Appendix A

Minority Language Coding Rules

Coding rules for what counts as a minority language

Last updated October 10, 2020

Coding Rules

Theory suggests that minority languages are either linguistically non-dominant (with <50% of the population speaking the language) OR politically non-dominant (with its speakers not having equitable access to political representation in the country). I try to capture these two dimensions in my coding rules. My two main data sources are Jacques Leclerc's database of language policy *L'aménagement linguistique dans le monde* ("Language Planning around the World," http://www.axl.cefan.ulaval.ca/index.html) and the 2019 version of the Ethnic Power Relations dataset (https://icr.ethz.ch/data/epr/).

I use the EPR dataset as my primary source, since EPR tracks both the numerical and political dominance of *ethnolinguistic* groups in all countries of the world. For ease and consistency in data collection, I make the assumption that a language is always tied to an ethnolinguistic group, even though this may or may not be true for different languages.[1] I use Leclerc's data if EPR's data is insufficient for a given country, and furthermore use resources such as the *Encyclopedia of World Cultures* and *Encyclopedia Britannica* as an extra layer of validation in determining which languages are minority languages.

1. This assumption seems to be valid; see May, "Rearticulating the Case for Minority Language Rights" for a good defense of language being an important marker of identity and cultural and political dimensions for ethnolinguistic groups at points in time. This holds even if how ethnicity is constructed is always in flux.

In some cases, distinguishing between a "language" and a "dialect" can be extremely tricky. For my codings, a "dialect" under question is a separate language if it is not significantly "mutually intelligible" with the "main" language, OR if the speakers of a certain "dialect" perceive themselves as having a distinct linguistic identity separate from the speakers of the main "language."[2] This is consistent with the criteria for one of my main sources (Ethnologue), which points out that this difference in perceived identity often means that the dialect has a long-standing standard form and body of literature distinct from the main language.[3] This is also consistent with previous work in sociolinguistics, for instance Haugen's seminal 1966 article on distinguishing language and dialect.[4]

This leads to the following coding rules:

1. A language spoken in any given country is considered as a **"minority language"** *if*:
   (a) It both corresponds to an ethnolinguistic group that makes up <50% of the population (as listed by EPR) AND appears on Leclerc's database as a language of a "minority group" (listed in the header).
   (b) Or, if a country has a "dominant" or "monopoly" group as listed by EPR, ALL other languages are coded as minority languages regardless of the size of their speaker base.
   (c) The ethnolinguistic group associated with the language has never been considered a "dominant" or "monopoly" group in that country, throughout the entire period of the country's history that EPR codes for.
   (d) In both cases, the language must not an immigrant language, as determined by research on Leclerc's database.

2. For instance, I code as separate the different varieties of Serbo-Croatian (Serbian, Croatian, etc.). Even though these varieties are mutually intelligible, the speakers perceive linguistic differences between their groups. These perceived differences must be *linguistic* for languages to be considered separate: e.g., the Hutu and Tutsi in Rwanda both speak Kinyarwanda, but the divisions between them are more based on ethnicity than language. As another example, I treat Hakka Chinese and Chinese as different languages because they are not mutually intelligible.

3. Ethnologue: Languages of the World, "The Problem of Language Identification," Ethnologue, 2021, https://www.ethnologue.com/about/problem-language-identification.

4. Einar Haugen, "Dialect, Language, Nation," *American Anthropologist* 68, no. 4 (1966): 922–935.

2. Only if EPR is inadequate in determining the linguistic groups in a country (eg, if it only lists groups divided by religious cleavages), I turn to Leclerc to determine the minority languages in a country. Here, I code for every non-immigrant language that Leclerc lists as a minority language in the country (under his header "*Groupes minoritaires*.")

3. If the country only contains one ethnolinguistic group as determined by EPR, there are two cases to consider:

   (a) The country is ethnically and linguistically homogeneous. In this case I determine that there are no minority languages in the country and all relevant columns in my Excel file are coded as n/a.

   (b) Ethnicity is not "politicized" in that country, but the country is ethnically and/or linguistically heterogeneous. If this happens, I code the languages under the "*Groupes minoritaires*" section of the Leclerc database as before.

   (c) To determine which of the previous two is the case, I turn to the description from the EPR documentation (https://growup.ethz.ch/atlas).

4. If EPR cites a mix of cleavages in a country (eg, linguistic and social groups such as mentioning Koreans in Japan but also the *Burakumin*), I code for all possible linguistic divisions and rely on Leclerc to code for any minority languages missing from EPR.