Geographic Birthplace as a Predictor of Primary Language:

A Cross-National Observational Study

OMXUS Research Initiative
research@omxus.com

February 2026

Preprint — Not peer reviewed

Abstract

Background: Language acquisition is a fundamental aspect of human development, yet the relative contributions of environmental versus biological factors remain underexplored in large-scale empirical studies. This observational study examines whether geographic birthplace predicts primary language spoken across multiple nations.

Methods: We analysed national census data from eight countries ($N = 1{,}811{,}487{,}320$ individuals) spanning six continents. The primary outcome was concordance between country of residence and dominant national language spoken. Chi-square goodness-of-fit tests and effect size calculations (Cohen's h) were conducted.

Results: Across all nations examined, geographic residence demonstrated strong concordance with national language acquisition (range: 72.0%–96.9%). Effect sizes ranged from $h = 0.46$ to $h = 1.22$ (mean = 0.93; classified as "medium" to "large" by conventional standards). The observed pattern held regardless of the specific language examined.

Conclusions: Geographic environment appears to be an extraordinarily strong predictor of language acquisition. Supplementary analysis of international adoption studies and twin research confirms that environment, not genetics, determines which language an individual speaks. The implications for understanding human behavioural acquisition more broadly warrant further investigation.

Keywords: language acquisition; environmental factors; cross-national study; census data; nature versus nurture

Key Points

- Geographic environment predicts primary language with 72–97% accuracy across eight nations
- Effect sizes (mean Cohen's $h = 0.93$) exceed conventional "large" thresholds
- International adoption studies show 100% language replacement regardless of genetic ancestry
- Twin studies confirm genetics affects language *ability*, not *which language* is spoken
- Findings have implications for understanding environmental determination of complex human behaviours

1 Introduction

1.1 Background

Human language is among the most complex cognitive abilities exhibited by any species. The average adult possesses a productive vocabulary of approximately 20,000–35,000 words, applies

grammatical rules unconsciously in real-time, and processes speech at rates exceeding 150 words per minute (Brysbaert et al., 2016). Despite this complexity, healthy children across all cultures acquire language with remarkable consistency.

The question of *how* language is acquired has been debated extensively. Nativist perspectives emphasise innate language acquisition devices (Chomsky, 1965), while empiricist perspectives highlight environmental exposure and social learning (Tomasello, 2003). However, large-scale empirical studies examining the actual distribution of language outcomes across populations remain surprisingly limited.

1.2 Research Question

This study addresses a straightforward empirical question:

> *To what extent does geographic birthplace predict the primary language an individual speaks?*

We performed systematic cross-national analyses quantifying this relationship with standardised effect size metrics.

1.3 Hypotheses

H_0 (Null Hypothesis): Geographic birthplace is not associated with primary language spoken, and language acquisition occurs independently of geographic environment.
H_1 (Alternative Hypothesis): Geographic birthplace is associated with primary language spoken, and language acquisition is related to geographic environment.
We set our significance threshold at $\alpha = 0.05$.

2 Methods

2.1 Design

Cross-sectional observational study using publicly available national census data.

2.2 Data Sources

We identified national statistical agencies with publicly available census data on language spoken. Countries were selected based on: (1) availability of recent census data (2011–2022); (2) data published in or translatable to English; (3) inclusion of language variables; and (4) geographic and linguistic diversity.

2.3 Variables

- **Predictor:** Country of residence at time of census
- **Outcome:** Primary language spoken (national language vs. other)

2.4 Statistical Analysis

For each country, we calculated:

1. Proportion speaking the dominant national language
2. Chi-square goodness-of-fit test against null expectation (50%)
3. Effect size (Cohen's h)

Cohen's h is calculated as:

$$h = 2\arcsin\left(\sqrt{p_1}\right) - 2\arcsin\left(\sqrt{p_2}\right) \tag{1}$$

Effect size interpretation (Cohen, 1988): Small = 0.20; Medium = 0.50; Large = 0.80.

3 Results

3.1 Sample Characteristics

The combined sample included $N = 1{,}811{,}487{,}320$ individuals from eight countries (Table 1).

Table 1: Sample characteristics by country.

Country	**Year**	**Population**	**Source**
Australia	2021	25,422,788	ABS
Canada	2021	36,991,981	Statistics Canada
China	2020	1,411,778,724	NBS
France	2021	67,390,000	INSEE
Germany	2022	82,700,000	Destatis
Mexico	2020	126,014,024	INEGI
New Zealand	2018	4,699,755	Stats NZ
United Kingdom	2021	56,490,048	ONS
Total		**1,811,487,320**	

3.2 Language Concordance

Table 2 presents the proportion speaking the dominant national language in each country.

Table 2: Proportion speaking dominant national language.

Country	**Language(s)**	**% Speaking**
Australia	English	72.0%
Canada	English/French	96.9%
China	Chinese	92.0%
France	French	91.2%
Germany	German	81.0%
Mexico	Spanish	93.8%
New Zealand	English	95.4%
United Kingdom	English	91.1%

3.3 Statistical Tests

All chi-square tests were significant at $p < .001$ (Table 3).

Table 3: Chi-square and effect size results.

Country	**%**	χ^2	**df**	p	**Cohen's** h
Australia	72.0	4,921,852	1	$< .001$	0.46 (S)
Canada	96.9	32,547,173	1	$< .001$	1.22 (L)
China	92.0	996,151,068	1	$< .001$	1.00 (L)
France	91.2	45,756,193	1	$< .001$	0.97 (L)
Germany	81.0	31,789,880	1	$< .001$	0.67 (M)
Mexico	93.8	96,700,138	1	$< .001$	1.07 (L)
New Zealand	95.4	3,874,779	1	$< .001$	1.14 (L)
United Kingdom	91.1	38,169,422	1	$< .001$	0.96 (L)

Note: S = Small, M = Medium, L = Large effect size.

3.4 Effect Size Summary

Mean Cohen's $h = 0.93$ (SD = 0.24), range 0.46–1.22. This exceeds the conventional "large" threshold of 0.80 by a factor of 1.2.

4 Discussion

4.1 Summary of Findings

Across eight nations representing 1.8 billion individuals:

1. Proportions speaking national language: 72.0%–96.9%
2. All p-values $< .001$
3. Mean effect size exceeds "large" threshold
4. Pattern consistent across languages and regions

We conclude, with considerable statistical confidence, that people tend to speak the language of wherever they grew up.

4.2 Supplementary Evidence

4.2.1 International Adoption Studies

Korean children adopted by Swedish families speak Swedish. Chinese children adopted by American families speak English. No study has ever found spontaneous birth-language acquisition without environmental exposure (Pallier et al., 2003; Hyltenstam et al., 2009).

Effect: 100% language replacement.

A Korean child adopted at age two by a family in Stockholm does not spontaneously produce Korean. The child produces Swedish. The genetic material is entirely Korean. The language is entirely Swedish. There is no ambiguity in this result.

4.2.2 Twin Studies

The Minnesota Study of Twins Reared Apart (Bouchard et al., 1990) found that genetics affects language *ability* (heritability estimates 25–70%), but *which language is spoken* shows 0% heritability.

The "Jim Twins"—identical twins separated at birth—both spoke English. This was because both were raised in Ohio, not because chromosome 7 encodes a preference for English.

4.2.3 Generational Language Shift

Hispanic immigrants to the United States show complete language shift within three generations (Portes & Rumbaut, 2001):

- 1st generation: 85% Spanish-dominant
- 2nd generation: balanced bilingualism
- 3rd generation: 92% English-dominant

The genetics did not change. The environment did.

4.3 The Obvious and Its Implications

The finding that "people speak the language of their environment" is, on its face, obvious. No serious person disputes it. It requires no p-values, no effect sizes, and certainly not a sample of 1.8 billion.

We performed this analysis anyway because of a curious asymmetry in how we reason about human behaviour.

When a child in Sydney speaks English, we attribute this to environment without hesitation. We do not speculate about genetic predispositions toward English. We do not suggest the child made a rational cost-benefit analysis and chose English over Mandarin. We do not convene a court to determine whether speaking English was a freely willed decision for which the child bears moral responsibility. We simply note that children learn the language they are exposed to.

Yet when a child raised in an environment saturated with aggression, instability, and trauma later exhibits aggressive, unstable, or traumatic behaviour, we reach for entirely different explanatory frameworks. We invoke rational choice. We invoke moral failing. We invoke "bad character." We build an entire justice system on the premise that this behaviour was *chosen* in a way that language apparently was not.

4.4 Aggression, Ability, and Expression

The twin studies point to an important distinction. Genetics influences language *ability*—some people are more verbally fluent than others, across all languages. But genetics does not determine *which language* is spoken. The capacity is innate; the expression is environmental.

We propose the same distinction applies to aggression. The capacity for intense emotional and behavioural drive—call it aggression, energy, intensity—likely has a heritable component. Some individuals run hotter than others. This is the "ability" analogue.

But how that intensity *manifests*—whether as violence, creativity, athletic drive, obsessive problem-solving, or writing satirical academic papers—is shaped by environment. The child with high intensity raised among artists may become an artist. The same child raised among fighters may become a fighter. The drive is the same. The output is environmentally determined.

This is not a claim that environment explains everything, nor a dissolution of responsibility. People are responsible for their behaviour, including learned behaviour—just as a fluent English speaker is responsible for what they say in English, even though they didn't choose to speak it. Understanding the origin of a pattern does not excuse the pattern. It does, however, suggest that punishment aimed at the *choice to offend* may be misdirected when no such choice occurred, and that intervention aimed at *reshaping the behavioural distribution* may be more effective.

4.5 Limitations

1. Census data measures language spoken, not language of first acquisition; some individuals may have shifted languages through migration.

2. The 50% null expectation is deliberately conservative (and absurd). A more realistic null would account for the global distribution of languages, but this would only strengthen the results.

3. The analogy from language to other behaviours is suggestive, not proven. Language may be uniquely environmentally determined because it is arbitrary—no language is biologically "correct." Behaviours with survival value may show more genetic variance.

4. This study cannot establish causation. It is possible that an unknown confound other than geographic environment explains why Australians speak English.

5 Conclusions

Geographic environment strongly predicts language acquisition (mean $h = 0.93$, all $p < .001$). This holds across eight nations, multiple languages, and 1.8 billion individuals.

Language is the most complex learned behaviour humans exhibit. It is acquired without explicit instruction, produces infinite novel outputs from finite rules, and is performed unconsciously at extraordinary speed. And it is 100% determined by environment.

If the most complex behaviour is environmentally determined, the default assumption for other behaviours—including those we classify as criminal—might reasonably be environmental as well. Not entirely. Not without nuance. But as a *default*, before we reach for character, choice, and punishment.

The data, at least, are clear. People speak the language they heard.

They may also feel the feelings they were shown.

References

[1] T. J. Bouchard, D. T. Lykken, M. McGue, N. L. Segal, and A. Tellegen, "Sources of human psychological differences: The Minnesota Study of Twins Reared Apart," *Science*, vol. 250, no. 4978, pp. 223–228, 1990.

[2] M. Brysbaert, M. Stevens, P. Mandera, and E. Keuleers, "How many words do we know? Practical estimates of vocabulary size," *Frontiers in Psychology*, vol. 7, p. 1116, 2016.

[3] N. Chomsky, *Aspects of the Theory of Syntax*. MIT Press, 1965.

[4] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Lawrence Erlbaum, 1988.

[5] K. Hyltenstam et al., "Dominant-language replacement: The case of international adoptees," *Bilingualism: Language and Cognition*, vol. 12, no. 2, pp. 121–140, 2009.

[6] C. Pallier et al., "Brain imaging of language plasticity in adopted adults," *Cerebral Cortex*, vol. 13, no. 2, pp. 155–161, 2003.

[7] A. Portes and R. G. Rumbaut, *Legacies: The Story of the Immigrant Second Generation*. University of California Press, 2001.

[8] M. Tomasello, *Constructing a Language*. Harvard University Press, 2003.