Linguistic Traits and Human Capital Formation

By Oded Galor, Ömer Özak and Assaf Sarid*

The origins of the vast inequality in the wealth of nations have been attributed to the persistent effect of existing variations in the distribution of geographical, cultural, institutional, and human characteristics across the globe.[1] In light of the coevolution of cultural and linguistic characteristics in the course of human history, the evolution of linguistic traits has conceivably reinforced the persistent effect of cultural factors on the inequality in the wealth of nations.

This research explores the impact of the coevolution of linguistic and cultural traits on the development process: Has this coevolution contributed to the persistence of cultural characteristics and their lasting effect on economic prosperity? Have linguistic traits merely reflected existing cultural characteristics or have they influenced human behavior and values and contributed directly to the development process?

In view of the pivotal role of languages in the transmission of knowledge and values, linguistic traits have plausibly reinforced the diffusion of cultural traits within and across generations. Natural selection across language structures favored those that fortified prevailing cultural traits since cultural characteristics which were manifested in language structures were more likely to persist across time and space. Moreover, linguistic traits per se may have directly influenced the individual mindset and thus human behavior, beyond their impact via cultural transmission.

* Galor: Department of Economics, Brown University; NBER, CEPR, IZA, CES-Ifo. E-mail: Oded_Galor@brown.edu. Özak: Department of Economics, Southern Methodist University; IZA. E-mail: ozak@smu.edu. Sarid: Department of Economics, University of Haifa. Email: asarid@econ.haifa.ac.il.

[1] Gallup, Sachs and Mellinger (1999), Guiso, Sapienza and Zingales (2006), Acemoglu, Johnson and Robinson (2001), Glaeser et al. (2004), and Ashraf and Galor (2013).

In particular, in a society characterized by distinct gender roles, and consequently by the existence of gender bias, grammatical gender, which has plausibly fortified these cultural norms, emerged and persisted over time. Similarly, in societies characterized by long-term orientation, a structure of the future tense, which has presumably reinforced the efficiency of future oriented behavior, emerged and persisted over time (Galor, Özak and Sarid, 2018).

Languages differ in the existence and the form of grammatical gender and the structure of their future tense. In particular, languages that are characterized by sex-based grammatical gender classify nouns according to biological gender. The presence of sex-based grammatical gender induces speakers to highlight gender distinctions even in situations in which gender may not play an intrinsic role. Thus, linguists as well as other scholars have argued that gender biases have been reinforced by languages characterized by sex-based grammatical gender systems.

Similarly, linguists distinguish between languages that are characterized by an *inflectional* versus *periphrastic* future tense. Inflectional future tense is associated with verbs that display morphological variation. In contrast, periphrastic future tense is characterized by roundabout or discursive phrases such as 'will', 'shall', 'want to', and 'going to', in the English language. Linguists suggest that "intention and prediction are most commonly expressed by the periphrastic future." (Bybee, Perkins and Pagliuca, 1994). Thus, scholars have argued that long-term oriented behavior have been reinforced periphrastic future tense.

This research examines the effects of these two language structures on contemporary human capital formation, conceivably via their indirect effect on the persistence of ancestral cultural traits, as well as their

direct effect on individual mindsets and behavior. In particular, the analysis explores the effect of (i) the presence periphrastic future tense on educational attainment, and (ii) the presence of sex-based grammatical gender on female educational attainment.

The study advances a novel identification methodology that resolves major limitation that characterize existing explorations of the association between language structures and economic outcomes, disentangling the impact of language from the persistent effect of other ancestral characteristics (Kashima and Kashima, 1998; Chen, 2013). In particular, it overcomes the shortcomings of the traditional epidemiological approach for the study of the persistence of culture, and develops an identification strategy that permits the isolation of the effect of linguistic traits, from the persistent effects of other cultural characteristics, on human behavior as well as the direct effect of linguistic traits from their potential indirect effects via their impact on the persistence of cultural traits.

I. Identification Strategy

The analysis exploits variations in language structures across individuals that are originated from the same ancestral homelands in order to identify the effect of language-embodied cultural traits on human behavior, transcending a major limitation of the influential epidemiological approach for the identification of the persistent effects of cultural traits on human behavior and economic outcomes (Giuliano, 2007; Fernandez and Fogli, 2009; Galor and Özak, 2016). While the epidemiological approach permits the exploration of the impact of the ancestral environment of children of migrants on cultural traits, accounting for the potential impact of geographical, institutional and cultural characteristics in the host country, it does not distinguish between the persistent effect of observed cultural characteristics and those of other unobserved ancestral characteristics that reflect the parental countries of origin.

In contrast, the proposed methodology isolates the direct effect of linguistic traits on human behavior, from the persistent effect of ancestral cultural characteristics. In particular, since some children of migrants from identical countries of origin speak different languages, one can disentangle the impact of linguistic traits from the ancestral environment, accounting for geographical, institutional and cultural characteristics that characterized the ancestral homelands and may partly govern the behavior of children of migrants (i.e., for the parental countries of origin fixed effects).

The analysis is conducted on individual data from the US Census and American Community Survey for the years 2000-2017 (Ruggles et al., 2019). It focuses on all children of migrants over the age of 24, who were either born in the US, or brought to the US before the age of 5. This sample consists of 747,062 individuals, whose parents migrated from 147 countries and speak 64 languages. The prevalence of sex-based grammatical gender and periphrastic future tense in these 64 languages is determined based on the classifications provided by Dryer (2013).

II. Periphrastic Future Tense and Education of Children of Migrants

This section explores the language-embodied effect of long-term orientation on human capital formation. In particular, it explores the effect of speaking a language with periphrastic future tense on the probability of college attendance among children of migrants in the US.

Table 1 establishes the positive effect of speaking a language with periphrastic future tense on college attendance. In particular, the estimates in columns (1)-(2) suggest that speaking a language with periphrastic future tense increases the probability of attending college by 23 percentage points, accounting for individual characteristics such as age, gender, marital status, state of residence, and year of interview, as well as the geographical characteristics of the historical homeland of the language. A sizable effect in comparison to a mean probability of 0.59 of college attendance in the sample as a whole. Column

Table 1—Periphrastic Future Tense and College Education of Children of Migrants

	College Attendance						
	All			Parental		No English	No Spanish
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Periphrastic Future Tense	0.232	0.226	0.053	0.038	0.035	0.053	0.029
	(0.057)	(0.057)	(0.020)	(0.012)	(0.012)	(0.024)	(0.015)
Geographical Controls (Language Homeland)	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State & Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Age, Gender, & Marital Status FE	No	Yes	Yes	Yes	Yes	Yes	Yes
Parental Country of Origin FE	No	No	Yes	Yes	Yes	Yes	Yes
Parental Education	No	No	No	Yes	Yes	Yes	Yes
Parental English Proficiency	No	No	No	No	Yes	Yes	Yes
R^2	0.05	0.13	0.16	0.21	0.21	0.23	0.24
Observations	735482	735482	735482	164722	164722	96738	96614

Notes: The table examines the effect of speaking a language with periphrastic future tense on the probability of college attendance among children of migrants in the US. Geographical characteristics in the historical homeland of the language include absolute latitude, mean elevation, mean ruggedness, coast length and pre-1500 crop return. Heteroskedasticity robust standard error estimates clustered at the country of origin, language and state levels are reported in parentheses.

(3) further accounts for the parental countries of origin, namely, the ancestral geographical, institutional, and cultural characteristics that may affect individual human capital formation. Thus, the estimated effect of periphrastic future tense isolates the effect of long-term orientation that is language-embodied from the persistent cultural effects of long-term orientation via non-linguistic channels. The estimate implies that speaking a language with periphrastic future tense increases the probability of attending college by 5.3 percentage points.

Columns (4)-(5) establish that the estimated effect is unaffected qualitatively if one further accounts for parental education levels and their command of the English language. Not surprisingly, parental education and their level of proficiency in English have a positive and sizable effect on their offspring's college attendance. Nevertheless, the estimates suggest that the effect of speaking a language with periphrastic future tense remains sizable and it is twice as large as the effect of having a parent that is proficient in the English language and nearly 1/3 of the effect of having a college educated parent. Finally, columns (6) and (7) establish that the results are unaffected qualitatively if English or Spanish speakers are excluded from the sample, accounting for augmented labor market opportunities, and greater incentives to invest in human capital, for individuals who are proficient in the two dominating languages in the US.

Thus, the analysis in Table 1 suggests that speaking a language with periphrastic future tense has a beneficial effect on college attendance, accounting for a host of individual, socio-economic and ancestral characteristics.

III. Sex-Based Grammatical Gender and Education of Female Children of Migrants

This section explores the effect of languages characterized by the existence of sex-based grammatical gender on female human capital formation. In view of the proposed hypothesis that in a society characterized by distinct gender roles the existence of grammatical gender have reinforced prevailing gender biases, the analysis explores whether languages characterized by the existence of sex-based grammatical gender have an adverse effect on female human capital formation. In particular, fol-

lowing the identification strategy exploited in the previous section, the analysis focuses on the effect of sex-based grammatical gender on college attendance of female children of migrants in the US.

In line with the proposed hypothesis, Table 2 establishes the negative effect of speaking a language with sex-based grammatical gender on female college attendance. In particular, columns (1) and (2) show that speaking a language with sex-based grammatical gender lowers the probability of females attending college by 23 percentage points, accounting for individual characteristics such as age, gender, marital status, state of residence, and year of interview, as well as the geographical characteristics of the historical homeland of the language. A sizable effect in comparison to a mean probability of 0.61 that a woman would attend college in the sample as a whole.

Nevertheless, this effect may capture the persistence of characteristics of the parental countries of origin of these women independently of grammatical gender. Thus, column (3) accounts for parental origins fixed effects, and therefore isolates the effect of gender bias that is language-embodied from the persistent cultural effects of gender bias via non-linguistic channels. The results suggest that sex-based grammatical gender itself has an adverse effect on the probability of attending college, lowering this probability by 5.5 percentage points.

Moreover, as established in columns (4) and (5), the adverse effect of speaking a language with sex-based grammatical gender on female college attendance is robust to the confounding effect of parental education and their proficiency in the English language. The estimated effect of sex-based grammatical gender on the probability of attending college is sizable and it amounts to nearly 50% of the estimated effect of having a college educated parent. Finally, columns (6) and (7) show that excluding individuals who speak the two main languages in the US (i.e., English and Spanish) does not affect the qualitative results.

Thus, the analysis in Table 2 suggests that speaking a language with sex-based grammatical gender has an adverse effect on female college attendance, accounting for a host of individual, socio-economic and ancestral characteristics.

IV. Conclusion

This research establishes the effects of linguistic traits on individual behavior. In particular, the analysis indicates that the presence of periphrastic future tense and its association with long-term orientation has a significant impact on educational attainment, while the presence of sex-based grammatical gender, and its association with gender bias, has a significant impact on female educational attainment.

The effect of linguistic traits on contemporary human capital formation, may a priori operate via their indirect effect on the persistence of ancestral cultural traits, as well as via their direct effect on individual mindsets and behavior. Thus, the study advances a novel identification methodology, that plausibly disentangles the direct effect of linguistic traits on human behavior from its indirect effect via its impact on the persistence of ancestral cultural traits that may govern contemporary human behavior. It exploits variations in language structures across individuals that are originated from the same ancestral homelands to isolate the linguistic channel from the cultural one.

REFERENCES

Acemoglu, Daron, Simon Johnson, and James A Robinson. 2001. "The Colonial Origins of Comparative Development: An Empirical Investigation." *The American Economic Review*, 91(5): 1369–1401.

Ashraf, Quamrul, and Oded Galor. 2013. "The out of Africa hypothesis, human genetic diversity, and comparative economic development." *The American Economic Review*, 103(1): 1–46.

Bybee, Joan L., Revere Perkins, and William Pagliuca. 1994. "The evolution of grammar."

Table 2—Sex-Based Grammatical Gender and College Education of Female Children of Migrants

	College Attendance						
	All			Parental		No English	No Spanish
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Sex-Based Grammatical Gender	-0.240	-0.230	-0.055	-0.059	-0.053	-0.086	-0.046
	(0.065)	(0.060)	(0.024)	(0.019)	(0.020)	(0.047)	(0.018)
Geographical Controls (Language Homeland)	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State & Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Age, Gender, & Marital Status FE	No	Yes	Yes	Yes	Yes	Yes	Yes
Parental Country of Origin FE	No	No	Yes	Yes	Yes	Yes	Yes
Parental Education	No	No	No	Yes	Yes	Yes	Yes
Parental English Proficiency	No	No	No	No	Yes	Yes	Yes
R^2	0.05	0.13	0.16	0.20	0.20	0.20	0.25
Observations	345778	345778	345778	66267	66267	38323	34731

Notes: This table examines the effect of speaking a language with sex-based grammatical gender on female college attendance. Geographical characteristics in the historical homeland of the language include absolute latitude, mean elevation, mean ruggedness, coast length, average caloric suitability index and the average caloric yield of plow-negative crops. Heteroskedasticity robust standard error estimates clustered at the parental countries of origin, language and state levels are reported in parentheses.

Chen, M Keith. 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.

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

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

Gallup, John Luke, Jeffrey D Sachs, and Andrew D Mellinger. 1999. "Geography and economic development." *International regional science review*, 22(2): 179–232.

Galor, Oded, and Ömer Özak. 2016. "The Agricultural Origins of Time Preference." *American Economic Review*, 106(10): 3064–3103.

Galor, Oded, Ömer Özak, and Assaf Sarid. 2018. "Geographical Roots of the Coevolution of Cultural and Linguistic Traits." *NBER Working Paper w25289.*

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

Glaeser, Edward L, Rafael La Porta, Florencio Lopez-de Silanes, and Andrei Shleifer. 2004. "Do institutions cause growth?" *Journal of economic Growth*, 9(3): 271–303.

Guiso, Luigi, Paola Sapienza, and Luigi Zingales. 2006. "Does Culture Affect Economic Outcomes?" *Journal of Economic Perspectives*, 20(2): 23–48.

Kashima, Emiko S, and Yoshihisa Kashima. 1998. "Culture and language the case of cultural dimensionsand personal pronoun use." *Journal of Cross-Cultural Psychology*, 29(3): 461–486.

Ruggles, Steven, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas, and Matthew Sobek. 2019. *IPUMS USA: Version 9.0 [dataset].* IPUMS.

Online Appendix

Robustness to Various Subsamples of Children of Migrants

This appendix establishes the robustness of results to alternative sample of children of migrants: (i) children who arrived to the US before the age of 5, ("one-and-a-half generation migrants"), (ii) children born in the US ("second generation migrants"), (iii) children of migrants over the age of 21.

The sample of "one-and-a-half generation migrants," consists of 524,774 individuals, who migrated into the US before the age of 5. They were born in 147 countries and speak 64 languages.

The sample of "second-generation migrants" consists 222,288 offspring who were born in the US to at least one foreign born parent. These individuals originated from 143 countries of origin of the mother and 140 countries of origin of the father and they speak 63 languages.

The use of second-generation migrants 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. For this reason, the analysis is performed using one-and-a-half or second generation migrants. Robustness of the results to higher order migrants, as well as to other potential concerns, is established in Galor, Özak and Sarid (2016).

Table B1—Periphrastic Future Tense and College Education: One-and-a-half Generation Migrants

	College Attendance							
	All			Parental			No ENG	NO SPA
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Periphrastic Future Tense	0.228	0.224	0.065	0.068	0.078	0.073	0.082	0.056
	(0.062)	(0.061)	(0.025)	(0.017)	(0.018)	(0.017)	(0.026)	(0.030)
Geographical Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Age FE	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Gender FE	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Marital Status FE	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Origin FE	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Parental Education	No	No	No	No	Yes	Yes	Yes	Yes
Parental English Proficiency	No	No	No	No	No	Yes	Yes	Yes
R^2	0.06	0.15	0.19	0.26	0.29	0.29	0.31	0.31
Observations	513028	513028	513028	30104	30104	30104	19664	17187

Notes: Heteroskedasticity robust standard error estimates clustered at the country of origin, language and state levels 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 B2—Periphrastic Future Tense and College Education of Second Generation Migrants

	College Attendance						
	All			Parental		No English	No Spanish
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Periphrastic Future Tense	0.229	0.223	0.026	0.027	0.025	0.047	0.027
	(0.054)	(0.052)	(0.010)	(0.010)	(0.010)	(0.024)	(0.013)
Geographical Controls (Language Homeland)	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State & Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Age, Gender, & Marital Status FE	No	Yes	Yes	Yes	Yes	Yes	Yes
Parental Country of Origin FE	No	No	Yes	Yes	Yes	Yes	Yes
Parental Education	No	No	No	Yes	Yes	Yes	Yes
Parental English Proficiency	No	No	No	No	Yes	Yes	Yes
R^2	0.06	0.14	0.18	0.21	0.21	0.22	0.24
Observations	131057	131057	131057	131057	131057	74968	76206

Notes: Heteroskedasticity robust standard error estimates clustered at the parental countries of origin, language and state levels 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 B3—Sex-Based Grammatical Gender and Female College Education: One-and-a-half Generation Migrants

	College Attendance							
	All			Parental			No ENG	NO SPA
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Sex-Based Grammatical Gender	-0.238	-0.233	-0.069	-0.063	-0.106	-0.096	-0.139	-0.086
	(0.067)	(0.061)	(0.025)	(0.031)	(0.033)	(0.034)	(0.068)	(0.043)
Geographical Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Age FE	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Gender FE	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Marital Status FE	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Origin FE	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Parental Education	No	No	No	No	Yes	Yes	Yes	Yes
Parental English Proficiency	No	No	No	No	No	Yes	Yes	Yes
R^2	0.06	0.16	0.19	0.26	0.28	0.29	0.29	0.34
Observations	250910	250910	250910	11619	11619	11619	7425	5705

Notes: Heteroskedasticity robust standard error estimates clustered at the parental countries of origin, language and state levels are reported in parentheses.

Table B4—Sex-Based Grammatical Gender and Female College Education of Second Generation Migrants

	College Attendance						
	All			Parental		No English	No Spanish
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Existence of Sex-Based Gender System	-0.201	-0.181	-0.014	-0.040	-0.036	-0.067	-0.036
	(0.048)	(0.046)	(0.021)	(0.017)	(0.017)	(0.018)	(0.022)
Geographical Controls (Language Homeland)	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State & Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Age, Gender, & Marital Status FE	No	Yes	Yes	Yes	Yes	Yes	Yes
Parental Country of Origin FE	No	No	Yes	Yes	Yes	Yes	Yes
Parental Education	No	No	No	Yes	Yes	Yes	Yes
Parental English Proficiency	No	No	No	No	Yes	Yes	Yes
R^2	0.07	0.14	0.17	0.20	0.20	0.20	0.26
Observations	52955	52955	52955	52955	52955	29998	27534

Notes: Heteroskedasticity robust standard error estimates clustered at the parental countries of origin, language and state levels 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.

*

Appendix References

Duncan, Brian, and Stephen J Trejo. 2011. "Intermarriage and the intergenerational transmission of ethnic identity and human capital for Mexican Americans." *Journal of Labor Economics*, 29(2): 195.

Duncan, Brian, and Stephen J Trejo. 2016. "The complexity of immigrant generations: Implications for assessing the socioeconomic integration of Hispanics and Asians." *NBER Working Paper Series*, , (w21982).

Galor, Oded, Ömer Özak, and Assaf Sarid. 2016. "Geographical Origins and Economic Consequences of Language Structures." Institute for the Study of Labor (IZA).