Migration as a Window into the Coevolution between Language and Behavior

VICTOR GAY
Department of Economics, University of Chicago,
Chicago, USA
victorgay@uchicago.edu

DANIEL L. HICKS
Department of Economics, University of Oklahoma
Norman, USA
hicksd@ou.edu

ESTEFANIA SANTACREU-VASUT
Department of Economics, ESSEC Business School
Cergy-Pontoise, France
santacreuvasut@essec.edu

Understanding the causes and consequences of language evolution in relation to social factors is challenging as we generally lack a clear picture of how languages coevolve with historical social processes. Research analyzing the relation between language and socio-economic factors relies on contemporaneous data. Because of this, such analysis may be plagued by spurious correlation concerns coming from the historical co-evolution and dependency of the relationship between language and behavior to the institutional environment. To solve this problem, we propose migrations to the same country as a microevolutionary step that may uncover constraints on behavior. We detail strategies available to other researchers by applying the epidemiological approach to study the correlation between sex-based gender distinctions and female labor force participation. Our main finding is that language must have evolved partly as a result of cultural change, but also that it may have directly constrained the evolution of norms. We conclude by discussing implications for the coevolution of language and behavior, and by comparing different methodological approaches.

1. Introduction

1.1. *The Methodological Challenge*

Disentangling whether language influences the evolution of society, whether social factors impact language evolution, or whether they are independent of each other, is a daunting challenge. Indeed, it requires ruling out spurious correlations in cross-cultural linguistic analysis and addressing a fundamental problem of identification. As Roberts and Winters (2013) highlight, it may be inappropriate to simply treat languages as independent data points because they are related by both vertical and horizontal transmission mechanisms.

For instance, sharing a common ancestor (language families) or spillovers via contact with neighboring languages in the past (linguistic areas) may generate spurious correlations between language and behavior. More concretely, it hinders our understanding of whether linguistic characteristics reflect changes in socio-economic relations and culture, whether they evolve independently, or even if they constraint and influence directly behavior. Roberts et al. (2015) demonstrate that cross-cultural correlations involving languages may be spurious once these language dependencies are accounted for and propose a series of empirical tests to help address these features of language.

In this paper, we consider a further methodological complication, which arises in when studying relationships between language structure and socioeconomic behavior: the potential for these associations to depend on the surrounding environment. That is, that language may co-evolve with institutional constraints.

An example is illustrative. Consider the correlation between future time reference (FTR) in language and the propensity to save as examined in Chen (2013) and Roberts et al (2015). Assume that a correlation between the two exists. That is, speakers of languages that exhibit a stronger FTR have a higher propensity to save. A task such as saving does not occur in vacuum. Rather, observed saving behaviors are dependent on the existence of a liquid and stable financial system regardless of individual preferences. Should such a system not exist (or should it be highly inefficient), a higher propensity to save may translate into higher investment in non-financial assets such as cattle – which may not be observable to the researcher. An empirical analysis of the relationship between languages' FTR and (financial) savings behavior could then falsely conclude that there is no relationship. Hence, it is possible for the estimated magnitude and significance of observed correlations between linguistic and socioeconomic behaviors to depend on the institutional environment within which individuals operate.

1.2. *Our Proposal*

We propose a new methodology to address this component of the identification problem: the application of the epidemiological approach. This approach takes its origin from epidemiologists who compare immigrants to natives in order to isolate the contribution of genetic factors from the influence of correlated environmental factors. The idea is to use immigrant populations to study the relationship between linguistic features and non-linguistic choices or individual outcomes that may evolve under a common institutional environment.

As an example, we study the labor market decisions of immigrants in the US. These migrants speak languages that exhibit varying levels of grammatical gender distinction. Theory suggests that we should expect women speaking languages that contain genders based on biological sex to participate less intensively in formal labor markets and instead to adopt more traditional gender roles such as work within the home (Hicks et al. 2015).

The empirical strategy we propose allows researchers to control for linguistic co-evolution, the institutional set up of the host country, and for unobservable cultural influences obtained in the origin country. This strategy draws its identification from migrants originating from the same country, but speaking languages with varying structure.

We empirically test this hypothesis on a sample of 675,000 immigrants in the U.S. from 156 countries and speaking 47 languages. We show that this approach is compatible to that of Roberts et al. (2015), which controls for language relatedness. In particular, allowing the intercept as well as the slope of the relationship to vary, as a function of language structure and behavior, is feasible. The rest of the paper is organized as follows. Section 2 presents the epidemiological approach. Section 3 presents an application. Section 4 concludes.

2. The Epidemiological Approach

Epidemiologists rely on the comparison of immigrant and native populations in order to isolate the contribution of genetic factors from the influence of environmental factors. This approach has been extensively applied within the economics research (Fernandez, 2007). Fundamentally, this approach implies studying variations across first and second-generation migrants to investigate the impact of their culture and disentangle its effect from the institutional and political environment of the host country. We propose that extending this approach to study language correlations with cultural and socio-economic outcomes is a fruitful avenue for future research.

Studying the behavior of migrants allows the researcher to compare individuals that evolve in a common institutional environment. As a result of the shared environment, incentives regarding their socioeconomic behavior are held constant across individuals. For linguistics specifically, it is possible to undertake a comparison of individuals who share the same country of origin, but speak different languages. Exploiting this source of heterogeneity allows researchers to control for a wide range of unobservable factors from both the home and the host country.

We provide an example to illustrate the set of strategies available to researchers when using this methodology. In particular, the next section presents an analysis of female labor participation among immigrants to the U.S and its correlation with sex-based grammatical distinctions in language. This application also highlights the richness of available census data concerning linguistic diversity both across and within countries of origin.

3. Application: Gender Marking and Female Labor Participation

3.1. *Data*

Our sample comes from the US in the American Community Survey 2007-2011 (ACS, 5% sample) and consists of migrants who report speaking a language other than English in their own home. This provides 675,000 observations from 156 countries and speaking 47 different languages. For each migrant, we have information about their labor market status, country of origin, language spoken in the home, and various other socioeconomic indicators such as income, education, marital status, level of English proficiency, and time since migration. Our outcome variable is a dummy variable equal to 1 if the individual is in the labor force and 0 otherwise. To quantify the presence of gender distinctions in language, we assign a dummy variable equal to 1 if the language has a gender system based on biological sex, and 0 if not. We obtain this information from the World Atlas of Language Structures (Dryer & Haspelmath 2013). While most languages around the world have a sex-based gender system, migrants to the US are from sufficiently diverse countries that the sample offers a wide variation in language structure. In particular, the average value of our linguistic dummy is 0.81 with a standard deviation of 0.39.

3.2. *Empirical Strategies available in the Epidemiological Approach*

A further key advantage of the epidemiological approach is that it allows the researcher to employ fixed effects strategies, which we illustrate in the following example. As a benchmark, we start by assessing the simple correlation between labor participation and sex-based grammatical distinctions in language. Because we are interested in the gap in participation between women speaking languages with different grammatical structure, we include an indicator variable equal to 1 if the individual is a woman, and an interaction between that indicator variable and our language variable.

The coefficient of interest is this interaction term: it measures the additional impact on labor participation of being a female migrant speaking a language with a sex-based gender system compared to being a female migrant speaking a language without a sex-based gender system. This effect is in addition to the estimated impact of being a female compared to being a male (captured by the female coefficient alone), and in addition to the direct impact of speaking sex-based language alone (regardless of gender).

Additionally, we control throughout the analysis for the individual's income and education levels, English proficiency, marital status and state of residence, as these factors may influence economic participation rates. In this setting, the interaction term compares women who have the same socioeconomic profile and live in the same state, but who speak a language that has a different grammatical structure. We use a simple OLS regression model. This simplifies the interpretation of the results, which are virtually the same as with a logit regression model. Column (1) in Table 1 presents these results.

A first strategy when using the epidemiological approach is to use country of origin fixed effects. This allows us to capture the role of norms of behavior related such as gender roles acquired prior to migration that are specific to an immigrant's country of origin. These fixed effects capture unobservable cultural influence on the migrants' behavior. Such a strategy allows us to effectively compare labor participation of women with similar socioeconomic background, living in the same US state, and coming from the same country, but speaking a language with a different grammatical structure. The results are presented in column (2) of Table 1.

Second, the epidemiological approach permits the use a set of fixed effects to address language relatedness. Indeed, languages may be related in two ways: a common ancestor (vertical dependence) and language contact (horizontal dependence) as discussed by Roberts et al. (2015). To account for the impact of language relatedness, we include a set of fixed effects for each language's family and linguistic area (Nichols et al. 2013). This allows the correlation between gender in language and labor market participation to have a different intercept across languages that pertain to a different language family and linguistic area. Column (3) of Table 1 includes language family and language area fixed effects.

Third, Roberts et al. (2015) argue that the strength of the correlation between a linguistic trait and a non-linguistic variable may itself be dependent on language relatedness. We can control for this dependence by including a set of interactions between each language's family and linguistic area, and the linguistic feature of interest itself. This allows the correlation to have a different slope across languages that pertain to a different language family and linguistic area. Column (4) of Table 1 presents the results.

A final strategy is to include fixed effects of the country of origin interacted by the subpopulation that the linguistic trait is supposed to affect. This approach depends on the particular nature of such a trait. In our example, the main assumption is that women speaking a language with a sex-based gender system are less likely to participate in the labor market, due to gender roles embedded in and/or caused by the language structure. If so, it should also be the case that these women behave differently than man in the country of origin. Therefore, an even more stringent strategy is to control for country of origin interacted with female fixed effects. With this strategy, we can control for characteristics of the origin country that are specific to women, thereby encapsulating the origin country characteristics that are most relevant to the question at hand. Column (5) of Table 1 presents the results.

3.3. *Results*

Table 1: Correlations between female labor participation and sex-based gender system

	Dependent variable: Female Labor Participation				
	(1)	(2)	(3)	(4)	(5)
Sex-Based	0.060***	0.037***	0.058***	0.086**	0.063
	[0.002]	[0.005]	[0.013]	[0.042]	[0.043]
Female	-0.173***	-0.167***	-0.166***	-0.166***	0.026
	[0.002]	[0.002]	[0.002]	[0.002]	[0.060]
Female x Sex-Based	-0.063***	-0.069***	-0.069***	-0.069***	-0.029***
	[0.003]	[0.003]	[0.003]	[0.003]	[0.010]
Socioeconomic Controls	Yes	Yes	Yes	Yes	Yes
Country of Origin FE	No	Yes	Yes	Yes	Yes
Language Fam. FE	No	No	Yes	Yes	Yes
Language Area FE	No	No	Yes	Yes	Yes
Language Fam FE x SB	No	No	No	Yes	Yes
Language Area FE x SB	No	No	No	Yes	Yes
Country x Female FE	No	No	No	No	Yes
Observations	674,476	669,739	669,720	669,720	669,720
R-squared	0.296	0.304	0.304	0.304	0.312

Notes: Estimates are survey weighted. Sample includes all immigrants aged 16 and above who report speaking a language other than English in the home. Additional controls include time since immigration, household income, household size, age, age squared, number of children, log wages, and indicators for survey wave, level of English language proficiency, marital status, student status, race and ethnicity, education level, and state of residence. Robust standard errors are in brackets. Source: Results calculated using the 2007-2011 ACS. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

The results in column (1) show that compared to male migrants speaking a language lacking a sex-based gender system, male migrants speaking a language that has a sex-based gender system are 6.0 percentage points more likely to be in the labor force. In comparison, similar female migrants are 6.3 percentage points less likely to be in the labor force. This discrepancy is in addition to the average gap in labor force participation between male and female migrants of 17.3 percentage points.

Controlling for the country of origin (column (2)) and language relatedness (columns (3) and (4)) alters the magnitude slightly but does not remove the significance of the results, suggesting that there is not much heterogeneity in the relationship between labor participation and language across origin countries, linguistic families and linguistic areas in this context. Finally, controlling for the interaction between country of origin and female reduces the magnitude of the coefficient of interest. Women speaking a language with a sex-based gender system are 2.9 percentage points less likely to be in the labor force than similar women speaking a language without a sex-based gender system. The coefficient on the interaction term is still significant at the 1% level.

4. Discussion

4.1. *Implications for the coevolution of language and behavior*

Our application and analysis has centered on presenting a set of simple yet powerful strategies that the epidemiological approach makes possible. Studying migrant populations has several additional advantages that researchers interested in the study of language evolution and its relation to non-linguistic phenomena may find useful. Our example demonstrates that the correlation between gender in language and female labor force participation is robust to controlling for country of origin and for language relatedness. Yet, the magnitude of the coefficient is substantially reduced when controlling for female specific country fixed effects. This implies that language must have evolved partly as a result of cultural change, but also that it may have directly constrained the evolution of norms, even if to a smaller extent.

4.2. *External versus Internal Validity of Different Approaches*

While they propose a series of series of empirical tests to be applied to cross-cultural data, Roberts et al. (2015) conclude that "experiments or case-studies would be more fruitful avenues for future research on this specific topic, rather than further large-scale cross-cultural correlational studies.'' We agree that there is much promise in experimental research. At the same time, while laboratory experiments arguably have a strong internal validity, they may not perform well in terms of external validity. The non-generalization of the results from lab experiments has been the subject of intensive research in economics (e.g., Stoop et al, 2012, Abeler & Nosenzo, 2014).

At the other extreme, cross-cultural studies perform well in terms of external validity by nature, but they are more likely to suffer from internal validity problems, as Roberts et al. (2015) makes clear. We thus place the epidemiological approach in the middle ground in terms of both external and internal validity. While the environment is not perfectly controlled by the researcher, migrants speaking different languages are observed within the same institutional environment. On the other hand, while findings are more generalizable than for lab or even framed field or natural experiments, migrants are a selected pool that may differ from the native populations.

While all approaches have advantages and disadvantages, the epidemiological approach provides researchers with an opportunity that should not be neglected. This is because (1) it provides a middle ground between cross-cultural correlations and experiments in terms of validity and (2) because it provides a rich new setting with which to test the relation between language evolution and non-linguistic phenomena.

Acknowledgments

All authors equally contributed. The feedback of two anonymous reviewers and of Sean Roberts has been crucial to improve the quality of the manuscript significantly. We are grateful to Ramon Ferrer-i-Cancho for helpful comments. This research has been conducted as part of the project Labex MME-DII (ANR11-LBX-0023-01).

References

Abeler, J., & Nosenzo D. (2014). Self-selection into laboratory experiments: pro-social motives versus monetary incentives. *Experimental Economics, 18(2),* 195-214.

Chen, M.K. (2013). The Effect of Language on Economic Behavior: Evidence from Savings Rates, Health Behaviors, and Retirement Assets. *American Economic Review, 103(2),* 690-731.

Dryer, M. S., & Haspelmath, M. (2013). WALS Online, Max Planck Institute for Evolutionary Anthropology, Leipzig.

Fernández, R. (2007). Alfred Marshall lecture women, work, and culture. *Journal of the European Economic Association, 5(2-3)*, 305–332.

Hicks, D. L., Santacreu-Vasut, E. & Shoham, A. (2015). Does mother tongue make for women's work? Linguistics, household labor, and gender identity. *Journal of Economic Behavior & Organization, 110*, 19–44.

Nichóls, J., Witzlack-Makarevich, A. & Bickel, B. (2013). The autotyp genealogy and geography database: 2013 release.

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

Roberts, S. & Winters, J. (2013), Linguistic diversity and traffic accidents: Lessons from statistical studies of cultural traits. *PloS one, 8(8),* e70902.

Stoop, J., Noussair, C. & van Soest, D. (2012). From the Lab to the Field : Cooperation among Fishermen. *Journal of Political Economy, 120(6),* 1027-1056.