AI-powered analysis of language diversity using U.S. Census/ACS data

Tomonori Nagano, LaGuardia Community College (CUNY) | tnagano@lagcc.cuny.edu

Abstract

U.S. Census and ACS data provide a long historical record of language use in the U.S., but longitudinal analysis is difficult because language questions, categories, and coding systems have changed over time. This project explores how AI-assisted coding workflows can help harmonize those inconsistencies and support large-scale analysis of language diversity.

Language data in the Census/ACS

- The U.S. Census started collecting information about language use in 1890, but its approach has not been consistent. There are three main phases of language items in the census (Stevens, 1999).

Phase	Language item
1890-1910 English Proficiency	**1890**: "Able to speak English. If not, the language or dialect spoken." **1910**: "Whether able to speak English; or, if not, give language spoken."
1920-1970 Mother Tongue	**1920**: "If of foreign birth, give the place of birth and, in addition, the mother tongue (mother tongue is the language of customary speech before coming to the United States)." **1940**: English proficiency question was dropped **1950**: No question about language at all **1960**: Mother tongue question was reinstated
1980-2025 Home Language	**1980 (to present)**: "Does this person speak a language other than English at home?" → "What is this language?" → "How well does this person speak English?"

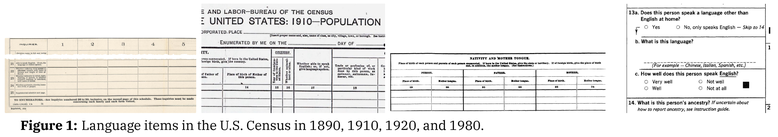

Figure 1: Language items in the U.S. Census in 1890, 1910, 1920, and 1980.

Issues in the language data in the Census

- Language items in the Census have never been as consistent as one might think. How language is conceptualized, coded, and categorized depends on the sociopolitical reality of each time period.
- **Language is not consistently conceptualized across years**
  - Language as a barrier to assimilation (1890-1910), as a genealogical attribute (1920-1970), and as a means of communication (1980-).
- **Language is not consistently coded across years**
  - Open-ended responses to the language item must be coded.
  - 42 languages in 1910, 63 languages in 1920, 64-382 languages between 1980-2005, and the adoption of ISO-639-3 (between 1,333 and 7,000 languages) in 2016 (U.S. Census, 1910; and others)

Issues in the language data (cont.)

- **Language is not consistently categorized**
  - There is no consensus on how "language" should be defined.
  - e.g., Chinese (vs. Mandarin, Cantonese, etc.) and Arabic (MSA and regional Arabic dialects)
  - e.g., Other languages ("Other languages of Asia" = Burmese, Karen, Turkish, Uzbek, etc.)

This Study

- Despite substantial inconsistencies, some studies attempt to employ census language data in their research.
- **Nagano (2015)**: Demographics of heritage language (HL) speakers in the U.S. between 1980-2010
  - The number of HL speakers⚠ grew much faster than the U.S. population from 1980⚠-2010 (26.98% vs. 10.88% per decade)
  - Spanish and Chinese⚠ remain the two largest HL groups and have grown rapidly over 30 years, while newer HLs (e.g., Arabic⚠, Hindi⚠, Dravidian⚠, Vietnamese, Russian⚠, Amharic, Tibetan) have also seen substantial gains.
  - Languages like French⚠, German⚠, Italian, Greek, Yiddish, and Dutch have declined.
- There were numerous challenges (⚠) in this study, and it took over 6 months to complete the first manuscript.

Research Question

Can a recent AI model (codex with GPT-5.4 for this study) resolve these challenges in the Census data? If so, how long does it take to replicate my study (human standard: 6 months)?

Methods

- AI tool: `Codex CLI` v.0.122.0 with GPT-5.4 (medium reasoning)
- The same data set (Census data from IPUMS [Ruggles et al., 2018]) that I used for my 2015 study
- A copy of my article for the 2015 study

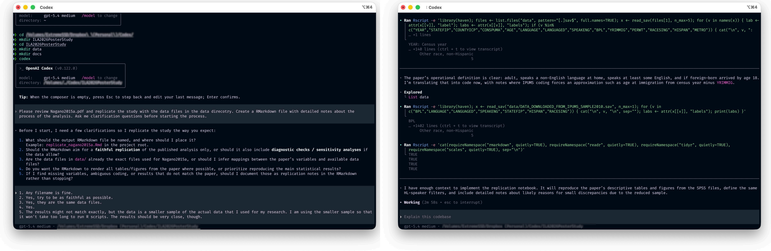

Figure 2: Open AI's codex on Terminal

Results

- Human baseline: 6 months (with countless sleepless nights and episodes of depression after working on the same data for a long time)
- AI: 3 minutes 7 seconds
- Accuracy: Mostly identical to the original study. See https://github.com/tnagano22/ILA2026UseAICensus for the AI response and the generated R script.

Follow-up Analysis

- It is possible that AI did so well because the analysis had already been done and published. AI could simply refer to it and replicate it without much effort.

Follow-up Question

Can AI do equally well with a novel problem with new analysis?

I asked whether I could include data from 1960 and 1970, which I wasn't able to accomplish in my original study.

- After 6 minutes 49 seconds, the AI agent responded with a reasonable approach.

AI's Response

You should treat 1960/1970 as a separate, pre-1980 proxy analysis, not as directly comparable additions to the current adult-HL series. The current notebook defines adult HL speakers with LANGUAGE "spoken at home," SPEAKENG, and YRIMMIG. That definition cannot be applied to 1960/1970 U.S. data because IPUMS LANGUAGE is unavailable for U.S. 1960 and 1970; it starts again in 1980 for the modern home-language question. ....

- I asked the AI to complete the suggested analysis. It took 1 minute 15 seconds to complete it. See https://github.com/tnagano22/ILA2026UseAICensus for the full response.

Conclusion

AI was able to reproduce a census-based language study with a substantially reduced time. It also successfully addressed a novel problem (i.e., pre-1980 data) with a reasonable approach. AI can be a valuable tool to the research community when it's used with human oversight and expert validation. The implications of this new technology are still not clear and how it will affect our research methodology at large remains to be seen.

Reference

- Nagano, T. (2015). Demographics of Adult Heritage Language Speakers in the United States: Differences by Region and Language and their Implications. *The Modern Language Journal*, 99(4), 771-792

- 4 example languages in 1890
  - German; Portuguese; Canadian French; Pennsylvania Dutch
- 42 “principal foreign languages spoken in the United States” in 1910 (United States Census Bureau, 1910)
  - Albanian; Armenian; Basque; Bohemian; Breton; Bulgarian; Chinese; Danish; Dutch; Finnish; Flemish; French; German; Greek; Gypsy; Irish; Italian; Japanese; Lappish; Lettish; Little Russian; Lithuanian; Magyar; Moravian; Norwegian; Polish; Portuguese; Rhaeto-Romanish (including Ladin and Friulan); Roumanian; Russian; Ruthenian; Scotch; Serbian or Croatian (Including Bosnian Daimatian Bohemian Herzegovinian and Montenegrin); Slovak; Slovenian; Spanish; Swedish; Syrian; Turkish; Welsh; Wendish; Yiddish
- 63 “principal foreign languages” in 1920 (United States Census Bureau, 1920)
  - Albanian; Arabian; Armenian; Basque; Bohemian (Czech); Breton; Bulgarian; Chinese; Croatian; Dalmatian; Danish; Dutch; English; Esthonian; Finnish; Flemish ; French ; Frisian; Friulan; Gaelic; Georgian; German; Great Russian; Greek; Gypsy; Hebrew; Hindu; Icelandic; Irish; Italian; Japanese; Korean; Kurdish; Lappish; Lettish; Lithuanian; Little Russian; Macedonian; Magyar; Montenegrin; Moravian (Czech); Norwegian; Persian; Polish; Portuguese; Romansh; Rumanian; Russian; Ruthenian; Scotch; Serbian; Slovak; Slovene; Spanish; Swedish; Syrian; Turkish; Ukrainian; Walloon; Welsh; Wendish; White Russian; Yiddish
- Adoption of *Classification and index of the world’s languages* (Voegelin & Voegelin, 1977)
  - 4 group classification:
    - Spanish, Other Indo-European languages , Asian and Pacific Island languages, All other languages
  - 42 group classification:
    - Amharic, Somali, or other Afro-Asiatic languages; Arabic; Armenian; Bengali; Chinese (incl. Mandarin, Cantonese); French (incl. Cajun); German; Greek; Gujarati; Haitian; Hebrew; Hindi; Hmong; Italian; Japanese; Khmer; Korean; languages; Ilocano, Samoan, Hawaiian, or other Austronesian languages; Malayalam, Kannada, or other Dravidian; Navajo; Nepali, Marathi, or other Indic languages; Other and unspecified languages; Other Indo-European languages; Other languages of Asia; Other Native languages of North America; Persian (incl. Farsi, Dari); Polish; Portuguese; Punjabi; Russian; Serbo-Croatian; Spanish; Swahili or other languages of Central, Eastern, and Southern Africa; Tagalog (incl. Filipino); Tamil; Telugu; Thai, Lao, or other Tai-Kadai languages; Ukrainian or other Slavic languages; Urdu; Vietnamese; Yiddish, Pennsylvania Dutch or other West Germanic languages; Yoruba, Twi, Igbo, or other languages of Western Africa;
  - 64 languages in IPUMS in 2010
    - African (not specified); Albanian; Aleut Eskimo; American Indian (not specified); Amharic, Ethiopian, etc.; Arabic; Armenian; Athapascan; Burmese, Lisu, Lolo; Celtic; Chinese; Czech; Danish; Dravidian; Dutch; Filipino Tagalog; Finnish; French; German; Greek; Hamitic; Hawaiian; Hebrew (Israeli); Hindi and related; Indonesian; Iroquoian; Italian; Japanese; Keres; Korean; Lithuanian; Magyar; Hungarian; Micronesian; Polynesian; Native; Navajo Navaho; Near East Arabic dialects; Norwegian; Persian; Iranian; Farsi; Polish; Portuguese; Rumanian; Russian; Serbo-Croatian, etc.; Siouan languages; Slovak; Spanish; Sub-Saharan Africa; Swedish; Thai, Siamese, Lao; Tibetan; Turkish; Ukrainian, etc.; Vietnamese; Yiddish, Jewish; Zuni; Other Balto-Slavic; Other East Southeast (ES) Asia; Other Malayan; Other or not reported; Other Persian dialects
- Adoption of International Organization for Standardization’s ISO-639-3 (Gambino, 2017)
  - The ISO-639 shows over 7,000 languages (https://www.iso.org/iso-639-language-code)
  - 1,333 languages from the ISO-639 were adopted for the U.S. Census coding
  - In practice, only 384 languages are reported.

References

Dietrich, S., & Hernandez, E. (2022). *American Community Survey report: Language use in the United States 2019* (Tech. Rep.). Washington, D.C.: US Census Bureau.

Gambino, C. (2017). *Inside the american community survey: 2016 language data overhaul.* Random Samplings (blog). Retrieved from https://www.census.gov/newsroom/blogs/random-samplings/2017/09/inside_the_american.html

Nagano, T. (2015). Demographics of adult heritage language speakers in the United States: Differences by region and language and their implications. *The Modern Language Journal*, *99*(4), 771-792. doi: 10.1111/modl.12272

OpenAI. (2025). *Introducing codex* (Tech. Rep.). New York, NY: Author. https://openai.com/index/introducing-codex/.

Ruggles, S., Flood, S., Goeken, R., Grover, J., Meyer, E., Pacas, J., & Sobek, M. (2018). *IPUMS USA: Version 8.0 [dataset].* Minneapolis, MN: IPUMS. (https://doi.org/10.18128/D010.V8.0)

Stevens, G. (1999). A century of us censuses and the language characteristics of immigrants. *Demography*, *36*(3), 387-397.

United States Census Bureau. (1910). *1910 census instructions to enumerators* (Tech. Rep.). Washington, D.C.: United States Census Bureau. (https://www.census.gov/content/dam/Census/programs-surveys/decennial/technical-documentation/questionnaires/1910-instructions.pdf)

United States Census Bureau. (1920). *1920 census instructions to enumerators* (Tech. Rep.). Washington, D.C.: United States Census Bureau. (https://www.census.gov/programs-surveys/decennial-census/technical-documentation/questionnaires/1920/1920-instructions.html)

US Census Bureau. (2017). *American community survey language code list.* Retrieved from https://www2.census.gov/programs-surveys/demo/about/language-use/primary_language_list.pdf

US Census Bureau. (2019). *About language use in the U.S. population.* Retrieved February 23, 2020, from https://www.census.gov/topics/population/language-use/about.html

Voegelin, C. F., & Voegelin, F. M. (1977). *Classification and index of the world’s languages*. New York, NY: Elsevier.