local communities using crop return and the length of the unproductive period. Reassuringly, the OLS estimates are similar to the ones in Table 10. On the other hand, the IV estimates suggest a much larger causal effect of the size of local communities on the existence of a future tense, with a one standard deviation increase in the size of local communities decreasing the existence of a future tense by 0.29 percentage points. Moreover, the overidentification test in Table A.14 suggests the instruments are valid.

3.7 Crop Return, Long-Term Orientation and Future Tense

This section analyzes the empirical relation between the share of speakers of languages with a future tense and long-term orientation across countries in the contemporary era. In particular, the previous sections have provided historical evidence that the same forces that gave rise to long-term orientation

Table 9: Agricultural Intensity, Crop Return and Size of Local Community

	Size of Local Community							
	Agricultural Intensity				Both			
	Full Sample		Future Sample		Full Sample		Future Sample	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Agricultural Intensity	0.61***	0.67***	0.62***	0.69***	0.58***	0.62***	0.55***	0.56***
	(0.03)	(0.04)	(0.05)	(0.08)	(0.03)	(0.04)	(0.06)	(0.10)
Crop Return (pre-1500CE)					0.11***	0.15***	0.20***	0.24***
					(0.03)	(0.04)	(0.06)	(0.08)
Regional FE	No	Yes	No	Yes	No	Yes	No	Yes
All Geographical Controls	No	Yes	No	Yes	No	Yes	No	Yes
Adjusted-R^2	0.37	0.46	0.38	0.45	0.38	0.47	0.41	0.49
Observations	614	614	163	163	614	614	163	163

Notes: This table establishes the positive statistically and economically significant effect of the level of agricultural intensity of a pre-modern society and its homeland return to agricultural investment on the size of the local community. Standardized coefficients. Heteroskedasticity robust standard error estimates are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.

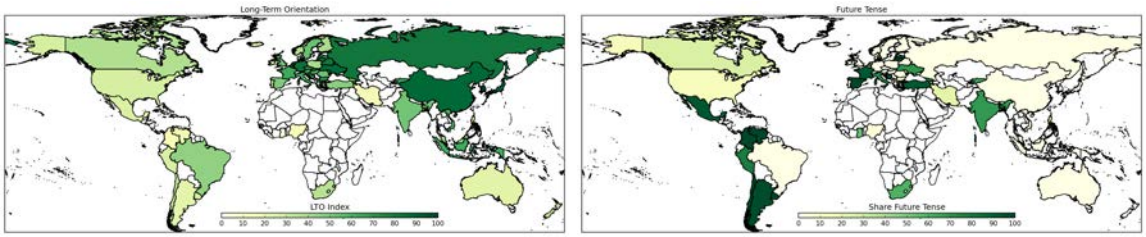
Table 10: Size of Local Community and Existence of Future Tense

	Existence of Future Tense				
	(1)	(2)	(3)	(4)	(5)
Size of Local Communities	-0.07*	-0.10**	-0.09**	-0.08*	-0.09**
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Continental FE	No	Yes	Yes	Yes	Yes
Main Geographic Controls	No	No	Yes	Yes	Yes
Main Precipitation Controls	No	No	No	Yes	Yes
Main Temperature Controls	No	No	No	No	Yes
Pseudo-R^2	0.02	0.09	0.12	0.14	0.18
Observations	163	163	163	163	163

Notes: This table establishes the negative, statistically and economically significant effect of the size of the community and the existence of future tense, accounting for regional fixed-effects and geographical characteristics as in previous tables. Heteroskedasticity robust standard error estimates are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.

also affected the existence of a future tense. But, if as proposed by the theory, language structures encode cultural traits and have a persistent effect on economic development, then one should expect individual's long-term orientation to be associated with the existence of a future tense in the language they speak.

In order to explore this association, the analysis uses two measures of long-term orientation at the country level (Galor and Özak, 2016). In particular, it examines the effect of the existence of future tense on the cultural dimension identified by Hofstede (1991) as Long-Term Orientation (LTO)



(a) Long-Term Orientation (b) Future Tense

Figure 8: Long-Term Orientation and Future Tense

as depicted in Figure 8.[29] Additionally, it uses a measure of LTO based on the share of individuals in a country that reported having LTO in the WVS.[30] In order to explore the relation between the share of speakers of languages with a future tense and Long-Term Orientation, the following empirical specification is estimated via ordinary least squares (OLS):

$$LTO_i = \beta_0 + \beta_1 \text{Future}_i + \beta_2 \text{Crop Return}_i + \sum_j \gamma_{0j} X_{ij} + \sum_c \gamma_c \delta_c + \epsilon_i, \tag{4}$$

where LTO_i is the level of Long-Term Orientation in country i, Future_i is the share of speakers of languages with a future tense in country i,[31] Crop Return_i is the pre-1500CE crop return of country i, $\{X_{ij}\}_j$ is a set of geographical characteristics of country i, $\{\delta_c\}$ is a complete set of regional fixed-effects, and ϵ_i is the error term of country i. The results of the previous section suggest that one should expect $\beta_1 < 0$ and $\beta_2 > 0$.

Table 11 presents the results of this analysis. In particular, column (1) shows the negative unconditional correlation between the share of speakers of languages with a future tense and LTO (Hofstede et al., 2010). The coefficient is statistically significant at the 1% and suggests that a one standard deviation increase in a country's share of population who speak a language with a future tense is associated with a 0.32 standard deviation decrease in LTO. Column (2) additionally accounts for the pre-1500CE crop return to which ancestors of the current population of the country had been exposed to (Galor and Özak, 2016). Reassuringly, both the share of speakers of languages with a future tense and pre-1500CE crop return are significant at the 1%, and their signs follow the pattern suggested by the theory and the historical evidence of previous sections. In particular, the coefficient on the future

[29] Hofstede et al. (2010) define Long-Term Orientation as the cultural value that stands for the fostering of virtues oriented toward future rewards, perseverance and thrift. Hofstede (1991) based his original analysis on data gathered from interviews of IBM employees across the world. This original data was later expanded using the data from the Chinese Values Survey and from the World Values Survey. The Long-Term Orientation (LTO) measure varies between 0 (short-term orientation) and 100 (long-term orientation). This measure is positively correlated with the importance ascribed future profits, savings rates, investment in real estate, and math and science scores (Hofstede et al., 2010).

[30] The measure of Long-Term Orientation is based on the following question in the WVS: "Here is a list of qualities that children can be encouraged to learn at home. Which, if any, do you consider to be especially important?" Individuals are considered to have Long-Term Orientation if they answered "Thrift, saving money and things".

[31] The share of speakers of languages with a future tense in country i is determined by the number of speakers of each language according to the Ethnologue. Since there are many languages for which the data on the existence of the future tense is missing, the analysis is restricted only to countries for which the data available covers at least 50% of the population.

Table 11: Long Term Orientation, Pre-1500 Crop Return and the Future Tense

	Long Term Orientation				
	Hofstede				WVS
	(1)	(2)	(3)	(4)	(5)
Future	-0.32***	-0.32***	-0.25**	-0.01	-0.26**
	(0.11)	(0.09)	(0.10)	(0.09)	(0.12)
Crop Return (Pre-1500, Ancestors)		0.54***	0.50***	0.46***	0.35**
		(0.10)	(0.08)	(0.09)	(0.14)
Main Geographic Controls	No	No	Yes	Yes	Yes
Regional FE	No	No	No	Yes	Yes
Adjusted-R^2	0.09	0.38	0.43	0.56	0.32
Observations	69	69	69	69	76

Notes: This table establishes the negative, statistically and economically significant association between the share of speakers of languages with a future tense and Long-Term Orientation. The analysis accounts for ancestors exposure to pre-1500CE crop return, regional fixed-effects and geographical characteristics as in previous tables. The analysis uses the measures of Long-term Orientation from Hofstede et al. (2010) and the World Values Survey. Standardized coefficients. Heteroskedasticity robust standard error estimates are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.

tense is negative, and its magnitude is hardly affected by the inclusion of pre-1500 crop return into the analysis. On the other hand, the coefficient on pre-1500CE crop return is positive and significant, suggesting that a one standard deviation increase in pre-1500CE crop return is associated with 0.54 standard deviation increase in LTO. It is important to highlight that in this analysis, pre-1500CE crop return is associated to the location of the ancestors of populations of contemporary countries, and not with languages' homelands, although in some cases the two might be identical. This allows the analysis to (partially) disentangle the association between these two historical components of LTO.

Column (3) additionally accounts for countries' geographical characteristics (absolute latitude, terrain ruggedness, mean elevation above sea level and coast length) without affecting the qualitative results. Additionally, column (4) accounts for regional time invariant unobserved heterogeneity. While the qualitative effect of pre-1500CE crop return remains unchanged, the coefficient on the future tense becomes insignificant, reflecting in part the lack of variation in the existence of future tense among languages spoken within many regions. E.g., due to their colonial history, most people in Latin America speak Spanish, and thus very little variation in the share of speakers with languages with a future tense remains in this region. Interestingly, the results in column (5), which replicate the analysis for the WVS measure in a larger sample, suggest that the association between LTO and future tense remains significant even when accounting for regional fixed-effects.

These results suggest that the future tense is associated with long-term orientation. Furthermore, the analysis provides suggestive evidence that the future tense is associated with long-term orientation in the contemporary era, above and beyond the effect of the crop return experienced by populations' ancestors. In particular, in a country where languages spoken today originated in regions different from the ones where the country's ancestors came from, the established association hints to a potential

direct effect of the future tense on contemporary levels of long-term association (and thus potentially on development). This might reflect the encoding of crop return in a language's homeland and its effect on long-term orientation into languages in the distant past.

4 The Origins of Other Language Structures

This section provides evidence that the theory presented above applies to other language structures as well. In particular, the analysis establishes the origins of two additional language structures: sex-based grammatical gender systems and politeness distinctions in pronouns.

4.1 Sex-Based Grammatical Gender Systems

The proposed hypothesis suggests that in a society characterized by distinct gender roles and consequently by the existence of gender bias, sex-based grammatical gender systems that could have fortified the existing social structure and cultural norms may have emerged and persisted over time. Moreover, agricultural characteristics that were conducive to a gender gap in agricultural productivity (e.g., crops and soil characteristics that were complementary for the usage of the plow (Alesina et al., 2013; Pryor, 1985)), and thus to distinct gender roles in society, may have fostered the emergence and the prevalence of sex-based grammatical gender systems.

Table 12: Geographic Origins of Plow Usage and Sex-Based Grammatical Genders

	Reduced Form		Mechanism			
	Grammatical Gender		Plow		Grammatical Gender	
	(1)	(2)	(3)	(4)	(5)	(6)
Plow Negative CSI (pre-1500CE)	-0.12**	-0.20***	-0.25***	-0.06***		
	(0.05)	(0.07)	(0.02)	(0.02)		
Caloric Suitability Index (pre-1500CE)	0.15***	0.21***	0.25***	0.10***		
	(0.05)	(0.06)	(0.02)	(0.02)		
Plow Usage					0.37***	0.20
					(0.10)	(0.12)
All Geographic Controls	No	Yes	No	Yes	No	Yes
Regional FE	No	Yes	No	Yes	No	Yes
Adjusted-R^2	0.03	0.21	0.20	0.47	0.13	0.28
Observations	217	217	1178	1178	114	114

Notes: This table establishes the positive statistically and economically significant effect of the geographical determinants of and actual plow usage on the existence of sex-based grammatical gender in a language. The first two columns provide the results of the reduced form, and columns (3)-(6) provide evidence on the mechanism. The table shows the estimates of an OLS regression. Heteroskedasticity robust standard error estimates are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.

Table 12 explores the relation between plow suitability, the usage of the plow, and the existence of sex-based grammatical genders. Columns (1) and (2) report the unconditional and conditional relation between plow and agricultural suitabilities and the existence of sex-based grammatical gender. The

results suggest a significant negative association between plow negative suitability and the existence of sex-based grammatical gender. In particular, the estimates suggest that a one standard deviation increase in plow negative suitability decreases the probability of having a sex-based grammatical gender by 17 percentage points. Columns (3) and (4) report the correlation between the usage of the plow across ethnic groups as reported in the Ethnographic Atlas and its determinants, without any controls and with the main geographical controls and regional fixed-effects. In line with the theory of Pryor (1985) and Alesina et al. (2013) negative plow suitability affects the adoption of the plow. Finally, columns (5) and (6) provide evidence of a positive and marginally significant association between the usage of the plow and the existence of sex-based grammatical gender systems in a language. Thus, as suggested by the proposed hypothesis, sex-based grammatical gender systems and the existence of gender roles in society have common geographical roots.

Table 13: Persistent Effect of Homeland vs. Urheimat Characteristics on Gender: Languages Outside Urheimat

	Existence of Sex-Based Gender System			
	Homeland		Urheimat	
	(1)	(2)	(3)	(4)
Plow Negative CSI (pre-1500CE)	-0.10	-0.17	0.22	-0.42**
	(0.12)	(0.11)	(0.19)	(0.17)
Caloric Suitability Index (pre-1500CE)	0.06	0.09	0.32***	1.07***
	(0.11)	(0.08)	(0.07)	(0.12)
Regional FE	No	Yes	No	Yes
Homeland Geographical Characteristics	No	Yes	No	No
Urheimat Geographical Characteristics	No	No	No	Yes
Adjusted-R^2	-0.00	0.19	0.38	0.68
Observations	100	100	100	100
Language Families	19	19	19	19

Notes: This table explores the relative contributions of agricultural productivity in the homeland vs. the Urheimat to the presence of future tense in a daughter language. Heteroskedasticity robust standard error estimates clustered at the language family level are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.

Tables 13 and C.2 further explore the relative contributions of pre-1500CE plow negative suitability in the homeland vs. the Urheimat to the presence of a sex-based grammatical gender system in a daughter language. In particular, Table 13 establishes that the existence of a sex-based grammatical gender system among daughter languages located outside the Urheimat of their proto-language is negatively significantly associated only with plow negative suitability in the Urheimat. Thus, the results further suggest the persistence of deep-historical origins of sex-based grammatical gender systems and their association with plow suitability. In particular, a one standard deviation increase in plow negative suitability in the Urheimat is associated with 42 percentage points decrease in the probability of existence of a sex-based grammatical system in a daughter language. Importantly, as before, by focusing on languages located outside the Urheimat of their proto-language and accounting for regional

fixed effects, the analysis mirrors the epidemiological approach to cultural diffusion, thus addressing potential concerns regarding omitted variables at the host-region level and providing support to the view that the future tense was formed mostly in the proto-language.

4.2 Politeness Distinctions in Pronouns

The theory suggests that in a hierarchical society characterized by obedience, conformity, and power distance, language structures that reinforced the existing hierarchical structure and cultural norms were likely to emerge and persist in this unequal society. In particular, politeness distinctions in pronouns (e.g., "tu" and "usted" in Spanish, "Du" and "Sie" in German, and "tu" and "vous" in French) were likely to appear and endure in this hierarchical society. Thus, geographical characteristics that were conducive to the development of hierarchical societies, (e.g., agricultural suitability and ecological diversity (Depetris-Chauvin and Özak, 2016; Diamond, 1997; Fenske, 2014; Litina, 2014)) would be expected to be associated with the emergence of politeness distinctions as well.

Table 14: Geographic Origins of Politeness Distinctions and Jurisdictional Hierarchy

	Reduced Form		Mechanism			
	Politeness		Jurisdictional Hierarchy		Politeness	
	(1)	(2)	(3)	(4)	(5)	(6)
Ecological Diversity	0.14***	0.09**	0.17***	0.10***		
	(0.03)	(0.04)	(0.03)	(0.03)		
Caloric Suitability Index (pre-1500CE)	0.11***	0.12***	0.17***	0.23***		
	(0.03)	(0.03)	(0.03)	(0.03)		
Jurisdictional Hierarchy					0.23***	0.18***
					(0.02)	(0.04)
All Geographic Controls	No	Yes	No	Yes	No	Yes
Regional FE	No	Yes	No	Yes	No	Yes
Adjusted-R^2	0.15	0.31	0.05	0.32	0.37	0.49
Observations	198	198	1169	1169	113	113

Notes: This table establishes the positive statistically and economically significant effect of the geographical determinants of statehood, as measured by jurisdictional hierarchy beyond the local level, and politeness distinctions in a language. The first two columns provide the results of the reduced form, and columns (3)-(6) provide evidence on the mechanism. The table shows the estimated coefficients in an OLS regression as the dependent variable in columns (3) and (4) is not binary. Heteroskedasticity robust standard error estimates are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.

Table 14 uses a linear probability model (OLS) to explore the relation between ecological diversity, caloric suitability, jurisdictional hierarchy beyond the local level and politeness distinctions.[32] Columns (1) and (2) present the correlation between ecological diversity, caloric suitability and politeness distinctions. The relations are positive and economically and statistically significant, suggesting in particular that a one standard deviation increase in ecological diversity increases the probability

[32]The Ethnographic Atlas reports the level of jurisdictional hierarchy beyond the local level (v33), which captures the level of statehood of an ethnicity.

of having politeness distinctions in the language by 9 percentage points. Columns (3) and (4) report the relation between the level of jurisdictional hierarchy beyond the local level and ecological diversity and caloric suitability. As in Fenske (2014) for ethnic groups in Africa and Depetris-Chauvin and Özak (2016) for ethnic groups in the world as a whole, ecological diversity has a positive statistically and economically significant effect on the emergence of jurisdictional hierarchy beyond the local level. Finally, columns (5) and (6) present the relation between jurisdictional hierarchy beyond the local level and the existence of politeness distinctions. As suggested by the hypothesis, jurisdictional hierarchy has a positive and economically and statistically significant effect at the 1% level on the emergence of politeness distinctions in a language. These results provide evidence that the origins of politeness distinctions are indeed generated by the same factors as hierarchical societies.

Table 15: Persistent Effect of Homeland vs. Urheimat Characteristics on Politeness: Languages Outside Urheimat

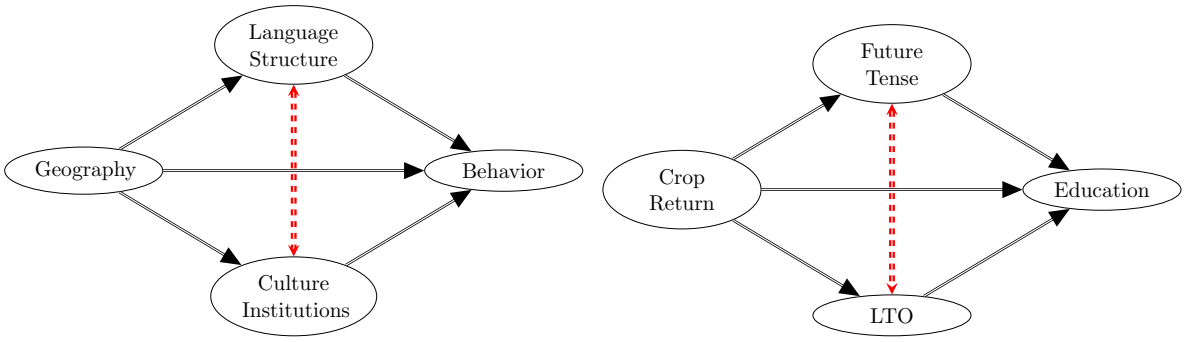
	Existence Politeness Distinctions			
	Homeland		Urheimat	
	(1)	(2)	(3)	(4)
Ecological Diversity	0.14***	0.13***	0.04	0.35**
	(0.03)	(0.03)	(0.15)	(0.15)
Caloric Suitability Index (pre-1500CE)	0.16***	0.13***	0.18	-0.28**
	(0.05)	(0.04)	(0.14)	(0.12)
Regional FE	No	Yes	No	Yes
Homeland Geographical Characteristics	No	Yes	No	No
Urheimat Geographical Characteristics	No	No	No	Yes
Adjusted-R^2	0.15	0.31	0.12	0.40
Observations	116	116	116	116
Language Families	19	19	19	19

Notes: This table explores the relative contributions of ecological diversity in the homeland vs. the Urheimat to the presence of future tense in a daughter language. Heteroskedasticity robust standard error estimates clustered at the language family level are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests; All regressions include a constant.

Tables 15 and C.3 further explore the relative contributions of ecological diversity and agricultural suitability in the homeland vs. the Urheimat to the presence of politeness distinctions in a daughter language. In contrast to the previous two language structures, and consistent with evidence about the greater adaptability of politeness distinctions, Table 15 and C.3 suggests that the existence of politeness distinctions among daughter languages is positively significantly associated with ecological diversity and agricultural suitability in their homeland as well as their change in the transition from the Urheimat to the homeland.

5 Language Structures & Contemporary Behavior

This section explores whether cultural traits encoded in language structures affect contemporary behavior (Figure 9(a)). In order to do so, this analysis exploits individual-level data on contemporary (economic) behavior and languages spoken by individuals. In particular, it analyzes the effect of languages with a future tense on long-term oriented behavior of its speakers above and beyond the effect of geography, institutions, as well as other cultural and ancestral characteristics. As Figures 9(a) and 9(b) suggest, the identification of this effect is fraught with complications due to the potential confounding effects of geography, culture and institutions.



(a) General Case (b) Special Case

Figure 9: Language and Contemporary Behavior

5.1 Identification Strategy

The analysis surmounts significant hurdles in the identification of the causal effect of language structures, particularly the existence of a future tense, on contemporary behavior. First, the research analyzes the behavior of second-generation migrants; thus, addressing the confounding effects of geography, institutions and culture in the (common) country-of-birth (Fernandez and Fogli, 2009; Galor and Özak, 2016; Giuliano, 2007). Second, it accounts for year and locality fixed-effects, accounting for the potential confounding effects of local geography, culture, institutions, and socio-economic conditions. Third, it accounts for individual characteristics like age, gender, marital status, which might affect behavior and correlate with the language spoken by the individual.

These three strategies define the epidemiological approach to the identification of cultural effects (Fernández, 2012; Galor and Özak, 2016; Giuliano, 2007). While this strategy allows the identification of the effects of culture, as opposed to institutions and geography, and aims to overcome the concerns regarding potential biases due to omitted variables, it is still potentially subject to the latter. In particular, the traditional epidemiological approach cannot overcome the potential concerns due omitted ancestral characteristics from the country of origin of individuals's parents.

In order to address this major deficiency of the epidemiological approach, the analysis *extends* it, by exploiting cultural variations encoded in language, while accounting for country-of-origin of parents fixed-effects. Thus, the analysis accounts for *all common* ancestral factors that affect an individual's behavior, i.e., geography, institutions and other cultural characteristics in the parental country of

origin. Hence, the analysis exploits variations in language structures of the languages spoken by individuals with the same parental ancestry in order to identify the cultural effects of language.[33]

Finally, after accounting for all these location, period and ancestry fixed-effects, one potential concern is that omitted parental characteristics might bias the estimated effect of language structures on individual's behavior. In order to address this concern, the analysis accounts for parental characteristics like education and the level of proficiency in the local language of the parents.

5.2 Crop Return, Future Tense, and Long-Term Oriented Behavior of Second-Generation Migrants

This section explores the effect of languages with a future tense on the long-term oriented behavior of its speakers above and beyond the effect of other cultural characteristics. Given the data requirements of the identification strategy exposed in the previous section, the analysis focuses on the effect of the future tense on human capital accumulation of second-generation migrants into the US. In particular, it explores the effect of language on the probability of college attendance of these second-generation migrants.[34,35]

In order to analyze the effect of the future tense on college attendance, the following general specification is estimated via ordinary least squares (OLS):

$$\text{College}_{istlp} = \beta_0 + \beta_1 \text{Future}_{istlp} + \beta_2 \text{Return}_{istlp} + \sum_j \gamma_{0j} X_{istlpj} + \sum_{stpj} \gamma_{stpj} \delta_{stpj} + \epsilon_{istlp}, \tag{5}$$

where College_{istlp} indicates whether individual i in state s in period t who speaks language l with parental ancestry p has attended college or not, Return_{istlp} is the pre-1500CE crop return in the homeland of language l spoken by the individual, $\{X_{istlpj}\}_j$ is a set of additional geographical characteristics of the homeland of the language spoken by the individual, $\{\delta_{stpj}\}_j$ is a set of fixed-effects that account jointly for individual characteristics j (sex, age, marital status), state s, year t, and parental ancestry p, and ϵ_{istlp} is the error term. Thus, the fixed-effects ensure that only individuals

[33]This strategy can be employed in other analyses as long as languages spoken by individuals are available as part of the data. Thus, it should allow the analyses to account for fixed-effects in conditions where normally it might not have been possible to. Additionally, by using variations in language structures or geographical characteristics associated with its homeland, the analysis can account for the ancestral composition of a population in a manner similar to Putterman and Weil (2010).

[34]Data is taken from the US Census and American Community Survey for the years post-2000 based on IPUMS (Ruggles et al., 2015). Second-generation migrants include all US-born individuals with at least one foreign born parent. The data on second generation migrants include 165250 offsprings of parents who migrated to the United States from 138 different countries – 137 countries of origin of the mother and 136 countries of origin of the father; these individuals speak 62 different languages. The sample of second-generation migrants in the US is constrained to include only individuals over 24 years of age in order to ensure they are old enough to have attended college. As shown in the appendix, similar results are obtained if the age is constrained to be over 21 or 18.

[35]By focusing on second-generation migrants in the US Census and American Community Surveys (ACS), the analysis overcomes a potential concern due to ethnic attrition bias (Duncan and Trejo, 2016). In particular, previous analyses that have employed the US census or ACS to study the effects of culture using migrants, have focused on all US-born individuals and tried to identify migrants and their ancestry by using individual's self-reported ancestry. Thus, these analyses have included all descendants of migrants that still identify with the country of origin of their ancestors. But, as Duncan and Trejo (2011, 2016), among others, have shown, individuals tend to self-identify differently depending on their generation, their true ancestry, and their socio-economic background. Thus, using second-and-higher-generation migrants can bias the results due to misidentification of ancestry.

that are similar in their individual characteristics, their location and ancestry are compared to each other. The results on the origins of the future tense presented in the historical analysis in section 3 suggest that future tense should have a negative effect and pre-1500 crop return a positive effect on college attendance (i.e. $\beta_1 < 0$ and $\beta_2 > 0$).

Table 16: Future Tense and College Attendance of Second-Generation Migrants

	College Attendance					
	(1)	(2)	(3)	(4)	(5)	(6)
Future Tense	-0.201***	-0.207***	-0.201***	-0.046***		-0.041***
	(0.013)	(0.007)	(0.007)	(0.011)		(0.012)
Crop Return (pre-1500CE)					0.010***	0.006*
					(0.003)	(0.003)
Main Geographical Controls	Yes	Yes	Yes	Yes	Yes	Yes
State FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Age FE	No	Yes	Yes	Yes	Yes	Yes
Gender FE	No	Yes	Yes	Yes	Yes	Yes
Marital Status FE	No	No	Yes	Yes	Yes	Yes
Parental Origin FE	No	No	No	Yes	Yes	Yes
Adjusted-R^2	0.05	0.07	0.08	0.13	0.13	0.13
R^2	0.05	0.11	0.17	0.45	0.45	0.45
Observations	165250	165250	165250	165250	165250	165250

Notes: This table establishes the negative significant effect of future tense on college attendance of second-generation migrants in the US. Heteroskedasticity robust standard error estimates clustered at the level of the included fixed-effects are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.

Table 16 establishes the negative effect of speaking a language with a future tense on college attendance of its speakers. In particular, columns (1)-(3) show that individuals who speak a language with a future tense have 20 percentage points lower probability of attending college than individuals with similar observable characteristics living in the same state and interviewed the same year, even after accounting geographical characteristics of the homeland of the language.

As mentioned in the identification strategy, one potential concern with the results of columns (1)-(3) is that the estimated effect of language also captures additional cultural effects due to the ancestry of the individual. In order to overcome this potential concern, column (4) additionally accounts for the parental country-of-origin. Thus, the estimated effect of future tense in column (4) captures the effect of language that is not explained by other ancestral traits. Moreover, by comparing individuals with similar observable characteristics, living in the same state in the same year, who have the same parental ancestry, but speak languages that differ in their future tense, the analysis isolates the cultural effect of the future tense. The results suggests that speaking a language with a future tense decreases the probability of attending college by 4.6 percentage points.

Additionally, Column (5) establishes that the pre-1500 crop return in a language's homeland has a positive effect on the accumulation of human capital of its speakers, even after accounting for all

other ancestral characteristics of an individual and other geographical characteristics of the language's homeland. Column (6) provides supportive evidence to the view that the future tense encoded the cultural effect of crop return. In particular, it suggests that the effect of crop return is mediated by a language's future tense. Thus, columns (5) and (6) provide supporting evidence that the effect of future tense captures the effect of the encoded cultural trait of long-term orientation, which was determined by crop return during the formation of a language (section 3).

There are various potential concerns with the results of Table 16. First, second-generation migrants in the US Census and ACS can only be identified for individuals who live with their parents. Although this is a representative sample of this subpopulation, which overcomes concerns due to ethnic attrition (see footnote 35) and allows for the control of parental characteristics in the analysis, it might potentially bias the results. Appendix B.8 explores the differences in observables between various samples of migrants. Reassuringly, it shows that only age and marital status differ between the full sample of second-generation migrants and the subsample that lives with their parents. Moreover, the sample of second-and-higher generation migrants, that has previously been employed in the literature, and which is subject to ethnic attrition bias, is more similar to the true third-and-higher-generation migrant sample.

In order to assess the potential bias due to the sample, Table B.1 replicates the basic results (without ancestry fixed-effects given the potential for ethnic attrition bias) for the sample of second-and-higher generation migrants. Additionally, Table B.2 replicates the analysis in Table 16 using the sample of one-and-a-half-generation migrants, i.e., migrants who were born in another country, but arrived to the US before age 5. The benefits of using this sample is that (i) it has similar properties for cultural analysis as second-generation migrants, and (ii) it overcomes the potential concerns due to both ethnic attrition and living arrangements.[36] Reassuringly, the results remain unchanged and suggest that speaking a language with a future tense lowers the probability of college attendance by 5 percentage points, above and beyond the effect of other ancestral traits.

Second, individuals' education levels are potentially determined by the education level of their parents. Similarly, parents' command of the English language, which is the official language in the US, might potentially affect individual's education levels as well as the language spoken at home. Table 17 explores the effect of accounting for parents' education levels and their command of the English language. Additionally, the analysis accounts now for fixed-effects for both parents' country-of-origin. Reassuringly, the results of Table 16 remain qualitatively unchanged. In particular, the effect of speaking a language with future tense remains negative and significant. Additionally, parental education and English levels have a positive effect on their offspring's college attendance, suggesting that college educated parents who have a good command of English have a higher probability of having their children attend college. The estimates suggest that not speaking a language with a future tense has an effect that is about a third of the effect of having a college educated mother or above half the effect of having a college educated father.

[36]The sample of the one-and-a-half-generation migrants includes 422081 individuals who migrated from 141 different countries to the United States when they were five years old or younger and speak 64 different languages. One-and-a-half-generation migrants are similar to second-generation migrants, since they were not the ones who made the decision to migrate and grew up in the US, so that they received their K-12 education in the United States.

Table 17: Future Tense and College Education of Second Generation Migrants Accounting for Parental Education and English Levels

	College Attendance								
	Parental Education			Parental English			Both		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Future Tense	-0.047***		-0.043***	-0.038***		-0.038***	-0.035***		-0.034***
	(0.006)		(0.007)	(0.008)		(0.008)	(0.008)		(0.008)
Crop Return (pre-1500CE)		0.010***	0.005***		0.004*	0.000		0.005**	0.002
		(0.002)	(0.002)		(0.002)	(0.002)		(0.002)	(0.002)
Mom's College Attendance	0.130***	0.130***	0.130***				0.134***	0.134***	0.134***
	(0.003)	(0.003)	(0.003)				(0.004)	(0.004)	(0.004)
Dad's College Attendance	0.073***	0.073***	0.073***				0.146***	0.147***	0.146***
	(0.003)	(0.003)	(0.003)				(0.004)	(0.004)	(0.004)
Mom's English Level				0.012***	0.012***	0.012***	0.013***	0.014***	0.013***
				(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Dad's English Level				-0.001	-0.001	-0.001	0.006***	0.006***	0.006***
				(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Main Geographical Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Origin FE for Both Parents	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Age FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Gender FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Marital Status FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted-R^2	0.14	0.14	0.14	0.14	0.14	0.14	0.18	0.18	0.18
R^2	0.23	0.23	0.23	0.23	0.23	0.23	0.26	0.26	0.26
Observations	165250	165250	165250	98623	98623	98623	98623	98623	98623

Notes: This table establishes the robustness of the negative effect of the future tense on college attendance to the inclusion of parental educational and English levels. Heteroskedasticity robust standard error estimates clustered at the level of the included fixed-effects are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.

Third, although the analysis accounts for parental origin fixed-effects and controls, and language level geographical characteristics, the effect of future tense might still reflect a general (ancestral) cultural trait associated with an individual's language, which might be unrelated to the existence of a future tense and long-term orientation. Table 18 explores this possibility by additionally accounting for other language structures. Reassuringly, the effect of future tense is unaffected by the inclusion of these additional language structures, which are mostly insignificant.

Finally, individual's educational choices can be affected by local socio-economic conditions. In particular, local labor market conditions and opportunities might be affected by ethnic networks, racial or ethnic discrimination, among others. The previous results addressed this issue partially by comparing observationally equivalent second-generation migrants within states. Table B.3 further establishes that the results are qualitatively similar if instead within-county level variation is exploited. Moreover,

Table 18: Future Tense and College Education of Second Generation Migrants
Accounting for other Linguistic Structures

	College Attendance						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Future Tense	-0.043***	-0.049***	-0.041***	-0.041***	-0.045***	-0.047***	-0.048***
	(0.007)	(0.008)	(0.008)	(0.014)	(0.008)	(0.008)	(0.008)
Crop Return (pre-1500CE)	0.005***	0.009***	0.008***	-0.003	0.008**	0.011***	0.011***
	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Mom's College Attendance	0.130***	0.131***	0.131***	0.133***	0.132***	0.132***	0.132***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Dad's College Attendance	0.073***	0.075***	0.075***	0.076***	0.076***	0.075***	0.075***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Past Tense		0.015					
		(0.014)					
Perfect Tense			-0.011				
			(0.007)				
Existence of Gender System				-0.030*			
				(0.018)			
Evidentiality					0.018**		
					(0.008)		
Consonant Inventories						0.001	
						(0.007)	
Consonant-Vowel Ratio							0.001
							(0.004)
Main Geographical Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Origin FE for Both Parents	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Age FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Gender FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Marital Status FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted-R^2	0.14	0.14	0.14	0.14	0.14	0.14	0.14
R^2	0.23	0.23	0.23	0.23	0.23	0.23	0.23
Observations	165250	158239	158239	153996	155905	157002	157002

Notes: This table establishes the robustness of the negative effect of future tense on college attendance to accounting for other language structures. The analysis accounts for parental ancestry fixed-effect, as well as for parental college attendance. Heteroskedasticity robust standard error estimates clustered at the level of the included fixed-effects are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.

labor market opportunities might be affected by speaking one of the two main languages in the US, namely English and Spanish. Additionally, the recent increase in (Spanish speaking) immigrants from Latin-America, many with lower levels of human capital, may bias the results. Reassuringly, Table 19 establishes that the results remain qualitatively unchanged if English or Spanish speakers are excluded from the analysis.

The previous results suggest that speaking a language with a future tense directly decreases the probability of attending college. One potential interpretation of these results is that using the future tense in itself affects behavior. On the other hand, it could be capturing within-country-of-origin variations in time preference. In particular, if parents come from the same country of origin, but differ in their culture and language, the effect of future tense might just be capturing these cultural differences.

Table 19: Future Tense and College Education of Second Generation Migrants
Accounting for Local Labor Market Conditions

	College Attendance					
	No English			No Spanish		
	(1)	(2)	(3)	(4)	(5)	(6)
Future Tense	-0.021**		-0.022**	-0.029***		-0.027***
	(0.009)		(0.009)	(0.006)		(0.006)
Crop Return (pre-1500CE)		0.000	-0.001		0.004***	0.002
		(0.002)	(0.001)		(0.002)	(0.001)
Mom's College Attendance	0.114***	0.114***	0.114***	0.124***	0.123***	0.124***
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Dad's College Attendance	0.135***	0.135***	0.135***	0.131***	0.130***	0.131***
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Main Geographical Controls	Yes	Yes	Yes	Yes	Yes	Yes
Origin FE for Both Parents	Yes	Yes	Yes	Yes	Yes	Yes
County FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Age FE	Yes	Yes	Yes	Yes	Yes	Yes
Gender FE	Yes	Yes	Yes	Yes	Yes	Yes
Marital Status FE	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted-R^2	0.19	0.19	0.19	0.19	0.19	0.19
R^2	0.34	0.34	0.34	0.34	0.34	0.34
Observations	52537	52537	52537	55176	55176	55176

Notes: This table establishes the negative significant effect of future tense on college attendance excluding English and Spanish speakers. The analysis accounts for parental ancestry fixed-effect, as well as for parental college attendance. Heteroskedasticity robust standard error estimates two-way clustered by country of origin of both parents are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.

Table B.4 shows the results of splitting the sample of second-generation migrants among those whose parents come from the same country and those whose parents come from different countries. The table establishes that the future tense has no effect in the sample of individuals whose parents come from the same country.[37] On the contrary, the effect of future tense remains qualitatively unchanged in the sample of migrants whose parents come from different countries. Although this could still capture some within-country-of-origin variation, it is less probable to do so. While the effect of the future tense may reflect the (encoded) long-term orientation of the parent whose language is spoken at home, the analysis cannot refute the presence of a direct effect of this language structure on college attendance.

[37]This result is driven by the fact that there is almost no variation in the future tense of the language spoken at home, and thus, the parental-country-of-origin fixed-effects absorb all the potential explanatory power of the future tense.

6 Conclusion

This research explores some of the most fundamental and intriguing mysteries about the origins of the coevolution of linguistic and cultural traits and their impact on the development process: Has the coevolution of linguistic and cultural traits contributed to the stability and the persistence of cultural characteristics and their lasting effect on economic prosperity? Has the evolution of languages reflected economic incentives, promoting an efficient economic exchange? Have language structures merely encoded existing cultural traits or have they influenced human behavior and values and contributed directly to the development process? What are the geographical roots of the coevolution of linguistic and cultural traits? Are the geographical characteristics that triggered the coevolution of culture and language critical for the understanding of the contribution of cultural and linguistic characteristics for the wealth of nations?

It advances the hypothesis and establishes empirically that variations in pre-industrial geographical characteristics that were conducive to higher return to agricultural investment, larger gender gap in agricultural productivity, and more hierarchical society, are at the root of existing cross-language variations in the presence of the future tense, grammatical gender, and politeness distinctions. Moreover, the research suggests that while language structures have been largely a reflection of the coding of past human experience and in particular the range of ancestral cultural traits in society, they had an independent effect on human behavior and economic outcomes.

The empirical methodology that is advanced in the course of this research augments the epidemiological approach and permits the study of the persistent effect of cultural factors, while accounting for other ancestral characteristics. In particular, it suggests that variations in the languages spoken by second-generation migrants originated from the same ancestral region can be exploited to account for country of origin fixed-effects and thus to overcome the potential biases that could be generated by omitted ancestral characteristics; an inevitable shortcoming in the existing implementation of the epidemiological approach.

References

Acemoglu, D., Johnson, S. and Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation, *The American Economic Review* **91**(5): 1369–1401.

Alesina, A., Devleeschauwer, A., Easterly, W., Kurlat, S. and Wacziarg, R. (2003). Fractionalization, *Journal of Economic growth* **8**(2): 155–194.

Alesina, A. and Ferrara, E. L. (2005). Ethnic diversity and economic performance, *Journal of economic literature* **43**(3): 762–800.

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

Altonji, J. G., Elder, T. E. and Taber, C. R. (2005). Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools, *Journal of Political Economy* **113**(1): 151–184.

Ashraf, Q. and Galor, O. (2013a). Genetic diversity and the origins of cultural fragmentation, *The American Economic Review* **103**(3): 528–533.

Ashraf, Q. and Galor, O. (2013b). The out of africa hypothesis, human genetic diversity, and comparative economic development, *The American Economic Review* **103**(1): 1–46.

Bellows, J. and Miguel, E. (2009). War and local collective action in sierra leone, *Journal of Public Economics* **93**(11): 1144–1157.

Bisin, A. and Verdier, T. (2000). Beyond the melting pot: cultural transmission, marriage, and the evolution of ethnic and religious traits, *The Quarterly Journal of Economics* **115**(3): 955–988.

Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S. J., Alekseyenko, A. V., Drummond, A. J., Gray, R. D., Suchard, M. A. and Atkinson, Q. D. (2012). Mapping the origins and expansion of the indo-european language family, *Science* **337**(6097): 957–960.

Brown, P. and Levinson, S. C. (1987). *Politeness: Some universals in language usage*, Vol. 4, Cambridge university press.

Brown, R. and Gilman, A. (1989). Politeness theory and shakespeare's four major tragedies, *Language in society* **18**(02): 159–212.

Cavalli-Sforza, L. L. (2000). *Genes, peoples, and languages*, 1st ed edn, North Point Press, New York.

Cavalli-Sforza, L. L., Menozzi, P. and Piazza, A. (1994). *The history and geography of human genes*, Princeton University Press, Princeton, N.J.

Chen, M. K. (2013). The effect of language on economic behavior: Evidence from savings rates, health behaviors, and retirement assets, *The American Economic Review* **103**(2): 690–731.

Conley, T. G. (1999). GMM estimation with cross sectional dependence, *Journal of econometrics* **92**(1): 1–45.

Corbett, G. G. (2013). Sex-based and non-sex-based gender systems, *in* M. S. Dryer and M. Haspelmath (eds), *The World Atlas of Language Structures Online*, Max Planck Institute for Evolutionary Anthropology, Leipzig.

Dahl, Ö. (1985). *Tense and aspect systems*, Basil Blackwell.

Dahl, Ö. (2000). The grammar of future time reference in european languages, *Tense and Aspect in the Languages of Europe* p. 309.

Dahl, Ö. and Velupillai, V. (2013). The future tense, *in* M. S. Dryer and M. Haspelmath (eds), *The World Atlas of Language Structures Online*, Max Planck Institute for Evolutionary Anthropology, Leipzig.

Depetris-Chauvin, E. and Özak, Ö. (2016). Population diversity, division of labor and comparative development, *Division of Labor and Comparative Development (April 4, 2016)* .

Desmet, K., Ortuño-Ortín, I. and Wacziarg, R. (2012). The political economy of linguistic cleavages, *Journal of development Economics* **97**(2): 322–338.

Deutscher, G. (2010). *The unfolding of language*, Random House.

Diamond, J. M. (1997). *Guns, germs, and steel: the fates of human societies*, 1st ed edn, W.W. Norton & Co., New York.

Dryer, Matthew S & Haspelmath, M. e. (2013). *The World Atlas of of Language Structures Online.*, Leipzig: Max Planck Institute for Evolutionary Anthropology.

Duncan, B. and Trejo, S. J. (2011). Intermarriage and the intergenerational transmission of ethnic identity and human capital for mexican americans, *Journal of Labor Economics* **29**(2): 195.

Duncan, B. and Trejo, S. J. (2016). The complexity of immigrant generations: Implications for assessing the socioeconomic integration of hispanics and asians, *NBER Working Paper Series* (w21982).

Easterly, W. and Levine, R. (1997). Africa's growth tragedy: policies and ethnic divisions, *The Quarterly Journal of Economics* pp. 1203–1250.

Fearon, J. D. (2003). Ethnic and cultural diversity by country, *Journal of Economic Growth* **8**(2): 195–222.

Fenske, J. (2014). Ecology, trade, and states in pre-colonial africa, *Journal of the European Economic Association* **12**(3): 612–640.

Fernández, R. (2012). Does culture matter?, *in* J. Benhabib, A. Bisin and M. O. Jackson (eds), *Handbook of Social Economics*, Vol. 1B, Elsevier, Amsterdam.

Fernandez, R. and Fogli, A. (2009). Culture: An empirical investigation of beliefs, work, and fertility, *American Economic Journal: Macroeconomics* **1**(1): 146–177.

Gallup, J. L., Sachs, J. D. and Mellinger, A. D. (1999). Geography and economic development, *International regional science review* **22**(2): 179–232.

Galor, O. and Moav, O. (2002). Natural selection and the origin of economic growth, *The Quarterly Journal of Economics* **117**(4): 1133–1191.

Galor, O. and Özak, Ö. (2015). Land productivity and economic development: Caloric suitability vs. agricultural suitability, *Working Paper, Southern Methodist University, Department of Economics* .

Galor, O. and Özak, Ö. (2016). The agricultural origins of time preference, *American Economic Review* **106**(10).

Giuliano, P. (2007). Living arrangements in western europe: Does cultural origin matter?, *Journal of the European Economic Association* **5**(5): 927–952.

Glaeser, E. L., La Porta, R., Lopez-de Silanes, F. and Shleifer, A. (2004). Do institutions cause growth?, *Journal of economic Growth* **9**(3): 271–303.

Guiso, L., Sapienza, P. and Zingales, L. (2004). The role of social capital in financial development, *American Economic Review* **94**(3): 526–556.

Guiso, L., Sapienza, P. and Zingales, L. (2006). Does culture affect economic outcomes?, *Journal of Economic Perspectives* **20**(2): 23–48.

Harutyunyan, A. and Özak, Ö. (2016). Culture, diffusion, and economic development, *Available at SSRN 2759127* .

Helmbrecht, J. (2003). Politeness distinctions in second person pronouns, *Pragmatics and beyond New Series* pp. 185–202.

Helmbrecht, J. (2005). Politeness distinctions in pronouns, *The world atlas of language structures* pp. 186–190.

Helmbrecht, J. (2013). Politeness distinctions in pronouns, *in* M. S. Dryer and M. Haspelmath (eds), *The World Atlas of Language Structures Online*, Max Planck Institute for Evolutionary Anthropology, Leipzig.

Hofstede, G. H. (1991). *Cultures and organizations: software of the mind*, McGraw-Hill, London.

Hofstede, G. H., Hofstede, G. J. and Minkov, M. (2010). *Cultures and organizations: software of the mind : intercultural cooperation and its importance for survival*, 3rd ed edn, McGraw-Hill, New York.

Lewis, M. P., Simons, G. F. and Fennig, C. D. (2009). *Ethnologue: Languages of the world*, Vol. 16, SIL international Dallas, TX.

Litina, A. (2014). The geographical origins of early state formation, *Technical report*, Center for Research in Economic Analysis, University of Luxembourg.

Lupyan, G. and Dale, R. (2010). Language structure is partly determined by social structure, *PloS one* **5**(1): e8559.

Michalopoulos, S. (2012). The origins of ethnolinguistic diversity, *The American Economic Review* **102**(4): 1508.

Miguel, E., Satyanath, S. and Sergenti, E. (2004). Economic shocks and civil conflict: An instrumental variables approach, *Journal of political Economy* **112**(4): 725–753.

Murdock, G. P. (1967). Ethnographic atlas: a summary, *Ethnology* pp. 109–236.

Murdock, G. P. and White, D. R. (1969). Standard cross-cultural sample, *Ethnology* pp. 329–369.

Nunn, N. and Wantchekon, L. (2011). The slave trade and the origins of mistrust in africa, *American Economic Review* **101**(7): 3221–52.

Olson, D. M., Dinerstein, E., Wikramanayake, E. D., Burgess, N. D., Powell, G. V., Underwood, E. C., D'amico, J. A., Itoua, I., Strand, H. E., Morrison, J. C. et al. (2001). Terrestrial ecoregions of the world: A new map of life on earth a new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity, *BioScience* **51**(11): 933–938.

Oster, E. (2014). Unobservable selection and coefficient stability: Theory and validation.

Pagel, M., Atkinson, Q. D., Calude, A. S. and Meade, A. (2013). Ultraconserved words point to deep language ancestry across eurasia, *Proceedings of the National Academy of Sciences* **110**(21): 8471–8476.

Perkins, R. D. (1992). *Deixis, grammar, and culture*, Vol. 24, John Benjamins Publishing.

Pryor, F. L. (1985). The invention of the plow, *Comparative Studies in Society and history* **27**(04): 727–743.

Putterman, L. and Weil, D. N. (2010). Post-1500 population flows and the long-run determinants of economic growth and inequality*, *The Quarterly journal of economics* **125**(4): 1627–1682.

Richerson, P. J., Boyd, R. and Henrich, J. (2010). Gene-culture coevolution in the age of genomics, *Proceedings of the National Academy of Sciences* **107**(Supplement 2): 8985–8992.

Roberts, S. G., Winters, J. and Chen, K. (2015). Future tense and economic decisions: controlling for cultural evolution, *PloS one* **10**(7): e0132145.

Ruggles, S., Genadek, K., Goeken, R., Grover, J. and Sobek, M. (2015). *Integrated Public Use Microdata Series: Version 6.0 [Machine-readable database]*, University of Minnesota, Minneapolis.

Spolaore, E. and Wacziarg, R. (2013). Long-term barriers to economic development, *Handbook of Economic Growth*, Vol. 2, Elsevier, p. 121.

Tabellini, G. (2010). Culture and institutions: economic development in the regions of europe, *Journal of the European Economic Association* **8**(4): 677–716.

Appendix (Not for publication)

A The Origins of Future Tense

A.1 Crop Return and Future Tense

Table A.1: Crop Return and Future Tense (OLS)

	Existence of Future Tense								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Crop Return	-0.06**	-0.08**	-0.08**	-0.08**	-0.09**	-0.08**	-0.09**	-0.09**	-0.12***
(pre-1500CE)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)
Absolute Latitude			-0.10*	-0.10*	-0.10*	-0.08	-0.07	-0.10	-0.15
			(0.05)	(0.05)	(0.05)	(0.05)	(0.07)	(0.11)	(0.11)
Elevation				0.00	-0.02	-0.03	-0.01	-0.05	-0.03
				(0.03)	(0.04)	(0.04)	(0.04)	(0.05)	(0.05)
Ruggedness					0.04	0.04	0.02	0.02	0.02
					(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Coast Length						-0.06***	-0.05***	-0.05***	-0.05***
						(0.02)	(0.02)	(0.01)	(0.02)
Precipitation							0.01	0.01	-0.01
							(0.09)	(0.09)	(0.08)
Precipitation							-0.09***	-0.05	-0.05
(std)							(0.03)	(0.05)	(0.05)
Precipitation							0.04	0.04	0.04
Volatility							(0.09)	(0.09)	(0.08)
Precipitation							-0.02	-1.06***	-1.01***
Spatial Correlation							(0.04)	(0.31)	(0.32)
Temperature (Daily								-0.08	-0.08
Mean)								(0.09)	(0.08)
Temperature (Daily								-0.05	-0.06
Mean) (std)								(0.05)	(0.05)
Temperature								0.04	0.09
Volatility								(0.09)	(0.09)
Temperature Spatial								1.04***	0.99***
Correlation								(0.31)	(0.31)
Unproductive Period									-0.09***
(pre-1500CE)									(0.03)
Regional FE	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted-R^2	0.01	0.03	0.04	0.04	0.04	0.05	0.06	0.09	0.11
Observations	275	275	275	275	275	275	275	275	275

Notes: This table replicates the analysis of Table 1 using an OLS estimation. The results are similar to the ones reported in Tble 1, and thus the table provides evidence that the analysis is robust to different estimation methods. Heteroskedasticity robust standard error estimates are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.

A.2 Persistent Effect of Urheimat's Crop Return on Future Tense

Table A.2: Persistent Effect of Urheimat Characteristics: Share of Daughter Languages with Future Tense

	Share of Daughter Languages with Future Tense				
	(1)	(2)	(3)	(4)	(5)
Crop Return (pre-1500CE)	-0.12**	-0.16***	-0.16***	-0.17***	-0.19***
	(0.05)	(0.06)	(0.05)	(0.06)	(0.06)
Regional FE	No	Yes	Yes	Yes	Yes
Main Geographical Controls	No	No	Yes	Yes	Yes
Precipitation Controls	No	No	No	Yes	Yes
Temperature Controls	No	No	No	No	Yes
Unproductive Period	Yes	Yes	Yes	Yes	Yes
Observations	74	74	74	74	74

Notes: This table replicates Table 6 without weighting the observations. It establishes the robust negative statistically and economically significant relation between the Urheimat's crop return and the share of daughter languages that have a future tense. The coefficients in the table are the average marginal effects of increasing crop return in the Urheimat in a zero-inflated fractional regression. Heteroskedasticity robust standard error estimates are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.

Table A.3: Crop Return and Future Tense across Language Families

	Existence of Future Tense				
	(1)	(2)	(3)	(4)	(5)
Crop Return (pre-1500CE)	-0.13**	-0.19***	-0.18***	-0.19***	-0.25***
	(0.05)	(0.06)	(0.05)	(0.06)	(0.07)
Regional FE	No	Yes	Yes	Yes	Yes
Main Geographical Controls	No	No	Yes	Yes	Yes
Precipitation Controls	No	No	No	Yes	Yes
Temperature Controls	No	No	No	No	Yes
Unproductive Period	Yes	Yes	Yes	Yes	Yes
Pseudo-R^2	0.06	0.12	0.19	0.23	0.39
Observations	74	74	74	74	74

Notes: This table establishes the negative economically and statistically significant effect of the Urheimat's crop return and the existence of a future tense in a proto language, assuming that the existence of future tense in a majority of daughter languages in a language family represents the existence of the future tense in the proto language. Observations are not weighted. Heteroskedasticity robust standard error estimates are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.

Table A.4: Crop Return and Future Tense across Language Families

	Existence of Future Tense					
	(1)	(2)	(3)	(4)	(5)	(6)
Crop Return (pre-1500CE)	-0.21***	-0.28***	-0.27***	-0.26***	-0.24***	-0.27***
	(0.06)	(0.04)	(0.05)	(0.05)	(0.06)	(0.07)
Regional FE	No	Yes	Yes	Yes	Yes	Yes
Main Geographical Controls	No	No	Yes	Yes	Yes	Yes
Precipitation Controls	No	No	No	Yes	Yes	Yes
Temperature Controls	No	No	No	No	Yes	Yes
Unproductive Period	No	No	No	No	No	Yes
Pseudo-R^2	0.15	0.35	0.41	0.52	0.57	0.57
Observations	74	74	74	74	74	74

Notes: This table establishes the negative economically and statistically significant effect of the Urheimat's crop return and the existence of a future tense in a proto language, assuming that the existence of future tense in a majority of daughter languages in a language family represents the existence of the future tense in the proto language. Observations are weighted to account for missing languages and future tense data.. Heteroskedasticity robust standard error estimates are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.

Table A.5: Crop Return and Future Tense across Language Families

	Existence of Future Tense					
	(1)	(2)	(3)	(4)	(5)	(6)
Crop Return (pre-1500CE)	-0.28***	-0.39***	-0.38***	-0.42***	-0.47***	-0.44***
	(0.07)	(0.07)	(0.07)	(0.06)	(0.07)	(0.08)
Unproductive Period (pre-1500CE)						0.04
						(0.09)
Regional FE	No	Yes	Yes	Yes	Yes	Yes
Main Geographical Controls	No	No	Yes	Yes	Yes	Yes
Precipitation Controls	No	No	No	Yes	Yes	Yes
Temperature Controls	No	No	No	No	Yes	Yes
Pseudo-R^2	0.24	0.43	0.50	0.58	0.64	0.64
Observations	73	73	73	73	73	73

Notes: This table establishes the negative, statistically, and economically significant effect of pre-1500CE potential crop return on the existence of future tense, accounting for regional fixed-effects and the set of geographical characteristics in Table 1. All independent variables have been normalized by subtracting their mean and dividing by their standard deviation at the language level, and the median at the language level is used in the analysis. Thus, all coefficients can be compared and show the effect of a one standard deviation in the independent variable on the probability of having a future tense in the language. Coefficients estimated using a weighted Probit, where weights correct representativeness of each language family in sample. For all variables the language family mean value is used in the analysis. Heteroskedasticity robust standard error estimates are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.

Table A.6: Crop Return and Future Tense across Language Families

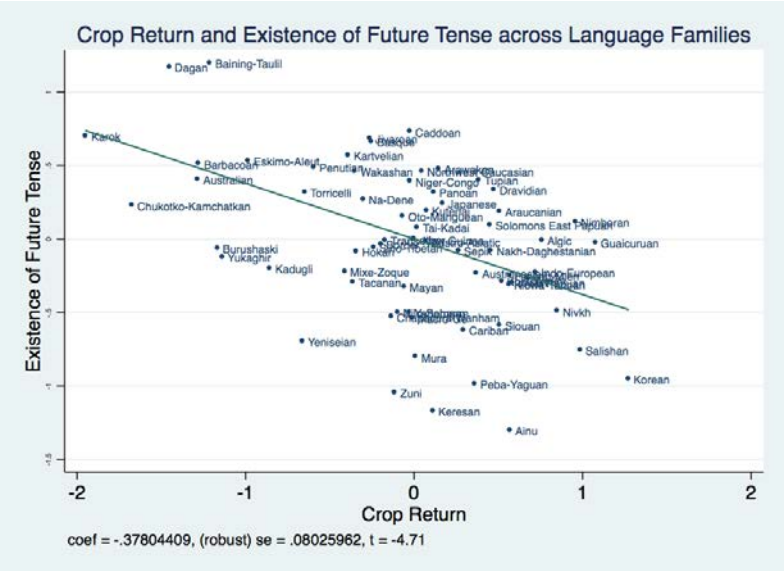
	Existence of Future Tense (Median)					
	(1)	(2)	(3)	(4)	(5)	(6)
Crop Return (pre-1500CE)	-0.28***	-0.39***	-0.38***	-0.41***	-0.40***	-0.43***
	(0.03)	(0.05)	(0.05)	(0.06)	(0.06)	(0.07)
Unproductive Period (pre-1500CE)						-0.10
						(0.11)
Regional FE	No	Yes	Yes	Yes	Yes	Yes
Main Geographical Controls	No	No	Yes	Yes	Yes	Yes
Precipitation Controls	No	No	No	Yes	Yes	Yes
Temperature Controls	No	No	No	No	Yes	Yes
Pseudo-R^2	0.27	0.50	0.53	0.62	0.72	0.72
Observations	66	66	66	66	66	66

Notes: This table establishes the negative, statistically, and economically significant effect of pre-1500CE potential crop return on the existence of future tense, accounting for regional fixed-effects and the set of geographical characteristics in Table 1. All independent variables have been normalized by subtracting their mean and dividing by their standard deviation. Thus, all coefficients can be compared and show the effect of a one standard deviation in the independent variable on the probability of having a future tense. Coefficients estimated using a weighted Probit, where weights correct representativeness of each language family in sample. For all variables the language family median value is used in the analysis. Heteroskedasticity robust standard error estimates are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.

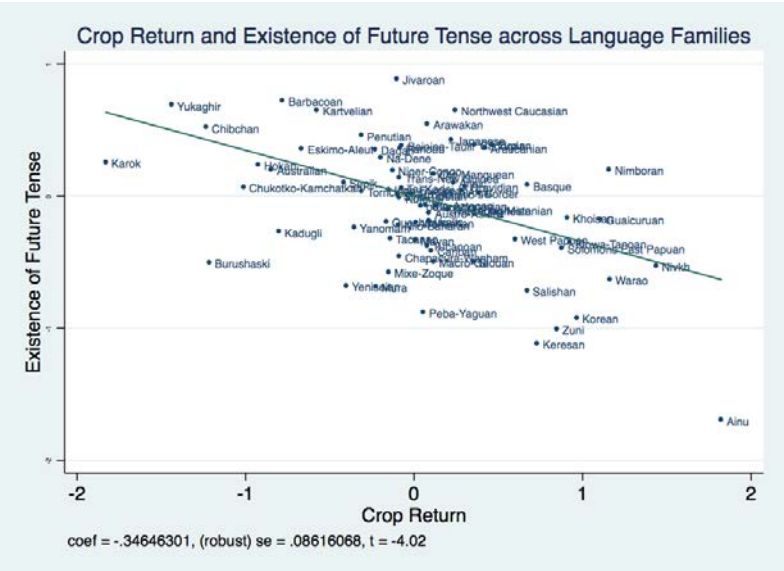
Table A.7: Crop Return and Future Tense across Language Genera

	Existence of Future Tense					
	(1)	(2)	(3)	(4)	(5)	(6)
Crop Return (pre-1500CE)	-0.21***	-0.20***	-0.23***	-0.22***	-0.22***	-0.26***
	(0.04)	(0.06)	(0.07)	(0.06)	(0.06)	(0.06)
Unproductive Period (pre-1500CE)						-0.11
						(0.07)
Regional FE	No	Yes	Yes	Yes	Yes	Yes
Main Geographical Controls	No	No	Yes	Yes	Yes	Yes
Precipitation Controls	No	No	No	Yes	Yes	Yes
Temperature Controls	No	No	No	No	Yes	Yes
Pseudo-R^2	0.16	0.17	0.24	0.31	0.31	0.32
Observations	142	142	142	142	142	142

Notes: This table establishes the negative, statistically, and economically significant effect of pre-1500CE potential crop return on the existence of future tense, accounting for regional fixed-effects and the set of geographical characteristics in Table 1. All independent variables have been normalized by subtracting their mean and dividing by their standard deviation at the language level, and the median at the language level is used in the analysis. Thus, all coefficients can be compared and show the effect of a one standard deviation in the independent variable on the probability of having a future tense in the language. Coefficients estimated using a weighted Probit, where weights correct representativeness of each language genus in sample. For all variables the language genus mean value is used in the analysis. Heteroskedasticity robust standard error estimates are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.



(a) Median



(b) Mean

Figure A.1: Crop Return and Future Across Language Families

Table A.8: Crop Return and Future Tense across Language Genera

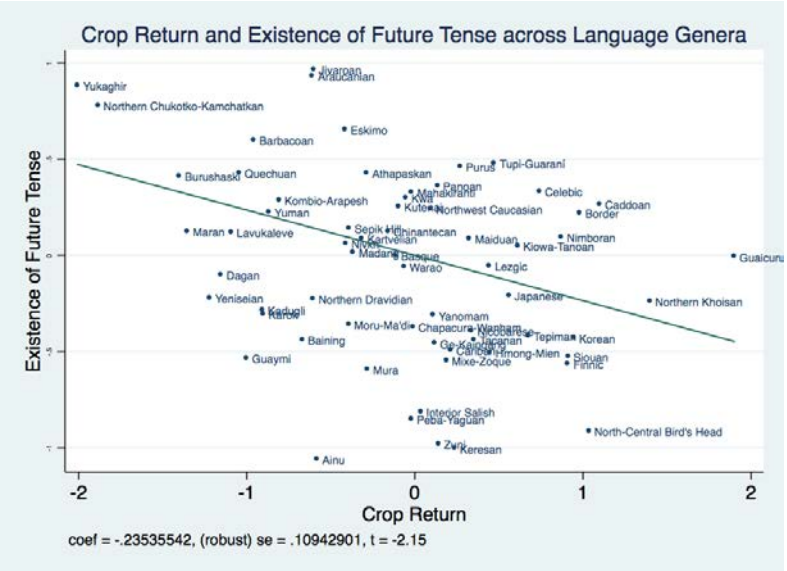
	Existence of Future Tense (Median)					
	(1)	(2)	(3)	(4)	(5)	(6)
Crop Return (pre-1500CE)	-0.20***	-0.24***	-0.29***	-0.28***	-0.27***	-0.30***
	(0.05)	(0.07)	(0.08)	(0.07)	(0.08)	(0.08)
Unproductive Period (pre-1500CE)						-0.07
						(0.08)
Regional FE	No	Yes	Yes	Yes	Yes	Yes
Main Geographical Controls	No	No	Yes	Yes	Yes	Yes
Precipitation Controls	No	No	No	Yes	Yes	Yes
Temperature Controls	No	No	No	No	Yes	Yes
Pseudo-R^2	0.14	0.21	0.25	0.31	0.32	0.33
Observations	147	147	147	147	147	147

Notes: This table establishes the negative, statistically, and economically significant effect of pre-1500CE potential crop return on the existence of future tense, accounting for regional fixed-effects and the set of geographical characteristics in Table 1. All independent variables have been normalized by subtracting their mean and dividing by their standard deviation. Thus, all coefficients can be compared and show the effect of a one standard deviation in the independent variable on the probability of having a future tense in the language. Coefficients estimated using a weighted Probit, where weights correct representativeness of each language genus in sample. For all variables the language genus median value is used in the analysis. Heteroskedasticity robust standard error estimates are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.

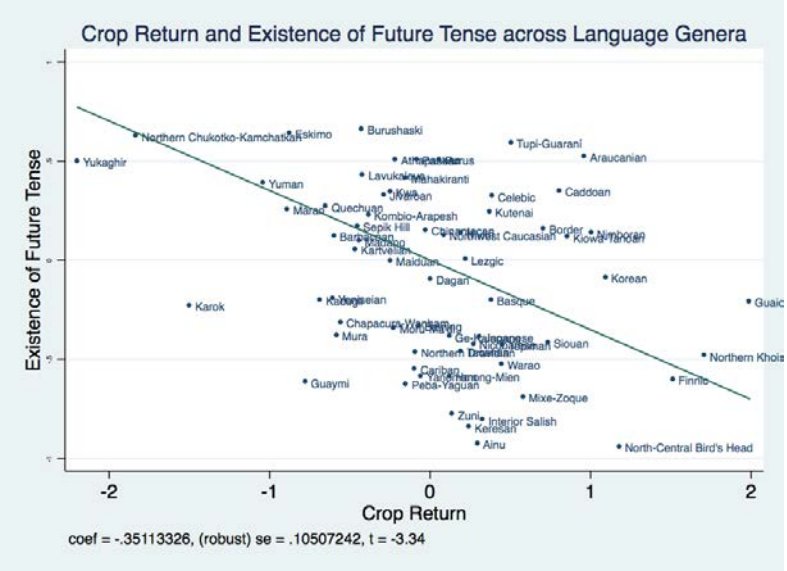
Table A.9: Crop Return and Language Structures across Language Families

	Language Structure								
	Temporal Structures			Non-Temporal Structures					
	Future	Past	Perfect	Gender	Possessive	Evidentiality	Consonants	C/V Ratio	Colors
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Crop Return (pre-1500CE)	-0.35***	-0.16	-0.08	0.01	-0.11	-0.01	0.29*	0.15	0.30
	(0.09)	(0.14)	(0.13)	(0.13)	(0.14)	(0.07)	(0.16)	(0.19)	(1.01)
All Geographic Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Regional FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted-R^2	0.47	0.23	0.08	0.19	0.27	0.69	0.67	0.48	0.11
Observations	73	73	73	59	64	71	73	73	29

Notes: This table establishes the negative, statistically, and economically significant effect of pre-1500CE potential crop return on the existence of future tense in a language family, and not with other language structures. The analysis accounts for regional fixed-effects and other geographical characteristics as in previous tables. Other language structures include the existence a past tense, a perfect tense, the number of genders, the existence of obligatory possessive inflections, semantic distinctions of evidentiality, the number of consonants, the ratio of consonants to vowels and the number of colors. Coefficients estimated using a weighted Probit, where weights correct representativeness of each language family in sample. For all variables the language family mean value is used in the analysis. Heteroskedasticity robust standard error estimates are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.



(a) Median



(b) Mean

Figure A.2: Crop Return and Future Across Language Genera

Table A.10: Crop Return and Language Structures across Language Families

	Language Structure								
	Temporal Structures			Non-Temporal Structures					
	Future	Past	Perfect	Gender	Possessive	Evidentiality	Consonants	C/V Ratio	Colors
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Crop Return (pre-1500CE)	-0.38***	-0.12	-0.08	0.19	-0.10	0.09	0.44***	0.28	1.61
	(0.08)	(0.11)	(0.13)	(0.14)	(0.13)	(0.06)	(0.15)	(0.17)	(2.90)
All Geographic Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Regional FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted-R^2	0.49	0.36	0.19	0.06	0.18	0.79	0.79	0.58	0.38
Observations	66	66	66	52	57	64	66	66	25

Notes: This table establishes the negative, statistically, and economically significant effect of pre-1500CE potential crop return on the existence of future tense in a language family, and not with other language structures. The analysis accounts for regional fixed-effects and other geographical characteristics as in previous tables. Other language structures include the existence a past tense, a perfect tense, the number of genders, the existence of obligatory possessive inflections, semantic distinctions of evidentiality, the number of consonants, the ratio of consonants to vowels and the number of colors. Coefficients estimated using a weighted Probit, where weights correct representativeness of each language family in sample. For all variables the language family median value is used in the analysis. Heteroskedasticity robust standard error estimates are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.

Table A.11: Crop Return and Future Tense Across Language Families (Monte Carlo)

	Existence of Future Tense					
	(1)	(2)	(3)	(4)	(5)	(6)
Crop Return	-0.15***	-0.23***	-0.22***	-0.19***	-0.17**	-0.19**
(pre-1500CE)	(0.04)	(0.05)	(0.05)	(0.07)	(0.08)	(0.08)
Unproductive Period						-0.10
(pre-1500CE)						(0.06)
Regional FE	No	Yes	Yes	Yes	Yes	Yes
Main Geographical Controls	No	No	Yes	Yes	Yes	Yes
Precipitation Controls	No	No	No	Yes	Yes	Yes
Temperature Controls	No	No	No	No	Yes	Yes

Notes: This table establishes the negative, statistically, and economically significant effect of pre-1500CE potential crop return on the existence of future tense, accounting for regional fixed-effects and the set of geographical characteristics in Table 1. Estimates based on Monte Carlo simulations sampling one language for each language family in each simulation. The results are based on 76 Language families and 5000 simulations. All independent variables have been normalized by subtracting their mean and dividing by their standard deviation at the language level. Thus, all coefficients can be compared and show the effect of a one standard deviation in the independent variable on the probability of having a future tense in the language. Heteroskedasticity robust standard error estimates are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.

Table A.12: Crop Return and Future Tense Across Language Families (Monte Carlo)

	Existence of Future Tense					
	(1)	(2)	(3)	(4)	(5)	(6)
Crop Return (pre-1500CE)	-0.17***	-0.24***	-0.23***	-0.20***	-0.19**	-0.20***
	(0.04)	(0.04)	(0.05)	(0.07)	(0.08)	(0.07)
Unproductive Period (pre-1500CE)						-0.10
						(0.07)
Regional FE	No	Yes	Yes	Yes	Yes	Yes
Main Geographical Controls	No	No	Yes	Yes	Yes	Yes
Precipitation Controls	No	No	No	Yes	Yes	Yes
Temperature Controls	No	No	No	No	Yes	Yes
Adjusted-R^2	0.18	0.25	0.26	0.25	0.24	0.26
Observations	76	76	76	76	76	76

Notes: This table establishes the negative, statistically, and economically significant effect of pre-1500CE potential crop return on the existence of future tense, accounting for regional fixed-effects and the set of geographical characteristics in Table 1. Estimates based on Monte Carlo simulations sampling one language for each language family in each simulation. The analysis assumes that the median level of future in the family represents the proto-language. The results are based on 76 Language families and 5000 simulations. All independent variables have been normalized by subtracting their mean and dividing by their standard deviation at the language level. Thus, all coefficients can be compared and show the effect of a one standard deviation in the independent variable on the probability of having a future tense in the language. Heteroskedasticity robust standard error estimates are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.

A.3 Mechanisms

Table A.13: Agricultural Intensity and Existence of Future Tense

	Existence of Future Tense					
	OLS					IV
	(1)	(2)	(3)	(4)	(5)	(6)
Agricultural Intensity	-0.07**	-0.11***	-0.10**	-0.09**	-0.10**	-0.34***
	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)	(0.10)
Regional FE	No	Yes	Yes	Yes	Yes	Yes
Main Geographic Controls	No	No	Yes	Yes	Yes	Yes
Main Precipitation Controls	No	No	No	Yes	Yes	Yes
Main Temperature Controls	No	No	No	No	Yes	Yes
First-stage F-statistic						18.78
Hansen's J-statistic						5.04
J-stat p-value						0.02
Adjusted-R^2	0.02	0.06	0.07	0.08	0.13	0.03
Observations	264	264	264	264	264	264

Notes: This table establishes the positive statistically and economically significant effect of agricultural intensity on the existence of future tense in the language spoken by a pre-modern society. The table replicates Table 8 in columns (1)-(5) using OLS and also instruments agricultural suitability with crop return and the length of the unproductive period, both of which affect a society's agricultural intensity. Reassuringly, the OLS estimates are similar to the Probit ones, while the IV estimates are 3.4 times larger, suggesting that a one standard deviation increase in agricultural suitability would decrease the probability of the existence of a future tense by 0.34 percentage points. While this hints that the estimates in Table 8 might be biased towards zero, the overidentification test in Table A.13 suggests that the IV does not satisfy the exclusion restriction. In part, this could be explained by the proposed mechanism in Figure 7(b). Heteroskedasticity robust standard error estimates are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.

Table A.14: Size of Local Community and Existence of Future Tense

	Existence of Future Tense					
	OLS					IV
	(1)	(2)	(3)	(4)	(5)	(6)
Size of Local Communities	-0.07*	-0.10**	-0.09**	-0.08*	-0.08*	-0.29***
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.08)
Regional FE	No	Yes	Yes	Yes	Yes	Yes
Main Geographic Controls	No	No	Yes	Yes	Yes	Yes
Main Precipitation Controls	No	No	No	Yes	Yes	Yes
Main Temperature Controls	No	No	No	No	Yes	Yes
First-stage F-statistic						24.46
Hansen's J-statistic						2.09
J-stat p-value						0.15
Adjusted-R^2	0.01	0.08	0.09	0.09	0.11	-0.02
Observations	163	163	163	163	163	163

Notes: This table replicates the analysis in Table 10 using OLS and additionally instrumenting the size of local communities using crop return and the length of the unproductive period. Reassuringly, the OLS estimates are similar to the ones in Table 10. On the other hand, the IV estimates suggest a much larger causal effect of the size of local communities on the existence of a future tense, with a one standard deviation increase in the size of local communities decreasing the existence of a future tense by 0.29 percentage points. Moreover, the overidentification test in Table A.14 suggests the instruments are valid. Heteroskedasticity robust standard error estimates are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.

Table A.15: Crop Return, Agricultural Suitability and Future Tense

	Existence of Future Tense					
	(1)	(2)	(3)	(4)	(5)	(6)
Crop Return (pre-1500CE)	-0.12***			-0.13**	-0.15***	-0.12**
	(0.03)			(0.05)	(0.03)	(0.05)
Caloric Suitability (pre-1500CE)		-0.08**		0.01		-0.04
		(0.03)		(0.05)		(0.05)
Agricultural Suitability			0.09		0.28**	0.32**
			(0.12)		(0.12)	(0.13)
All Geographic Controls	Yes	Yes	Yes	Yes	Yes	Yes
Regional FE	Yes	Yes	Yes	Yes	Yes	Yes
Pseudo-R^2	0.14	0.12	0.11	0.14	0.15	0.16
Observations	275	275	275	275	275	275

Notes: Heteroskedasticity robust standard error estimates are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.

B Language Structures & Contemporary Behavior

B.1 Crop Return, Future Tense, and Long-Term Oriented Behavior of Second-Generation Migrants

Table B.1: Pre-1500CE Crop Return, Future Tense, and College Education of Second and Higher Generation Migrants

	College Attendance				
	(1)	(2)	(3)	(4)	(5)
Future Tense	-0.100***	-0.132***	-0.125***		-0.111***
	(0.014)	(0.005)	(0.004)		(0.004)
Crop Return (pre-1500CE)				0.034***	0.019***
				(0.001)	(0.001)
Main Geographical Controls	Yes	Yes	Yes	Yes	Yes
State FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Age FE	No	Yes	Yes	Yes	Yes
Gender FE	No	Yes	Yes	Yes	Yes
Marital Status FE	No	No	Yes	Yes	Yes
Adjusted-R^2	0.02	0.07	0.08	0.08	0.08
R^2	0.02	0.07	0.09	0.09	0.09
Observations	12206839	12206839	12206839	12206839	12206839

Table B.2: Pre-1500CE Crop Return, Future Tense, and College Education
One-and-a-Half Generation Migrants (Who Arrived at Age ≤ 5)

	College Attendance					
	Language				Crop Return	Both
	(1)	(2)	(3)	(4)	(5)	(6)
Future Tense	-0.205***	-0.208***	-0.204***	-0.056***		-0.054***
	(0.013)	(0.007)	(0.005)	(0.007)		(0.007)
Crop Return (pre-1500CE)					0.009***	0.003
					(0.002)	(0.002)
Main Geographical Controls	Yes	Yes	Yes	Yes	Yes	Yes
State FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Age FE	No	Yes	Yes	Yes	Yes	Yes
Gender FE	No	Yes	Yes	Yes	Yes	Yes
Marital Status FE	No	No	Yes	Yes	Yes	Yes
Parental Origin FE	No	No	No	Yes	Yes	Yes
Adjusted-R^2	0.06	0.09	0.10	0.15	0.15	0.15
R^2	0.06	0.11	0.17	0.48	0.48	0.48
Observations	422081	422081	422081	422081	422081	422081

Notes: Heteroskedasticity robust standard error estimates clustered at the level of the included fixed-effects are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.

Table B.3: Future Tense and College Education of Second Generation Migrants Accounting for Parental Education and English Levels, and Local Socio-Economic Conditions (County Level)

	College Attendance								
	Parental Education			Parental English			Both		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Future Tense	-0.032***		-0.029***	-0.030***		-0.030***	-0.029***		-0.027***
	(0.005)		(0.005)	(0.006)		(0.006)	(0.005)		(0.005)
Crop Return (pre-1500CE)		0.006***	0.003**		0.003*	0.000		0.005***	0.003**
		(0.002)	(0.001)		(0.002)	(0.001)		(0.001)	(0.001)
Mom's College Attendance	0.131***	0.131***	0.131***				0.133***	0.133***	0.133***
	(0.006)	(0.006)	(0.006)				(0.006)	(0.006)	(0.006)
Dad's College Attendance	0.141***	0.142***	0.141***				0.143***	0.143***	0.143***
	(0.006)	(0.006)	(0.006)				(0.006)	(0.006)	(0.006)
Mom's English Level				0.013***	0.013***	0.013***	0.014***	0.014***	0.014***
				(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
Dad's English Level				-0.001	-0.001	-0.001	0.006***	0.006***	0.006***
				(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
Main Geographical Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Origin FE for Both Parents	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
County FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Age FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Gender FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Marital Status FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted-R^2	0.18	0.18	0.18	0.14	0.14	0.14	0.18	0.18	0.18
R^2	0.30	0.30	0.30	0.27	0.27	0.27	0.30	0.30	0.30
Observations	91613	91613	91613	91613	91613	91613	91613	91613	91613

Notes: Heteroskedasticity robust standard error estimates two-way clustered by country of origin of both parents are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.

Table B.4: Pre-1500CE Crop Return, Future Tense, and College Education of Second Generation Migrants
Effect of Parents Origin

	College Attendance					
	Same			Different		
	(1)	(2)	(3)	(4)	(5)	(6)
Future Tense	-0.009		-0.010	-0.028***		-0.014*
	(0.006)		(0.006)	(0.008)		(0.008)
Crop Return (pre-1500CE)		-0.000	-0.001		0.018***	0.016***
		(0.001)	(0.001)		(0.003)	(0.003)
Mom's College Attendance	0.112***	0.112***	0.112***	0.155***	0.155***	0.155***
	(0.005)	(0.005)	(0.005)	(0.006)	(0.006)	(0.006)
Dad's College Attendance	0.121***	0.121***	0.121***	0.163***	0.163***	0.163***
	(0.006)	(0.006)	(0.006)	(0.007)	(0.007)	(0.007)
Main Geographical Controls	Yes	Yes	Yes	Yes	Yes	Yes
County FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Age FE	Yes	Yes	Yes	Yes	Yes	Yes
Gender FE	Yes	Yes	Yes	Yes	Yes	Yes
Marital Status FE	Yes	Yes	Yes	Yes	Yes	Yes
Parental Origin FE	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted-R^2	0.19	0.19	0.19	0.18	0.18	0.18
R^2	0.34	0.34	0.34	0.37	0.37	0.37
Observations	54252	54252	54252	42614	42614	42614

Notes: Heteroskedasticity robust standard error estimates three-way clustered by state and country of origin of both parents are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.

Table B.5: Pre-1500CE Crop Return, Future Tense, and College Education of Second Generation Migrants

	Educational Level Higher than High School					
	(1)	(2)	(3)	(4)	(5)	(6)
Future Tense	-0.218***	-0.225***	-0.221***	-0.056***		-0.053***
	(0.013)	(0.008)	(0.007)	(0.012)		(0.012)
Crop Return (pre-1500CE)					0.008***	0.003
					(0.003)	(0.003)
Main Geographical Controls	Yes	Yes	Yes	Yes	Yes	Yes
State FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Age FE	No	Yes	Yes	Yes	Yes	Yes
Gender FE	No	Yes	Yes	Yes	Yes	Yes
Marital Status FE	No	No	Yes	Yes	Yes	Yes
Parental Origin FE	No	No	No	Yes	Yes	Yes
Adjusted-R^2	0.05	0.12	0.20	0.49	0.49	0.49
R^2	0.05	0.12	0.20	0.49	0.49	0.49
Observations	18845303	18845303	18845303	18845303	18845303	18845303

Notes: Heteroskedasticity robust standard error estimates clustered at the level of the included fixed-effects are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.

Table B.6: Pre-1500CE Crop Return, Future Tense, and College Education of Second Generation Migrants
Accounting for Parental Education and English Levels

	Educational Level Higher than High School								
	Parental Education			Parental English			Both		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Future Tense	-0.056***		-0.052***	-0.041***		-0.042***	-0.037***		-0.038***
	(0.008)		(0.009)	(0.010)		(0.011)	(0.010)		(0.011)
Crop Return (pre-1500CE)		0.010***	0.004*		0.002	-0.002		0.003	-0.000
		(0.002)	(0.002)		(0.003)	(0.003)		(0.003)	(0.003)
Mom's Education Level (HS+)	0.137***	0.137***	0.137***				0.138***	0.138***	0.138***
	(0.004)	(0.004)	(0.004)				(0.005)	(0.005)	(0.005)
Dad's Education Level (HS+)	0.069***	0.069***	0.069***				0.147***	0.147***	0.147***
	(0.004)	(0.004)	(0.004)				(0.005)	(0.005)	(0.005)
Mom's English Level				0.012***	0.012***	0.012***	0.013***	0.013***	0.013***
				(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Dad's English Level				-0.002	-0.002	-0.002	0.003*	0.003*	0.003*
				(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Main Geographical Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Origin FE for Both Parents	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Age FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Gender FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Marital Status FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Parental Origin FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted-R^2	0.25	0.25	0.25	0.26	0.26	0.26	0.29	0.29	0.29
R^2	0.25	0.25	0.25	0.26	0.26	0.26	0.29	0.29	0.29
Observations	18845303	18845303	18845303	11187136	11187136	11187136	11187136	11187136	11187136

Notes: Heteroskedasticity robust standard error estimates clustered at the level of the included fixed-effects are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.

Table B.7: Pre-1500CE Crop Return, Future Tense, and College Education
One-and-a-Half Generation Migrants (Who Arrived at Age ≤ 5)

	Educational Level Higher than High School					
	(1)	(2)	(3)	(4)	(5)	(6)
Future Tense	-0.220***	-0.223***	-0.219***	-0.067***		-0.067***
	(0.011)	(0.006)	(0.006)	(0.007)		(0.007)
Crop Return (pre-1500CE)					0.007***	-0.000
					(0.002)	(0.002)
Main Geographical Controls	Yes	Yes	Yes	Yes	Yes	Yes
State FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Age FE	No	Yes	Yes	Yes	Yes	Yes
Gender FE	No	Yes	Yes	Yes	Yes	Yes
Marital Status FE	No	No	Yes	Yes	Yes	Yes
Parental Origin FE	No	No	No	Yes	Yes	Yes
Adjusted-R^2	0.06	0.12	0.20	0.55	0.55	0.55
R^2	0.06	0.12	0.20	0.55	0.55	0.55
Observations	42457006	42457006	42457006	42457006	42457006	42457006

Notes: Heteroskedasticity robust standard error estimates clustered at the level of the included fixed-effects are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.

B.2 Comparing Migrants Samples

Table B.8: Means across Generational Samples (Census vs. CPS)

	Means							
	1.5 Generation		2nd Generation				2+ Generations	
	Census	CPS	Census	CPS (living with Parents)	CPS (not living with Parents)	CPS (All)	Census	CPS (3+ Generation)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Education Level (HS+)	0.596***	0.648***	0.552***	0.600***	0.568***	0.571***	0.535***	0.572***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.000)	(0.000)	(0.000)	(0.000)
Age	43.742***	38.625***	33.913***	34.092***	55.963***	54.376***	51.685***	50.133***
	(0.022)	(0.024)	(0.022)	(0.032)	(0.017)	(0.017)	(0.004)	(0.004)
Gender	1.518***	1.518***	1.457***	1.462***	1.537***	1.531***	1.526***	1.527***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.000)	(0.000)	(0.000)	(0.000)
Marital Status	2.702***	2.737***	4.933***	5.099***	2.597***	2.779***	2.524***	2.489***
	(0.003)	(0.005)	(0.004)	(0.005)	(0.002)	(0.002)	(0.000)	(0.001)
Observations	429372	174094	181099	94331	1205633	1299964	20596324	14180541

Notes: Standard error estimates are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.

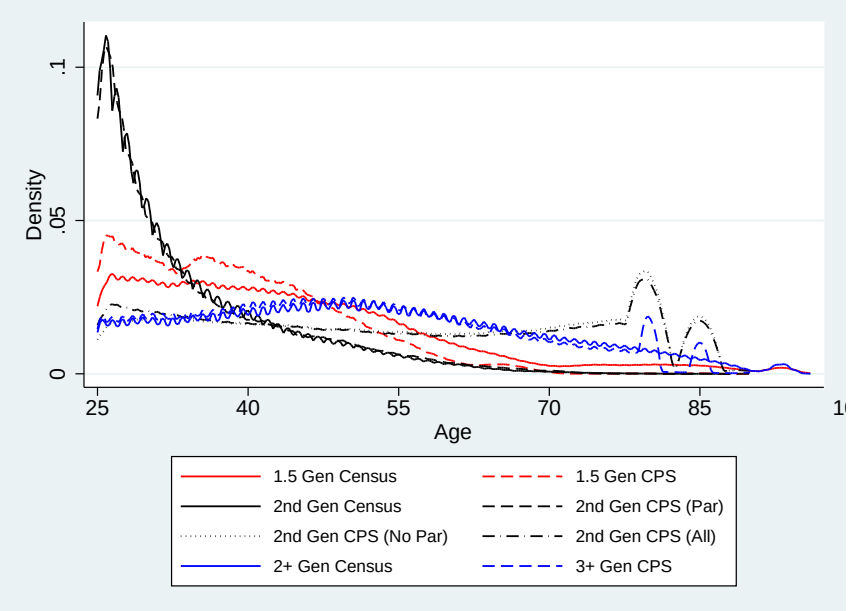


Figure B.1: Age Density in the Census and CPS

Table B.9: Means across Generational Samples (Census vs. CPS)

	Means							
	1.5 Generation		2nd Generation				2+ Generations	
	Census	CPS	Census	CPS (living with Parents)	CPS (not living with Parents)	CPS (All)	Census	CPS (3+ Generation)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Education Level (HS+)	0.590***	0.638***	0.545***	0.603***	0.581***	0.583***	0.542***	0.568***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Age	42.489***	38.163***	33.347***	33.505***	54.517***	52.666***	50.258***	49.930***
	(0.002)	(0.000)	(0.002)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)
Gender	1.510***	1.503***	1.445***	1.437***	1.530***	1.521***	1.522***	1.524***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Marital Status	2.855***	2.810***	5.000***	5.143***	2.624***	2.846***	2.679***	2.547***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Observations	43181154	403711034	20841131	245898566	2544423483	2790322049	1831557413	28887227869

Notes: Standard error estimates are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.

B.3 Future Tense and Education in the WVS

Table B.10: Pre1500 Crop Return, Future Tense, and Education – World Values Survey

	Education Level								
	Basic Controls			Income			Religion		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Crop Return (pre-1500CE)	0.18***		0.23***	0.18***		0.23***	0.25***		0.28***
	(0.04)		(0.04)	(0.04)		(0.04)	(0.04)		(0.04)
Future Tense		-0.43***	-0.47***		-0.40***	-0.44***		-0.25***	-0.30***
		(0.05)	(0.05)		(0.05)	(0.05)		(0.05)	(0.05)
Income FE	No	No	No	No	No	No	Yes	Yes	Yes
Religion FE	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Main Geographical Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Wave FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Gender FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Age FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted-R^2	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	0.09	0.09	0.09
Observations	108213	108213	108213	108213	108213	108213	108213	108213	108213

C Origins and Effects of Other Language Structures

Table C.1: Geographic Origins of Sex-Based Grammatical Genders and Plow Usage

	Results (Probit)					
	Reduced Form		Mechanism			
	Gender		Plow		Gender	
	(1)	(2)	(3)	(4)	(5)	(6)
Plow Negative CSI (pre-1500CE)	-0.12**	-0.19***	-0.19***	0.01		
	(0.05)	(0.07)	(0.02)	(0.02)		
Caloric Suitability Index (pre-1500CE)	0.14***	0.21***	0.17***	0.08***		
	(0.05)	(0.06)	(0.02)	(0.02)		
Plow					0.32***	0.23**
					(0.07)	(0.10)
All Geographic Controls	No	Yes	No	Yes	No	Yes
Continental FE	No	Yes	No	Yes	No	Yes
Pseudo-R^2	0.03	0.25	0.23	0.52	0.11	0.37
Observations	217	217	1178	824	114	101

Notes: This table establishes the positive statistically and economically significant effect of the geographical determinants of and actual plow usage on the existence of sex-based grammatical gender in a language. The first two columns provide the results of the reduced form, and columns (3)-(6) provide evidence on the mechanism. The table shows the average marginal effects of a Probit regression. Heteroskedasticity robust standard error estimates are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests.

Table C.2: Persistent Effect of Urheimat Characteristics on Gender

	Existence of Sex-Based Gender System						
	All Languages				Languages In/Near Urheimat		
					All	Avg.$\Delta <$ 0.5SD	Avg.$\Delta <$ 0.01SD
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Urheimat Plow Negative CSI (pre-1500CE)	-0.11	-0.33***	-0.11	-0.30***	-0.37***	-0.40***	-0.37**
	(0.10)	(0.08)	(0.09)	(0.09)	(0.12)	(0.12)	(0.14)
Urheimat Caloric Suitability Index (pre-1500CE)	0.25***	0.43***	0.23***	0.39***	0.46***	0.47***	0.41***
	(0.08)	(0.08)	(0.08)	(0.09)	(0.10)	(0.11)	(0.15)
Change in Plow Negative CSI			-0.07	-0.07	-0.04	0.09	-0.14
			(0.07)	(0.08)	(0.10)	(0.19)	(0.25)
Change in CSI			-0.03	0.06	-0.00	-0.10	0.13
			(0.06)	(0.06)	(0.06)	(0.16)	(0.26)
Regional FE	No	Yes	No	Yes	Yes	Yes	No
Urheimat Geographical Characteristics	No	Yes	No	Yes	Yes	Yes	No
Change in Geographical Characteristics	No	No	No	Yes	Yes	Yes	No
Adjusted-R^2	0.14	0.40	0.16	0.38	0.31	0.22	0.08
Observations	213	213	213	213	185	146	79
Language Families	70	70	70	70	69	67	48

Notes: Heteroskedasticity robust standard error estimates clustered at the language family level are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests; All regressions include a constant.

Table C.3: Persistent Effect of Urheimat Characteristics on Politeness

	Existence of Politeness Distinctions						
	All Languages				Languages In/Near Urheimat		
					All	Avg.$\Delta <$ 0.5SD	Avg.$\Delta <$ 0.01SD
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Ecological Diversity	0.16**	-0.07	0.24***	0.00	-0.01	0.01	0.11
	(0.08)	(0.08)	(0.07)	(0.08)	(0.08)	(0.09)	(0.18)
Caloric Suitability Index (pre-1500CE)	0.08*	0.11*	0.09**	0.08	0.04	0.01	0.07
	(0.04)	(0.06)	(0.04)	(0.05)	(0.05)	(0.05)	(0.07)
Change in Ecological Diversity (ΔED)			0.10***	0.09**	0.09*	-0.01	-0.00
			(0.04)	(0.04)	(0.04)	(0.08)	(0.05)
Change in CSI			0.08	0.08**	0.12***	-0.00	-0.01
			(0.05)	(0.04)	(0.04)	(0.10)	(0.07)
Regional FE	No	Yes	No	Yes	Yes	Yes	No
Urheimat Geographical Characteristics	No	Yes	No	Yes	Yes	Yes	No
Change in Geographical Characteristics	No	No	No	Yes	Yes	Yes	No
Adjusted-R^2	0.15	0.32	0.19	0.40	0.37	0.40	-0.14
Observations	196	196	196	196	169	109	21
Language Families	66	66	66	66	64	54	20

Notes: Heteroskedasticity robust standard error estimates clustered at the language family level are reported in parentheses; *** denotes statistical significance at the 1% level, ** at the 5% level, and * at the 10% level, all for two-sided hypothesis tests; All regressions include a constant.

D Variable Definitions, Sources and Summary Statistics

Table D.1: Summary Statistics of the Existence of Future Tense by Region

Region	Observations	Mean	Std. Dev.
Sub-Saharan Africa	66	0.47	0.503
Middle East and North Africa	8	0.5	0.53
Europe and Central Asia	56	0.52	0.50
South Asia	21	0.81	0.40
East Asia and Pacific	71	0.45	0.50
North America	22	0.59	0.50
Latin America	31	0.45	0.50
Total	275	0.51	0.50

Table D.2: Summary Statistics of the Existence of Sex-Based Grammatical Gender Systems by Region

Region	Observations	Mean	Std. Dev.
Sub-Saharan Africa	27	0.63	0.49
Middle East and North Africa	7	0.71	0.49
Europe and Central Asia	40	0.48	0.51
South Asia	16	0.63	0.50
East Asia and Pacific	70	0.27	0.45
North America	25	0.08	0.28
Latin America	32	0.28	0.46
Total	227	0.37	0.48

Table D.3: Summary Statistics of the Existence of Politeness Distinctions by Region

Region	Observations	Mean	Std. Dev.
Sub-Saharan Africa	36	0.14	0.35
Middle East and North Africa	4	0.25	0.50
Europe and Central Asia	34	0.71	0.46
South Asia	19	0.63	0.50
East Asia and Pacific	59	0.32	0.47
North America	18	0.00	0.00
Latin America	28	0.18	0.39
Total	207	0.34	0.48

Table D.4: Summary statistics

Variable	**Mean**	**Std. Dev.**	**Min.**	**Max.**	**N**
Absolute Latitude	0.096	1.025	-1.302	2.613	275
Elevation	0.027	1.026	-0.92	4.827	275
Ruggedness	-0.014	0.979	-0.877	6.162	275
Coast Length	0.024	1.154	-0.302	11.692	275
Precipitation	-0.078	0.928	-1.3	4.4	275
Precipitation (std)	-0.02	0.911	-0.667	8.314	275
Precipitation Volatility	-0.064	0.926	-1.531	4.665	275
Precipitation Spatial Correlation	0.064	0.939	-2.133	0.810	275
Temperature (Daily Mean)	-0.054	0.977	-2.996	1.176	275
Temperature (Daily Mean) (std)	-0.017	0.929	-0.877	4.876	275
Temperature Volatility	0.079	0.991	-1.641	3.504	275
Temperature Spatial Correlation	0.068	0.939	-2.161	0.683	275

Table D.5: Summary statistics

Variable	**Mean**	**Std. Dev.**	**Min.**	**Max.**	**N**
Intensity of Agriculture	8.890	3.061	2	12	264