LANGUAGE DOCUMENTATION & CONSERVATION

Vol. 4 (2010), pp. 97-151
http://nflrc.hawaii.edu/ldc/
http://hdl.handle.net/10125/4474

The Index of Linguistic Diversity: A New Quantitative Measure of Trends in the Status of the World's Languages

David Harmon
George Wright Society/Terralingua

Jonathan Loh
Zoological Society of London/Terralingua

The Index of Linguistic Diversity (ILD) is a new quantitative measure of trends in linguistic diversity. To derive the ILD we created a database of time-series data on language demographics, which we believe to be the world's largest. So far, the database contains information from nine editions of *Ethnologue* and five other compendia of speaker numbers. The initial version of the ILD, which draws solely on the *Ethnologue* subset of these data, is based on a representative random sample of 1,500 of the world's 7,299 languages (as listed in the 2005 edition). At the global level, the ILD measures how far, on average, the world's languages deviate from a hypothetical situation of stability in which each language is neither increasing nor decreasing its share of the total population of the grouping. The ILD can also be used to assess trends at various subglobal groupings. Key findings:

- Globally, linguistic diversity declined 20% over the period 1970–2005.
- The diversity of the world's indigenous languages declined 21%.
- Regionally, indigenous linguistic diversity declined over 60% in the Americas, 30% in the Pacific (including Australia), and almost 20% in Africa.

1. INTRODUCTION.[1] Concern about the future of the world's languages has been building for the better part of two decades. A large amount of qualitative evidence points to an impending mass extinction[2] of languages. The quality of this evidence ranges from merely

[1] We are grateful to The Christensen Fund for underwriting this work as part of a larger project on Global Indicators of the Status and Trends of Linguistic Diversity and Traditional Knowledge, which is being carried out by the NGO Terralingua. Luisa Maffi of Terralingua provided valuable comments throughout the project. We owe a large debt of thanks to M. Paul Lewis, editor of *Ethnologue*, for providing copies of the earliest editions; we also thank him for answering questions about the book's publishing history and reviewing the technical report upon which this paper was based. We are indebted to Margaret Florey for her comments on an earlier draft of the manuscript, to Ashbindu Singh for reviewing the technical report, and to Kenneth L. Rehg and two anonymous referees for their comments on the final draft.

[2] Outside of specialist discussions, the issue of language endangerment is almost always couched in terms of "extinction." Applying the extinction concept to language is fraught with theoretical difficulties—and, even more troublingly, can be used by unsympathetic authorities to thwart the interests of language communities. Still, the metaphor is firmly ensconced in the both the popular and professional literature, and the alternatives (such as "sleeping" or "silent" languages) also have problems. For a good discussion of the difficulties in determining the precise moment when a language goes extinct, see Evans 2001.

E-ISSN 1934-5275

anecdotal to very accurate narrative accounts based on firsthand knowledge of the language demographics of individual speech communities. It is a highly valuable body of evidence, leaving no room to doubt that the entirety of the world's languages—not just their number, but also the linguistic and cultural diversity they represent—is being severely diminished. For a host of complex reasons, people are abandoning their mother tongues and switching to other languages, almost always ones with larger numbers of speakers; thereby, more and more people are being concentrated into fewer and fewer languages.

However, there is much less quantitative evidence of a global linguistic diversity crisis. To help fill this gap we have created the Index of Linguistic Diversity (ILD), which we believe to be the first-ever quantitative index of trends in linguistic diversity based on time-series data on numbers of mother-tongue speakers. The ILD assesses trends in linguistic diversity by comparing changes in the relative distribution of mother-tongue speakers against a benchmark of the situation prevailing in 1970, the earliest year we could set the index based on the data available. The ILD measures how far, on average, the languages in a given geographical grouping deviate from a hypothetical situation of stability in which each language is neither increasing nor decreasing its share of the total population of the grouping. For example, ILD Global, an index of the world's overall linguistic diversity, measures the average deviation of the world's languages from a hypothetical situation in which each language is neither increasing nor decreasing its share of the global population. The index does this by measuring changes in the number of mother-tongue speakers from a globally representative sample of 1,500 languages over the period 1970–2005. (See Appendix A for a discussion of the ILD database.) The ILD can be calculated at different geographic scales and for different groupings of languages; each of these versions of the index uses the same methods.

The main finding of this research is that linguistic diversity has seriously declined since 1970. The overall linguistic diversity of the world, as measured by ILD Global, declined by 20% over the 35–year period (Figure 1). We also assessed the diversity of the world's indigenous languages—which make up 80–85% of the total number—on both global and regional levels. We did this because the status of the world's indigenous languages is important to global initiatives such as the Convention on Biological Diversity, as well as to indigenous communities themselves. ILD Global Indigenous, which measures the diversity of the world's indigenous languages, declined by 21% (Figure 2). The diversity of indigenous languages declined in all regions as well.

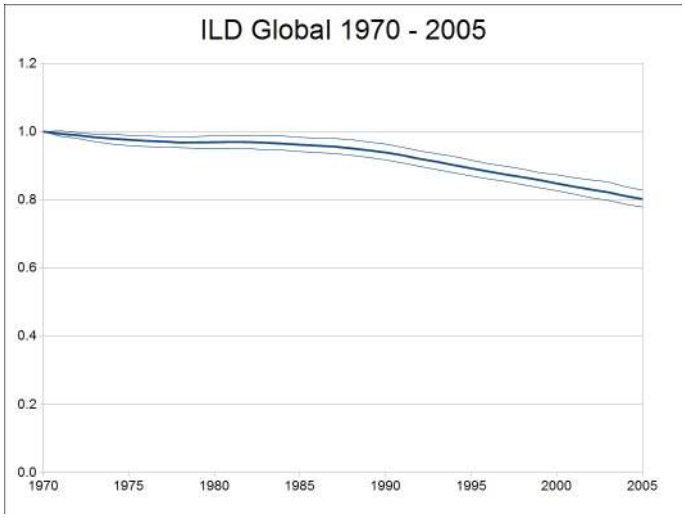


FIGURE 1: ILD Global, 1970–2005.
In Figures 1–7, The upper and lower confidence limits (CLs), showing the boundaries of the 95% confidence interval, are depicted as small lines above and below the main trendline.

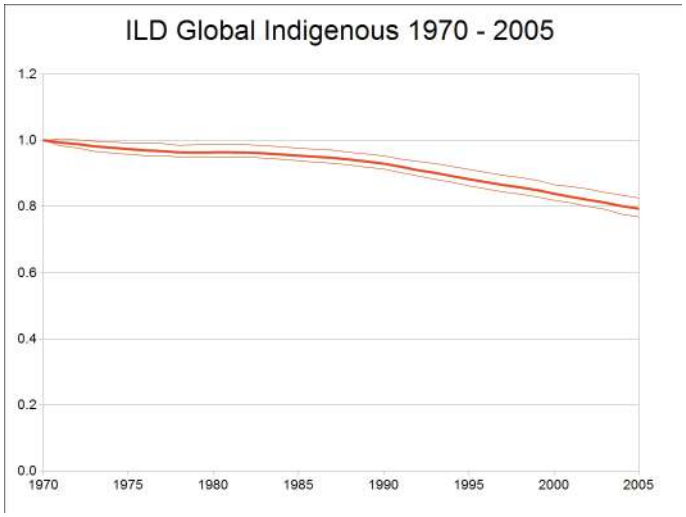


FIGURE 2: ILD Global Indigenous, 1970–2005.

2. WHAT IS LINGUISTIC DIVERSITY? Linguistic diversity is often viewed from three related (but not necessarily correlated) perspectives; this is the approach taken, for instance, by Daniel Nettle (1999). The first is what he calls *language diversity,* and we will call *language richness,* "the number of different languages in a given geographical area" (Nettle 1999:10). The term "language richness" encapsulates two points: first, that speech forms can be and routinely are classified as discrete languages, despite the well-known difficulties of distinguishing languages from dialects; and second, that these discrete languages are countable.

Another perspective on linguistic diversity is that of *phylogenetic diversity,* or variation above the level of languages, such as "the number of different lineages of languages found in an area." Nettle notes that phylogenetic groupings can be identified on many levels—language families, for example (Nettle 1999:10, 115). An area where many closely related languages are spoken therefore has greater language richness but less phylogenetic diversity than one with fewer languages belonging to several different families. The third perspective often used is *structural diversity,* which is the variation found among structures within languages, such as morphology, word order, phonology, and so on (Nettle 1999:130–148).

For the purposes of developing a quantitative measure such as the ILD, we depart slightly from the definitions of linguistic diversity outlined above, and borrow some related concepts from the field of ecology. Language richness can be thought of as being analogous to species richness, the number of species found in a given area. In addition to richness, a second component in species diversity is evenness, or the distribution of individual organisms among species. In the case of linguistic diversity, evenness is the distribution of individual speakers among languages. For example, two regions in both of which ten languages are spoken each have the same richness, but the region in which each language is spoken by 10% of the population has greater evenness, and therefore higher linguistic diversity, than one in which 91% of the population speaks one language and only 1% of the population speaks each of the other nine. We think that this concept is critical in measuring changes in linguistic diversity over comparatively short time scales. Relatively few of the world's languages have become extinct as mother tongues in the last few decades, so language richness in most areas of the world has declined only slightly. And yet, we would argue, diversity has declined much more than this because the distribution of mother-tongue speakers among extant languages has become more uneven: more speakers are becoming concentrated in fewer languages. While phylogenetic and structural diversity are important, these concepts are not currently incorporated into the index. In summary, for the purposes of the ILD, we define linguistic diversity as the number of languages and the evenness of distribution of mother-tongue speakers among languages in a given area.

3. THE NEED FOR A LINGUISTIC DIVERSITY INDEX. If there are already projections of the future magnitude of language extinctions, why is there a need for an index like the ILD? First, published estimates of the percentage of languages likely to die out during this century are, to date, little more than informed conjecture. Categorical statements of the

rate of extinction—"X number of languages are dying every year"—are widely quoted but almost never referenced to a rigorous estimate.[3]

Second, even if better estimates were available, merely tracking when particular languages go extinct does not account for the loss of linguistic diversity occurring during the course of pre-extinction language shift. A great deal of linguistic diversity is lost well before a declining language finally goes extinct, as speakers shift to other (usually larger) languages, intergenerational transmission declines, and usage becomes restricted to fewer speakers, domains, and functions. Quantifying changing distributions of mother-tongue speakers prior to extinction is therefore important.

Moreover, focusing on language extinction rates places undue emphasis on what is perceived to be the terminal state of linguistic diversity decline. If "language extinction" is to have any useful meaning, it must be specified that the term actually refers to the condition of a language no longer being spoken as a mother tongue. While there are several possible definitions of "mother tongue,"[4] what we mean by the term is that language which an individual would speak first (though not necessarily exclusively) if given free rein to choose. The term "first language" as used in *Ethnologue* (see Lewis 2009:13) captures the essence of what we mean. For the purposes of constructing the ILD, we assume that even multilingual people can have only one mother tongue.

Moreover, many languages, extinct as mother tongues in the sense just defined, continue to be spoken in everyday use as second languages or in one or more select domains (e.g., at home, as part of ceremonies, etc.). A language that is extinct as a mother tongue may live on as a *language of heritage* and, in some cases, might one day be revived as a mother tongue. Popular accounts usually gloss over or omit the fact that, with reference to language, the extinction metaphor does not necessarily imply absolute irreversibility.[5]

---

[3] It appears that many estimates originate in speculations made by Michael Krauss in his seminal 1992 *Language* paper, in which he said it is conceivable that as many as 90% of the world's languages could become extinct or irreversibly moribund by the end of the 21st century, and speculated that 50% of the world's 6,000 languages (his consensus figure of that time) were already moribund (Krauss 1992:6–7). Crystal (2000:19) notes this, and then proceeds to work through the math, deriving an estimate of 26 extinctions per year by extrapolating Krauss' estimate (6,000 languages, 50% loss over the next 100 years). This or a similar calculation appears to be the basis for statements such as that of the Living Tongues Institute for Endangered Languages, which says on its web site: "Every two weeks the last fluent speaker of a language passes on and with him/her goes literally hundreds of generations of traditional knowledge encoded in these ancestral tongues" (http://www.livingtongues.org/index.html; accessed June 2009). Similar statements can be found in journalistic accounts. Summarizing languages according to an endangerment typology (e.g., Krauss 2006; UNESCO 2009a) holds promise for a more accurate projection of likely extinctions; for an example see Table 10.2 in Evans 2010, and the accompanying discussion (pp. 211–216).

[4] See the discussion in Skutnabb-Kangas 2000:105–115 and the commentary thereon in Harmon 2002:56–58; see also Gunnemark and Kenrick 1985:242.

[5] For a somewhat different definition of "heritage language," see Golla 2007:8–9. In addition, a small number of languages, which we refer to as "auxiliary languages," were never spoken as mother tongues but instead always restricted to a particular domain. For these auxiliary languages, the term "extinction" simply refers to their no longer being spoken at all.

So, while obtaining accurate projections of mother-tongue language extinctions is important, they need to be augmented by a quantitative measure of current global trends in linguistic diversity. Clearly, the claims of those who tout the loss of linguistic diversity as a major problem for the world would be strengthened if there were quantitative evidence to support their arguments. Government officials, other decision-makers, and the general public will likely take the decline of linguistic diversity more seriously if there is a readily understandable global metric that captures the current magnitude of the problem. That is what the ILD is designed to provide.

4. WHAT THE ILD MEASURES. As stated earlier, the ILD uses language evenness in conjunction with language richness as a proxy for linguistic diversity. Because the goal of the index is to measure trends in linguistic diversity, it must account for changes in evenness and richness: that is, changes in the relative distribution of mother-tongue speakers among discrete languages within the total population, as measured from the starting point of the index to its ending point. The ILD indicates the rate of change in linguistic diversity by measuring how far, on average, the languages in a given grouping deviate from a hypothetical situation in which each language is neither increasing nor decreasing its share of the total population of that grouping.

To illustrate this, let us look at ILD Global, with measures the world's overall linguistic diversity. ILD Global tracks the trend in the world's linguistic diversity since 1970, the earliest year for which sufficient data are available to calculate the index. The index value is set equal to 1 in the baseline year, and in each subsequent year shows the trend in the share of the world population represented by the average[6] of all the languages in the sample relative to the baseline year. If the average is declining, it means that the distribution is becoming less even (i.e., more skewed), with a few large languages increasing their global share at the expense of many smaller languages. If the average is increasing, it means that the distribution is becoming more even, with many languages increasing their share at the expense of a few large languages. If somehow each language could maintain its initial proportion, the ILD Global trendline would be flat. Any increases and decreases in the index can also be thought of as changes in the relative abundance of the world's languages: a rising trendline means more people are shifting away from dominant languages to minority languages, while a falling trendline means more people are shifting to majority languages and away from minority ones.

To calculate an ILD for languages spoken in a given population, we track the proportion of the total population speaking each language in each year, and then take the average. The index measures how that average changes over time. Thus the ILD can be said to measure the concentration or distribution of mother-tongue speakers among the world's languages. What does it mean to say that ILD Global declined 20% over the period 1970–2005? It means that, for all languages spoken worldwide in 1970, their average share of the world's population declined by 20% over 35 years. (Appendix B contains a technical discussion of how the ILD tracks changes in the average share, and also provides some

[6] The average is calculated using the geometric mean rather than the arithmetic mean. See Appendix B for a more detailed explanation.

simple comparative scenarios with accompanying graphics that make clear how the ILD changes under different conditions of language shift.)

Although we have used ILD Global as an example, the same methods and reasoning apply to subglobal ILDs. For instance, ILD Americas measures the trend in the share of the population of the Americas represented by the average of all the region's languages in the sample relative to the baseline year.

The ILD is entirely retrospective, indicating changes that have taken place in the past, and is not designed to predict future changes. It is not a measure of the future viability of any one language or group of languages. Rather, it provides a snapshot of the trends in the distribution of speakers among the world's languages between the starting year of the index (1970) and the final year (2005 in the current version).

In the ILD each language carries equal weight, regardless of its relative size. While it is possible to produce a weighted index that would impart more importance to phylogenetic diversity—say, by giving extra weight to isolates—such weightings are always more or less arbitrary. Making the ILD be unweighted means that the phylogenetic uniqueness of any particular language does not differentially affect the calculation of the index. Neither does the mode of shift that any particular language may be undergoing, so that, for example, language attrition caused by rural-to-urban migration is, in terms of its effect on the ILD, no different than attrition caused by intergenerational transmission failure within a geographically homogeneous speech community, or language loss caused by a catastrophic decline of a community of speakers. This means the ILD is not useful for illuminating the sociolinguistic bases of language shift.

Finally, it is worth noting again that the ILD is not a measure of language extinction: a 10% decline in the index does not mean that 10% of languages went extinct over the period being measured. For example, it is possible that most of the world's languages could decline until only a few speakers of each are left, while a few languages become dominant with many millions of speakers: the ILD would show a marked decline and yet the total number of extant languages would remain constant. In that case the number of extinctions would remain zero, yet the ILD would indicate that almost all linguistic diversity had been lost.

5. THE ILD DATABASE. The ILD is based on a sample of 1,500 languages selected at random from the 7,299 languages listed in the 15th edition of *Ethnologue* (Gordon 2005). (The 16th edition, Lewis 2009, appeared too late for us to include in this study.) This sample size—representing just over 20% of the world's languages—is higher than is needed to constitute a statistically representative global sample. Having a sample size much larger than required for global analysis allows statistically valid analysis of subglobal samples. A larger-than-needed sample size also provides a cushion against sample attrition.

Our long-term aim is to base the ILD on a variety of data sources, not just *Ethnologue*. However, we decided to restrict the first version of the ILD to *Ethnologue* data to minimize potential inconsistencies in language-status assessment that could come from incorporating multiple sources of data into a single time series. Thanks to the assistance of M. Paul Lewis of SIL International, we were able to obtain copies of some extremely rare early editions, which allowed us to complete a collection of all 15 editions available at the time of analysis. This enabled us to move the ILD's starting date further back than

we initially anticipated. After reviewing all the editions, we selected the following nine on which to base the initial version of the ILD: 1st (WBT 1951), 5th (Canonge and Pittman n.d. [ca. 1958]), 9th (B.F. Grimes 1978), 10th (B.F. Grimes 1984), 11th (B.F. Grimes 1988); 12th (B.F. Grimes 1992a), 13th (B.F. Grimes 1996a), 14th (B.F. Grimes 2000a), and 15th (Gordon 2005). We chose these editions because (at the time) they spanned the entire history of *Ethnologue* while giving priority to later editions whose contents are much more comprehensive.

Our first step was to enter benchmark demographic information from the 15th edition of *Ethnologue* (Gordon 2005) into a Base Data Entry Form. Next, we reviewed the nine editions listed above looking for data on the number of mother-tongue speakers for our sample languages.[7] Within the time available to us we were able to examine six of the editions (1st, 5th, 9th, 12th, 14th, and 15th) for data on the full sample of 1,500 languages. For the remaining three editions (10th, 11th, and 13th), we were able to search for the 751 languages in our sample from Africa and the Americas only. Thus, we performed a total of 11,253 data searches. After eliminating duplicates, we were left with 2,703 unique datapoints; these form the basis for the first iteration of the ILD.

Our protocol was to enter results for each of these data searches into the database using a Mother-Tongue Speaker Trend Data Form, even though the vast majority of them did not produce unique datapoints. Doing this ensures that there is no ambiguity about whether a particular data source has been consulted with regard to any given language.

Before creating the ILD we analyzed the data for representativeness, eliminated duplicate datapoints and entries having no data, assessed and adjusted for data trend anomalies, removed discrepant datapoints, and dealt with apparent extinctions within time series. Technical detail about these steps, along with samples of the Base Data Entry Form and the Mother-Tongue Speaker Trend Data Form (and explanations of the fields in these forms) and a discussion of other points related to the creation of the database and data analysis, are in Appendix A. Readers who wish to review the data can do so at http://www.terralingua.org/projects/ild/ild.htm.

6. CALCULATING THE ILD. The following account describes a simplified method for calculating the ILD that requires data on each language for each year included in the index.

---

[7] As noted above, there are a small number of auxiliary languages that have never had any mother-tongue speakers. Five of these happened to fall within our sample (The 3-letter codes in square brackets are the languages' ISO 639-3 codes; for more, see Appendix A.): Amerax [aex], reputed to be spoken only as a second language by neo-Muslims in American prisons (and which, incidentally, is no longer listed in the 16th edition of *Ethnologue*); To [toz], an ancient secret male initiation language of the Gbaya people of Cameroon; Lucumi [luq], a secret language used for ritual by the Santeria religion; Yinglish [yib], a blend of Yiddish [yid] and English [eng] that is used as a second language only; and Europanto [eur], an artificial language mixing elements from major European languages, which is spoken in the European Union buildings in Belgium. Because these auxiliary languages do not currently have any mother-tongue speakers we excluded them from the calculation of the ILD. They could, conceivably, gain mother-tongue speakers in the future, as has the best-known intentionally constructed auxiliary language, Esperanto [esp], which now has 200–2,000 mother-tongue speakers (Gordon 2005).

For a detailed explanation of the method needed when there are gaps in the data, as with numbers of speakers, refer to Appendix B. The method has three steps that remain the same whether the global level or a regional grouping is being analyzed:

The fraction *F* of the total population (global or regional) represented by each datapoint (*N* speakers of language *l* in year *y*) was calculated.

$$F_{ly} = N_{ly}/P_y$$

where

N_{ly} is the number of speakers of language *l* in year *y*, and
P_y is the total population in year *y*.

The total populations from 1950 to 2005 of the world and five regions—Africa, Asia, Pacific, Europe, and the Americas—were taken from UN Population Division (2006 revision), downloaded from http://esa.un.org/unpp/index.asp.

The geometric mean of the *F* values in each year was calculated:

$$M = (F_1.F_2.F_3...F_n)^{1/n}$$

where

n = total number of languages.

Finally, the geometric means in each year were chained together to form an index, such that:

$$I_y = I_{y-1}(M_y/M_{y-1})$$

where

I_y = the Index of Linguistic Diversity in year *y*
M_y = the geometric mean *F* value in year *y*, and
M_{y-1} = the geometric mean *F* value the previous year

and the index value in 1970 was set to unity

$$I_{1970} = 1.0$$

In this way, the ILD shows the trend in the fraction of the total population that speaks a language that is average or typical of all languages in the sample.

7. RESULTS.

Global Linguistic Diversity. ILD Global (Figure 1), which covers all the languages in the sample, both indigenous and non-indigenous, shows a slow decline from 1.0 to 0.95 between 1970 and 1988, but a steeper decline from 0.95 to 0.80 between 1988 and 2005. The upper and lower confidence limits (CLs) show the boundaries of the 95% confidence interval, and are depicted in this and the other graphs as small lines above and below the main trendline.[8]

Global Indigenous Linguistic Diversity. ILD Global Indigenous (Figure 2), which covers only the indigenous languages in the sample, declined from 1.0 to 0.94 between 1970 and 1988, and from 0.94 to 0.79 between 1988 and 2005. It shows a marginally greater decline than the global ILD, but the two trends are largely similar as most of the languages in the global dataset are indigenous languages (see Appendix A for discussion).

Regional Indigenous Linguistic Diversity. Changes in indigenous linguistic diversity differ among regions. ILD Africa Indigenous increased from 1.00 to 1.07 between 1970 and 1985, and then declined rapidly from 1.07 to 0.83 in 2005 (Figure 3). The increase in the 1970s and early 1980s suggests that African indigenous languages were becoming more equally distributed in terms of speaker numbers during that period, but from the mid-1980s on the distribution became increasingly skewed, with many languages' share of the total African population declining.

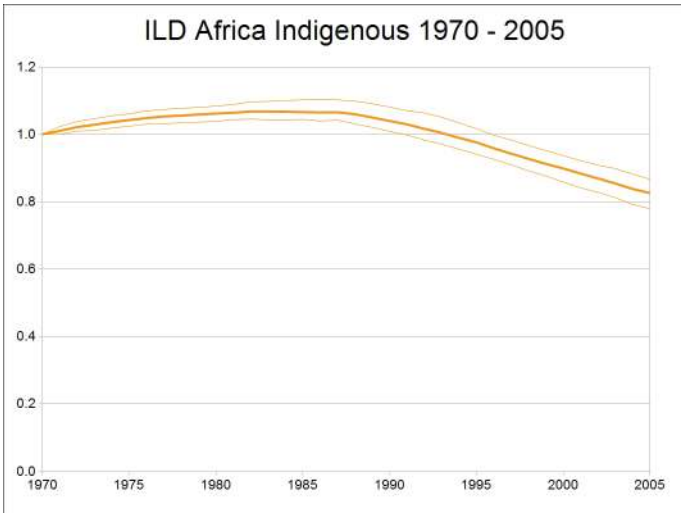


FIGURE 3: ILD Africa Indigenous, 1970–2005.

[8] Confidence limits were calculated by bootstrapping with 1,000 bootstraps.

ILD Americas Indigenous shows the steepest decline of any region, falling from 1.00 to 0.71 between 1970 and 1980, and from 0.71 to 0.36 between 1980 and 2005 (Figure 4).

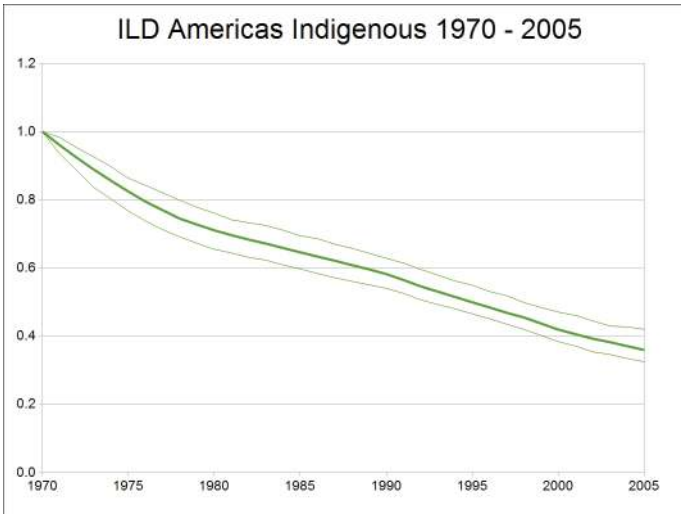


FIGURE 4: ILD Americas Indigenous, 1970–2005.

ILD Eurasia Indigenous, like its African counterpart, showed an initial increase from 1.00 to 1.10 between 1970 and 1981, suggesting that there was a slight gain in the proportion of the total population speaking an indigenous language. It flattened out for about a decade between 1981 and 1991, and then declined very slightly to 1.07 in 2005 (Figure 5). Overall the index shows little change in linguistic diversity in Eurasia.

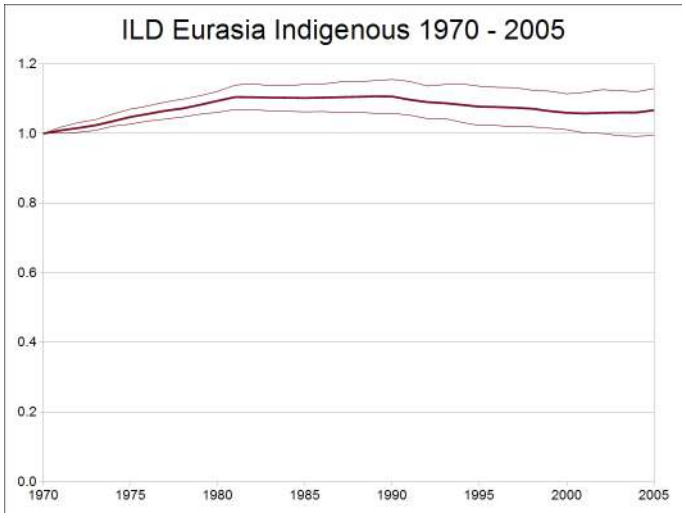


FIGURE 5: ILD Eurasia Indigenous, 1970–2005.

ILD Pacific Indigenous (which includes Australia) shows the second steepest decline after the Americas. The index fell steadily from 1.0 to 0.82 in 1999, then dropped steeply from 0.82 to 0.70 between 1999 and 2005 (Figure 6). The widening confidence intervals in the last few years of the index suggest a higher degree of uncertainty in the trend after 1999, which would be reduced with additional data.

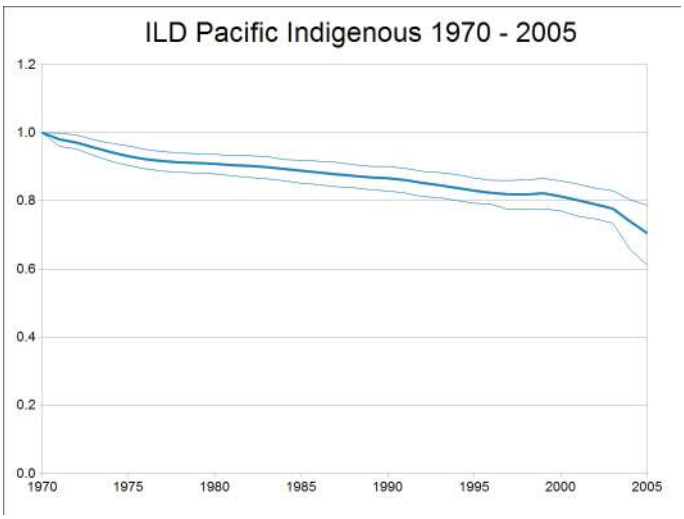


FIGURE 6: ILD Pacific Indigenous, 1970–2005.

Because the linguistic situation in Australia is distinctive within in the Pacific region, ILD Australia Indigenous (Figure 7) shows a national ILD for Australian Aboriginal/Torres Strait Islander languages alone. The graph includes a second, non-*Ethnologue* data source for comparison: the Australian Bureau of Statistics (ABS). Data from the two sources are shown both separately and combined. ABS data from on numbers of speakers of aboriginal languages from 1996 to 2006 were used to compare trends derived from these data (blue line) with those derived from *Ethnologue* data (red line). The *Ethnologue* data show a decline from 1.0 to 0.7 between 1970 and 1991, then a faster decline from 0.70 to 0.38 between 1991 and 2005. The ABS data show a decline from 1.0 in 1996 to 0.87 in 2006, which is similar to the rate of the *Ethnologue*-based ILD from 1970 to 1991. Combining data from both sources results in an index that declines from 1.0 to 0.53 between 1970 and 2006. Whichever data source[9] is used, Australia shows a more rapid loss of linguistic diversity than the rest of the Pacific region. The rate of loss is comparable to that of the Americas.

[9] As acknowledged by the compilers of *Ethnologue*, the data for Australia in editions immediately preceding the current one (i.e., the 16th, published in 2009) were quite out-of-date (M. Paul Lewis, pers. comm., 25 May 2009), and some linguists have raised concerns about the accuracy of ABS data. Nonetheless, the overall point—that Aboriginal languages are in sharp decline—does not seem to be dispute.

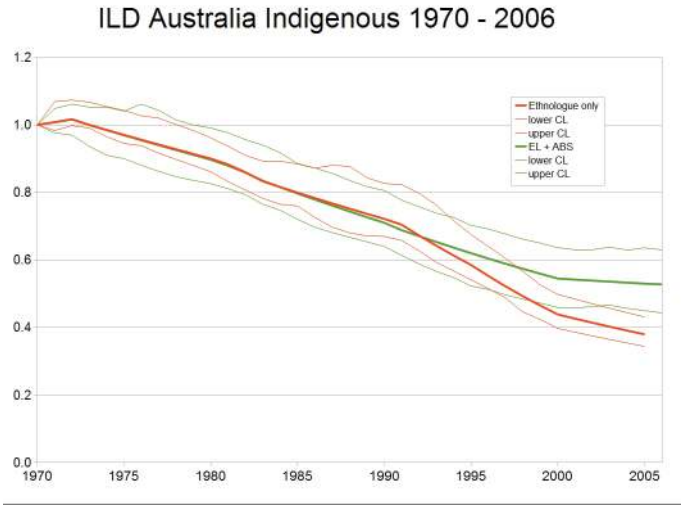


FIGURE 7: ILD Australia Indigenous, 1970–2005, and ABS Data, 1996–2006.

Figure 8 shows the four regional ILDs in one chart for ease of comparison.

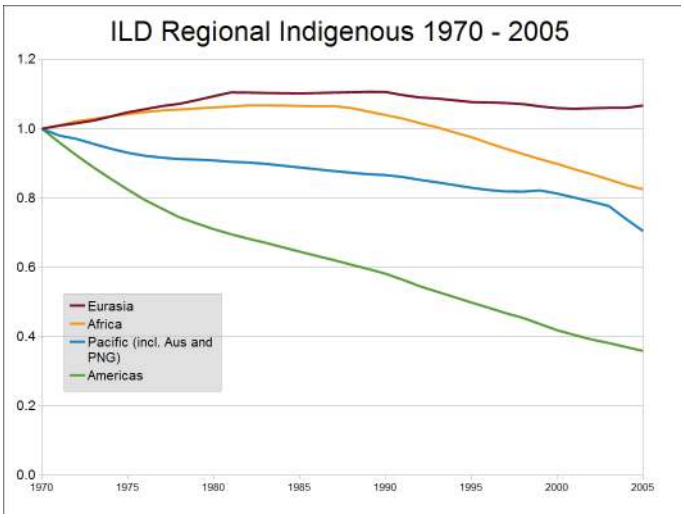


FIGURE 8: Regional Indigenous ILDs, 1970–2005.

8. DISCUSSION.

Decline in Global Linguistic Diversity. Figure 1 shows the global trendline for the ILD. ILD Global shows a slow decline from 1.0 to 0.95 between 1970 and 1988, but a steeper decline from 0.95 to 0.80[10] between 1988 and 2005. The overall decline of 20% in the space of 35 years shows that linguistic diversity is being lost at a significant rate, but even more importantly, the rate of loss has increased from about – 0.3% per year in the 1970s and 1980s to more than –1.0% per year in the 1990s and 2000s. This is a stark indication of the scale of the recent loss of global linguistic diversity. The rapid disappearance of one-fifth of the linguistic diversity that existed in the world in 1970 is a quantitative depiction of the continuing widespread shift from smaller languages to larger languages. The more the ILD Global declines, the more the world's mother-tongue speakers are concentrated into fewer languages.

Decline in Global Indigineous Linguistic Diversity. Figure 2 shows that the decline in the diversity of the world's indigenous languages has been similar, which is unsurprising in that most of the languages in the world (by our estimate, 80–85%) are indigenous languages. (See Appendix A for discussion.) ILD Global Indigenous declined from 1.0 to 0.79 between 1970 and 2005—a 21% decrease. The average annual rate of decline in indigenous linguistic diversity was slightly faster than the global average in the 1970s and 1980s, but only by a fraction of a percent per year.

Making judgments about whether particular languages are to be considered "indigenous" can be difficult, and this problem is discussed further in Appendix A. Suffice it to say here that it is indeed important to make such judgments. Indigenous communities themselves certainly want to know the status of indigenous languages; see, for example, the documents associated with the International Expert Group Meeting on Indigenous Languages (UNPFII 2008).

Moreover, the Convention on Biological Diversity has identified stemming the rate of loss of linguistic diversity and in the number of speakers of indigenous languages as one its indicators for assessing progress toward meeting its 2010 Biodiversity Target. The acceleration in the loss of linguistic diversity indicated by the ILD Global Indigenous implies that this particular CBD target will not be met.

Declines in Regional Indigenous Linguistic Diversity. A comparison of the various regional indigenous ILDs (Figure 8) shows some interesting results. Some regions are declining more rapidly than others, particularly the Americas, which declined by 64% over the period (Figure 4). The fact that the Americas showed the greatest overall decline should not necessarily be interpreted as meaning that linguistic diversity is, consequently, lower there than in other regions. It simply means that the Americas underwent the most rapid decline of all four regions between 1970 and 2005. It may well have been the case that the

[10] It has been suggested to us that this pattern may indicate, at a "macro" level, the phenomenon of what Nancy Dorian has called "abrupt transmission failure" or "tip," in contrast to "gradual shift." Dorian wrote: "In terms of possible routes toward language death, it would seem that a language which has been demographically highly stable for several centuries may experience a sudden 'tip,' after which the demographic tide flows strongly in favor of some other language" (1981:51).

Americas were much more linguistically diverse in 1970 compared with other regions, such as Europe for example, in which the majority of linguistic diversity was lost prior to 1970.

The Pacific region (Figure 6) shows the second greatest rate of decline, 30% over 35 years, while ILD Africa Indigenous (Figure 3) declined by nearly 20%. This suggests that indigenous languages are in very rapid decline in comparison to total population growth in the region as a whole in the Americas, and in rapid decline in Africa and the Pacific.

Eurasia was the only region to show an increase in its indigenous ILD (Figure 5). There, indigenous languages are growing at the same rate as the overall population.[11]

In addition to the regional analyses, we calculated a national ILD for indigenous Australian languages (Figure 7). We did this in two ways: first based on the data for the 20 Australian languages in the ILD database, and then using additional data from the Australian Bureau of Statistics. The ABS data are based on censuses conducted in 1996 and 2006, and show an average decline of 13% over ten years across 45 languages for which there were two datapoints. This gives a yardstick of trends in a relatively well-monitored group of indigenous languages with which to compare the trends reported in *Ethnologue* (over the last decade of the index at least). ILD Australia Indigenous, based on *Ethnologue* data only, showed a decline of over 60%, but with the addition of the ABS data from 1996 to 2006, this decline was reduced to less than 50% (Figure 7). Nevertheless, this reflects a severe and rapid loss of linguistic diversity in Australia since the 1970s.

There are aspects of these results which may change with further analysis. Africa, Asia, and Europe show increases in diversity in the 1970s and 1980s. These increases are possibly an artifact of some *Ethnologue* data which do not reflect genuine changes in diversity. Some of these data anomalies may be discovered with additional scrutiny of the dataset.

Starting Point of the Index. Another consideration of an index based on numbers of speakers is when to fix the initial starting point. A flat trendline describes a state in which richness is being maintained (i.e., not being lost) in relative chronological terms; that is, relative to the starting year of the analysis. It is important to understand that the initial starting point (in this case 1970) does not describe a maximal state of linguistic diversity in absolute terms. For any set of languages spoken in a given region, maximum diversity is reached when each language has an equal number of speakers. The starting point of the ILD for any given region is highly unlikely to be maximal. Qualitative estimates point toward global linguistic richness having reached its peak thousands of years ago, long before there were any quantitative data by which to measure it.

Ideally, the ILD's starting point should be as early as possible. *Ethnologue* has sufficient quantitative data to set the starting point at 1970, but prior to 1970 there are not enough datapoints from which to derive global numerical trends. Yet as just noted the

[11] We combined Europe and Asia into a single Eurasian region because the sample size in Europe is so small. In Europe, indigenous languages in our sample are in moderate decline in comparison to Europe's total population, but the sample size is not large enough to enable us to draw significant conclusions. Therefore, the ILD Eurasian Indigenous trendline can mostly be attributed to indigenous languages in Asia growing at the same rate as the overall population in Asia.

global peak in linguistic diversity was reached centuries before 1970. That context—the knowledge that most of the world's linguistic diversity was lost before we were even able to start measuring it—must always be kept in mind when interpreting the ILD.

Data Quality: Is the ILD Valid? *Ethnologue* is widely recognized as the most authoritative source of information on the number of speakers of the world's languages. In 1992, at the dawn of concern over language endangerment, linguist Michael Krauss called it "by the far the best single source available" on the number of languages and their speakers globally (Krauss 1992:4, n1). That assessment has not changed: in 2007, the editor of the *Encyclopedia of the World's Endangered Languages* referred to it as the "most comprehensive compendium of the world's languages that has yet been produced..." (Moseley 2007b:x).

Nonetheless, these experts—and *Ethnologue's* compilers themselves—also acknowledge that the quality of its data is uneven. *Ethnologue* draws its speaker data from a wide variety of sources, "everything from popular reference books to missionary field reports to specialist monographs by professional linguists" (Harmon 1995:12). Even data taken from government censuses, which might be taken as reliable on their face, are in fact often inaccurate when it comes to reporting language statistics (see, for example, Voegelin and Voegelin 1977:8; Garza Cuarón and Lastra 1991:94, 96; and esp. Skutnabb-Kangas 2000:30–32, and the cites thereunder). The problem is underlined by the current *Ethnologue* editor, M. Paul Lewis, who writes that calculating the number of speakers "is probably the most difficult component of the language information for us to stay on top of" (Lewis, pers. comm., 25 May 2009).[12]

Given these difficulties, it is reasonable to ask whether the underlying data are so inaccurate as to make the ILD (or any time-series linguistic diversity index) invalid. We think that the answer is no, for several reasons:

There is no reason to think that there is a systematic bias towards either overcounting or undercounting the number of mother-tongue speakers within the *Ethnologue* dataset upon which the ILD is based. There are many reasons why a particular datapoint may be an overcount (e.g., the enumerator simply reported, without investigation, the entire ethnic group as mother-tongue speakers) or an undercount (e.g., the enumerator was unaware of the existence of additional mother-tongue speakers elsewhere). We are not aware of any evidence that shows one of these types of error being more prevalent than the other within *Ethnologue* (or any other data source that we have consulted to date). Indeed, one might instead argue that precisely because there has been no systematic means for counting mother-tongue speakers, there is no reason to think that the results are systematically biased one way or the other. If it could be shown that enumerator errors consistently tended (or are likely to tend) toward either overcounting or undercounting, then the ILD trendline would indeed be invalid (absent statistical adjustment to correct for the bias). If there is no systematic bias one way or the other, then—given a large enough sample size, which we believe ours is—it is reasonable to assume that instances of overcounting and undercounting would, on average, cancel each other out.

---

[12] For further discussion of *Ethnologue* data quality and its ramifications for trend analyses, see Harmon 1995.

This caveat about sample size is important. The ILD methodology is designed to measure average trends in large groups of languages. This means that inaccuracies in the time series for any one language cannot unduly affect the overall trendline, for the reasons just given. There is no inherent reason why the ILD methodology could not be applied to small groups of languages, but the results would be valid only if it could assured that the data were gathered in a consistent way.

The ILD Global trendline aligns with the large and convincing body of qualitative evidence pointing to a decline in linguistic diversity. A decline of 20% over the period 1970–2005 is an entirely plausible outcome in view of this evidence. Had the ILD shown, say, an increase over the period, or a precipitous global decline, then that would be *prima facie* evidence calling into question the accuracy of the underlying data.

The ILD is premised on there being a one-to-one equivalence between the cumulative number of mother-tongue speakers of the world's languages and the global population; in other words, on the assumption that each person can have only one mother tongue. This premise depends for its validity on a precisely specified definition of "mother tongue"; as discussed earlier, our definition is "that language an individual would speak most often if given free rein to choose." Under this definition, even multilingual people can have only one mother tongue. With this in mind, let us imagine what a perfect global census of the number of mother-tongue speakers would look like:

- The census would be a true temporal snapshot, taking place worldwide over a very short time; say, a single day.
- It would query every single person in the world.
- Each census-taker would have exactly the same understanding of our definition of "mother tongue," and would have the ability to explain it with such fidelity that every respondent would have a 100% identical understanding of our meaning of the term.
- Every multilingual respondent would be able (and willing) to prioritize among his/her languages as to which one is his/her single mother tongue according to our "first preference" definition.
- Every respondent would feel free to answer truthfully, without fear of political, social, economic, or other repercussions, and would indeed answer truthfully.
- The census would be replicated at regular intervals to produce accurate time-series data.

The fact that our imaginary census is obviously unattainable does not mean that the less-than-perfect numbers available to us have no value. For example, an analysis of the 1992 *Ethnologue* data found a reasonably close correspondence between the cumulative number of mother-tongue speakers and the global population at that time (Harmon 1995:12–13). This strongly suggests an underlying plausibility in the *Ethnologue* speaker-totals data.

In the final analysis, it is always possible to dismiss quantitative assessments of complex global phenomena by claiming that the data aren't good enough—or to categorically rule out any quantitative representation of such phenomena on ideological or philosophical

grounds. Accepting such criticisms, however, leaves us in the position of likely never being able to say anything very precise about the global status of linguistic diversity.

Other Caveats and Limitations. In the course of developing the ILD, we had the opportunity to present it as a work-in-progress at two international meetings and in a variety of informal discussions with colleagues. In those exchanges, several points emerged repeatedly that are worth sharing here:

- Many people are skeptical of the validity and usefulness of global indices, often because they don't understand their purpose. The technical basis—and inherent limitations—of such indices not only must be carefully explained, but potential political misuses must be acknowledged and warned against.
- Key concepts that underlie the ILD, such as "language extinction" and "mother-tongue speaker," are nuanced and must be carefully qualified.
- Quantitative indicators are a complement to, not a substitute for, in-depth qualitative knowledge of linguistic and cultural diversity.
- Virtually all indigenous peoples who care about the continuation of their traditional culture believe that maintaining their native language is the linchpin (cf. UNDESA 2009:57–59).

While we expect the ILD to prove a useful tool to communities, analysts and academics, policymakers, and the general public, any index is only as good as the underlying data available at the time. *Ethnologue* is the best single source for data on the numbers of speakers of languages around the world, and information from its various editions is an indispensable part of any analysis of recent trends in language demographics. Nonetheless, as we discussed above, *Ethnologue* data come from a variety of primary and secondary sources and are, inevitably, uneven. We believe that *Ethnologue* time-series data are valid, but without question language demographic data in general can be improved. It should be borne in mind when using the initial version of the ILD that better data will, in the future, produce even more accurate trendlines.

It is also important to acknowledge that global indices such as the ILD should be used to provide broad contextual background for policy frameworks, rather than as guidance for on-the-ground policy decisions. No large-scale language index can hope to fully represent the complexities that must be accounted for in any policy affecting individual language communities. Nor can a global or regional index do more than outline the state of linguistic diversity at these levels; much more fine-grained analyses are required to get a complete picture.

As suggested above, quantitative analyses such as the ILD must be supplemented by knowledge derived through other methods. This is especially relevant with respect to languages because most linguistic diversity is tied to traditional knowledge systems of indigenous people. These systems primarily rely on non-quantitative observational science and narrative, often transmitted orally rather than in writing. Therefore, any global numerical index, including the ILD, runs the risk of being irrelevant (or, worse, antithetical) to the needs of indigenous communities if it is not properly qualified as noted above—and, in

addition, supplemented by other information that is generated by the communities themselves.

In short, the ILD and similar global indices that deal with potentially controversial phenomena, such as language policy, must carefully be placed in context whenever they are used as an educational or policy-orientation tool, and should never be used as a sole source of information.

Future Development of the ILD. As part of future work, we plan to add data from the 16th edition of *Ethnologue* (Lewis 2009) for our 1,500 sample languages; in fact, we would like to expand the database to achieve complete coverage of all the world's languages. We also hope to be able to enter into the ILD database all available speaker-numbers data from other global compendia of language statistics, such as Voegelin and Voegelin 1977, the series of monographs produced under the editorship of T. Sebeok in the 1960s and 1970s, and the recent *Encyclopedia of the World's Endangered Languages* (Moseley 2007a, and citations thereunder listed in the References), as well as information from UNESCO's *Atlas of the World's Languages in Danger* (UNESCO 2009b) and other UNESCO-led data-gathering efforts. All of these will provide data with which to compare, or add to, those from *Ethnologue*.

But the full potential of the ILD methodology won't be realized until we are able to expand it to include language demographic data in addition to counts of mother-tongue speakers. To fully understand the status of and trends in the world's linguistic diversity, we need to go beyond using language richness (the number of discrete languages) and language distribution as a proxy—although that is where we, of necessity, have had to begin our work with the ILD. For example, it may be possible to create versions of the ILD that address phylogenetic diversity by using data on language family affiliations that are already included in *Ethnologue*. The methodology could also be applied to certain special language categories, thus producing versions such as ILD Creoles or ILD Isolates. There may be scope for incorporating structural diversity into the ILD by drawing on data from the World Atlas of Language Structures (Haspelmath et al. 2005; http://wals.info). Even better understanding will come when we are able to augment speaker-numbers data with deeper knowledge about all the factors that determine language demographics and drive trends in linguistic diversity.

Appendix A. Technical discussion of the ILD database

ORIGINS. The ILD database has its origins in work done in the mid-1990s in which a shadow database of the 12th edition of *Ethnologue* (B.F. Grimes 1992a) was created and analyzed for demographic trends. The work involved entering into a FileMaker Pro database a variety of demographic information relevant to speaker trends and language viability on all 6,760 languages reported in the 12th edition. Each record represented a discrete language distinguished by a unique three-letter code assigned by *Ethnologue*. The information was used to produce a basic analysis of the demographic structure of the world's languages (Harmon 1995). The ILD database expands on the Harmon 1995 database and is organized on the same principle. It too is keyed to discrete languages as reported in *Ethnologue*: in this case, the 15th (Gordon 2005). It was in this edition that the unique three-letter language identifier codes assigned by the International Standards Organisation (ISO) first came into use.

STRUCTURE. The ILD database is structured around these ISO codes, which follow the ISO 639-3 standard. In terms of quality control, the ISO code is the most critical piece of information in the ILD database because it signifies a discrete language. A number of languages share the same name, many have variant names, and still others have self-names that are different from those that have become established in English (e.g., Magyar = Hungarian). ISO codes avoid confusion by assigning a unique three-letter code to each language that is considered discrete. In the 16th edition of *Ethnologue*, which appeared too late for use in calculating the initial ILD, ISO codes are also assigned to "macrolanguages," defined by ISO as "multiple, closely related individual languages that are deemed in some usage contexts to be a single language" (quoted in Lewis 2009:9). The 16th edition lists 55 such macrolanguages. Arabic [ara], Chinese [zho], Serbo-Croatian [hbs], and Kurdish [kur] are some prominent examples.

The purpose of ISO codes is the same as that of Linnean binomials for biological species: they serve to uniquely identify separate entities no matter what their vernacular names are in different languages. The ISO code is written in lowercase; when we refer to them in this paper, they appear in square brackets.

The ISO codes derive, in part, from *Ethnologue's* earlier proprietary three-letter codes, first published in the 10th edition (B.F. Grimes 1984), which were written in uppercase. As noted by the current *Ethnologue* editor, M. Paul Lewis, the "adoption of the ISO 639-3 standard both 'took over' the previously existing [*Ethnologue*] codes but also involved an alignment of those codes with the already existing [interim] ISO 639-2 code set." This resulted in some confusing re-assignments of the old *Ethnologue* codes to new languages under the ISO code set, "but the ongoing principle of the [ISO 639-3] standard that no code is ever re-assigned (henceforth) or re-used provides the immensely valuable benefit that we can now not only uniquely identify languages but also be able to trace their identification history (what they were split from, merged with, etc.) since all of the ISO codes retain their original denotations" (Lewis, pers. comm., 25 May 2009). *Ethnologue* now follows a prescribed process in which all changes to the language roster—whether additions, deletions, mergers, splits, or name changes—are recorded with ISO and published on the *Ethnologue* website.

The adoption of ISO 639-3 will go a long way toward ending confusion over language names. Moreover, henceforth it should be simple for anyone to see trace reclassifications in linguistic status (e.g., a dialect being elevated to consideration as a discrete language, or vice versa) made by *Ethnologue's* editors from edition to edition. These problems remain, of course, for those like us who wish to retrospectively analyze data from editions of *Ethnologue* prior to the 15th.

BUILDING THE DATABASE

The ILD database was built in a series of steps:

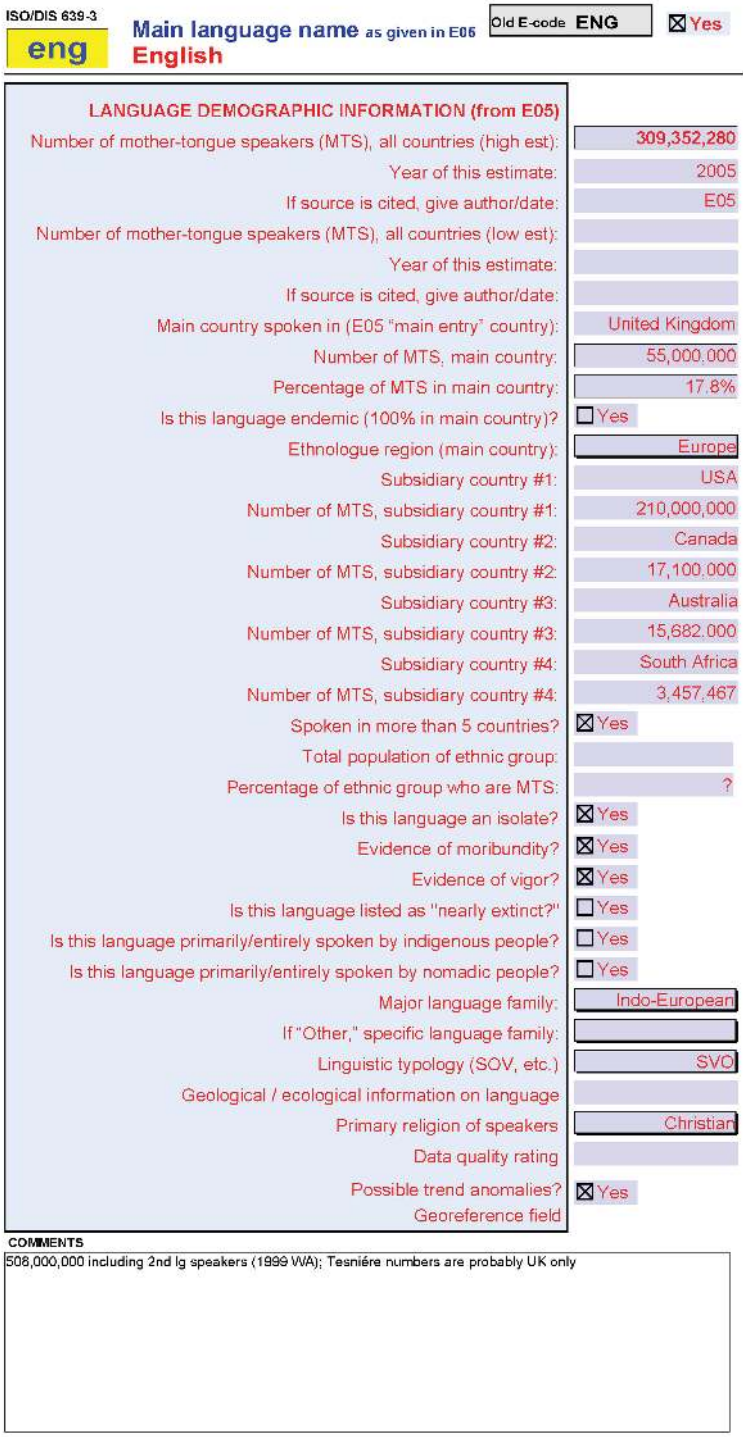
1. Select random sample;
2. Enter base demographic information from the 15th edition of *Ethnologue*;
3. Enter mother-tongue speaker numbers from earlier editions of *Ethnologue*;
4. Analyze sample for representativeness;
5. Eliminate duplicate datapoints;
6. Assess and adjust for possible data trend anomalies;
7. Remove discrepant datapoints; and
8. Deal with apparent extinctions within time series.

1. SELECTION OF RANDOM SAMPLE. The ILD is based on a random sample of 1,500 of the world's 7,299 languages. This sample size was chosen because we determined it to be the largest we could reasonably deal with over the period of project funding. A sample size of 1,500 is far higher than is needed to constitute a statistically representative global sample, but is also allows statistically valid analysis of subglobal samples and provides a cushion against sample attrition (more on this below).

To create the sample, we used the statistics program "R" to generate 1,500 random numbers between 1 and 7,299. An alphabetical list of the 7,299 ISO codes—[aaa] (Ghutuo) through [zyp] (Zyphe)—was imported from FileMaker Pro into an Excel spreadsheet, numbered consecutively, and then the random numbers matched to the ISO codes. The result was a random sample of 1,500 languages.

2. ENTRY OF BASE DEMOGRAPHIC INFORMATION. We extracted all demographic information from the 15th edition of *Ethnologue* for the 1,500 languages in our sample. Figure A-1 shows the form used to record the base demographic information; Table A-1 explains the fields in the form.

FIGURE A-1



ISO/DIS 639-3: eng
Main language name as given in E06: English
Old E-code: ENG ☒ Yes

LANGUAGE DEMOGRAPHIC INFORMATION (from E05)	
Number of mother-tongue speakers (MTS), all countries (high est):	309,352,280
Year of this estimate:	2005
If source is cited, give author/date:	E05
Number of mother-tongue speakers (MTS), all countries (low est):	
Year of this estimate:	
If source is cited, give author/date:	
Main country spoken in (E05 "main entry" country):	United Kingdom
Number of MTS, main country:	55,000,000
Percentage of MTS in main country:	17.8%
Is this language endemic (100% in main country)?	☐ Yes
Ethnologue region (main country):	Europe
Subsidiary country #1:	USA
Number of MTS, subsidiary country #1:	210,000,000
Subsidiary country #2:	Canada
Number of MTS, subsidiary country #2:	17,100,000
Subsidiary country #3:	Australia
Number of MTS, subsidiary country #3:	15,682,000
Subsidiary country #4:	South Africa
Number of MTS, subsidiary country #4:	3,457,467
Spoken in more than 5 countries?	☒ Yes
Total population of ethnic group:	
Percentage of ethnic group who are MTS:	?
Is this language an isolate?	☒ Yes
Evidence of moribundity?	☒ Yes
Evidence of vigor?	☒ Yes
Is this language listed as "nearly extinct?"	☐ Yes
Is this language primarily/entirely spoken by indigenous people?	☐ Yes
Is this language primarily/entirely spoken by nomadic people?	☐ Yes
Major language family:	Indo-European
If "Other," specific language family:	
Linguistic typology (SOV, etc.)	SVO
Geological / ecological information on language	
Primary religion of speakers	Christian
Data quality rating	
Possible trend anomalies?	☒ Yes
Georeference field	

COMMENTS

508,000,000 including 2nd lg speakers (1999 WA); Tesniére numbers are probably UK only

TABLE A-1

Field label as shown on Base Data Entry Form	**Type of field**	**Explanation**
ISO-DIS 639/3	text	The unique three-letter ISO code that identifies each discrete language.
Main language name as given in E[thnologue 20]05	text	The primary name for the language as given in *Ethnologue*. For languages spoken in more than one country, *Ethnologue* generally provides separate entries for each country, with cross-references back to a main entry, which is usually under the country where the language originated. In such instances, we took the primary language name as given in the main entry, and also took all the demographic information from the main entry.
Old E[thnologue] code	text	The unique three-letter code assigned in previous editions of *Ethnologue*. These have been superseded by the ISO codes.
(unlabeled check field, upper righthand corner)	check field	The sorting check field was used to demarcate languages that are part of the random sample of 1,500.
Number of mother-tongue speakers (MTS), all countries (high est):	number	The number of mother-tongue speakers reported for the language. If a range is given, this number is the high estimate.
Year of this estimate:	text	The year the above figure was estimated. If no year was given, "2005" was entered.
If source is cited, give author/ date:	text	If given, a source of the estimate and date for the source. If no source is given, "E05" is entered.
Number of mother-tongue speakers (MTS), all countries (low est):	number	If a range is given, the low estimate of the number of mother-tongues speakers. Otherwise left blank.
Year of this estimate:	text	The year the above figure was estimated. If no year was given, "2005" was entered.
If source is cited, give author/ date:	text	If given, a source of the estimate and date for the source. If no source is given, "E05" is entered.
Main country spoken in (E05 "main entry" country):	text	The country under which the main entry for the language is to be found. This is not always the country in which the most speakers live. For example, for English [eng] the main entry country is not that with the largest number of speakers (USA), but is instead the UK, the language's country of origin.
Number of MTS, main country:	number	If spoken in more than one country, the number of mother-tongue speakers given under the main entry.

Percentage of MTS in main country:	automatic calculation	[Number of MTS, main country]/[Number of mother-tongue speakers (MTS), all countries (high est)]
Is this language endemic (100% in main country)?	check field	If the above calculation is 100%, the language is considered endemic and this field is checked.
Ethnologue region (main country):	drop-down text menu	*Ethnologue* is organized according to five regions: Africa, Americas, Asia, Europe, and Pacific; for the ILD, we separated out Australia from the Pacific.
Subsidiary country #1:	text	If spoken in more than one country, the name of the first subsidiary country listed under the main entry.
Number of MTS, subsidiary country #1:	number	Number of mother-tongue speakers in first subsidiary country.
Subsidiary country #2:	text	The name of the second subsidiary country listed under the main entry.
Number of MTS, subsidiary country #2:	number	Number of mother-tongue speakers in second subsidiary country.
Subsidiary country #3:	text	The name of the third subsidiary country listed under the main entry.
Number of MTS, subsidiary country #3:	number	Number of mother-tongue speakers in third subsidiary country.
Subsidiary country #4:	text	The name of the fourth subsidiary country listed under the main entry.
Number of MTS, subsidiary country #4:	number	Number of mother-tongue speakers in fourth subsidiary country.
Spoken in more than 5 countries?	check field	Checked if "yes."
Total population of ethnic group:	number	If given by *Ethnologue*, the total number in the ethnic group.
Percentage of ethnic group who are MTS:	automatic calculation	[Percentage of ethnic group who are MTS]/[Number of mother-tongue speakers (MTS), all countries (high est)]
Is this language an isolate?	check field	Checked "yes" if language is considered an isolate (unrelated to any other language).
Evidence of moribundity?	check field	Checked "yes" if *Ethnologue's* description of the language shows any indications of moribundity. For further explanation, see text.
Evidence of vigor?	check field	Checked "yes" if *Ethnologue's* description of the language shows any indications of vigor. For further explanation, see text.
Is this language listed as "nearly extinct?"	check field	Checked "yes" if *Ethnologue* lists the language as "nearly extinct."

Is this language primarily/entirely spoken by indigenous people?	check field	Checked if our analysis determined that the language is spoken by an indigenous people. For further explanation, see text.
Is this language primarily/entirely spoken by nomadic people?	check field	Checked "yes" if *Ethnologue* description indicates that the speakers are nomads/mobile peoples.
Major language family:	drop-down text menu	*Ethnologue* assigns languages to one of the following categories: Afro-Asiatic, Austronesian, Indo-European, Language isolate, Niger-Congo, Sino-Tibetan, Trans-New Guinea, or Other.
If "Other," specific language family:	drop-down text menu	If the language falls into the "Other" category, it is assigned to one of the following subcategories: Alacalufan, Algic, Altaic, Amto-Musan, Andamanese, Araun, Araucanian, Arawakan, Artificial Language, Arutani-Sape, Australian, Austro-Asiatic, Aymaran, Barbacoan, Basque, Bayono-Awbono, Caddoan, Cahuapanan, Carib, Chapacura-Wanham, Chibchan, Chimakuan, Choco, Chon, Chukotko-Kamchatkan, Chumash, Creole, Deaf Sign Language, Dravidian, East Bird's Head, East Papuan, Eskimo-Aleut, Geelvink Bay, Guahiban, Harakmbet, Hmong-Mien, Hokan, Huavean, Iroquoian, Japanese, Jivaroan, Kartvelian, Katukinan, Keres, Khoisan, Kiowa Tanoan, Kwomtari-Baibai, Language Isolate, Left May, Lower Mamberamo, Lule-Vilela, Macro-Ge, Maku, Mascoian, Mataco-Guaicuru, Mayan, Misumalpan, Mixe-Zoque, Mixed Language, Mura, Muskogean, Na-Dene, Nambiquaran, Nilo-Saharan, North Caucasian, Oto-Manguean, Panoan, Peba-Yaguan, Penutian, Pidgin, Quechuan, Salishan, Salivan, Sepik-Ramu, Siouan, Sko, Subtiaba-Tlapanec, Tacanan, Tai-Kadai, Tarascan, Torricelli, Totonacan, Tucanoan, Tupi, Unclassified, Uralic, Uru-Chipaya, Uto-Aztecan, Wakashan, West Papuan, Witotoan, Yanomam, Yeniseian, Yukaghir, Zamucoan, Zaparoan
Linguistic typology (SOV, etc.)	drop-down text menu	For some languages, *Ethnologue* indicates the linguistic typology: OSV, OVS, SOV, SVO, VOS, or VSO.
Geological/ecological information on language	text	For some languages, *Ethnologue* indicates the general geological/ecological conditions of the main area inhabited by its speakers.

Primary religion of speakers	drop-down text menu	For some languages, *Ethnologue* indicates the primary religion of speakers, using the following list: Buddhist (unspecified), Buddhist (Lamaist), Christian, Confucianism, Daoist, Hindu, Jewish, Mandaism, Muslim (unspecified), Muslim (Alevi), Muslim (Shi'a), Muslim (Sunni), Polytheist, Shamanist, Syncretism, Traditional Religion, Zoroastrianism
Data quality rating	(undetermined)	Country entries in some pre-2005 editions of *Ethnologue* contained a simple A–D data quality rating. This was discontinued in the 2005 edition, but we are reserving such a field for possible future use in the database.
Possible trend anomalies?	check field	For further explanation, see text.
Georeference field	(undetermined)	We are reserving a field that would be tied to current work on mapping biocultural diversity being undertaken by Terralingua, which includes *Ethnologue's* GIS coordinates for the world's languages.

3. ENTRY OF MOTHER-TONGUE SPEAKER INFORMATION. We searched the nine editions of *Ethnologue* used as the basis of the initial ILD for the number of mother-tongue speakers of the 1,500 languages in our sample. Figure A-2 shows the form used to record the base demographic information; Table A-2 explains the fields in the form.

FIGURE A-2

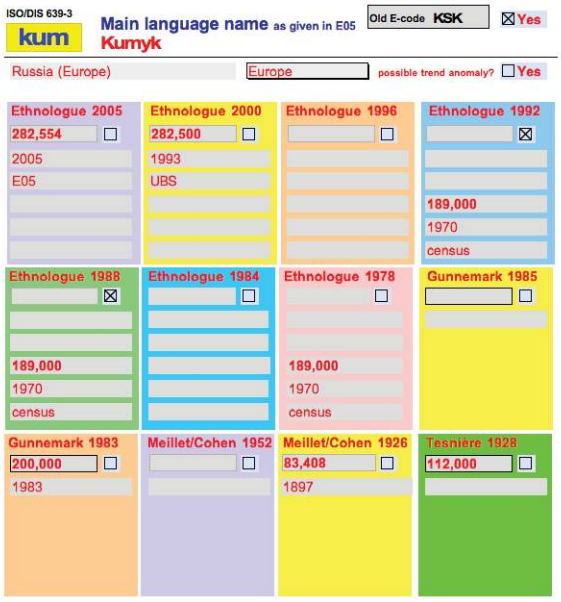


TABLE A-2

Field label as shown on form	Type of field	Explanation
ISO-DIS 639/3	text	The unique three-letter ISO code that identifies each discrete language. For further explanation, see text.
Main language name as given in E[thnologue 20]05	text	The primary name for the language as given in *Ethnologue*. For languages spoken in more than one country, *Ethnologue* generally provides separate entries for each country, with cross-references back to a main entry, which is usually under the country where the language originated. In such instances, we took the primary language name as given in the main entry, and also took all the demographic information from the main entry.
Old E[thnologue] code	text	The unique three-letter code assigned in previous editions of *Ethnologue*. These have been superseded by the ISO codes.
(unlabeled field, left side, just beneath horizontal line) Main country spoken in (E05 "main entry" country):	text	The country under which the main entry for the language is to be found. This is not always the country in which the most speakers live. For example, for English [eng] the main entry country is not that with the largest number of speakers (USA), but is instead the UK, the language's country of origin.
(unlabeled field, center, just beneath horizontal line) Ethnologue region (main country):	drop-down text menu	*Ethnologue* is organized according to five regions: Africa, Americas, Asia, Europe, and Pacific; for the ILD, we separated out Australia from the Pacific.
Possible trend anomaly?	check field	For further explanation, see text.
Description of fields in *Ethnologue* edition source blocks (Ethnologue 2005, Ethnologue 2000 ... Ethnologue 1951), from top to bottom		
Number of mother-tongue speakers (MTS), all countries (high est):	number	The number of mother-tongue speakers reported for the language. If a range is given, this number is the high estimate.
Year of this estimate:	text	The year the above figure was estimated. If no year was given, the year of the edition is entered.
Source of estimate:	text	If given, a source of the estimate.

Number of mother-tongue speakers (MTS), all countries (low est):	number	If a range is given, this is the low estimate of the number of mother-tongue speakers. If a number is given here but no high estimate is given above, it means that this number represents a minimum estimate. If neither of foregoing two conditions applies, the field is left blank.
Year of this estimate:	text	The year the above figure was estimated. If no year was given, the year of the edition is entered.
Source of estimate:	text	If given, a source of the estimate.
(unlabeled check box) Duplicate datapoint control check field:	check field	If the figures reported in this block are identical (in both number of speakers reported and in terms of source citation), then this datapoint is a duplicate and is omitted from the trend analysis.

4. ANALYSIS OF SAMPLE REPRESENTATIVENESS. The 15th edition of *Ethnologue* (Gordon 2005) provides global statistics for three language demographic variables that we used to assess our sample's representativeness: language size, language family, and main region of the language. We compared our sample to the global total for the three variables, and found that it is closely representative of the global distribution for all three variables (Table A-3).

TABLE A-3.1: Representativeness by language size

Representativeness by language size (extinct languages excluded)									
Number of mother-tongue speakers per language	**0**	**1–100**	**101–1,000**	**1,001–10,000**	**10,001–100,000**	**100,001–1,000,000**	**>1,000,000**	**No data**	**Total**
Number of languages, ILD sample	73	129	216	430	363	171	65	53	1,500
% of ILD sample	4.9	8.6	14.4	28.7	24.2	11.4	4.3	3.5	100.0
Number of languages, global total	387	548	1,071	1,967	1,779	892	347	308	7,299
% of global total	5,3	7.5	14.7	26.9	24.4	12.2	4.8	4.2	100.0

TABLE A-3.2: Representativeness by language family

Representativeness by language family (extinct languages excluded)				
	Ethnologue 2005	**% of global total**	**ILD sample**	**% of sample**
Major lg families				
Afro-Asiatic	353	5.11	67	4.70
Austronesian	1,246	18.03	242	16.96
Indo-European	430	6.22	102	7.15
Niger-Congo	1,495	21.63	295	20.67
Sino-Tibetan	399	5.77	93	6.52
Trans-New Guinea	561	8.12	112	7.85
	4,484	64.87	911	63.84
Other lg families & classifications				
Alacalufan	1	0.01	0	0.00
Algic	31	0.45	6	0.42
Altaic	64	0.93	17	1.19
Amto-Musan	2	0.03	0	0.00
Andamanese	4	0.06	0	0.00
Araun	7	0.10	1	0.07
Araucanian	2	0.03	1	0.07
Arawakan	49	0.71	14	0.98
Artificial Language	1	0.01	1	0.07
Arutani-Sape	2	0.03	0	0.00
Australian	224	3.24	46	3.22
Austro-Asiatic	169	2.45	36	2.52
Aymaran	3	0.04	1	0.07
Barbacoan	5	0.07	1	0.07
Basque	3	0.04	0	0.00
Bayono-Awbono	2	0.03	0	0.00
Caddoan	4	0.06	1	0.07
Cahuapanan	2	0.03	1	0.07
Carib	29	0.42	5	0.35
Chapacura-Wanham	4	0.06	0	0.00
Chibchan	21	0.30	6	0.42
Chimakuan	1	0.01	0	0.00
Choco	7	0.10	1	0.07
Chon	2	0.03	0	0.00

Chukotko-Kamchatkan	5	0.07	1	0.07
Creole	82	1.19	20	1.40
Deaf Sign Language	119	1.72	25	1.75
Dravidian	73	1.06	19	1.33
East Bird's Head	3	0.04	1	0.07
East Papuan	33	0.48	6	0.42
Eskimo-Aleut	10	0.14	1	0.07
Geelvink Bay	33	0.48	7	0.49
Guahiban	5	0.07	2	0.14
Harakmbet	2	0.03	0	0.00
Hmong-Mien	35	0.51	5	0.35
Hokan	19	0.27	4	0.28
Huavean	4	0.06	0	0.00
Iroquoian	7	0.10	1	0.07
Japanese	12	0.17	1	0.07
Jivaroan	4	0.06	0	0.00
Kartvelian	5	0.07	0	0.00
Katukinan	3	0.04	1	0.07
Keres	2	0.03	0	0.00
Khoisan	22	0.32	3	0.21
Kiowa Tanoan	5	0.07	2	0.14
Kwomtari-Baibai	6	0.09	0	0.00
Language Isolate	36	0.52	8	0.56
Left May	6	0.09	1	0.07
Lower Mamberamo	2	0.03	0	0.00
Lule-Vilela	1	0.01	1	0.07
Macro-Ge	24	0.35	3	0.21
Maku	6	0.09	1	0.07
Mascoian	4	0.06	0	0.00
Mataco-Guaicuru	11	0.16	1	0.07
Mayan	68	0.98	15	1.05
Misumalpan	2	0.03	0	0.00
Mixe-Zoque	17	0.25	8	0.56
Mixed Language	19	0.27	4	0.28
Mura	1	0.01	0	0.00
Muskogean	6	0.09	2	0.14

Na-Dene	41	0.59	9	0.63
Nambiquaran	3	0.04	0	0.00
Nilo-Saharan	197	2.85	40	2.80
North Caucasian	33	0.48	7	0.49
Oto-Manguean	172	2.49	35	2.45
Panoan	19	0.27	4	0.28
Peba-Yaguan	1	0.01	0	0.00
Penutian	23	0.33	7	0.49
Pidgin	5	0.07	3	0.21
Quechuan	45	0.65	9	0.63
Salishan	19	0.27	5	0.35
Salivan	3	0.04	0	0.00
Sepik-Ramu	100	1.45	24	1.68
Siouan	12	0.17	6	0.42
Sko	7	0.10	2	0.14
Subtiaba-Tlapanec	4	0.06	1	0.07
Tacanan	6	0.09	2	0.14
Tai-Kadai	74	1.07	14	0.98
Tarascan	2	0.03	0	0.00
Torricelli	53	0.77	12	0.84
Totonacan	11	0.16	1	0.07
Tucanoan	20	0.29	5	0.35
Tupi	60	0.87	13	0.91
Unclassified	43	0.62	11	0.77
Uralic	36	0.52	5	0.35
Uru-Chipaya	2	0.03	0	0.00
Uto-Aztecan	56	0.81	8	0.56
Wakashan	4	0.06	1	0.07
West Papuan	26	0.38	6	0.42
Witotoan	6	0.09	3	0.21
Yanomam	4	0.06	0	0.00
Yeniseian	2	0.03	0	0.00
Yukaghir	2	0.03	1	0.07
Zamucoan	2	0.03	0	0.00
Zaparoan	4	0.06	1	0.07
	2,428	35.13	516	36.16

TABLE A-3.3: Representativeness by region

Representativeness by region (extinct languages excluded)						
	Africa	**Americas**	**Asia**	**Europe**	**Pacific**	**Total**
Ethnologue 2005	2,092	1,002	2,269	239	1,310	6,912
% of global total	30.3	14.5	32.8	3.5	19.0	100.0
ILD sample	408	226	476	44	273	1,427
% of sample	28.6	15.8	33.4	3.1	19.1	100.0

5. ELIMINATION OF DUPLICATE DATAPOINTS. As we entered mother-tongue speaker information, we analyzed each datapoint to see if it was unique (i.e., represented new data) or a duplicate of an earlier datapoint. In terms of developing time-series data on speaker numbers, it would have been ideal if each of our 11,253 data searches had produced a unique datapoint. The reality is far from the ideal, however, and one reason to have a sample size much larger than the minimum required for statistical validity is to account for attrition: in our case, languages having to be excluded from the ILD because they have fewer than two unique datapoints from which to construct a trend. Because of the paucity of speaker statistics for many languages, this is not an uncommon occurrence. It is standard practice for *Ethnologue* to carry over estimates from earlier editions if a newer (and therefore presumably more current estimate) is unavailable. Out of our sample of 1,500 languages, 391 had to be excluded from the ILD calculation because they had either no speaker totals listed in any of the editions of *Ethnologue* we consulted, or else had only one unique datapoint. We can expect this situation to improve with future editions of *Ethnologue*, for there is now a sustained effort by the editors to report speaker totals from as many languages as possible. (A major gap is deaf languages, for which speaker totals are rarely reported.) After all steps of the data analysis were completed, we were left with 2,703 unique datapoints from which we created the initial version of the ILD.

6. ASSESSMENT AND TREATMENT OF POSSIBLE TREND ANOMALIES. It is not uncommon for successive estimates of speaker numbers for a particular language to vary widely. To account for this, we assessed the time series for all 1,500 languages in our sample for possible trend anomalies: large or rapid changes in the reported numbers of mother-tongue speakers within the chronological sequence of estimates for that language.

There are many reasons why a particular datapoint in a time series could possibly be anomalous. It might reflect some kind of major difference in the way the speakers were counted, or in interpretation of what constitutes the language itself, or in who qualifies as a mother-tongue speaker. Perhaps the datapoint in question could simply reflect an incomplete count of speakers despite the researcher having canvassed all known locations where speakers live. Or it could reflect an incomplete count of speakers because the researcher failed to canvass all known locations. It could even be (though it is more unlikely) that the seemingly discrepant datapoint is, in fact, accurate and all the others are wrong—for example, maybe all the other datapoints included second-language speakers.

Not all languages will show linear trends in their speaker numbers, and this in itself is not a reason to suspect a possible trend anomaly. It may be that a particular language's numbers truly are fluctuating. Similarly, some languages may show an unbroken upward or downward trend, but within that trend there will be huge jumps or declines that might lead one to question the accuracy of the numbers. In all these situations, the controlling questions are, what is the magnitude of the reported change, how quickly is it happening, and how plausible is it relative to the size of the language?

In terms of plausibility, a cardinal principle is that smaller percentage changes are more plausible across the board, no matter if the language has 100 speakers to start with or 1,000,000. We can easily imagine a small language going from 100 to 99 speakers over a 20-year period, and just as easily imagine a language with 1,000,000 going to 990,000 over the same period. However, as the percentage changes become larger, the plausibility of those changes depends on how large the language is initially. For a language starting out at 100 speakers, it is plausible to imagine a situation in which it declines 90% over the 20-year period, going from 100 to 10 speakers. Perhaps most or all of the speakers were old to begin with (a not uncommon occurrence in such cases) and they died over the period while at the same time no children were being brought up using the language as their mother tongue. Or perhaps there was a catastrophe that struck the village where all the speakers lived, causing most of the speakers to die. These are plausible scenarios. But it is far less plausible to see how a language could go from 1,000,000 speakers to 100,000 in just 20 years. So a corollary point is that the plausibility of changes in speaker numbers declines as (a) the percentage of change increases, (b) the language size increases, and (c) the time period over which the change is said to occur decreases. That is, large percentage changes in the size of large languages over short periods of time are the least plausible.

We identified possible trend anomalies by calculating the rate of change in number of speakers between one datapoint and the next. Three degrees of possible anomaly were identified: differences between successive datapoints that represented a rate of change equivalent to a doubling or halving in number over a period of (a) ten years, (b) five years, or (c) three years. Languages where numbers of speakers were below 1,000 were excluded because, as just noted, very small populations are liable to undergo rapid fluctuations.

For our assessment of possible trend anomalies within the ILD's 1,500-language sample, we analyzed all instances flagged by the 3-year doubling/halving filter. There were 157 languages having such instances (Table A-4). Our analysis consisted of:

- Identifying the flagged datapoint(s).
- Assessing the likelihood of that datapoint being anomalous. This involved a number of considerations, including size of the language, keeping in mind that, in general, the smaller the language, the easier it is to accurately count its number of speakers (Voegelin and Voegelin 1977:8); special qualities of certain data sources used by *Ethnologue*, based on our experience in working with the database (e.g., long-time observation has shown estimates from certain data sources cited by *Ethnologue* tend to run higher or lower than others cited in other editions of *Ethnologue*); and whether or not the trendline for the language contains one possible anomalous datapoint or several.
- Excluding the anomalous datapoint, if necessary.

For datapoints that we assessed as being "definitely anomalous" or "probably anomalous," the datapoint was excluded from the ILD. This has the effect of smoothing the trendline. For datapoints assessed as "possibly anomalous," or "may not be anomalous," the decision to exclude or not varied depending on our judgment, using the considerations outlined above: some were excluded, while some were left unchanged. When in doubt, our policy was to leave the data unchanged. See Table A-4.

TABLE A-4

ISO	**Language name**	**Is data trend anomalous?**	**Reason**	**Data treatment**
afb	Arabic, Gulf Spoken	probably anomalous	decline from E92 to later estimates too steep to be plausible in a language with millions of speakers	disregard E92 as outlier
ald	Alladian	probably anomalous	Taber's estimates generally run low	disregard E69 as outlier; all other estimates show steady rise in numbers
amx	Anmatyerre	possibly anomalous	Wurm and Hatori estimate very low in comparison to Black's	disregard Wurm and Hatori estimate (1981) because subsequent editions endorse Black's 1983 estimate
apl	Apache, Lipan	may not be anomalous	extremely low speaker numbers might explain pecentage decline	leave data unchanged
apm	Apache, Mescalero-Chiricahua	probably anomalous	unlikely that [apm] gained 800 speakers between 1969 and 1978	disregard E69 because subsequent editions endorse E78 estimate
arg	Aragonese	probably anomalous	unlikely that [arg] declined by 19,000 speakers between 1989 and 1993; prior to E92 had been lumped in with Spanish	disregard E92 because two subsequent editions give identical estimates from 2 different sources
asb	Assiniboine	probably anomalous	unlikely that [asb] declined from 1000-2000 in 1969 to 100 in 1977; also possible that it did not actually increase from 100 in 1977 to 150-200 in 1986	disregard E69 as outlier because subsequent editions give much more comparable estimates
ask	Ashkun	may not be anomalous	*Ethnologue* estimate of 7000 held from E78 through E2000; since no external data sources given, it appears that the E78 esimate was simply carried over to subsequent editions	disregard E84, E88, E92, E96, and E00; use 2-datapoint trendline: E78 and E05

aue	=Kx'aull´ein	probably anomalous	unlikely that [aue] went from 4890 in 1977 to 3000 in 1991 and then back up to 5000 in 1993	disregard E92 as outlier
bae	Baré	possibly anomalous	debatable that [bae] went from 263 speakers in 1988 to 0 in 2005, but not inconceivable	leave data unchanged
bis	Bislama	possibly anomalous	debatable that [bis] went from 1200 in 2000 to 6200 in 2005, but not inconceivable	leave data unchanged
bjl	Bulu (Papua New Guinea)	probably anomalous	unlikely that [bjl] went from 200 speakers in 1978 to 566 in 1982	disregard E78 as outlier
bjz	Baruga	probably anomalous	unlikely that [bjz] went from 4000-6000 in 1969 to 1051 in 1971	disregard E69 as outlier
bpp	Kaure	probably anomalous	unlikely that [bpp] spiked at 4000 in 1991 when other estimates are less than 1000	disregard E92 as outlier
bra	Braj Bhasha	probably anomalous	unlikely that [bra] declined from 11+million in 1977 to 44500 in 1997	disregard E78 in favor of more recent datapoint (which is repeated in E05)
brn	Boruca	probably anomalous	unlikely that [brn] went from 500 in in 1978 to 5 in 1986	disregard E78 as outlier
bsr	Bassa-Kontagora	may not be anomalous	datapoints in E00 and E05 revert to a count of 10 in 1987, and E05 says [bsr] is extinct, or nearly extinct; less conservatively, E92 give a count of 0	disregard E78, use 2-datapoint trendline: E00 (10 in 1987) and E92 (0 in 1992)
bwt	Bafaw-Balong	definitely anomalous	E78 and E69 estimates are for Balong only	disregard E78 and E69
bzp	Kemberano	probably anomalous	unlikely that [bzp] went from 150 in 1978 to 1500 in 1987	disregard E78 because subsequent editions endorse E92
caz	Canichana	may not be anomalous	E92 estimate of 25 is dated to 1958; E69 estimate of 25 is dated to 1968; E78 and E88 estimates of 25 are undated, so are listed in the database as being from 1978 and 1988, but this is likely misleading	disregard E78 and E88; use 3-datapoint trendline: E92, E69, and E00
cbb	Cabiyarí	may not be anomalous	fluctuations of this magnitude are conceivable with a small lg like [cbb]	leave data unchanged

cbn	Nyahkur	probably anomalous	discrepancies between E92 and E78, and between the high/low average of E92 and E00 and E05, appear to be of too great a magnitude to be true demographics	disregard E92 and E78
cbr	Cashibo-Cacataibo	probably anomalous	E00 estimate is undated and therefore attributed to 1998, but probably dates from earlier	disregard E00; use E69, E78, and E05 as trendline
cch	Atsam	probably anomalous	unlikely that [cch] increased from 8500 in 1969 to 35000 in 1972	disregard E69 as outlier
cku	Koasati	possibly anomalous	debatable increase to, and then decline from 1996 datapoint	leave data unchanged
cle	Chinantec, Lealao	possibly anomalous	debatable that [cle] went from 3500/5000 in 1978 t0 800/900 in 1982	disregard E78 as outlier
cod	Cocama-Cocamilla	probably anomalous	no discernable pattern: estimates vary widely and cannot be reconciled	disregard all datapoints
cun	K'iché, Cunén	may not be anomalous	E05 notes "significant monolingualism" and only a slight move toward Spanish; this suggests E78 estimate (repeated in E88) is an undercount	disregard E78 and E88
dal	Dahalo	probably anomalous	unlikely decrease from 1987 to 1992 datapoints	disregard E92
djm	Dogon, Jamsay	probably anomalous	unlikely decrease from 1995 to 1998 datapoints	disregard E00
dng	Dungan	probably anomalous	unlikely increase from 1969 to 1970 datapoints	disregard E69 as outlier
dor	Dori'o	probably anomalous	unlikely increase from 1998 to 1999 datapoints	disregard E00 as outlier
dyi	Senoufo, Djimini	probably anomalous	unlikely increase from 1991 to 1993 datapoints	disregard E92
eee	E	probably anomalous	unlikely increase from 1990 to 1992 datapoints	disregard E92
eot	Beti (Cote d'Ivoire)	probably anomalous	unlikely decrease from 1966 to 1977 datapoints	disregard E69
faf	Fagani	probably anomalous	unlikely increase from 1998 to 1999 datapoints	disregard E00
fgr	Fongoro	definitely anomalous	thought extinct in E92; more recent editions reference 1983 estimate	disregard E92
fip	Fipa	probably anomalous	unlikely increase from 1990 to 1992 datapoints	disregard E92

fri	Frisian, Western	probably anomalous	unlikely decrease from 1976 to 1978 datapoints	disregard E78; construct 2-datapoint trendline using E05 (1976) and E88 (1988)
gid	Gidar	probably anomalous	unlikely increase from 1966 to 1967 datapoints	disregard E69
glk	Gilaki	probably anomalous	unlikely increase from 1991 to 1993 datapoints	disregard E92
goa	Guro	probably anomalous	unlikely decrease from 1966 to 1967 datapoints	disregard E78
gvf	Golin	probably anomalous	unlikely increase from 1978 to 1981 datapoints	disregard E78
gvl	Gulay	probably anomalous	unlikely increase from 1990 to 1993 datapoints	disregard E92
gyi	Gyele	probably anomalous	unlikely increase from 2000 to 2005 datapoints	disregard E00
hae	Oromo, Eastern	possibly anomalous	debatable increase from 1978 to 1998 datapoints	leave data unchanged
hbn	Heiban	probably anomalous	unlikely increase from 1966 to 1972 datapoints	disregard E78 as outlier
hio	Tsoa	probably anomalous	unlikely increases between 1976 and 1977 datapoints and between 2000 and 2004 datapoints	disregard E88 and E00; construct 2-datapoint trend-line using E78 and E05
hmd	Hmong, Northeastern	probably anomalous	unlikely decrease from 1982 to 1987 datapoints	disregard E00
hsb	Sorbian, Upper	possibly anomalous	debatable decrease from 1991 to 1196 datapoints	leave data unchanged
huc	ǀHua	possibly anomalous	debatable increase from 1966 to 1978 datapoints	disregard E00 as a duplicate datapoint; otherwise leave data unchanged
huu	Huitoto, Murui	probably anomalous	unlikely increase from 1969 to 1976 datapoints and from 1976 and 1982 datapoints	disregard E69 and E78 as probable underestimations
igl	Igala	probably anomalous	unlikely increase from 1969 to 1973 datapoints	disregard E69
ilb	Ila	probably anomalous	unlikely decrease to, and then in-crease from, 1973 datapoint	disregard E78 as outlier
iru	Irula	probably anomalous	trendline highly variable; E05 vastly out of line with all previous estimates	disregard E78 and E05
itv	Itawit	probably anomalous	unlikely decrease from 1969 to 1973 datapoints	disregard E78 as outlier
izi	Izi-Ezaa-Ikwo-Mgbo	probably anomalous	unlikely increase from 1969 to 1973 datapoints	disregard E69

jai	Jakalteco, Western	probably anomalous	unlikely increases from 1988 to 1992 and again from 1992 to 2000	disregard E88 and E92
jeg	Jeng	probably anomalous	unlikely increase from 1978 to 1981 datapoints	disregard E78; construct 3-datapoint trendline using E69, E92, and E05
kav	Katukina	probably anomalous	unlikely decrease from 1969 to 1976 datapoints	disregard E69
kca	Khanty	possibly anomalous	debatable decrease from 1969 to 1970 datapoints	leave data unchanged
kdr	Karaim	possibly anomalous	debatable decrease from 1969 to 1977 datapoints	leave data unchanged
khg	Tibetan, Khams	probably anomalous	unlikely increase from 1977 to 1987 datapoints	disregard E78 as outlier
khy	Kele (Democratic Repulic of Congo)	probably anomalous	unlikely increase from 1971 to 1980 datapoints	disregard E78
kia	Kim	probably anomalous	unlikely increase from 1991 to 1993 datapoints	disregard E92
kll	Kalagan, Kagan	possibly anomalous	debatable decrease from 1969 to 1977 datapoints	disregard E78
kln	Kalenjin	probably anomalous	E92, E78, and E69 all appear to be underestimations	disregard E69, E78, and E92
kng	Koongo	probably anomalous	E78 and E88 appear to be overestimations (and E05 discards E88 estimate dating from 1987 in favor of a lower estimate dating from 1986)	disregard E78 and E88
kou	Koke	probably anomalous	unlikely decrease from 1969 to 1971 datapoints	disregard E69
krk	Kerek	possibly anomalous	debatable decrease from 1975 to 1991 datapoints	leave data unchanged
ksi	Krisa	probably anomalous	unlikely increase from 1969 to 1972 datapoints	disregard E69 as outlier
kum	Kumyk	probably anomalous	unlikely increase from 1969 to 1970 datapoints	disregard E69 as outlier
kwz	Kwadi	probably anomalous	E78 datapoint conflicts with note in E05 that says [kwr] had 3 speakers in 1971 and was considered by J.C. Winter in 1981 to have been extinct by then; since E00 listed as extinct	disregard E78; perhaps we could construct trendline with 3 speakers in 1971 and 0 in 1981?
kyr	Kuruáya	definitely anomalous	E78 datapoint obviously erroneous	disregard E78
kzw	Karirí-Xocó	possibly anomalous	debatable decrease to 0 (1978; E78) from 163 (1969; E69), but conceivable	leave data unchanged

lbo	Laven	probably anomalous	unlikely increase from 1978 to 1981 datapoints	disregard E78
leb	Lala-Bisa	probably anomalous	unlikely increase from 1969 to 1973 datapoints	disregard E69
lez	Lezghian	probably anomalous	unlikely increase from 1969 to 1970 datapoints	disregard E69
lga	Lungga	may not be anomalous	trendline looks plausible	leave data unchanged
lif	Limbu	probably anomalous	unlikely increase from 1967 to 1971 datapoints; debatable increase from 1998 to 2005 datapoints	disregard E69; leaves others unchanged
lma	Limba, East	probably anomalous	unlikely increase from 1991 to 1993 datapoints	disregard E92
lpa	Lelepa	possibly anomalous	unlikely decrease from 1983 to 1989 datapoints; but not inconceivable	leave data unchanged
mbb	Manobo, Western Bukidnon	probably anomalous	unlikely decrease from 1969 to 1977 datapoints	disregard E69
mco	Mixe, Coatlán	probably anomalous	trendline highly variable	disregard all datapoints??
mdt	Mbere	probably anomalous	unlikely increase from 1990 to 2005 datapoints	disregard E92
mei	Midob	probably anomalous	unlikely increase from 1977 to 1983 datapoints and from 1983 to 1993 datapoints	disregard E78 and E92?
mez	Menominee	possibly anomalous	debatable decrease from 1969 to 1977 datapoints	disregard E69
mit	Mixtec, Southern Puebla	possibly anomalous	unlikely decrease from 1977 to 1982 datapoints; but not inconceivable	leave data unchanged
mjj	Mawak	possibly anomalous	unlikely decrease from 1969 to 1981 datapoints; but not inconceivable	leave data unchanged
mjp	Malapandaram	probably anomalous	it appears E78 and E92 were simply repeating the 1961 census figures	disregard E78 and E92
mju	Manna-Dora	probably anomalous	it appears E78 and E92 were simply repeating the 1961 census figures	disregard E78 and E92
mnc	Manchu	probably anomalous	unlikely decrease from E78 high estimate to subsequent estimates	disregard E78
msm	Manobo, Agusan	probably anomalous	unlikely increase from 1977 to 1981 datapoints	disregard E78

mug	Musgu	probably anomalous	unlikely increase from 1969 to 1972 datapoints	disregard E69
mvk	Mekmek			
myx	Masaba	probably anomalous	unlikely increases from 1966 to 1978 datapoints and again from 1990 to 1991 datapoints	disregard E69 and E92
ngu	Náhuatl, Guerrero	definitely anomalous	unlikely increase from 1969 to 1977 datapoints; E78 and E88 contradict each other	disregard E69, E78, and E88
niv	Gilyak	probably anomalous	E78 high / low estimates too far out of line, as is E92 high estimate	disregard E78 and E92 high estimate; construct 3-datapoint trendline out of E92 low estimate and E00 and E05 estimates
nkf	Naga, Inpui	possibly anomalous	debatable increase to, and then decrease from, E92 and E00 estimates	disregard E92 and E00 as probable overestimations
nog	Nogai	probably anomalous	unlikely increase from 1969 to 1970 datapoints	disregard E69
now	Nyambo	probably anomalous	unlikely increase from 1987 to 2005 datapoints	disregard E05 (or else disregard all except E05, if we want to go with the latest estimate as being the most accurate)
nut	Nung (Viet Nam)	probably anomalous	unlikely increase from 1973 to 1981 datapoints	disregard E92 as outlier
nza	Mbembe, Tigon	may not be anomalous	trendline debatable but not implausible	leave data unchanged
nzm	Naga, Zeme	probably anomalous	unlikely increase from 1961 to 1971 datapoints and from 1990 to 1994 datapoints	disregard E69 and E92
ojw	Ojibwa, Western	may not be anomalous	trendline looks plausible	leave data unchanged (keeping in mind E00 is a duplicate datapoint)
ots	Otomi, Estado fr Mexico	probably anomalous	unlikely increase to, and then decrease from, 1978 datapoint	disregard E78 as outlier
paf	Paranawat	possibly anomalous	debatable decrease from 1969 to 1978 datapoints	disregard E69
pcg	Paniya	probably anomalous	unlikely decrease to, and then increase from, 1971 datapoint	disregard E78 as outlier
pch	Pardhan	probably anomalous	unlikely decrease to, and then increase from, 1971 datapoint	disregard E78 as outlier

pis	Pijin	may not be anomalous	debatable increase from 1975 to 1997 datapoints	leave data unchanged
pmu	Panjabi, Mirpur	probably anomalous	unlikely increase from 2000 to 2005 datapoints	disregard E00
pou	Poqomam, Southern	probably anomalous	unlikely increase from 1982 to 1991 datapoints	disregard E88
quu	K'iché, Eastern	probably anomalous	unlikely increase from 1978 to 1982 datapoints; E05 notes that it is spoken by "all ages"	disregard E88 as outlier
rej	Rejang	probably anomalous	unlikely increase from 1977 to 1981 datapoints	disregard E78
she	Sheko	probably anomalous	unlikely increase from 1966 to 1972 datapoints	disregard E69
sih	Zire	definitely anomalous	E00 datapoint erroneous	disregard E00; also, I suspect E69 datapoint (which didn't give a year and so is listed as 1969) is the same as the E78 datapoint, which is sourced to 1939
smn	Inari Sami	possibly anomalous	debatable decrease from 1978 to 1983 datapoints, but conceivable	leave data unchanged
soe	Songomeno	possibly anomalous	unlikely increase from 1971 to 1972 datapoints	disregard E78
syc	Syriac	probably anomalous	unlikely decrease from 1978 to 2000 datapoints	disregard E78
tab	Tabassaran	probably anomalous	unlikely increase from 1969 to 1970 datapoints	disregard E69
tan	Tangale	probably anomalous	unlikely increase from 1969 to 1973 datapoints	disregard E69
tbe	Tanimbili	possibly anomalous	unlikely decrease from 1998 to 1999 datapoints, but not inconceivable	leave data unchanged
tbx	Kapin	probably anomalous	unlikely decrease from 1979 to 1980 datapoints	disregard E92
tcc	Datooga	probably anomalous	trendline highly variable; E05 notes lg use vigorous, so overall upward trendline seems most likely possibility	disregard E92 and E00 as probable overestimations
tdg	Tamang, Western	probably anomalous	unlikely increase from 1989 to 1991 datapoints	disregard E92

thr	Tharu, Rana	probably anomalous	unlikely increase from 1985 to 2000 datapoints; debatable increase from 2000 to 2005 datapoints	disregard E92
thu	Thuri	probably anomalous	unlikely increase from 1966 to 1971 datapoints	disregard E78 as outlier
thv	Tamahaq, Tahaggart	probably anomalous	unlikely decrease from 1976 to 1987 datapoints; 1987 datapoint is low-end estimate (probably for Algeria only) and out of line with subsequent estimates	disregard E78 and E88; use 2-datapoint trendline: E96 and E00
tic	Tira	probably anomalous	unlikely increase from 1977 to 1982 datapoints; 1977 appears to be repetition (from a different source) of 1966 datapoint)	disregard E78
tou	Tho	probably anomalous	unlikely increase from 1996 to 1999 datapoints	disregard E00
tpi	Tok Pisin	probably anomalous	unlikely increase from 1977 to 1982 datapoints	disregard E78
tqu	Touo	probably anomalous	unlikely increase from 1976 to 1981 datapoints; debatable increase from 1998 to 1999 datapoints	disregard E78
trr	Taushiro	may not be anomalous	trendline looks plausible (very small language)	leave data unchanged
tsg	Tausug	may not be anomalous	increases are large, but not totally implausible	leave data unchanged
tsi	Tsimshian	possibly anomalous	trendline debatable but not implausible	leave data unchanged
tsr	Akei	possibly anomalous	trendline debatable but not implausible	leave data unchanged
tud	Tuxá	may not be anomalous	trendline looks plausible (very small language)	leave data unchanged
tzc	Tzotzil, Chamula	probably anomalous	non-census datapoints make trendline highly variable	disregard E69, E88, and E92; construct 2-datapoint trendline from E78 and E00 (both based on census data)
urd	Urdu	probably anomalous	unlikely increase from 1969 to 1971 datapoints	disregard E69
waz	Wampur	probably anomalous	unlikely decrease from 1969 to 1970 datapoints	disregard E69 as outlier
wic	Wichita	may not be anomalous	trendline looks plausible (very small language)	leave data unchanged

wir	Wiraféd	may not be anomalous	trendline looks plausible (very small language)	leave data unchanged
wll	Wali (Sudan)	probably anomalous	unlikely increase from 1977 to 1978 datapoints	disregard E92
xrw	Karawa	possibly anomalous	trendline highly variable, but E05 comments suggest that it may be plausible	leave data unchanged
xsy	Saisiyat	possibly anomalous	debatable increase from 1969 to 1973 datapoints	disregard E69
yee	Yimas	possibly anomalous	debatable decrease from 1977 to 1981 datapoints, but possible	leave data unchanged
yig	Yi, Guizhou	possibly anomalous	E00 datapoint appears to supersede E92	disregard E92
ykm	Yakamul	possibly anomalous	debatable increase from 1978 to 1981 datapoints	leave data unchanged
yra	Yerakai	possibly anomalous	debatable decrease from 1969 to 1971 datapoints	leave data unchanged
yuy	Yugur, East	probably anomalous	unlikely decrease from 1990 to 1991 datapoints	disregard E92
ywt	Yi, Western	probably anomalous	unlikely decrease from 1990 to 1991 datapoints	disregard E92
zat	Zapotec, Tabaa	possibly anomalous	trendline highly variable	disregard E88 as outlier
zav	Zapotec, Yatzachi	probably anomalous	unlikely increase from 1969 to 1977 datapoints	disregard E69
zeg	Zenag	possibly anomalous	debatable decrease from 1979 to 1980 datapoints	disregard E92
zen	Zenaga	probably anomalous	unlikely decrease from 1992 to 1998 datapoints	disregard E78 and E92
zkp	Kaingáng, São Paulo	may not be anomalous	trendline looks plausible (very small language)	leave data unchanged
zro	Záparo	may not be anomalous	trendline looks plausible (very small language)	leave data unchanged

7. REMOVAL OF DISCREPANCIES IN ESTIMATES FROM THE SAME YEAR. In a few instances, different editions of *Ethnologue* report different estimates for a language, but attribute them to the same year. As an example, for Guerrero Nahuatl [ngu], the 1978 edition gives high and low figures of 180,000 and 160,000, respectively (average = 170,000), attributing the estimates to SIL 1977. However, the 1988 edition gives figures of 90,000 and 80,000 (average = 85,000) and also attributes the estimates to SIL 1977. In such instances we used the most recent estimate in the calculation and dropped the older.

8. TREATMENT OF EXTINCT LANGUAGES WITHIN TIME SERIES. Languages that are extinct as mother tongues are shown as zero values in the database of numbers of speakers. A time series of zeros would imply no overall trend in the status of that language and therefore those languages were taken out of the sample from the year after that in which they were recorded as going extinct. Eleven languages in our sample were reported as having zero speakers before the end of a non-zero time series, i.e., subsequent editions of *Ethnologue* reported one or more speaker. This may have occurred because a language was believed to have gone extinct, but later found to be still in use among a small community of speakers. In such cases the zero values were removed from the ILD. In instances where the zero value was the first of only two datapoints, leaving only a single datapoint, then both were removed from the ILD.

9. TREATMENT OF SPLITS AND MERGERS. If pre-2005 editions of *Ethnologue* considered a language in our sample as a dialect of a larger language, we excluded any datapoints for that larger language from the database. However, if these editions gave separate speaker totals for these putative dialects, those totals were included as datapoints. Conversely, if pre-2005 editions of *Ethnologue* considered a language in our sample as comprising two or more distinct languages, and if separate speaker totals were available for those putatively distinct languages, we aggregated the totals into a single datapoint.

10. SPECIAL CHALLENGES OF *ETHNOLOGUE* DATA ANALYSIS. Any retrospective analysis of language demographic data presents certain challenges. Next, we discuss three that are particular to *Ethnologue*.

Changes in the number of languages. Since its inception in 1951, each new edition of *Ethnologue* has reported a higher number of languages. In the earliest editions, some of this increment could be explained by the addition of "new" languages of which Western linguistic science was previously unaware. In more recent editions, the editors explain that the increment is, except for a very few cases, no longer due to such "discoveries" but rather to dialects of single languages being reclassified as separate languages. At the same time, as part of the editorial process of preparing each new edition, a number of entries are dropped from the roster. Prior to the advent of ISO codes and roster-change documentation, it may be presumed that most often such entries were expunged because the speech form in question was redefined as a dialect, with its old entry being merged with that of the parent language in the new edition. (See above for our treatment of such cases.) Additionally, in some instances entries appear to have been dropped because they were determined to simply be alternative names for another language already on the roster, or because they were names for an entire ethnic group, not a language.

The 15th edition of *Ethnologue* listed 7,299 languages for the world. The sifting of speech forms to determine whether or not they should be considered discrete languages will no doubt continue, but, significantly, the 16th edition lists 7,296 languages—essentially the same number as the previous edition. As M. Paul Lewis writes, "the rate of languages being split off from existing ones and previously separate languages being re-classified as a single larger language is about equal. In the 16th edition, the count of living languages has diminished (by 3) for the first time. This is largely the result of mergers of existing lan-

guages, though we raise the possibility (in the Introduction) that it could also be the result of our having re-classified a good number of 'Nearly Extinct' languages as 'Extinct'..." (Lewis, pers. comm., 25 May 2009).

Evidence of moribundity and vigor. Two important components of the ILD database are the fields that record *Ethnologue's* qualitative assessments of whether use of a particular language is moribund or vigorous. *Ethnologue* uses a number of standard locutions to indicate that a language has or may become moribund (i.e., it is no longer being learned by young people). *Ethnologue* descriptions were considered to be indicative of moribundity if they point to a decline in the use of the language by, or its disfavor among, young people, or if they make some general reference to language loss (e.g., "It is reported that the language appears to be dying out"). In such cases the Moribundity checkbox was checked.

Ethnologue also has a number of standard locutions that indicate that a language is vigorous. The most common is the simple notation "Language use vigorous." *Ethnologue* descriptions were considered to be indicative of vigor if they point to robustness in the use of the language, its acceptance by young people, its being taught in school, there being a language revitalization program in place, etc., or if they make some general reference to language vigor (e.g., "The people have a positive attitude toward the language"). In such cases the Vigor checkbox was checked.

It is important to understand that these characterizations of language moribundity and vigor are descriptive, not diagnostic. Not infrequently, a language entry may contain evidence of both moribundity and vigor. Examples include cases where indigenous languages are still suffering declines in acceptance/use by young people, but for which language revitalization efforts have begun.

Is the language spoken primarily/entirely by indigenous people? This question, which is an important concern of the ILD, is problematic because there is no standard list of indigenous peoples/languages. In fact, as the United Nations' State of the World's Indigenous Peoples concludes, there is no standard definition of "indigenous peoples," no definition of the term has ever been adopted by a U.N.-system body, and indigenous peoples themselves have "rejected the idea of a formal definition of indigenous peoples at the international level to be adopted by states. Similarly, government delegations expressed the view that it was neither desirable nor necessary to elaborate a universal definition of indigenous peoples" (UNDESA 2009:4–5).

Absent definitive guidance, we used our experience with the dataset and knowledge of the ethnographic literature to determine which languages to check as "indigenous." We used the definition given in the International Labour Organisation's Convention 169 on Indigenous and Tribal Peoples (1989) as a general guide to which groups should be considered indigenous:

> (a) tribal peoples in independent countries whose social, cultural and economic conditions distinguish them from other sections of the national community, and whose status is regulated wholly or partially by their own customs or traditions or by special laws or regulations;

> (b) peoples in independent countries who are regarded as indigenous on account of their descent from the populations which inhabited the country, or a geographical region to which the country belongs, at the time of conquest or colonisation or the establishment of present state boundaries and who, irrespective of their legal status, retain some or all of their own social, economic, cultural and political institutions.

The process was straightforward for the Americas and Europe, where our knowledge of the ethnographic literature made identification of indigenous groups relatively simple. For example, the only European languages in our sample that we marked as "indigenous" were Ume Sami [sju], spoken in Sweden, and a handful of North Caucasian and Altaic languages whose *Ethnologue* "Main Region" is European Russia. Some minority languages in our sample, such as Welsh [gym], although they might also be considered indigenous, were not so marked because it would have meant the inclusion of most European languages. In Africa, although some have questioned the application of "indigenous" to any of the continent's languages (UNDESA 2009:6), we feel the situation is fairly clear-cut: almost every small language group in sub-Saharan Africa was marked as indigenous (the exceptions being sign languages and creoles). In Australia, it is easy to identify the Aboriginal/Torres Strait Islander languages, and in the Pacific (including Papua New Guinea) it is also obvious which languages are indigenous.

In Asia, our relative lack of ethnographic knowledge made the process more difficult (and again, some have challenged the use of the concept in at least some areas of Asia). In India, for example, we generally marked as "indigenous" only those languages that *Ethnologue* listed as being spoken by a Scheduled Tribe, or those that are well-known to be indigenous (such as Andamanese languages). In China, we marked those languages spoken by groups listed as official minority nationalities; in Japan, the Ryukyuan languages; in Taiwan, the small non-Chinese languages; and so forth. Virtually all languages in Indonesia with fewer than 10,000 speakers were so marked.

Admittedly, this process is ad hoc and inevitably we will have made mistakes. However, given the strong interest on the part of various international bodies in the status of indigenous languages globally (e.g., UNPFII, the Convention on Biological Diversity), we feel it is important to make a start at identifying them.

In the final analysis, of the 1,500 languages in our sample we considered 1,285, or 85.6%, to be indigenous. This estimate is supported by the fact most of the world's languages are endemic to a single country (i.e., spoken there and nowhere else). In our ILD sample, 1,453 out of the 1,500 languages had a speaker-number estimate; of these, 1,187, or 81.6%, were endemic. This is very close to the results of Harmon's earlier study of the 1992 edition of *Ethnologue*, in which he found that 83.3% of the world's languages are endemic (Harmon 1995:10). It seems logical to assume that there is a very large overlap, probably on the order of 95%, between indigenous and endemic languages. If that assumption is correct, and if we conservatively posit our figure of 85.6% to represent a high-end estimate of the proportion of indigenous languages in the world, then we can derive a low estimate of 81% by multiplying 85.6 by the 95% overlap. Rounding off this range, we therefore believe it reasonable to estimate that 80–85% of the world's languages are spoken by indigenous people.

Appendix B. Calculating the Index of Diversity

MEASURING THE GEOMETRIC MEAN SHARE OF A POPULATION IN TERMS OF NUMBERS OF SPEAKERS OF LANGUAGES. The ILD uses language evenness in conjunction with language richness as a proxy for linguistic diversity. Because the goal of the index is to measure trends in linguistic diversity, it must account for changes in richness and evenness: that is, changes in the relative distribution of mother-tongue speakers among discrete languages within the total population, as measured from the starting point of the index to its ending point. For any given grouping of languages at a particular starting point in time—call it Time 0—the way we measure their relative distribution is to calculate each one's share of the total population of the grouping and then find the average of those shares; this average share becomes the numerical benchmark for relative distribution at Time 0. We then move to a subsequent point in time—Time 1—and redo the calculations. This yields a new average share. We then compare the change in average share from Time 0 to Time 1, thus producing a trendline of changes in the relative distribution of the languages in that grouping.

As an example, consider languages grouped at the global level. In any given year, each language in the world holds a particular share of the global population: languages with a large number of mother-tongue speakers have greater shares, while languages with a smaller number have lesser ones. With each passing year the shares held by individual languages change—and thus the average share changes—because (1) the world's languages are growing at different rates and (2) speakers are shifting between languages. Tracking those changes in average share across the years produces a trend in the distribution of the world's languages, and the simplest way to show the trend graphically is by depicting changes in the average share as a single line that goes either up or down from one year to the next. That is what the ILD Global trendline does.

It is important to specify what we mean by "average," because in mathematics there are actually several kinds of averages, some of which are more appropriate for analyzing certain sets of numbers than others. When most people use the word "average," they usually mean a simple calculation in which one adds a set of numbers and then divides by a count of numbers in the set; thus, the average of 2 and 8 is 5 (2 + 8 = 10 divided by 2). Technically, this calculation is called the arithmetic mean, and it works fine for simple sets of numbers. But another kind of average, the geometric mean, is more appropriate for analyzing data sets with skewed distributions such as the size distribution of languages. Wikipedia happens to have a very clear explanation:

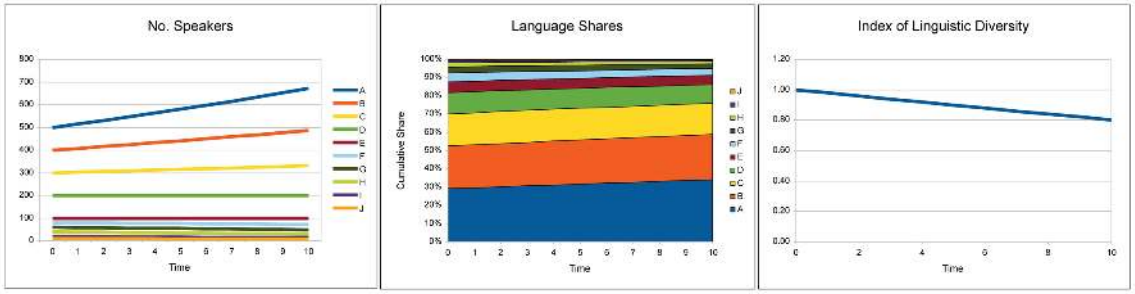
> The geometric mean, in mathematics, is a type of mean or average, which indicates the central tendency or typical value of a set of numbers. It is similar to the arithmetic mean, which is what most people think of with the word "average," except that instead of adding the set of numbers and then dividing the sum by the count of numbers in the set, n, the numbers are multiplied and then the nth root of the resulting product is taken.... The geometric mean ... is ... often used for a set of numbers whose values are meant to be multiplied together or are *exponential in nature, such as data on the growth of the human population* or interest rates of a financial investment. (emphasis added; http://en.wikipedia.org/wiki/Geometric_mean, accessed March 2010)

In our simple example above, the arithmetic mean was 5, but the geometric mean is 4 (2 x 8 = 16, and thence the square root of 16, since there are 2 numbers in the set; if there had been 3 numbers one would take the cube root, etc.).

When we use the term "average" with respect to the ILD, we refer to the geometric mean, not the arithmetic mean. The reason we use the geometric mean is precisely because of the consideration that we have highlighted in the quoted definition above: we are analyzing language data in a world where the numbers of speakers are unevenly distributed among languages: more than 94% of the world's people speak one of the 389 largest languages, each of which has more than a million speakers, while the other 6,520 non-extinct languages account for the fewer than 5% of the world's population (http://www.ethnologue.com/ethno_docs/distribution.asp?by=size, accessed March 2010; Lewis 2009 includes the same analysis but with slightly differing figures). If we were to construct an index using the conventional notion of "average"—i.e., by calculating trends in the arithmetic mean—we would be unable to accurately reflect shifts in evenness because the arithmetic mean would not give a meaningful measure of the extremely skewed distributions. Indeed, the arithmetic mean share of the world's population is constant over time for any distribution of speaker numbers as long as they remain greater than zero.

Figures B-1 provides an example that illustrates this, and also shows how the ILD is calculated. We have set up the example as a simplified model of the real world where there are a certain number of languages with different numbers of mother-tongue speakers. In our example the world consists of 10 languages, A through J, each having a different number of speakers at Year 0, the starting point of the index calculation. We then set a different growth rate for each language—just as in the real world each language is growing (or declining) at a different rate—through Year 10, the endpoint of the index. In Figure B-1, we made Languages A, B, and C grow; Languages D and E stay the same; and Language F, G, H, I, and J decline. And, just like in the real world, the overall population is growing, going from 1,710 in Year 0 to 1,962 in Year 10.

FIGURE B-1



Language	0	1	2	3	4	5	6	7	8	9	10	Annual Rate	Percent change
A	500	515	530	546	563	580	597	615	633	652	672	1.03	3%
B	400	408	416	424	433	442	450	459	469	478	488	1.02	2%
C	300	303	306	309	312	315	318	322	325	328	331	1.01	1%
D	200	200	200	200	200	200	200	200	200	200	200	1.00	0%
E	100	100	100	100	100	100	100	100	100	100	100	1.00	0%
F	80	79	78	78	77	76	75	75	74	73	72	0.99	-1%
G	60	59	58	56	55	54	53	52	51	50	49	0.98	-2%
H	40	39	38	37	35	34	33	32	31	30	29	0.97	-3%
I	20	19	18	18	17	16	16	15	14	14	13	0.96	-4%
J	10	10	9	9	8	8	8	8	7	7	7	0.96	-4%
Total (N)	**1710**	**1732**	**1754**	**1777**	**1801**	**1826**	**1851**	**1878**	**1905**	**1933**	**1962**		
pA	0.29	0.30	0.30	0.31	0.31	0.32	0.32	0.33	0.33	0.34	0.34		
pB	0.23	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.25	0.25	0.25		
pC	0.18	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17		
pD	0.12	0.12	0.11	0.11	0.11	0.11	0.11	0.11	0.10	0.10	0.10		
pE	0.06	0.06	0.06	0.06	0.06	0.05	0.05	0.05	0.05	0.05	0.05		
pF	0.05	0.05	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04		
pG	0.04	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.02		
pH	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02		
pI	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01		
pJ	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		
Arithmean	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10		
Geomean	**0.054**	**0.053**	**0.052**	**0.051**	**0.050**	**0.049**	**0.048**	**0.047**	**0.045**	**0.044**	**0.043**		
ILD	**1.00**	**0.98**	**0.96**	**0.94**	**0.92**	**0.90**	**0.88**	**0.86**	**0.84**	**0.82**	**0.80**		

However, when we look at the geometric mean we see a different story: the average is decreasing, going from 0.54 in Year 0 to 0.43 in Year 10. This is because the geometric mean is not just indicating a raw average, as the arithmetic mean does; rather, it indicates the average share of the global population held by each language in a world where the size distribution of languages is highly skewed and languages are growing at different rates (positive or negative). That is why we use the geometric mean to measure the situation of languages in the real world. And in this example, the geometric mean correctly indicates that a loss of distributional diversity is occurring.

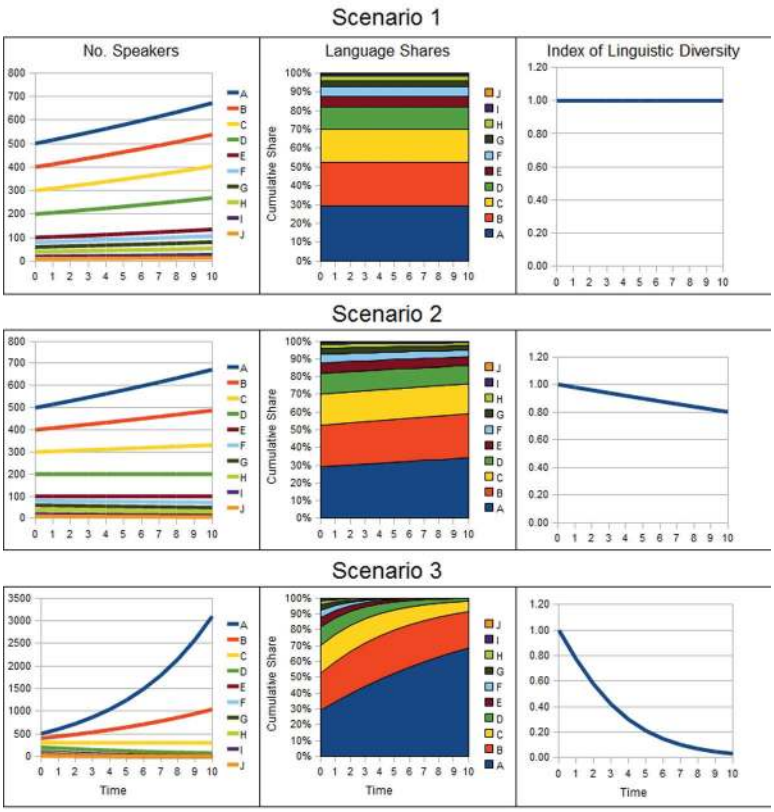
The ILD simply compares changes in the geometric mean—the average share—over time by dividing the geometric mean at the endpoint of the index by that at the starting point. Here, the calculation is 0.43 divided by 0.54 = 0.80. So in Figure B-1 the ILD declines from 1.00 to 0.80—just as ILD Global did in the real world over the period 1970–2005.

This was by design, of course: we set up the growth rates such that Figure B-1 would approximate the situation in the real world. This gives us a point of comparison to which we can add two other scenarios that show how the ILD behaves under different extremes.

In Figure B-2 we have taken the graphs from Figure B-1 and flanked them by similar graphs that illustrate these additional scenarios. On the top row (Scenario 1), we begin with the same simplified world as before: the same ten languages with the same starting populations as in Figure B-1. But this time all the languages grow at the same rate—that

is, each one holds its share of the population. This is the hypothetical situation of stability that produces a perfectly flat ILD trendline, as can be seen from the upper righthand graph. On the bottom row (Scenario 3), we have the same starting conditions, but this time the annual growth rates reflect a sharp decline in most of the ten languages. Here, we see that the ILD trendline declines sharply, reflecting the steep loss of diversity under this scenario. If one compares the three middle graphs (Language Shares), we see that as diversity declines as we go from the top to the bottom row, the area of the graph taken up by the largest languages shifts or begins to "bulge," with the area taken up by the smaller languages being "squeezed out." This is a graphical depiction of shifts in the distribution (or concentration) of the world's speakers.

FIGURE B-2



CALCULATING THE ILD IN A REAL WORLD OF MISSING DATAPOINTS. The calculation of the ILD works in three steps. The description below differs from the simplified version given in the main text in that one does not need to know the number of speakers of every language in every year in order to calculate the index. It allows for missing datapoints by interpolating between datapoints, assuming a constant annual rate of growth (or decline). This is the simplest assumption one can make in the absence of data. However,

no datapoints were extrapolated using this method, as the assumption of a constant annual rate of change beyond the first and last data years is not always reasonable. Therefore missing datapoints remained prior to the first data year and after the last data year for each language. To allow for this, the index was calculated by finding the average change in share from one year to the next for all languages for which actual or interpolated datapoints existed, and then chaining together the average changes for each year into an index starting at one in the baseline year. This method is adapted from that of the Living Planet Index (Loh et al. 2005).

1. The fraction F of the total population (global or regional) represented by each datapoint (a datapoint means N speakers of language *l* in year *y*) was calculated.

$$F_{ly} = (N_{ly}+1)/P_y$$

where N_{ly} is the number of speakers of language *l* in year *y*, and
P_y is the total population in year *y*.

To avoid taking the log of zero or dividing by zero in step 2, each *N* value was increased by 1. The total populations from 1950 to 2005 of the world and five regions—Africa, Asia, Pacific, Europe and the Americas—were taken from UN Population Division (2006 revision). Downloaded from http://esa.un.org/unpp/index.asp.

Missing annual values between consecutive N_{ly} values were interpolated. This was done by assuming a constant annual rate of change between two datapoints. The intermediate values were calculated using a simple log-linear interpolation.

$$N_i = N_p(N_q/N_p)^{(i-p)/(q-p)}$$

where *i* = year of intermediate datapoints,
p = year of the preceding datapoint, and
q = year of the subsequent datapoint.

For example if $N_{1980} = 1000$ and $N_{2000} = 100$,
then $N_{1990} = 1000 \times (100/1000)^{(10/20)} = 1000 \times 0.1^{1/2} = 316$

2. The geometric mean of the ratio of fraction of speakers from one year to the next across all languages in the sample was calculated. This was done by log-transforming the ratio of consecutive F values such that:

$$d_y = log_{10}(F_{ly}/F_{ly-1})$$

F_{ly} = fraction of population speaking language *l* at year *y*,
F_{ly-1} = fraction of population speaking language *l* the preceding year.

The mean d value for all languages with data in a single year was then calculated

$$\bar{d}_y = \frac{1}{n_y} \sum_{i=1}^{n_y} d_{iy}$$

where

n_y = number of languages with some value (actual or interpolated) for F in the year y (not all languages in the sample have data for every year of the index because the earliest datapoint may be after 1970 or the most recent before 2005, and no values were extrapolated).

3. Finally, the geometric means in each year were antilogged and chained together to form an index, such that

$I_y = I_{y-1}\, 10^{dy}$
and the index value in 1970 was set to unity.

$I_{1970} = 1.0$

where I_y = the Index of Linguistic Diversity in year y.

In this way, the ILD shows the trend in the fraction of the total population that speaks a language that is average or typical of all languages in the sample. Note that the interpolation was not linear but log-linear, and that the average change in numbers across all languages was taken as the geometric mean and not the arithmetic mean. This means that increases and decreases in the ILD are equivalent to each other for the purpose of calculating the index. For instance, using log-linear interpolation and log-transforming all the data in this way, a doubling of the fraction of a population speaking language A between 1970 and 2005 would be cancelled out by a halving of the fraction of the population speaking language B over the same period. This is because doubling means multiplying by 2, whereas halving represents multiplying by 0.5. The arithmetic mean of 2 and 0.5 is 1.25, whereas the geometric mean is 1.

References

Adelaar, Willem. 2007. Latin America. In Moseley 2007a, 97–100.
Adelaar, Willem & J. Diego Quesada. 2007. Meso-America. In Moseley 2007a, 197–209.
Black, Paul. 1983. Aboriginal languages of the Northern Territory. Darwin, NT, Australia: School of Australian Linguistics, Darwin Community College.
Bradley, David. 2007. East and Southeast Asia. In Moseley 2007a, 349–422.
Canonge, Elliott & Dick Pittman. N.d. [ca. 1958]. *The fifth edition of the Ethnologue of Bibleless tribes for prayer intercessors, Bible translators, missionaries, prospective missionaries, mission councils*. Glendale, CA: Wycliffe Bible Translators.
Crevels, Mily. 2007. South America. In Moseley 2007a, 103–194.
Dimmendaal, Gerrit J. & F. K. Erhard Voeltz. 2007. Africa. In Moseley 2007a, 579–634.
Dorian, Nancy C. 1981. *Language death: The life cycle of a Scottish Gaelic dialect*. Philadelphia: University of Pennsylvania Press.
Evans, Nicholas. 2001. The last speaker is dead—long live the last speaker! In Paul Newman and Martha Ratcliff (eds.). *Linguistic fieldwork*, 250–281. Cambridge: Cambridge University Press.
Evans, Nicholas. 2010. *Dying words: Endangered languages and what they have to tell us*. Chichester, UK: Wiley-Blackwell.
Garza Cuarón, Beatriz & Yolanda Lastra. 1991. Endangered languages in Mexico. In Robert H. Robins and Eugenius Uhlenbeck (eds.), *Endangered languages,* 93–134. Oxford and New York: Berg.
Golla, Victor. 2007. North America. In Moseley 2007a, 1–95.
Gordon, Raymond G., Jr. (ed.). 2005. *Ethnologue: Languages of the world*. 15th edn. Dallas: SIL International.
Grimes, Barbara F. (ed.). 1974. *Ethnologue*. 8th edn. Huntington Beach, CA: Wycliffe Bible Translators.
Grimes, Barbara F. (ed.). 1978. *Ethnologue*. 9th edn. Huntington Beach, CA: Wycliffe Bible Translators.
Grimes, Barbara F. (ed.). 1984. *Ethnologue: Languages of the world*. 10th edn. Dallas: Wycliffe Bible Translators.
Grimes, Barbara F. (ed.). 1988. *Ethnologue: Languages of the world*. 11th edn. Dallas: Summer Institute of Linguistics.
Grimes, Barbara F. (ed.). 1992a. *Ethnologue: Languages of the world*. Vol. 1. 12th edn. Dallas: Summer Institute of Linguistics.
Grimes, Barbara F. (ed.). 1992b. *Ethnologue index*. Vol. 2. 12th edn. Dallas: Summer Institute of Linguistics.
Grimes, Barbara F. (ed.). 1996a. *Ethnologue: Languages of the world*. 13th edn. Dallas: Summer Institute of Linguistics.
Grimes, Barbara F. (ed.). 1996b. *Ethnologue language family index*. 13th edn. Dallas: Summer Institute of Linguistics.
Grimes, Barbara F. (ed.). 1996c. *Ethnologue name index*. 13th edn. Dallas: Summer Institute of Linguistics.

Grimes, Barbara F. (ed.). 2000a. *Ethnologue, Volume 1: Languages of the world*. 14th ed. Dallas: SIL International.

Grimes, Barbara F. (ed.). 2000b. *Ethnologue, Volume 2: Maps and indexes*. 14th ed. Dallas: SIL International.

Gunnemark, Erik & Donald Kenrick. 1985. *A geolinguistic handbook*. Gothenburg, Sweden: The authors.

Gunnemark, Erik & Donald Kenrick. 1983. *What language do they speak? A geolinguistic handbook covering languages, countries and peoples in the whole world*. Gothenburg, Sweden: The authors.

Harmon, David. 1995. The status of the world's languages according to *Ethnologue*. *Southwest Journal of Linguistics* 14:1/2. 1–28.

Harmon, David. 2002. *In light of our differences: How diversity in nature and culture makes us human*. Washington, DC: Smithsonian Institution Press.

Haspelmath, Martin, Matthew S. Dryer, David Gil, and Bernard Comrie (eds.). 2005. *The world atlas of language structures*. Oxford: Oxford University Press.

ILO [International Labour Organisation]. 1989. *C169 Indigenous and Tribal Peoples Convention, 1989*. http://www.ilo.org/ilolex/english/convdisp1.htm.

Krauss, Michael. 1992. The world's languages in crisis. *Language* 68. 4–10.

Krauss, Michael. 2006. Classification and terminology for degrees of the language endangerment. In Matthias Brenzinger (ed.), *Language diversity endangered,* 1–8. Berlin: Mouton de Gruyter.

Lewis, M. Paul (ed.). 2009. *Ethnologue: Languages of the world*. 16th edn. Dallas: SIL International.

Loh, Jonathan, R.E. Green, T. Ricketts, J. Lamoreaux, M. Jenkins, V. Kapos & J. Randers. 2005. The Living Planet Index: Using species population time series to track trends in biodiversity. *Philosophical Transactions of the Royal Society of London B* 360. 289–295.

Maffi, Luisa. 2005. Linguistic, cultural, and biological diversity. *Annual Review of Anthropology* 29. 599–617.

Meillet, A[ntoine] & M[arcel] Cohen. 1952. *Les langues du monde*. Revised edn. Paris: Centre National de la Recherche Scientifique / H. Champion.

Meillet, A[ntoine] & M[arcel] Cohen. 1924. *Les langues du monde*. Collection Linguistique publiée par La Société de Linguistique de Paris 16. Paris: Librairie Ancienne Édouard Champion.

Moseley, Christopher (ed.). 2007a. *Encyclopedia of the world's endangered languages*. London and New York: Routledge.

Moseley, Christopher. 2007b. General introduction. In Moseley 2007a, vii–xvi.

Nettle, Daniel. 1999. *Linguistic diversity*. Oxford: Oxford University Press.

Pittman, R[ichard] S. (ed.). 1965. *Ethnologue*. 6th edn. Santa Ana, CA: Wycliffe Bible Translators.

Pittman, Richard S. (ed.). 1970. *Ethnologue*. 7th edn. Santa Ana, CA: Wycliffe Bible Translators.

Salminen, Tapani. 2007. Europe and North Asia. In Moseley 2007a, 211–280.

Skutnabb-Kangas, Tove. 2000. *Linguistic genocide in education—or worldwide diversity and human rights?* Mahwah, NJ: Lawrence Erlbaum Associates.

Tesnière, L. 1928. *Statistique des langues de l'Europe*. Appendix in A[ntoine] Meillet, *Les langues dans l'Europe nouvelle*, 291–484. Paris: Payot.
UNDESA [United Nations Department of Economic and Social Affairs]. 2009. *State of the world's indigenous peoples*. New York: United Nations.
UNESCO [United Nations Educational, Scientific, and Cultural Organisation]. 2009a. A methodology for assessing language vitality and endangerment. http://www.unesco.org/culture/ich/index.php?pg=00142.
UNESCO. 2009b. *UNESCO interactive atlas of the world's languages in danger.* http://www.unesco.org/culture/ich/index.php?pg=00206. (Accessed February 2010.)
UNPFII [United Nations Permanent Forum on Indigenous Issues]. 2008. *International Expert Group Meeting on Indigenous Languages*. http://www.un.org/esa/socdev/unpfii/en/EGM_IL.html. (Accessed February 2010.)
van Driem, George. 2007. South Asia and the Middle East. In Moseley 2007a, 283–347.
Voegelin, C. F. & F. M. Voegelin. 1977. *Classification and index of the world's languages*. New York: Elsevier.
WBT [Wycliffe Bible Translators]. 1951. *Missionary ethnologue for intercessors, translators, missionaries, and mission councils*. 1st edn. Berwick, Victoria, Australia: Wycliffe School of Linguistics. (9 pp., mimeograph.)
WBT [Wycliffe Bible Translators]. 1952. *Translator's ethnologue for intercessors, translators, missionaries, and mission councils*. 2nd edn. Grand Forks, ND, and Glendale, CA: Wycliffe Bible Translators. (25 pp., mimeograph.)
WBT [Wycliffe Bible Translators]. 1953. *Missionary ethnologue for intercessors, translators, missionaries, and mission councils*. Revised and extended edn. of 1st (1951) edn. Berwick, Victoria, Australia: Wycliffe School of Linguistics. (3 pp., mimeograph.)
WBT [Wycliffe Bible Translators]. N.d. [ca. 1953]. *Translator's ethnologue for intercessors, translators, missionaries, and mission councils*. 4th edn. Norman, OK, and Glendale, CA: Wycliffe Bible Translators. (28 pp., mimeograph)
Whalen, D.H. & Gary F. Simons. 2009. Endangered language families. Paper presented at the 1st International Conference on Language Documentation and Conservation, University of Hawai´i, 12–14 March.
Winter, J.C. 1981. Die Khoisan-Famille. In Bernd Heine, Thilo C. Schadeberg, and Ekkehard Wolff, eds., *Die Sprachen Afrikas: Ein Handbuch*, 329-374. Hamburg: Helmut Buske Verlag.
Wurm, Stephen A. 2007. Australasia and the Pacific. In Moseley 2007a, 425–577.
Wurm, S. A. & Shirô Hattori (eds.) 1981. Language atlas of the Pacific area. Part 1, New Guinea area, Oceania, Australia. Canberra: Pacific Linguistics C-66.
WWF [Worldwide Fund for Nature] 2008. *Living Planet Report 2008*. Gland, Switzerland; WWF.

David Harmon
dharmon@georgewright.org

Jonathan Loh
jonathan@livingplanet.org.uk