The Domestic and International Common Language Database

Tamara Gurevich, Peter R. Herman, Farid Toubal, Y. Yotov

Drexel Economics
Working Paper Series

School of Economics
LeBow College of Business
Drexel University
WP 2024-02

The Domestic and International Common Language Database

Tamara Gurevich
U.S International Trade Commission

Peter R. Herman
U.S International Trade Commission

Farid Toubal
University of Paris-Dauphine

Yoto V. Yoto
Drexel University

Abstract

We construct a new global database on common language. The data cover 242 countries and territories and are based on information about the speakers of 6,674 languages. We provide 8 bilateral measures reflecting different dimensions of linguistic connections, including common official languages, common native and acquired languages, and linguistic proximity across different languages. A key novelty of the dataset is that it includes consistently defined information on linguistic relationships not only between different countries but within the administrative borders of countries as well.

JEL classification: D60; F14; F19; C54; Z13
Keywords: Common language; Ethno-linguistic diversity; Language Data

The Domestic and International Common Language (DICL) Database

Tamara Gurevich Peter R. Herman Farid Toubal Yoto V. Yotov

March 2024

Abstract We construct a new global database on common language. The data cover 242 countries and territories and are based on information about the speakers of 6,674 languages. We provide 8 bilateral measures re-flecting different dimensions of linguistic connections, including common official languages, common native and acquired languages, and linguistic proximity across different languages. A key novelty of the dataset is that it includes consistently defined information on linguistic relationships not only between different countries but within the administrative borders of countries as well.

Tamara Gurevich
Research Division, Office of Economics
U.S. International Trade Commission
tamara.gurevich@usitc.gov

Farid Toubal
University of Paris Dauphine – PSL, CEPII and CEPR
farid.toubal@dauphine.psl.eu

Peter R. Herman
Research Division, Office of Economics
U.S. International Trade Commission
peter.herman@usitc.gov

Yoto V. Yotov
Drexel University
School of Economics
yvy23@drexel.edu

Background & Summary

The Domestic and International Common Language (DICL) Database provides new data on linguistic relationships within and between countries. Using extensive data on 6,674 languages in 242 countries and territories, we have constructed a collection of 8 indices measuring different dimensions of linguistic connections. These indices provide information about common official languages, speakers of common native and acquired (secondary) languages, and linguistic proximity between different languages. Each index provides measures of the linguistic connections between the populations in different countries (international) as well as within each country (domestic).

The 8 indices are each able to capture different dimensions of linguistic relationships.[1] We construct two measures of common official language based on a liberal (COL) and a more restrictive definition of official languages (COR). COL and COR capture languages that hold an official status within each country and therefore represent commonalities in the typical languages of business, education, and legal institutions, for example. The index of common native language (CNL) captures the languages spoken as a mother tongue by populations and can reflect both an ability to communicate as well as a wide range of other cultural similarities between populations, such as ethnic ties. The index of common acquired languages (CAL) captures connections between non-native speakers and can be useful in identifying communication through common regional or global languages. The index of common spoken language (CSL) combines native and acquired language speakers into a single measure capturing all linguistic ties. Finally, the linguistic proximity indices capture populations speaking different but closely related languages that may be mutually or partially intelligible. We construct 3 proximity indices—calculated for native language speakers (LPN), acquired language speakers (LPA), and all speakers (LPS). They allow for the consideration of communication between different languages and across language families.

The data provided in the DICL database expands upon the existing common language information used extensively throughout the literature, such as the data provided by Gurevich and Herman (2018), Melitz and Toubal (2014) and Mayer and Zignago (2011). The DICL data are based on a wider collection of languages and more detailed information on the status of these languages throughout the world. The richness of the underlying data allows for the creation of multiple new continuous measures, which are better able to capture the diversity in linguistic ties between populations than the indicator variables present in most other databases.[2] The database is also the first to provide *consistently defined* measures of both international and

[1] A detailed description of the variables included in the database is presented in section 1.2.

[2] The primary exception is the data of Melitz and Toubal (2014), which also includes continuous measures but is based on more limited information about language use throughout the world. Melitz and Toubal (2014) do not provide domestic measures.

domestic language diversity.[3] Ultimately, the DICL database is a more comprehensive and diverse source of information on country-level and bilateral linguistic connections then any existing public database of its type.

The DICL indices present a detailed picture of the extensive and diverse linguistic relationships around the world. Figure 1 depicts 8 heatmaps of the values of each index structured as a matrix with columns and rows representing each of the 242 countries. Darker shades denote closer ties while lighter shades denote weaker ties. As is clear from the heatmaps, the two indices reflecting official languages (COL and COR) present a fairly balanced mix of relationships. Meanwhile, the remaining 6 indices of suggest that close linguistic ties are relatively uncommon but vary considerably in intensity. As a further illustration of this rich variation, Figure 2 depicts the CNL index as a network in which the higher the CNL value between two countries (nodes), the larger and darker the link between them. This visualization in particular highlights the diverse extent to which countries tend to be connected to others linguisticly. Some are moderately connected to many other countries, as is often the case with populations speaking major global languages. Some exhibit strong connections with a smaller cluster of other countries, often representing groups of regional neighbors or colonial ties. Others oriented along the periphery of the network exhibit very few strong connections, reflecting their linguistic remoteness compared to much of the rest of the world.

Measures of linguistic commonality or fractionalization have been commonly used in empirical studies to identify the direct and indirect impacts of linguistic connections between populations. In some cases, they have been used to examine direct implications of language, such as the ability to communicate. In others, they have been used as proxies for otherwise difficult to observe social connections, such as cultural affinities and identity. Similar measures have been used extensively in anthropology and linguistics (c.f. Greenberg (1956, 1987),), sociology (c.f. Lieberson (1964)), political sciences (c.f. Fearon and Laitin (1999); Fearon (2003)) and economics (c.f. Ginsburgh and Weber (2020); Ginsburgh et al. (2016); Gazzola and Wickström (2016) for a review of the literature).

1 Methods

1.1 Description of the Source Data

The DICL indices are built using a single and unique source of data, which is the *Ethnologue: Languages of the World* dataset (21st edition) (Simons and Fennig, 2018). This is the most comprehensive linguistic database available. It describes in detail more than seven thousand languages in 242 countries and territories. We make

[3] Desmet et al. (2009) provide measures of within-country language diversity that draw on many of the same linguistic concepts as the DICL database but they do not provide any corresponding international measures.

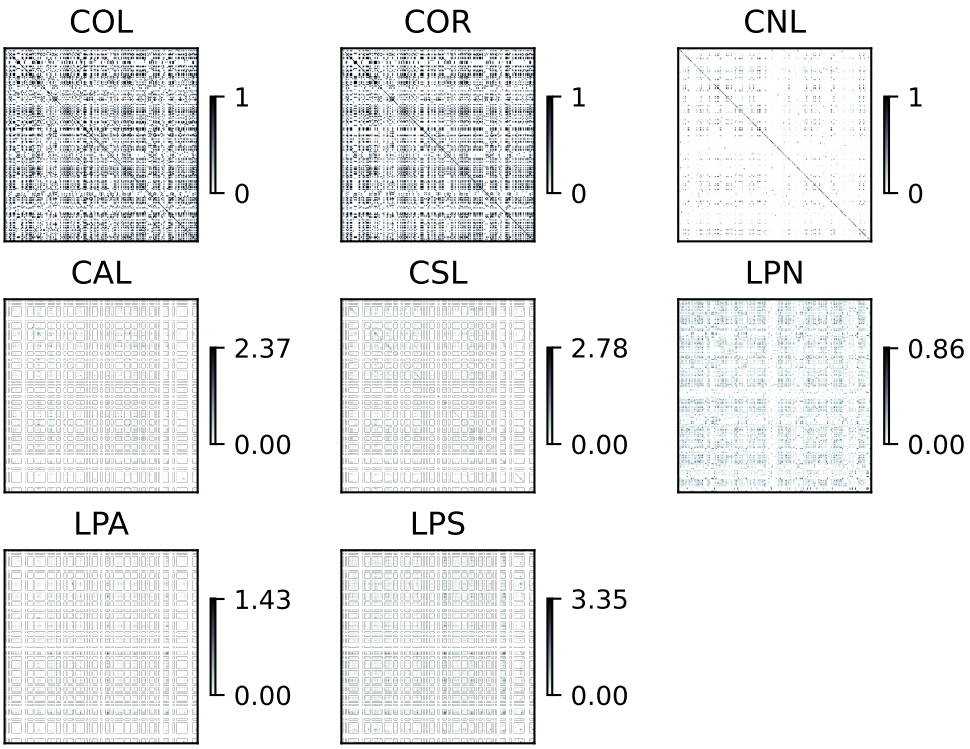

Figure 1: Heatmaps of the DICL language indices

use of 6,674 languages for which data on native speakers (L1-Users) is reported by Ethnologue. Ethnologue also reports the number of acquired language speakers (L2-Users) for 649 languages. As mentioned in Lewis (2005), the distinction between languages and dialects is very difficult to draw.[4]

There is a large variation in the number of languages spoken in each country. Some countries have only one or two languages reported in the dataset while others have several hundred languages listed. On average, there are 47 languages spoken in a country, although the median number is about a third of that (17). Papua New Guinea has the largest number of native and total spoken languages. Meanwhile, there are 9 countries with speakers of only 1 language.[5] Our data covers a large number of speakers, of which most are considered native speakers. As an illustration, Table 1 presents the most widely-spoken languages based on the total number of speakers (native or acquired) across all countries. English and Mandarin Chinese are the two most widely-spoken languages, both representing more than 1.1 billion speakers globally. Notably, the English-speaking population is primary comprised of acquired speakers across 165 countries while the

[4]There are highly detailed and overlapping subgroups within each language community, each representing different dialects. Languages and dialects can come together or split apart. The linguistic proximity variables connect each language and dialect to other languages and dialects, rather than assuming that there exists only a distinct ("meso") language to connect with. This means that our conventions of language use link us to a broader cluster of similar languages and dialects rather than just a single, unique language.

[5]The countries are the British Indian Ocean Territory; Democratic People's Republic of Korea; Faeroe Islands; Maldives; Pitcairn; Saint Helena, Ascension, and Tristan da Cunha; Saint Pierre and Miquelon; San Marino; and Vatican City.

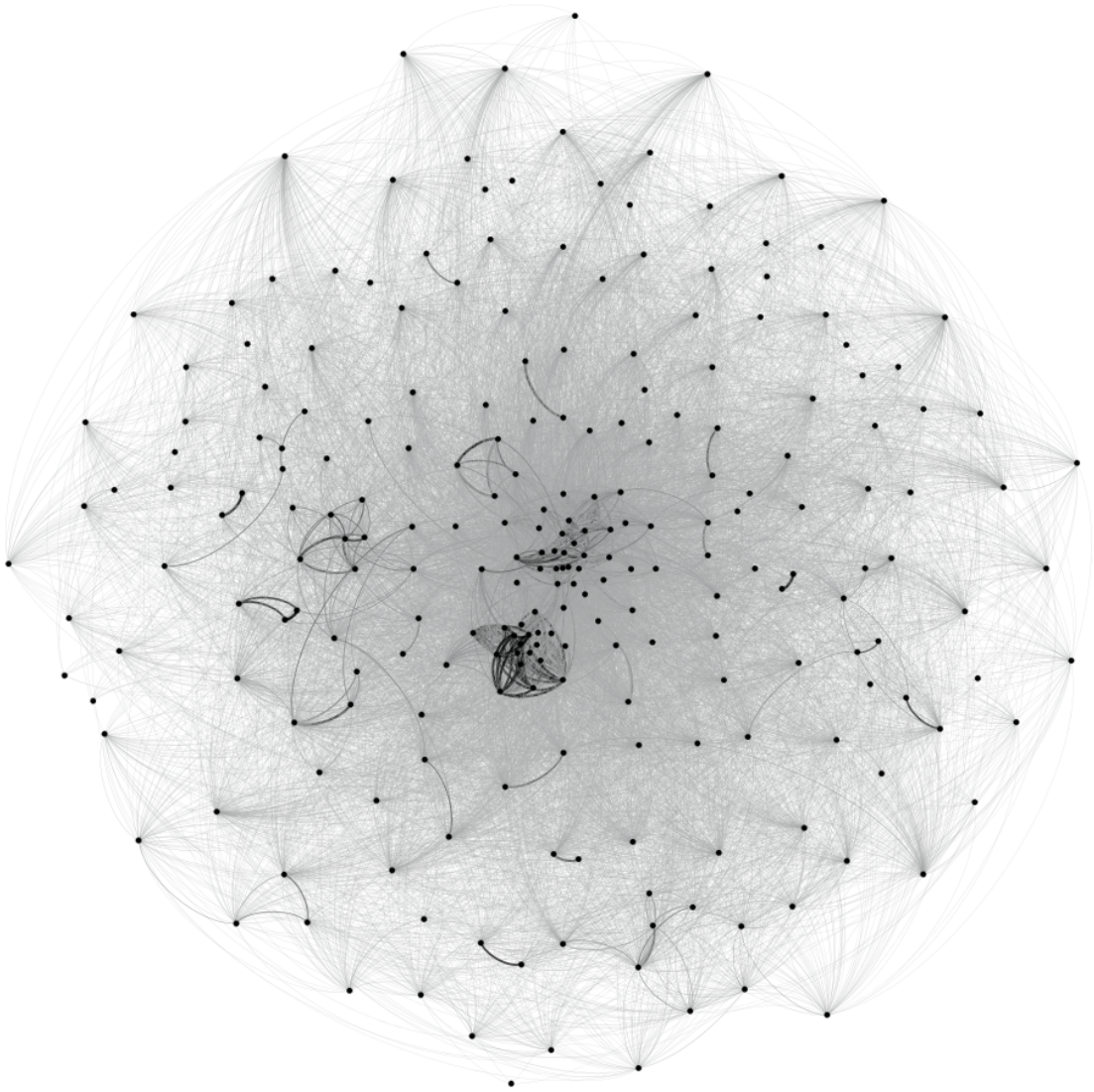

Figure 2: Visualization of the Common Native Language (CNL) index

Table 1: Top languages by speaker number and the share that are native speakers

Language	Total users (M)	Native speakers (%)
English	1,122	33.7
Mandarin Chinese	1,107	82.1
Hindi	534	48.7
Spanish	513	86.2
French	285	27.0
Russian	264	58.2
Bengali	262	92.7
Portuguese	237	94.2
Indonesian	198	11.6
Urdu	163	42.4

Mandarin-speaking population is largely made up of native speakers.

1.2 Construction of Common Language Indices

To construct the DICL language indices, we gather four types of information: the official language(s) of each country; the native and acquired language(s) spoken in each country; and linguistic trees, which are used to construct cladistic distances between different languages. This information is used to construct 8 separate indices describing different dimensions of linguistic relationships within and between countries. These indices include two measures describing common official languages, three measures describing common native and/or acquired languages, and three measures describing the linguistic proximity between distinct languages.

For the two official languages indices, we follow the Ethnologue definition of official language, which specifies the function for which each language is recognized in a country.[6] We derive two measures based on different definitions for official language. The first (COL_{ij}) uses a liberal definition that includes languages identified by Ethnologue as serving any official function—whether national or provincial, statutory or de facto. The second is a "restricted" version (COR_{ij}) that is based on a narrower subset of official languages representing only *national* statutory or de facto languages.[7] These official languages are the ones in which government functions and business are conducted throughout the country. Both indices represent a major difference with existing datasets that only consider a limited subset of statutory official languages. The two indices are defined such that they equal 1 if the two countries share a common language that is recognized as having an official function under each respective definition, and 0 otherwise. Within a country ($i = j$), the indices are set equal to 1 if the country has at least 1 official language.[8] Of the 29,403 unique pairs of

[6]Ethnologue's definitions of language status are described in Table 3 of the Methodology section of https://www.ethnologue.com/methodology/.

[7]Specifically, COR_{ij} utilizes only languages designated with the function codes "SNL" (statutory national language), "SNW" (statutory national working language), "DNL' (de facto national language), and "DNW" (de facto national working language). It does not include languages that are official at the state or provincial but not national level, for example.

[8]Most (but not all) countries exhibit at least 1 official national language, so most domestic values are 1 for both indices. Malawi is an unusual example that has a broadly defined official language but not a narrowly defined one, so that $COL_{ii} = 1$

countries in the data, 6,386 (21.7 percent) share a common official language under the definition of COL. Meanwhile, 5,790 (19.7 percent) pairs do under the definition of COR.

For the common native language index (CNL_{ij}), we base our measure on the products of the percentages of native speakers in each country pair. The product represents the probability that two people chosen at random from each pair of countries share a common native language k. The common native language index is computed as

$$CNL_{ij} = \sum_{k \in K} (l_{ki} \times l_{kj}), \quad \forall i, j \in \Omega.$$

Ω denotes the set of countries in our dataset and l_k is the percentage of native speakers of a specific language k from the set of all languages K in a country i or j. We make full use of the 6,674 languages with speakers available in the dataset to compute the index. The common native language index between populations of the same country is given by $CNL_{ii} = \sum_{k \in K} l_{ki}^2$, which is the opposite of the commonly employed ethno-linguistic fractionalization (ELF) index (Alesina et al., 2003).

For the common acquired languages index (CAL_{ij}), we base our measure on the products of the percentages of speakers of acquired languages in each country pair. The acquired language index is computed as

$$CAL_{ij} = \sum_{k \in K} (a_{ki} \times a_{kj}), \quad \forall i, j \in \Omega.$$

The only difference in this computation compared to CNL is that the share of acquired language speakers (a_{ki}) is used instead of the share of native speakers. There are 649 languages with acquired speakers available in the dataset to compute the index. The domestic version of the index is similarly defined according to $CAL_{ii} = \sum_{k \in K} a_{ki}^2$. It is often the case that individuals speak multiple acquired languages, which can result in index values that exceed 1. These larger values imply that not only can most people speak a common language, many share multiple common languages. Given this duplicity, the interpretation of the CAL index as a probability is not entirely appropriate. A corresponding measure bounded between 0 and 1 could readily be constructed via normalization if desired.[9] In total, 116 CAL indices (0.3 percent) are greater than 1. It should also be noted that this index captures ties between acquired language speakers only and does not consider connections between acquired speakers and native speakers. The CAL index is slightly more limited than the official or native language indices described above because there are 46 countries that are lacking any information on acquired language speakers. No CAL index values are given for these 46 countries, resulting in a total of 38,416 CAL records for the other 196 countries.[10]

but $COR_{ii} = 0$.

[9] For example, $CAL/max(CAL)$.

[10] The 46 countries are AFG, ALA, ANT, ARE, BHR, BIH, BLM, BMU, CCK, CXR, CYM, ERI, FLK, FRO, GGY, GRD, GRL, IOT, ISL, JEY, KNA, KWT, LBY, LIE, MAF, MDV, MSR, NFK, OMN, PCN, PRK, PSE, SDN, SHN, SMR, SOM, SPM, SXM, SYR, TCA, TJK, TKL, VAT, VCT, VGB, and VNM (ISO3 codes).

The common spoken language index (CSL_{ij}) is derived in the same way as the CNL and CAL indices but is based on speakers of both native and acquired languages. It is computed as

$$CSL_{ij} = \sum_{k \in K} \left[(l_{ki} + a_{ki}) \times (l_{kj} + a_{kj}) \right], \quad \forall i, j \in \Omega.$$

Similarly, the domestic values are computed as $CSL_{ii} = \sum_{k \in K} (l_{ki} + a_{ki})^2$. Of the three common language measures, CSL provides the broadest coverage as it captures native speakers sharing a common language with acquired speakers and, therefore, reflects language ties between all speakers. As with CAL, the CSL index can exceed 1 in cases where populations widely share more than 1 common language. 162 of the values (0.4 percent) are greater than 1. The CSL index is also limited by the unavailability of acquired speaker information for 46 countries and does not include values for these countries.

Finally, for the linguistic proximity indices (LPN_{ij}, LPA_{ij}, and LPS_{ij}), we turn to linguistic trees. The indices are inspired by ideas from Laitin (2000) and Fearon (2003), who used the Ethnologue classification of languages into trees, branches, and sub-branches as the basis to compare different languages. Within the classification, languages are broadly grouped into 152 language families with common characteristics based on root proto-languages. Stemming from the proto-languages, languages are repeatedly split into different branches and sub-branches with their closest relatives. These linguistic trees provide a means for evaluating the similarities between two languages based on the proximity of their respective branches. For example, languages on a common branch within a family may share many words, characters, and other systematic features—often rendering them mutually intelligible. The further apart two languages are, however, the fewer features they share in common and the less likely it is that any parts of them are mutually intelligible.

We develop an algorithm that makes use of all languages in our dataset and the full structure of the language trees. It measures the cladistic distance between languages by how far two languages are from their closest common proto-language. Using the linguistic family classification provided by Ethnologue, we compare the proportion of the linguistic tree that each pair of languages shares to the proportion in which they diverge. Linguistic proximity between languages k and h is given by the number of common branches (b_{kh}) the two languages share starting at the proto-language divided by the average length of the branches that terminate in each language (b_k or b_h). For example, if two languages split after the seventh branch, one of those languages is 10 steps removed from its proto-language, and the other language is 11 steps removed, then the proximity for these two languages is $P_{kh} = b_{kh}/[0.5(b_k + b_h)] = 7/10.5$.

To account for the fact that most countries have populations speaking multiple languages, we use the speaker share in a country to aggregate across languages. This results in a population-weighted measure of

linguistic proximity. We compute the native language proximity index as

$$LPN_{ij} = \sum_{k \neq h \in K} (l_{ki} \times l_{hj} \times P_{kh}), \quad \forall i, j \in \Omega.$$

Notably, the aggregation is based solely on pairings of different languages ($k \neq h$) as same language pairs are already described by the the CNL index. A similar index is constructed for acquired languages (LPA_{ij}) and all spoken languages (LPS_{ij}) using each country's acquired language share (a_{ik}) and native plus acquired language share ($l_{ik} + a_{ik}$), respectively.

Using this algorithm, we construct an index that takes a value of zero if two languages do not originate from a common proto-language and share no part of a language tree. For languages that stem from the same proto-language, a value between zero and one is assigned. Languages that diverge early in their family tree receive lower proximity score than languages that share many common branches prior to splitting. Low index values signal that the populations largely do not speak any closely related languages. High index values may be indicative of two types of connections. In all cases, larger values suggest the presence of widely spoken, closely related languages. For the LPA and LPS indices, larger values may also reflect multilingualism in related languages, which allows for higher values than LPN.

2 Data Records

The DICL database is available from the USITC Gravity Portal at https://www.usitc.gov/data/gravity/dicl.htm.

In total, the database contains 12 columns and 58,564 rows comprised of indices for 29,403 unique country pairs. For convenience, the international records are mirrored so that there is a record for both the pair (i, j) and (j, i). The domestic records appear once for each country.[11] The 12 columns contain each of the language measures described in the previous section as well as names and ISO 3-digit alpha identifiers for each country. The first row of each column contains a column label. The columns are defined as follows.

- **iso3_i:** Country i ISO 3-digit alpha identifier
- **country_i:** Country i name
- **iso3_j:** Country j ISO 3-digit alpha identifier
- **country_j:** Country j name
- **col:** Common official language indicator

[11] 2 * 29,161 mirrored international records + 242 domestic measures = 58,564 total records.

- **cor:** Restricted official language indicator based on a narrower definition of what qualifies as an official language
- **cnl:** Common native language index
- **cal:** Common acquired language index
- **csl:** Common spoken language index (native and acquired)
- **lpn:** Linguistic proximity index for different native languages
- **lpa:** Linguistic proximity index for different acquired languages
- **lps:** Linguistic proximity index for different spoken languages (native and acquired)

3 Technical Validation

3.1 Index Characteristics

The language data paints a vivid picture of the complex ways in which countries are interconnected via language. Table 2 provides summary statistics of each variable in the DICL database. As demonstrated by the mean and percentiles, most pairs of countries exhibit relatively weak linguistic ties and low index values compared to the countries with the greatest ties and largest values.

Table 2: Summary statistics of the DICL indices. Type, count of non-missing values, mean, standard deviation, minimum, maximum, and 25th, 50th, and 75th percentiles are reported for each index.

Index	Type	Count	Mean	Standard Deviation	Minimum	25%	50%	75%	Maximum
COL	Indicator	58564	0.214	0.410	0.000	0.000	0.000	0.000	1.000
COR	Indicator	58564	0.194	0.395	0.000	0.000	0.000	0.000	1.000
CNL	Continuous	58564	0.019	0.101	0.000	0.000	0.000	0.000	1.000
CAL	Continuous	38416	0.062	0.143	0.000	0.000	0.001	0.053	2.375
CSL	Continuous	38416	0.114	0.213	0.000	0.000	0.012	0.125	2.780
LPN	Continuous	58564	0.058	0.115	0.000	0.000	0.005	0.059	0.866
LPA	Continuous	38416	0.035	0.076	0.000	0.000	0.006	0.037	1.437
LPS	Continuous	38416	0.183	0.253	0.000	0.017	0.086	0.251	3.357

Looking at the mean value of each country's indices, we can examine which countries tend to be closely connected to the rest of the world linguistically and which are not. Table 3 presents the 5 most and least closely connected countries based on each index. The mean official language indices, COL and COR, tend to be high for countries with multiple recognized languages—especially those that recognize globally or regionally prominent languages. Such countries include India and Rwanda, for example. The least

Table 3: Countries with the highest and lowest average indices. For CAL, CSL, LPA, and LPS, countries lacking information on acquired languages are not considered

Index	Ranking	Countries
CNL	Highest	Saint Helena, Ascension, and Tristan da Cunha; Isle of Man; British Indian Ocean Territory; Guernsey; Falkland Islands
CNL	Lowest	Cameroon; South Sudan; Congo, Democratic Republic of the; Central African Republic; Chad
CAL	Highest	Northern Marianas; Dominica; Luxembourg; Cameroon; American Samoa
CAL	Lowest	Iran; Isle of Man; Turkey; Bhutan; Japan
CSL	Highest	Luxembourg; Cameroon; Dominica; United Kingdom; U.S. Virgin Islands
CSL	Lowest	East Timor; Bhutan; Ethiopia; Iran; Myanmar
COL	Highest	India; Jersey; Seychelles; Mauritius; Rwanda
COL	Lowest	Christmas Island; Saint Helena, Ascension, and Tristan da Cunha; Cocos (Keeling) Islands; Aland Islands; Falkland Islands
COR	Highest	Seychelles; Mauritius; Rwanda; Canada; Vanuatu
COR	Lowest	Christmas Island; Saint Helena, Ascension, and Tristan da Cunha; Cocos (Keeling) Islands; Aland Islands; Falkland Islands
LPN	Highest	Brazil; Portugal; Italy; San Marino; Vatican City
LPN	Lowest	Korea, North; Korea, South; Japan; Mongolia; Taiwan
LPA	Highest	Luxembourg; Croatia; Equatorial Guinea; Slovenia; Denmark
LPA	Lowest	Iran; Turkey; Bhutan; Japan; China
LPS	Highest	Luxembourg; Croatia; Netherlands; Slovenia; Andorra
LPS	Lowest	Japan; Thailand; Taiwan; China; Laos

connected via official languages are simply the countries that have no officially recognized languages, such as Saint Helena, Ascension, and Tristan da Cunha and the Aland Islands.

The common language indices highlight numerous different types of connections. The countries with the highest average CNL indices tend to be countries that are linguistically homogeneous and have large native English-speaking populations, such as Saint Helena, Ascension, and Tristan da Cunha; Isle of Man; and Guernsey. The least connected on average tend to be countries with native speakers of many different languages, such as Cameroon (nearly 170 different languages) or the Democratic Republic of the Congo (more than 200 languages). The CAL index highlights a very different set of counties. The most connected tend to be countries with very large populations of English acquired speakers, such as Northern Marianas and Dominica. The CAL index also highlights countries like Luxembourg, which has a diverse set of native languages but it's speakers also speak multiple acquired languages in large numbers, such as French, German, English, Italian, and Spanish. The CSL index similarly highlights countries with large populations of speakers of major global languages, including those that speak these major languages naively, like the United Kingdom. It also highlights countries with rampant multilingualism in prominent languages, such as Luxembourg.

Meanwhile, the least connected based on the CSL index are countries that largely speak uncommon languages and have relatively small multilingual populations, such as East Timor and Bhutan.

The linguistic proximity indices identify countries with many extensive connections between closely related languages. The highest LPN countries are generally those with large populations speaking a language that is itself not especially prominent globally but is closely related to major international languages. For example, these countries include majority Portuguese and Italian speaking countries like Brazil, Portugal, and Italy. By comparison, the countries with the smallest LPN indices are largely east Asian countries with majority languages that are not closely related to any major global languages. For LPA, the most connected countries are largely European countries with large populations of speakers that have learned the major languages of Europe and elsewhere. Notably, several central European countries, such as Croatia and Slovenia, have especially high average LPA indices due to their populations of both western language and Russian speakers. The countries with the lowest LPA indices are mainly countries with few speakers of any major acquired language, such as Iran and Turkey. The LPS index highlights many of the same types of countries as the LPN and LPA indices, although not always in the same order.

Countries also differ in the distribution of their language indices. Some countries have close ties with relatively few other countries while others have similar ties to many other countries. Herfindahl–Hirschman-type Indices (HHIs) produced using the language indices can be used to examine these distributions. For each country i in the set of countries N, the HHI is defined as $HHI_i = \sum_{j \in N} (100 * w_{ij}/s_i)^2$ for language index w and $s_i = \sum_{j \in N} w_{ij}$. Countries with the highest HHI values are those with the most concentrated language ties, and vice versa.

For CNL, the countries with the most concentrated native language connections are the Maldives, Iceland, Faeroe Islands, Cape Verde, and Myanmar. Meanwhile, the least concentrated countries are Belize, Gibraltar, Sint Maarten, Netherlands Antilles, and Vanuatu. Among the largest economies, many of the majority-English or Spanish countries like the United States, United Kingdom, and Spain have relatively low concentration. By comparison, Japan, Indonesia, and Poland have relatively high concentration. Meanwhile, China, Germany, and India are roughly in the middle.

For CAL, the countries with the most concentrated acquired language connections are Myanmar, South Sudan, Bhutan, and Iran. Meanwhile, Portugal, Switzerland, Belgium, and Spain arise as having the least concentrated language connections. The United States, India, and the United Kingdom are moderately concentrated; Germany and Poland are not very concentrated; and China, Indonesia, and Japan are highly concentrated.

As demonstrated by these descriptions of the database, each index reflects distinct dimensions of linguistic connections. To further emphasize these differences, Table 4 presents a correlation matrix for the 8 measures.

Table 4: Correlations between the DICL language indices

	COL	COR	CNL	CAL	CSL	LPN	LPA	LPS
COL	1.00	0.94	0.27	0.32	0.52	0.04	-0.02	0.02
COR	0.94	1.00	0.29	0.31	0.53	0.02	-0.06	-0.05
CNL	0.27	0.29	1.00	0.02	0.56	-0.02	-0.04	-0.03
CAL	0.32	0.31	0.02	1.00	0.66	0.07	0.37	0.28
CSL	0.52	0.53	0.56	0.66	1.00	0.09	0.22	0.21
LPN	0.04	0.02	-0.02	0.07	0.09	1.00	0.09	0.72
LPA	-0.02	-0.06	-0.04	0.37	0.22	0.09	1.00	0.58
LPS	0.02	-0.05	-0.03	0.28	0.21	0.72	0.58	1.00

With the exception of the two official language measures, the correlations between the indices are relatively low, implying that each is capturing different types of variation in language patterns. These differences in the information embodied in each index may be of considerable importance for users and should be considered carefully when using the data.

3.2 Relationships between Language and Economic Indicators

To further demonstrate the properties and possible uses of the data, we perform a series of regressions using a variety of commonly studied economic indicators. We do so for two categories of indicators. First, we regress a selection of our bilateral language variables against several bilateral economic outcomes: international trade, foreign direct investment (FDI), and migration. Second, we regress a selection of our within-country language measures against several domestic outcomes: per capita real GDP growth, government quality, and civil conflict. In both cases, the regressions are meant to demonstrate correlations between language and economic outcomes and are not necessarily intended to identify causal relationships. They also present ways in which the DICL indices can be further used to produce other types of language variables, such as combined measures or indices of linguistic diversity.

3.2.1 Bilateral Language Relationships

To examine bilateral relationships with common language, we evaluate the correlation between native language and international trade, foreign direct investment, and migration. We construct a single measure of native language connections that is equal to the average value of the CNL and LPN measures ($LANG_{ij} = 0.5 * [CNL_{ij} + LPN_{ij}]$). Data on foreign trade was sourced from Felbermayr et al. (2020) and covered the period from 1950–2020. Data on foreign direct investment was sourced from Larch and Yotov (2022) and covered 1990–2011. Data on migration was sourced from the United Nations Department of Economic and Social Affairs, Population Division (2020) and covered 2015. A Poisson Pseudo Maximum

Likelihood (PPML) estimator was used in all three cases (Santos Silva and Tenreyro, 2006).[12] In addition to the measure of language, additional controls were included for bilateral distance, contiguous borders, joint World Trade Organization (WTO) membership, joint European Union (EU) membership, preferential trade agreements (PTA), colonial ties, and religion. These data were sourced from the Dynamic Gravity Dataset (Gurevich and Herman, 2018).

Table 5 presents the results from these regressions. Columns (1), (4), and (7) reflect specifications including language with the two geographic controls for distance and contiguity. In all three cases, common language is positively correlated with trade, FDI, and migration at conventional levels of statistical significance. The addition of the international policy controls for WTO, EU, and PTA membership in columns (2), (5), and (8) has a small impact on the magnitude of each relationship but all remain positive and significant. Finally, the relationships remain robust to the inclusion of additional cultural controls for colonial ties and common religion in columns (3), (6), and (9). Based on these findings, common language is associated with greater trade, FDI, and immigration. These findings are also consistent with the work of Gurevich et al. (2021), which used the DICL data to study international trade flows. Notably, Gurevich et al. (2021) also found significant positive effects of within-country native language on domestic trade.

[12] PPML offers several potential advantages. It allows for the inclusion of zeros in the dependent variable and provides superior treatment of heteroskedasticity.

Table 5: Estimated relationships between native language and bilateral economic outcomes

	A. International Trade			B. Foreign Direct Investment			C. Migration		
	(1) Geogr	(2) Policy	(3) Cultr	(4) Geogr	(5) Policy	(6) Cultr	(7) Geogr	(8) Policy	(9) Cultr
Native language	0.576**	0.735***	1.031***	3.524***	3.601***	2.473***	3.058***	2.903***	1.366**
	(0.252)	(0.256)	(0.236)	(0.581)	(0.567)	(0.528)	(0.639)	(0.620)	(0.597)
Distance	-0.865***	-0.761***	-0.892***	-0.684***	-0.544***	-0.289***	-1.391***	-1.311***	-1.185***
	(0.028)	(0.036)	(0.041)	(0.063)	(0.086)	(0.104)	(0.079)	(0.079)	(0.078)
Contiguity	0.446***	0.404***	0.258***	0.375**	0.344**	0.434***	1.488***	1.300***	0.933***
	(0.065)	(0.064)	(0.067)	(0.157)	(0.146)	(0.144)	(0.149)	(0.143)	(0.137)
WTO		0.532***	0.514***		0.923	0.508		1.508***	1.624***
		(0.171)	(0.186)		(0.597)	(0.592)		(0.367)	(0.515)
EU		0.514***	0.540***		0.307	0.789***		-0.466*	0.052
		(0.093)	(0.090)		(0.255)	(0.224)		(0.247)	(0.229)
PTA		0.228***	0.254***		0.542***	0.628***		0.899***	0.700***
		(0.060)	(0.058)		(0.170)	(0.134)		(0.117)	(0.095)
Colony			0.163**			0.450***			1.355***
			(0.077)			(0.135)			(0.152)
Religion			1.109***			-1.919***			-2.816***
			(0.195)			(0.574)			(0.434)
Obs.	734164	734164	508739	36674	36674	36074	52441	52441	30800

Notes: This table presents findings regarding three categories of outcomes: (i) international between 1950 and 2020 (Columns 1-3), (ii) foreign direct investment between 1990 and 2011 (Columns 4-6) , and (iii) migration in 2015 (Columns 7-9). "Native language" represents the average of the DICL CNL and LPN indices. Additional controls include bilateral distance, contiguity, WTO membership (WTO), European Union membership (EU), preferential trade agreements (PTA), colonial ties, and religion. Standard errors are reported in parentheses and were clustered at country-pair level. ***, 1%, **, 5%, *, 10% levels of significance.

3.2.2 Domestic Language Relationships

To examine the domestic relationships with language, we evaluate the correlation between several dimensions of the DICL language data and per capita GDP growth, government quality, and civil conflict. Data on GDP growth from 2008 to 2017 was sourced from Bolt and Van Zanden (2020) and covered 139 countries. Data on the quality of government was sourced from Teorell et al. (2023) and reflected governments in 2017.[13] Data on the onset of civil conflicts was sourced from Arbatli et al. (2020) and spanned the period 1960–2017.[14] Additional controls for genetic diversity, ethnic diversity, geography, institutions, population, and lagged per capita GDP were included as well.[15] In the case of civil conflict, controls for time period and intertemporal spillovers were also included. These additional controls were sourced from Bolt and Van Zanden (2020) (GDP per capita) and Arbatli et al. (2020) (all other controls). Finally, the domestic regressions were conducted using ordinary least squares (OLS) to analyze per capita real GDP growth and government quality. Logistic regressions were used to examine the onset of civil conflict. The reported marginal effects were evaluated at the sample means.

Table 6 presents the results from the domestic regressions. For narrative purposes and in keeping with the past literature, we use the inverse of the DICL measures to reflect linguistic diversity instead of commonality. Specifically, we define a measure of native language diversity equal to 1-CNL and a measure of cladistic diversity equal to 1-LPN. A third measure, average language diversity, is equal to the average of these two measures. Columns (1), (4), and (7) regress the average measure against per capita GDP growth, government quality, and civil conflict, respectively. In all three cases, language diversity is positively correlated with the respective dependent variable, suggesting that linguistically diverse populations tend to have higher per capita GDP growth, higher quality of government, and higher incidents of civil conflict, *ceteris paribus*. Columns (2), (5), and (8) include just the native diversity measure and produce similar positive and significant results, although the magnitude of the relationship is smaller in all three cases. Finally, columns (3), (6), and (9) include only the cladistic diversity measure and similarly result in smaller but otherwise consistent correlations.

Together, these bilateral and country-level regressions demonstrate that the new DICL data effectively reflect the economic relationships associated with common languages or linguistic diversity. The correlations

[13] Government quality was quantified by the average ICRG index encompassing Corruption, Law and Order, and Bureaucracy Quality in 2017. The measure is scaled from 0 to 1, where higher values indicate superior government quality.

[14] The conflict data were constructed from the UCDP/PRIO Armed Conflict Dataset (Gleditsch et al., 2002), capturing the occurrence of civil conflict in any given year within the 1960–2017 period.

[15] Geographical controls include absolute latitude, ruggedness, distance to the nearest waterway, and an indicator for small island nations. Institutional controls consist of 1-year lagged legal origin dummies (British and French legal origins) and covariates related to executive constraints, political regime type (democracy and autocracy), and colonial history. The controls for population and per capita GDP were 1-year lagged, log transformed values. In the long-term growth regression, the lagged per capita GDP is replaced by the initial level of per capita GDP. Furthermore, there is a control for the presence of oil, indicated by a time-invariant indicator for the discovery of a petroleum (oil or gas) reserve by the year 2003 in the regressions on the onset of civil conflict.

present with the DICL are generally consistent with those among other sources of linguistic information and expand upon their scopes in terms of the languages and countries included, gradation and variability, and domestic coverage.

Table 6: Estimated relationships between linguistic diversity and domestic economic outcomes

	A. Per Capita Real GDP Growth			B. Quality of the Government			C. Onset of Civil Conflict		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Average language diversity	0.317***			0.184***			0.015***		
	(0.116)			(0.066)			(0.004)		
Native diversity		0.247**			0.158***			0.008**	
		(0.101)			(0.054)			(0.004)	
Cladistic diversity			0.290***			0.153**			0.020***
			(0.104)			(0.068)			(0.005)
Genetic diversity	-1.722**	-1.940**	-1.353*	0.014	-0.136	0.186	0.063*	0.056	0.077**
	(0.734)	(0.753)	(0.745)	(0.407)	(0.417)	(0.407)	(0.034)	(0.035)	(0.034)
Ethnic diversity	-0.247**	-0.231*	-0.188**	-0.097	-0.100	-0.060	-0.013***	-0.009*	-0.013***
	(0.112)	(0.117)	(0.094)	(0.064)	(0.064)	(0.058)	(0.005)	(0.005)	(0.005)
Controls:									
Geography	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Institution	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Population	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
P.c GDP (Lagged)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Temporal spillovers	No	No	No	No	No	No	Yes	Yes	Yes
Time dummies	No	No	No	No	No	No	Yes	Yes	Yes
Number of Countries	139	139	139	119	119	119	144	144	144
Obs.	139	139	139	119	119	119	5678	5678	5678

Notes: This table presents findings regarding three categories of outcomes: (i) Per Capita Real GDP Growth between 2008 and 2017 (Columns 1-3), (ii) the quality of government (Columns 4-6) , and (iii) the onset of civil conflict spanning the period from 1960 to 2017 (Columns 7-9). "Average language diversity" reflects the average value of "native diversity" $(1 - CLN)$ and "cladistic diversity" (1-LPN). Geographical controls include absolute latitude, ruggedness, distance to the nearest waterway, and an indicator for small island nations. Institutional controls consist of 1-year lagged legal origin dummies (British and French legal origins) and covariates related to executive constraints, political regime type (democracy and autocracy), and colonial history. Additionally, 1-year lagged log-transformed values of total population and per capita GDP are included. In the long-term growth regression, the lagged per capita GDP is replaced by the initial level of per capita GDP. Furthermore, there is a control for the presence of oil, indicated by a time-invariant indicator for the discovery of a petroleum (oil or gas) reserve by the year 2003 in the regressions on the onset of civil conflict. To address duration dependence and temporal spillovers in conflict outcomes, Columns (7-9) account for the lagged incidence of conflict. For this analysis on repeated cross-sectional data, we include time dummies. Heteroscedasticity-robust standard errors are reported in parenthesis and clustered at country level in Columns (7-9). ***, 1%, **, 5%, *, 10% levels of significance.

4 Conclusion

The DICL uses the same country identifiers as the Dynamic Gravity dataset (Gurevich and Herman, 2018) and International Trade and Production dataset (Borchert et al., 2021). Combining the DICL with the economic, geographic, and international trade information in these datasets should be seamless.

The DICL provides data that is comparable with many of the measures commonly used throughout the literature. Table 7 presents the correlation between each DICL index and the closest measures from 3 sources. First, the "common_language" (both countries share a commonly spoken language based on the CIA World Factbook) indicator of the Dynamic Gravity dataset. Second, the "comlang_off" (common official language) and "comlang_ethno" (common language spoken by at least 9% of population in each country) indicators of the GeoDist dataset (Mayer and Zignago, 2011). Third, the "col" (common official language) indicator, "cnl" (common native language) index, "csl" (common spoken acquired language) index, "lp1" (linguistic proximity based on language trees) index, and "lp2" (linguistic proximity based on common words) index of Melitz and Toubal (2014), hereafter "MT 2014". In each case, the measures differ in the data they are based upon, definition, and language coverage. The measures of MT 2014 represent the closest comparison as they are defined similarly (except the LP measures). Notably, the country coverage of the DICL (242 countries) is generally greater than that of the GeoDist (224) or MT 2014 (195) datasets, but fewer than the Dynamic Gravity dataset (254). Nonetheless, the correlation coefficients are based solely on the common country-pairs.

Some of the new DICL indices are highly correlated with earlier measures but others diverge substantially. The similarities primarily occur with the official language indicators ("comlang_off" and "col") as well as the "cnl" index of MT 2014, which all feature similar technical definitions. In most other cases, the correlations are relatively low, emphasizing the extent to which the DICL database provides new information. These differences are primarily explained by differences in how the variables are defined and the DICL database's greater coverage of more languages.

Table 7: Correlations between the DICL indices and similar measures from the Dynamic Gravity, GeoDist, and Melitz and Toubal (MT 2014) databases

Source	Variable	Type	Count	COL	COR	CNL	CAL	CSL	LPN	LPA	LPS
Dynamic Gravity	common_language	Indicator	64516	0.56	0.56	0.21	0.23	0.39			
GeoDist	comlang_off	Indicator	50176	0.72	0.75						
GeoDist	comlang_ethno	Indicator	50176			0.29	0.24	0.46			
MT 2014	col	Indicator	37830	0.73	0.76						
MT 2014	csl	Continuous	37830			0.50	0.41	0.85			
MT 2014	cnl	Continuous	37830			0.69	-0.03	0.53			
MT 2014	lp1	Continuous	28950						0.59	0.12	0.52
MT 2014	lp2	Continuous	28950						0.50	0.11	0.45

References

Alesina, A., A. Devleeschauwer, W. Easterly, S. Kurlat, and R. Wacziarg (2003). Fractionalization. *Journal of Economic Growth 8*(2), 155–194. https://doi.org/10.1023/A:1024471506938.

Arbatlı, C. E., Q. H. Ashraf, O. Galor, and M. Klemp (2020). Diversity and conflict. *Econometrica 88*(2), 727–797. https://doi.org/10.3982/ECTA13734.

Bolt, J. and J. L. Van Zanden (2020). Maddison style estimates of the evolution of the world economy. a new 2020 update. *Maddison-Project Working Paper WP-15, University of Groningen, Groningen, The Netherlands*. http://reparti.free.fr/maddi2020.pdf.

Borchert, I., M. Larch, S. Shikher, and Y. V. Yotov (2021). The international trade and production database for estimation (ITPD-E). *International Economics 166*, 140–166. doi: https://doi.org/10.1016/j.inteco.2020.08.001.

Desmet, K., I. Ortuño-Ortín, and S. Weber (2009). Linguistic diversity and redistribution. *Journal of the European Economic Association 7*(6), 1291–1318. https://doi.org/10.1162/JEEA.2009.7.6.1291.

Fearon, J. D. (2003). Ethnic and cultural diversity by country. *Journal of economic growth 8*(2), 195–222. https://doi.org/10.1023/A:1024419522867.

Fearon, J. D. and D. D. Laitin (1999). Weak states, rough terrain, and large-scale ethnic violence since 1945. In *Annual Meetings of the American Political Science Association, Atlanta, GA*, Volume 8, pp. 19–99. https://web.stanford.edu/group/fearon-research/cgi-bin/wordpress/wp-content/uploads/2013/10/Weak-States-Rough-Terrain-and-Large-Scale-Ethnic-Violence-Since-1945.pdf.

Felbermayr, G., C. Syropoulos, E. Yalcin, and Y. Yotov (2020, May). On the Heterogeneous Effects of Sanctions on Trade and Welfare: Evidence from the Sanctions on Iran and a New Database. School of Economics Working Paper Series 2020-4, LeBow College of Business, Drexel University. https://ideas.repec.org/p/ris/drxlwp/2020_004.html.

Gazzola, M. and B.-A. Wickström (2016). *The economics of language policy*. MIT Press.

Ginsburgh, V. and S. Weber (2020). The economics of language. *Journal of Economic Literature 58*(2), 348–404. https://www.jstor.org/stable/10.2307/27030435.

Ginsburgh, V., S. Weber, and P. Macmillan (2016). *The Palgrave handbook of economics and language*. Springer.

Gleditsch, N. P., P. Wallensteen, M. Eriksson, M. Sollenberg, and H. Strand (2002). Ucdp/prio armed conflict dataset codebook version 18.1. https://ucdp.uu.se/downloads/replication_data/2018_c_666956-l_1-k_ucdp-prio-acd-181-codebook.pdf.

Greenberg, J. H. (1956). The measurement of linguistic diversity. *Language 32*(1), 109–115. https://doi.org/10.2307/410659.

Greenberg, J. H. (1987). *Language in the Americas*. Stanford University Press.

Gurevich, T. and P. Herman (2018). The dynamic gravity dataset: 1948-2016. USITC Working Paper 2018-02-A. https://www.usitc.gov/data/gravity/dgd.htm.

Gurevich, T., P. R. Herman, F. Toubal, and Y. V. Yotov (2021, January). One nation, one language? domestic language diversity, trade, and welfare. CEPR Discussion Paper Series DP15701. https://repec.cepr.org/repec/cpr/ceprdp/DP15701.pdf.

Laitin, D. D. (2000). What is a language community? *American Journal of political science*, 142–155. https://doi.org/10.2307/2669300.

Larch, M. and Y. Yotov (2022, September). Deep Trade Agreements and FDI in Partial and General Equilibrium: A Structural Estimation Framework. School of Economics Working Paper Series 2022-7, LeBow College of Business, Drexel University. https://ideas.repec.org/p/ris/drxlwp/2022_007.html.

Lewis, B. (2005). *From Babel to dragomans: interpreting the Middle East*. Oxford University Press.

Lieberson, S. (1964). An extension of greenberg's linguistic diversity measures. *Language 40*(4), 526–531. https://doi.org/10.2307/411935.

Mayer, T. and S. Zignago (2011). Notes on CEPII's Distances Measures: The GeoDist Database. *CEPII Working Paper 2011 - 25*. https://dx.doi.org/10.2139/ssrn.1994531.

Melitz, J. and F. Toubal (2014). Native language, spoken language, translation and trade. *Journal of International Economics 93*(2), 351–363. https://doi.org/10.1016/j.jinteco.2014.04.004.

Santos Silva, J. M. C. and S. Tenreyro (2006). The log of gravity. *The Review of Economics and Statistics 88*(4), 641–658. https://doi.org/10.1162/rest.88.4.641.

Simons, G. F. and C. D. Fennig (2018). *Ethnologue: Languages of the World* (21 ed.). Dallas, Texas: SIL International.

Teorell, J., A. Sundström, S. Holmberg, B. Rothstein, N. P. Alvarado, C. M. Dalli, and Y. Meijers (2023). The quality of government standard dataset, version jan23. *University of Gothenburg: The Quality of Government Institute*. http://www.qog.pol.gu.se.

United Nations Department of Economic and Social Affairs, Population Division (2020). International migrant stock 2020. https://www.un.org/development/desa/pd/content/international-migrant-stock.