Domestic and International Common Language (DICL) Database

Technical Note

Tamara Gurevich[*], Peter R. Herman[*], Farid Toubal[†], Yoto V. Yotov[‡]

February 11, 2021

Abstract

This document provides a technical description of common language variables in the Domestic and International Common Language Database. To download data and view the associated working paper, see the dataset page at `https://www.usitc.gov/data/dicl.htm`.

[*]U.S. International Trade Commission, Office of Economics. The views expressed in this paper are strictly those of the authors and do not represent the opinions of the United States International Trade Commission or any of its Commissioners.

[†]University of Paris-Dauphine – PSL, LEDa, CEPII, CESifo and CEPR.

[‡]Drexel University.

About: The database contains index measures of linguistic similarity between 242 countries both domestically and internationally. The domestic measures capture linguistic similarities present among populations within a single country while the international indexes capture language similarities between two different countries. The indexes reflect three aspects of language: common official languages, common native languages, and linguistic proximity across languages. This database has many uses, such as in in models of bilateral flows—including FDI, migration, and international trade—as well as in regional or country level analyses.

Recommended citation:
Gurevich, Tamara, Peter Herman, Farid Toubal, and Yoto Yotov, (2021), "One Nation, One Language? Domestic Language Diversity, Trade and Welfare", USITC Economics Working Paper 2021–01–B.

File structure and identification: the file contains six variables—two country identifiers and four measures of linguistic commonality. Each of the 58,564 observations is identified by a pair of ISO codes (ISO3 and ISO3_2).

Source Data: The DICL indexes are derived from from the 21^{st} edition of Ethnologue. For additional information, see:

Simons, Gary F. and Charles D. Fennig (eds.), (2018), "Ethnologue: Languages of the World, Twenty-first edition," Dallas, Texas: SIL International. Online: `http://www.ethnologue.com`.

Linguistic similarity variables: In what follows, country-pairs are denoted by subscripts i and j. Each instance where $i \neq j$ describes an international linguistic relationship (i.e., a relationship between populations in two different countries). Each instance where $i = j$ describes a domestic linguistic relationship (i.e., a relationship between the population within a single country).

The DICL database contains four measures of linguistic commonality.

- $\boldsymbol{COL_{ij}}$: a binary measure of common official language. For $i \neq j$, the variable equals 1 if two countries share a *de jure* or *de facto* regional or national official language, as defined in Ethnologue. It is 0 otherwise. For $i = j$, the variable is defined to be 1.
- $\boldsymbol{CNL_{ij}}$: a continuous index in $[0, 1]$ reflecting the likelihood that two people selected at random from populations i and j will speak the same native language. For each common native language k spoken in a country i or j, the share of native speakers of that language is denoted l_i^k or l_j^k, respectively, and K denotes the set of languages spoken in both countries. The index is defined as follows:

$$CNL_{ij} = \sum_{k \in K} (l_i^k \times l_j^k) \quad \forall i \neq j,$$

$$CNL_{ii} = \sum_{k \in K} (l_i^k)^2.$$

- $\boldsymbol{LP_{ij}}$: a continuous index in $[0, 1]$ measuring how similar the languages spoken by two populations i and j are. It is computed using two components: a measure of linguistic proximity, described below, and the populations of speakers in each country.

The first component, denoted $PROX_{mn}$, measures the linguistic proximity between two languages m and n. Each language can be described by its place within a language family using the linguistic notion of language trees. Linguistic trees begin with a root proto-language and split sequentially into multiple branches as language groups diverge from their ancestral proto-language. $PROX_{mn}$ is defined as the number of common branches the two languages share, starting from the proto-language, divided by the average length of branches that terminate in each language. For example, French and Italian both stem from the proto-language Indo-European. French is 7 branches removed from Indo-European while Italian 6 branches removed. The two languages belong to 4 common levels of language families before diverging to different branches. Therefore, $PROX_{French,Italian} = 5/(0.5*(7+6)) = 0.78$. Put more intuitively, French and Italian share an average 78 percent of their respective language trees. In general,

$$PROX_{mn} = \frac{b_{mn}}{0.5(b_m + b_n).}$$

where b_{mn} is the number of common branches between languages m and n, and b_m and b_n are the branch lengths of languages m and n, respectively. For two languages that do not originate in the same proto-language, $PROX_{mn} = 0$.

The second component accounts for the fact that most countries have multiple native languages with different numbers of speakers. The complete LP_{ij} measure is constructed as a population weighted aggregate of the $PROX_{mn}$ measures:

$$LP_{ij} = \sum_{m \in K} \sum_{n \in K} (l_i^m \times l_j^n) * PROX_{mn}.$$

As before, l_q^p denotes the share of native speakers of language p in country q and K denotes the set of all languages spoken in both countries.

- $\boldsymbol{CL_{ij}}$: a continuous index in $[0, 1]$ computed as the simple average of COL_{ij}, CNL_{ij}, and LP_{ij}.