REVIEW ARTICLE

Ethnologue 16/17/18th editions: A comprehensive review

Harald Hammarström

Max Planck Institute for Psycholinguistics, Nijmegen

Ethnologue: Languages of the world. 16th edn., ed. by M. Paul Lewis, 2009. 17th edn., ed. by M. Paul Lewis, Gary F. Simons, and Charles D. Fennig, 2013. 18th edn., ed. by M. Paul Lewis, Gary F. Simons, and Charles D. Fennig, 2015. Dallas: SIL International.

Ethnologue (http://www.ethnologue.com) is the most widely consulted inventory of the world's languages used today. The present review article looks carefully at the goals and description of the content of the *Ethnologue*'s 16th, 17th, and 18th editions, and reports on a comprehensive survey of the accuracy of the inventory itself. While hundreds of spurious and missing languages can be documented for *Ethnologue*, it is at present still better than any other nonderivative work of the same scope, in all aspects but one. *Ethnologue* fails to disclose the sources for the information presented, at odds with well-established scientific principles. The classification of languages into families in *Ethnologue* is also evaluated, and found to be far off from that argued in the specialist literature on the classification of individual languages. *Ethnologue* is frequently held to be splitting: that is, it tends to recognize more languages than an application of the criterion of mutual intelligibility would yield. By means of a random sample, we find that, indeed, with confidence intervals, the number of mutually unintelligible languages is on average 85% of the number found in *Ethnologue*.*

Keywords: *Ethnologue*, number of languages, mutual intelligibility, language classification, definition of language

* This review article was originally written for the 16th edition of *Ethnologue*. Since it took many years to complete the research needed to write the review, it was not submitted until February 23, 2013, that is, four years after the appearance of the 16th edition. Only weeks after, in March 2013, the 17th edition was released. Given that a review of an outdated edition would be of much less value, this review was subsequently updated (in October 2013 to December 2014) to also cover the 17th edition. During the editorial process in early 2015, the 18th edition of *Ethnologue* was released. The 18th edition differs less from the 17th edition than the 17th differs from the 16th, and so this review was updated once again (in July 2015) to also cover the 18th edition. Wherever relevant, the text reviews all three editions in parallel, allowing the reader to appreciate the differences between them.

Over 250 individuals helped me on an ad hoc basis with clarificatory and confirmatory information about the language situation in their area of expertise. Those whose help was of special value are cited by name in the corresponding place in the text. I also wish to acknowledge a special debt of gratitude to Roger Blench for answering all questions Nigerian and beyond and to Bonny Sands for extraordinary help with access to hard-to-find source materials. None of these people are responsible for any misinterpretations I may have added.

I also wish to thank the following libraries for granting access and services: Centralbiblioteket (Gothenburg), Institutionen för orientaliska och afrikanska språk (Gothenburg), Etnografiska Muséet (Göteborg), LAI (Göteborg), Carolina Rediviva (Uppsala), Nordiska Afrikainstitutet (Uppsala), Karin Boye (Uppsala), Kungliga Biblioteket (Stockholm), Stockholms Universitets Bibliotek (Stockholm), Latin-Amerika Institutet (Stockholm), Universiteitsbibliotheek (Leiden), KITLV (Leiden), Universiteitsbibliotheek (Amsterdam), Institute for Asian and African Studies (Helsinki), Max Planck Institute for Evolutionary Anthropology (Leipzig), Max Planck Institute for Psycholinguistics (Nijmegen), Universitätsbibliothek (Leipzig), Butler/Columbia University (New York City), Institut für Afrikanistik (Cologne), Bibliothèque Nationale Française (Paris), INALCO (Paris), SOAS (London), ILPGA (Paris), Sprachwissenschaft (Zürich), Radboud Universiteit (Nijmegen), Ibero-Amerikanisches Institut (Berlin), Asien-Afrika Institut (Hamburg), Museo Nacional de Antropología (Mexico City), Biblioteca Daniel Cosío Villegas (of El Colegio de México, Mexico City), and Völkerkundliche Bibliothek (Frankfurt).

This research was made possible thanks to the financial support of the Language and Cognition Department at the Max Planck Institute for Psycholinguistics, Max-Planck Gesellschaft, and a European Research Council's Advanced Grant (269484 'INTERACT') to Stephen C. Levinson.

1. GENERALITIES. The *Ethnologue* is a work aiming to catalogue all known living languages of the world. The 16th edition (henceforth E16) was released in 2009, and its entries were taken over as ISO 639-3 standard for language identification. The 17th edition (henceforth E17) was released in 2013, but the dependence was now reversed, and E17 explicitly states that it reproduces the inventory rendered by ISO 639-3. The 18th edition (henceforth E18) was released in 2015 and continues the latter relationship to ISO 639-3. The 16th and 17th editions come as hardcover books covering over 1,000 pages, but the full contents of the books are also freely available online at http://www.ethnologue.com (for the most recent version), http://www.ethnologue.com/17/ (for the 17th edition), and http://archive.ethnologue.com/16/ (for an archived 16th edition). The web availability greatly facilitates access and searchability, providing an enormous service to the linguistic community on behalf of the SIL.

E16, E17, and E18 are organized similarly: introduction, statistical summaries, language entries, maps, and finally a bibliography and indices. I concentrate on the bulk of the work, that is, on the language entries and information about them in the introduction. Inasmuch as they are correct, there is little to say about indices, statistical summaries, and maps.

The review is organized as follows. I first review the information provided in the E16/E17/E18 introductions, including notes and numbers on the kinds of languages (pidgin, sign, speech registers, etc.) listed (§2). The accuracy of the E16/E17/E18 language inventory compared to that which can be gauged from the literature is measured in §3. Actual lists of spurious and missing languages can be found in the online appendices along with references to the literature that substantiate their claimed status. Section 4 provides empirical data on the relation between mutual intelligibility and the language/dialect divisions actually found in the E16/E17/E18 entries, and discusses the implications this has for the number of languages in the world. The E16/E17/E18 classifications of languages into families are addressed in §5, and the merits of E16/E17/E18 vis-à-vis alternative listings are discussed in §6. The review concludes with overall impressions (§7). Additional detailed information is provided in online appendices, which are available at http://muse.jhu.edu/journals/language/v091/91.3.hammarstrom01.pdf. Appendix A lists languages missing from E16/E17/E18, and Appendix B lists entries in E16/E17/E18 that are spurious. Appendix C contains examples of erroneous classifications in E16. Appendix D contains an assessment of language/dialect divisions on a sample of 100 languages from E16/E17/E18.

2. THE INTRODUCTIONS. The introductions are concise but provide a good explanation of the principles behind E16/E17/E18.[1] This is not an easy task, and many comparable works resort to smoke-screening the fact that they do not know (or care) about the principles actually used in language listings. In E16/E17/E18 we are given an explanation of what the aims and limits of inclusion are, what different kinds of entries there are (pidgin, sign, etc.), and what information various fields contain (population, region, map projection), as well as a fairly extensive discussion of levels of language endangerment. Examples of descriptions that became clearer in the 16th edition compared to the 15th are on the systematic information about Bible translation and on the occasional inclusion of extinct languages. (Information on Bible translation is said to be included because the Bible is the most widely translated of all books.) Examples of descriptions

[1] The introduction chapter in the book version corresponds to the information in the About tab in the online version.

that became clearer in the 17th edition compared to the 16th are the more elaborate explanations of the population, typology, location, and dialects fields. Examples of descriptions that became clearer in the 18th edition compared to the 17th are the more elaborate explanations of the Language status field (which covers language endangerment), the website, and the nature of updates.

A significant difference between E16, E17, and E18 concerns the listing of extinct languages. In the introduction to E16, it is stated that the aim is to include SOME extinct languages (as a bonus on the set of living languages, where the aim is to include ALL), namely:

- extinct languages that were listed as living in some previous *Ethnologue* edition but subsequently went extinct,[2] and
- extinct languages that are in current use in the scriptures or liturgy of a faith community.

In E17, there is no such passage. E17 is explicitly declared to follow ISO 639-3, which does aim to include all types of extinct languages,[3] and indeed, many Australian languages extinct before 1951 and absent from E16, for example, were carried over from ISO 639-3 into E17. Moreover, new for E17 is a Language status field, which (in addition to political recognition) encodes extinctness, level of endangerment, and degree of vitality if revitalized. There is thus no stated policy in E17 to only cover living languages,[4] or to only cover post-1951 living languages plus extinct liturgical languages. In E18, the corresponding text of the introduction has reverted back to the E16 stance.

While, for the most part, the E16/E17/E18 introduction does not hide pertinent information, in a number of cases it does, and in a number of other cases it does not accurately describe the language listing in E16/E17/E18. I highlight the most important such problems here.

2.1. THE DEFINITION OF LANGUAGE. Perhaps the most important paragraph concerns the definition of language, which is therefore worth quoting and discussing in full:

> The ISO 639-3 standard applies the following basic criteria for defining a language in relation to varieties which may be considered dialects:
> - Two related varieties are normally considered varieties of the same language if speakers of each variety have inherent understanding of the other variety at a functional level (that is, can understand based on knowledge of their own variety without needing to learn the other variety).
> - Where spoken intelligibility between varieties is marginal, the existence of a common literature or of a common ethnolinguistic identity with a central variety that both understand can be a strong indicator that they should nevertheless be considered varieties of the same language.
> - Where there is enough intelligibility between varieties to enable communication, the existence of well-established distinct ethnolinguistic identities can be a strong indicator that they should nevertheless be considered to be different languages.

The definition is the same in all of E16, E17, and E18. I am concerned only with the descriptive standards of this definition of language, that is, whether it is understandable and, if so, whether the application of the criteria to raw data yields the listing actually found in E16/E17/E18. I do not address the question of whether this definition is the

[2] The first edition of the *Ethnologue* appeared in 1951.

[3] See http://www-01.sil.org/iso639-3/types.asp, accessed 6 October 2013.

[4] The only hint in this direction is the first sentence of the E17 introduction, which reads '*Ethnologue: Languages of the World* is a comprehensive reference work cataloging all of the world's known living languages'. The qualification 'living' here is not matched by the contents of the introduction. Thus, the phrasing is presumably a remnant from earlier editions.

most appropriate one vis-à-vis other possible definitions, since this is not something argued for in the book under review. Readers who want answers to the latter question will have to look elsewhere than E16/E17/E18.

Strictly speaking, the last two criteria of the definition do not meet the requirements for being criteria that define something because the phrasing 'can be' allows the reader to disregard them as he/she pleases. If this is intended, one cannot reproduce E16/E17/E18's list of languages based on raw data on varieties. Arguably, to make things clear, E16/E17/E18 should therefore indicate, for each language, by which of the three criteria the language in question made it onto the list. If this is not intended—that is, if the 'can be's should read as 'is'—then they should be so rephrased. If so, it would be feasible, in principle, to reproduce the E16/E17/E18 listing based on raw data. It would be advisable, however, to indicate the instantiated criteria for every language anyway, since the existence of a 'common ethnolinguistic identity' is possibly more obscure than the obscurity it obviates ('marginal intelligibility').

The phrasing of the first criterion is also infelicitous. By an often-highlighted chain of inferences, it implies that all varieties in a dialect chain constitute one language. A typical dialect situation might have A mutually intelligible with B, and B mutually intelligible with C, but A and C not mutually intelligible. By the first criterion, A and B are the same language, and B and C are the same language, which implies that all three are the same language (the latter step because of the meaning of *same*). Even the quickest glance at the actual listings in E16/E17/E18 reveals that dialect chains are not treated this way; that is, it is not the case that each dialect chain has been collapsed into one language each. In E16/E17/E18, what appears to be the case is that dialect situations, such as A, B, C above, fall out as two language entries (placing B arbitrarily), with more than two language entries in more complex dialect chains involving more separate varieties. Therefore, the mutual-intelligibility-based criterion that E16/E17/E18 ACTUALLY appear to be using is the converse of the first criterion: 'For each language entry, all varieties that belong to it are mutually intelligible'. This criterion is not operationally phrased. To make it operational (though not necessarily practical) one can prepose: 'find a grouping of varieties into languages such that … '.

2.2. MACROLANGUAGES. New for the 16th edition, and kept in the 17th and 18th, is the concept of MACROLANGUAGES (which also have three-letter ISO 639-3 codes). Macrolanguages are defined as (emphasis and list formatting added):

- MULTIPLE,
- CLOSELY RELATED individual languages that
- are deemed in SOME USAGE CONTEXTS to be a SINGLE LANGUAGE.

An arbitrary group of languages—for example, 'South American indigenous languages' or 'languages whose names begins with the letter "A" '—does not qualify as a macrolanguage because of the requirement that the languages in question should be closely related. We are not told whether E16/E17/E18 aims to be complete with respect to macrolanguages. If the definition given is to be taken literally, then the listing of fifty-five (E16) or sixty (E17/E18) macrolanguages is very incomplete, as almost any set of closely related individual languages is deemed to be a single language in SOME context; for example, this is often the case in historical classification. The motivation for introducing macrolanguages is given in the (one) line: that it 'provides us with a way to represent the fact that linguistic varieties function simultaneously as both individual units and within a larger functional matrix' (E17). Possibly, this means that the intention is for macrolanguages to serve a purpose in the sociopolitical sphere, rather than just any usage context.

Since macrolanguages do not replace ordinary languages in E16/E17/E18 and are relatively few in number, I do not discuss them further here.

2.3. LANGUAGE CLASSIFICATION. According to E16:

> Language classification information comes from a variety of sources. Generally, the organization of linguistic relationships outlined in the *International Encyclopedia of Linguistics* (Frawley 2003) is followed for most language families. For Austronesian languages, the *Comparative Austronesian Dictionary* (Tryon 1995) is followed most frequently. Departures from these primary sources are included based on more recent comparative studies as they are reported to us.

As I pointed out in a review of the 15th edition (Hammarström 2005), the reference to the *International Encyclopedia of Linguistics*, 2nd edn. (*IEL*, Frawley 2003), is an empty self-reference since the *IEL* follows *Ethnologue*'s 14th edition in its classification (Frawley 2003:xiv):

> These lists [of language families and their members] were compiled by Barbara Grimes—not by authors of the articles—using the *Ethnologue* … There remain great controversies in the field over which languages belong to which families, and, indeed, some of the groupings in the lists are at odds with the positions of the authors of the articles. The goal of including the lists was not to resolve controversies—or promote them!—but to ensure that the user has maximum information.

The *IEL* adds no further substance to the classification, and surely one can provide the user with better 'maximum information' than arbitrariness and contradiction, which the passage boils down to. Similarly, the classification in the *Comparative Austronesian dictionary* (Tryon 1995) says (Grimes et al. 1995:122) it follows the *Ethnologue* 11th edition (Grimes 1988) for all but the higher-level nodes, without adding or committing any extra substance to this classification.

Furthermore, the E16 claim that 'departures from these primary sources are included based on more recent comparative studies as they are reported to us' is not accurate. In reality, SIL has a team of subarea editors who prepare reports to the general editor. The present reviewer has seen such reports. These reports cover classification and combine opinions from SIL area experts and advice actively solicited (by the subarea editors) from non-SIL specialist linguists. The subarea editors find compromises for differing opinions within their respective areas, but there is no evidence in the macrolevel classification of any attempt at unifying the (widely differing) principles for classification current in the subareas. Beyond these subarea reports, according to testimonies from many colleagues in linguistics, it appears that classification information submitted voluntarily by non-SIL linguists to the editor is set aside if not cosubmitted with a supporting SIL member.

While it would be inappropriate to ask that the SIL embark on a large-scale enterprise of historical linguistics, it is perfectly appropriate to request that the procedure underlying the E16 language classification should be described correctly, regardless of whether this procedure is justified. A procedure that gives credence to SIL members over nonmembers obviously could not survive scientific scrutiny, but it would nevertheless prevent misunderstandings about the E16 classification, which is increasingly being cited as 'compromise' classification.

The corresponding section in E17 (a similar passage is retained in E18) has improved in its descriptive accuracy and no longer contains the circular justification:

> Language classification information comes from a variety of sources. The Ethnologue attempts to report the generally accepted consensus of scholars working in the language family based on published works and scholarly review. For this edition, the language classifications for several major families have undergone thorough review and revision. The sources on which the classifications are based are not overtly cited in the language entry but may be included in the list of general references listed at the country level. The sources used for classifications are available on request by contacting the Editor; see Contact us.

However, the actual procedure for the 'attempts to report the generally accepted consensus' is still not declared. Whatever the attempts were (and were not) is highly relevant information, and the failure to disclose it runs counter to scientific principles. Similarly, if there are sources explaining the basis for classification, why not cite them overtly? Lastly, the statement that sources are available on request appears to be nominally correct, but the underlying sources for various languages appear not to be in order. I asked (Nov. 2013) for the source of the classification of five languages chosen for their peculiar E16/E17 classification: Kamar [keq], Phimbi [phm], Santa Maria La Alta Nahuatl [nhz], Enwan (Edu State) [env], and Eastern Ngad'a [nea]. For Kamar [keq] and Enwan (Edu State) [env], the classification sources were not known. For Phimbi [phm], the source was said to be Maho 2009, but this source actually follows E16 and does not have any independent evidence for the language Phimbi [phm] or its classification (Maho 2009 and p.c., November 2013). For Santa Maria La Alta Nahuatl [nhz], the source was said to be Campbell 1997, but Campbell does not mention Santa Maria La Alta Nahuatl [nhz] and makes no subdivisions of Nahuatl varieties at all (Campbell 1997:134), so this source gives no information on how to classify Santa Maria La Alta Nahuatl [nhz] against the dozens of other Nahuatl entries in E16/E17. Nor does Campbell (1997), in turn, cite any other sources that treat the classification of Santa Maria La Alta Nahuatl [nhz] (Lastra de Suárez 1986 is cited but does not cover Santa Maria La Alta Nahuatl [nhz], while Lastra 1990 does, but is not cited). For Eastern Ngad'a [nea], the sources for classification were said to be Blust 2008 and Gray et al. 2009, but neither of these sources mentions or cites any work (beyond *Ethnologue*) that mentions Eastern Ngad'a [nea]. Thus, out of the five queries for classification sources, none provided any noncircular information on the classification of the languages in question. In §5 we review the actual outcome classification (rather than the description of it).

2.4. Sources and updates. A large number of sources for individual data items are cited properly. A welcome novelty since the E16 edition is that '[c]itations of published sources in the text of *Ethnologue* follow the conventional format of author surname followed by publication year. Personal communications, unpublished, and more general sources such as censuses, are identified by placing the year before the name of the source'. For most items of data, however, no source is cited; in particular, most of the time no source is cited to justify the entry itself, or to at least explain where the data came from.

From a scientific perspective, the lack of systematic sourcing is the biggest weakness of E16/E17/E18. The lack is somewhat puzzling. After all, no data is made up of thin air—it all comes from somewhere[5]—so why not declare it? E16/E17/E18 gives only one reason, namely, space: 'Lamentably, space does not permit a listing of [every correspondent who has communicated with us since [the fifteenth edition was released in 2005 (E16)/the sixteenth edition was released in 2009 (E17)]/every contributor since *Ethnologue* came into existence (E18)]. Moreover, the list of contributors over the nearly six decades of Ethnologue publication, whose contributions can still be seen, defies documentation' (E16). Possibly this is a valid reason for the book version, but for the internet version there are no space limitations.

[5] Fortunately, it has not been the general practice of the E16/E17/E18 editorial team throughout the years to discard the source or its name once the information from it has been integrated. As I have experienced myself, it is occasionally possible to find out where a certain entry actually comes from via the help of a willing SIL member.

According to E16/E17, 'this edition contains nearly 60,000 updates and corrections from the previous one' (curiously, the claimed number of updates between the 15th and 16th editions turned out to be the same as that between the 16th and 17th editions). The meaning of this number is mysterious since it gives an average of eight updates per entry, or, on average, more than one update per field for every entry. But the updates are not evenly distributed, and whatever counts as an update is something very lightweight. Some quick computational comparisons of the 16th edition with the 15th gives the following. About 2,500 entries have not been changed at all in the name, dialects, population, and comment fields (whether explicitly indicated or not). At most, 1,350 entries have been updated and indicated as such (as evidenced by the occurrence of the tokens '2005', '2006', '2007', or '2008'). It would have been more informative if E16/E17 reported the number of updated entries or the number of updated fields, rather than the obviously diluted number of 'updates' (characters?). E18 has improved on exactly this point, reporting on the number of updated entries and the size of the update (at least one field).

2.5. FEEDBACK. New for the 17th edition[6] was the ability to register and thus be able to provide feedback to the *Ethnologue* editors directly from a specific page. Making it easier to provide feedback is certainly a step in the right direction.

2.6. THE LANGUAGE INVENTORY ACCORDING TO E16, E17, AND E18. The 7,412 (E16), 7,561 (E17), and 7,532 (E18) entries are categorized as per Table 1.

E16	LIVING	EXTINCT	NO EST	TOTAL
Macrolanguages	55	—	—	55
Canonical spoken languages	6,682	373	155	7,210
Deaf sign languages	57	1	71	129
Artificial/constructed languages	0	0	1	1
Pidgin languages	4	3	10	17
TOTAL				7,412
E17				
Macrolanguages	60	—	—	60
Canonical spoken languages	6,857	408	81	7,346
Deaf sign languages	71	3	63	137
Artificial/constructed languages	0	0	1	1
Pidgin languages	12	4	1	17
TOTAL				7,561
E18				
Macrolanguages	60	—	—	60
Canonical spoken languages	6,954	363	—	7,317
Deaf sign languages	137	1	—	138
Artificial/constructed languages	1	0	—	1
Pidgin languages	13	3	—	16
TOTAL				7,532

TABLE 1. The language inventory in numbers, as of E16, E17, and E18.

The column 'living' counts the number of entries for which E16/E17 lists a speaker number greater than zero in the population field. The column 'extinct' counts the number of entries in E16/E17 for which the population field lists zero speakers (or a phrase to this effect). The column 'no est.' counts the number of entries where 'no estimate

[6] See http://www.ethnologue.com/ethnoblog/mpl/check-out-new-ethnologue.

available' or an equivalent phrase occupies the population field. Impressionistically, most of the entries in E16/E17 with 'no estimate' are living languages for which no population estimate is given, rather than languages whose living/extinct status cannot be inferred. For E18, there is a language status field, and the columns 'living'/'extinct' then simply count the cases marked as extinct or not.

Esperanto [epo] is the sole language included as an artificial (E16)/constructed (E17/E18) language, presumably because it is the only(?) such language known to have native speakers (Bartlett 2006). A few nonnatively spoken languages—for example, Callawaya [caw] (Muysken 2009), Gail [gic] (Cage 2003), Leti (Cameroon) [leo] (Dieu & Renaud 1983), and La'bi [lbi] (Moñino 1977)—are included, but most such known languages, for example, Urban Youth languages (Kießling & Mous 2004), are not included.

It is clear that a large number of attested pidgin languages are missing. Due to the transient nature of pidgins, however, information as to the existence of a pidgin is typically more ambiguous than the corresponding information about a language with native speakers. I therefore refrain from discussing the E16/E17/E18 pidgin entries in detail, and refer to the comprehensive listing of pidgins by Bakker and Parkvall (2010). The Bakker & Parkvall 2010 listing differentiates different levels of evidence for the existence of a pidgin, rather than a strictly binary decision of existence or not.

I am not qualified to judge the sign language entries, so they are left unreviewed here. The remainder of this review is restricted to languages spoken as a first language.

3. Spurious and missing languages. A number of extant languages are missing from E16/E17/E18, and a number of entries in E16/E17/E18 are spurious, that is, do not exist as languages or duplicate other existing entries. In order to systematically enumerate missing and spurious languages from E16/E17/E18, the following method was pursued. First, a very large collection of bibliographical references[7] to descriptive work on the languages of the world was annotated as to the language(s) described, causing, for example, any reference to a language missing from E16/E17/E18 to become apparent. Second, the classification according to the research literature was reviewed for every E16/E17/E18 language, causing, for example, duplicate entries to become apparent by competing for the same slot in the classification. Third, a survey of one specific grammatical characteristic was carried out across the research literature for every E16/E17/E18 language, causing, for example, duplicate entries to become apparent by being grounded in the same source.

3.1. Missing languages. The languages missing in E16/E17/E18 are listed in Appendix A. To be more precise, a language is listed there as missing if:

- extant published literature can make a convincing case that the language exists (or existed; see below), and,
- extant published literature can make a convincing case that the language is not intelligible to any language already listed in E16/E17/E18.

An important note is that I do not list languages that are missing solely by virtue of the interpretation of a dialect situation correctly understood (but interpreted differently) in E16/E17/E18. This matter is separately treated in §4.[8] For example, if an E16/E17/E18 entry subsumes a number of varieties with borderline intelligibility, and the facts are correctly indicated (e.g. the names of the varieties given as dialects, and the comments about intelligibility), such cases are not listed here, even if there are good reasons to interpret

[7] See http://www.glottolog.org (accessed 20 January 2012) for more information.

[8] A large number of other such cases are taken up in the list of scheduled updates to the ISO 639-3 inventory, traceable via http://www.sil.org/iso639-3/changes.asp.

the same facts as yielding different entries. However, if an E16/E17/E18 entry shows signs of misunderstanding (missing the existence of a variety, having an erroneous indication of intelligibility level, or giving a blanket statement with no indicated basis, etc.), any variety that is arguably not intelligible is listed as a missing language in Appendix A.

In all cases, references are provided to the literature that support the argument made regarding the missing language in question.

Some 236 (E16), 477 (E17), and 198 (E18) missing languages were encountered. More than half of the 477 missing languages for E17 represent languages known to be extinct by 1951, which were not intended to be included in E16/E18 but were, at least according to its introduction, intended to be included in E17. (The corresponding number of missing languages in E16/E18, including those extinct by 1951, would have been 501 (E16) and 468 (E18).) The exact numbers of missing languages divided by macroarea are shown in Table 2.

3.2. SPURIOUS LANGUAGES. Appendix B lists entries in E16/E17/E18 that are spurious. To be more precise, an entry is listed here as spurious if:

- it duplicates another extant E16/E17/E18 entry, or
- it cannot be asserted that the entity denoted in the entry was a language different from every other entry in E16/E17/E18.

Again, I do not list languages that are spurious solely by virtue of the interpretation of a dialect situation correctly understood (but interpreted differently) in E16/E17/E18, and in all cases references are provided to the literature that support the argument made about the spurious language in question.

Some 191 (E16), 168 (E17), and 141 (E18) spurious languages were encountered. The numbers of spurious languages divided by macroarea are shown in Table 2.

E16	MISSING A1951	(MISSING B1951)	SPURIOUS
Africa	64	(9)	47
Australia	50	(35)	4
Eurasia	56	(93)	71
North America	13	(39)	6
Pacific	29	(5)	22
South America	24	(84)	41
TOTAL	236	(265)	191
E17			
Africa	55	11	41
Australia	40	32	6
Eurasia	52	91	59
North America	11	49	4
Pacific	25	5	17
South America	22	84	41
TOTAL	205	272	168
E18			
Africa	49	(10)	25
Australia	40	(32)	5
Eurasia	52	(90)	51
North America	11	(49)	4
Pacific	24	(5)	16
South America	22	(84)	40
TOTAL	198	(270)	141

TABLE 2. Numbers of missing and spurious languages in E16/E17/E18. The actual languages are detailed in Appendix A and B. The column marked B1951 signifies that the languages in question were extinct by 1951, while that marked A1951 signifies that the languages in question were not known to be extinct by 1951.

4. THE LANGUAGE/DIALECT DIVISION. Many blanket statements have appeared regarding the (too high?) number of languages in E16/E17/E18 and the language/dialect division. To take a few recent examples, Gippert (2012:21), with an example involving Germanic languages, declares that 'How dubious the calculation of languages in "Ethnologue" is … the number of 6,500 languages world-wide, consistently repeated in both scientific and popular publications … is nothing but a popular myth'. Similarly, Dixon (2012:463–64), citing a few examples of politically motivated language splits, argues that 'two modes of speaking are regarded as dialects of a single language if they are mutually intelligible … even the figure of 5,445 "languages" [from the tenth edition of *Ethnologue—HH*] is far too high … my estimate is that the figure is not more than 4,000, and probably a good deal less than this'. Indeed, it is easy to come up with examples of overcounting from the E16/E17/E18 listing, or, given the leeway in the E16/E17/E18 definition, to come up with examples of inconsistencies. It is also easy to come up with examples where there is no overcounting and, less easy but still not difficult, to come up with examples of undercounting (see e.g. the review of the 15th edition for examples that are all retained in E16; Hammarström 2005). However, examples are only examples and do not necessarily generalize.

I wish to point out here that defining languages on purely linguistic grounds is not necessarily fraught with THEORETICAL problems. A widespread belief holds that one cannot define language vs. dialect in any consistent and intuition-preserving way based solely on the binary (yes/no) criterion of mutual intelligibility. This view is premature: Hammarström 2008 shows that, for any set of varieties and a yes or no relation of intelligibility between each member of a pair, it is possible to define language/dialect in a consistent way, that is, such that all varieties that belong to the same language are mutually intelligible, and such that language entries are not unnecessarily multiplied. A second widespread idea holds that intelligibility between languages as a binary property (rather than gradient) is necessarily an arbitrary decision, that is, 77% lexicostatistical similarity, 87% in a sentence-repetition test, or some other threshold percentage in a text-comprehension test. This too may be premature, as a binary intelligibility without thresholds is definable on formal languages that mimic essential properties of natural languages (Hammarström 2010).

To seriously address the question of whether there is overcounting IN GENERAL in E16/E17/E18, and to obtain a sharper estimate of the number of mutually intelligible languages (henceforth MI-languages) in the world, I have sampled 100 entries from E16 AT RANDOM, checked each, and labeled it with one of the following:

- −1: represents varieties intelligible to speakers of some other entry
- OK: represents varieties intelligible to all of its own speakers but not to those of some other entry, or
- +1: represents varieties not intelligible to all of its own speakers nor to those of some other entry.[9]

The languages sampled and the individual assessment (plus source and comments) for each is given in Appendix D. In all cases, the information in the cited sources is preferable to E16 since the sources explain how and where the information presented was obtained.

[9] This indicates that the entry, based on unintelligibility, should be split. In cases encountered in the sample, the entry should be split in two, rather than some higher number.

Of the 100 entries, on the criterion of intelligibility, twenty-one should be merged with another existing entry, six entries should be split (in two), and the other seventy-three entries should remain. This boils down to a proportion of (73 + 6 * 2)/100 = 0.85 mutually intelligible languages to E16 entries. Since the sample was random, with high probability, the results do generalize (Cochran 1963).

The sample was 100 out of 6,969 entries of mother-tongue spoken languages not already deemed spurious. 0.85 * 7054 entries is 5995.9. With a confidence interval of 99%, the number of L1 spoken languages in E16 is between 5,092 and 6,899. With a confidence interval of 95%, the number of L1 spoken languages in E16 is between 5,324 and 6,668.

Given that there are something like 5,996 L1 spoken MI-languages in E16, adding the number of MI-languages not in E16 should give us the total number of known languages in the world. There are 236 MI-languages not extinct by 1951 and 265 extinct by 1951 (see Appendix A). Thus, a good estimate of the total number of known MI-languages is 6,497 (with a confidence interval of 99% it is between 5,593 and 7,400, and with a confidence interval of 95%, it is between 5,825 and 7,169). These figures are summarized in Table 3.

	ESTIMATE	95% INTERVAL		99% INTERVAL	
		LOWER	HIGHER	LOWER	HIGHER
In E16	5,996	5,092	6,899	5,324	6,668
MI-languages A1951 not in E16	236				
MI-languages B1951 not in E16	265				
TOTAL number of MI-languages	6,497	5,593	7,400	5,825	7,169

TABLE 3. Figures on the estimated number of attested assertable MI-languages spoken as a first language, based on the E16 figures with missing languages added (A1951 signifies missing MI-languages not known to be extinct before 1951, and B1951 signifies missing MI-languages extinct before 1951).

Thus, a total number of living languages around 6,000 or of known languages around 6,500 is far from being 'a popular myth'. It is a fairly well-justified estimate.

5. CLASSIFICATION. In §2, we reviewed the description of the principles said to be behind the E16/E17/E18 classification of languages into families and subfamilies. The present section addresses the actual outcome. Of spoken mother-tongue languages, *Ethnologue* recognizes 121 (E16), 140 (E17), or 132 (E18) language families, 50 (E16), 82 (E17), or 96 (E18) language isolates, and 73 (E16), 65 (E17), or 62 (E18) unclassified languages, as well as a number of mixed languages and creoles. While language classification is not the primary focus of E16/E17/E18, it is worthwhile to evaluate it properly, in order for it not to be mischaracterized and misapplied inside and/or outside the field of linguistics. For example, Pompei and colleagues (2011) call the *Ethnologue* classification an 'expert classification'. Whalen and Simons (2012:161–62) interpret E16/E17's unclassified languages as being independent linguistic stocks[10] and lament the loss of diversity if these 'unclassified' languages go extinct. Are these inferences justified?

In fact, the E16/E17/E18 classification contains a large number of languages that are not (sub)classified in harmony with experts. The first category of errors are of an elementary kind: bookkeeping, name confusion, misunderstanding of linguistic vs. nonlinguistic classification, not checking relevant research, and not keeping up with relevant

[10] A stock is defined (Whalen & Simons 2012:156) as 'the largest grouping of languages for which relatedness can be demonstrated and for which a plausible protolanguage can be reconstructed'.

research. The second category is where expert publications provide contradictory or insufficient information, and E16/E17/E18 have chosen to follow one or the other expert inconsistently, rather than attempting to find out which expert has the most/least convincing argument.

The first type of error seems to occur uniformly in all areas, except perhaps in North America. Appendix C gives some examples of errors of this kind in order to illustrate the point (for E16; the situation is not much different in E17/E18). In the interest of space, this is not (in fact, it is far from) an exhaustive list.

At the end of the day, how 'expert'-like is the E16/E17/E18 classification overall? Hammarström et al. 2014 has a complete classification and subclassification of the languages of the world based on a consistent weighing of the arguments of experts, where the justification for each node is traceable to the relevant publication. A standard way to measure the difference between two trees T_1 and T_2 is the Robinson-Foulds distance, which, in essence, counts the number of nodes found in T_1 but not in T_2 plus the number of nodes found in T_2 but not in T_1 (Day 1985). We restrict the comparison to the 6,794 (E16)/6,812 (E17)/6,835 (E18) languages that are classified as part of a family, as an isolate, or left unclassified (i.e. excluding mixed languages, creoles, pidgins, sign languages, and speech registers) and that are not spurious (as per the listing in this review).

The E16 classification thus has 2,242 nodes, of which 1,265 are also found in the classification of Hammarström et al. 2014. The Hammarström et al. 2014 classification has a total of 3,596 nodes concerning E16 languages, of which, again, 1,265 are found in E16. This amounts to an unnormalized Robinson-Foulds distance of $\frac{2242-1265+3596-1265}{2} = 1654$ and a normalized distance of $\frac{3308}{3308+1265-1} = 0.723$. This can be taken to mean that only 56.4% (1,265/2,242) of the E16 nodes are expert-like, and that only 35.2% (1,265/3,596) of expert-like nodes are recognized in E16, yielding a total expert-like-ness of only $1 - 0.723 = 0.276$ or 27.6%.

The E17 classification thus has 2,198 nodes, of which 1,337 are also found in the classification of Hammarström et al. 2014. The Hammarström et al. 2014 classification has a total of 3,617 nodes concerning E17 languages, of which, again, 1,337 are found in E17. This amounts to an unnormalized Robinson-Foulds distance of $\frac{2198-1337+3617-1337}{2} = 1570.5$ and a normalized distance of $\frac{3141}{3141+1337-1} = 0.702$. This can be taken to mean that only 60.8% (1,337/2,198) of the E17 nodes are expert-like, and that only 37.0% (1,337/3,617) of expert-like nodes are recognized in E17, yielding a total expert-like-ness of only $1 - 0.702 = 0.298$ or 29.8%.

The E18 classification thus has 2,200 nodes, of which 1,354 are also found in the classification of Hammarström et al. 2014. The Hammarström et al. 2014 classification has a total of 3,654 nodes concerning E18 languages, of which, again, 1,354 are found in E18. This amounts to an unnormalized Robinson-Foulds distance of $\frac{2200-1354+3654-1354}{2} = 1573$ and a normalized distance of $\frac{3146}{3146+1354-1} = 0.699$. This can be taken to mean that only 61.5% (1,354/2,200) of the E18 nodes are expert-like, and that only 37.1% (1,354/3,654) of expert-like nodes are recognized in E18, yielding a total expert-like-ness of only $1 - 0.699 = 0.301$ or 30.1%.

Thus, although E17 and E18 come marginally closer than E16, in no sense can E16/E17/E18 be approximated to an 'expert'-classification.

6. Discussion. Apart from the languages listed as missing/spurious and apart from extinct languages that went extinct before 1951, as far as I have been able to tell, the remaining entries in E16/E17/E18 exist in a one-to-one relationship with speech communities recognizable from the literature. However, the literature itself does not cover the world entirely. There are various regions of the world that are inhabited, but the linguis-

tic literature cannot fully account for which languages are spoken there and how they relate to other known varieties. Thus, in all likelihood, there are further languages extant in the world that neither E16/E17/E18 nor the literature can argue for convincingly.

A few trends seem, impressionistically, to be present in the list of spurious languages:

- Cross-border languages counted twice
- Both an overarching language with considerable variation and its subvarieties
- Merging of different raw lists of languages, for example, old vs. new listings or census lists vs. linguistic survey lists, without deep checking for duplicates
- Duplication of the ancestral or new language of an ethnic group who have shifted language in near-historical times
- Thin entities, for example, a people are said to have lived on a certain island without much further information

One and the same problem underlies these kinds of errors: the lack of explicit sources for the justification of a language. If there had been a source for every entry detailing what the entry is based on (location, name, linguistic data, or whatever is thought to constitute the evidence for the language), it would be a near-mechanical task to merge different lists by matching the data at hand. At present, one has to search the entire literature and second-guess the justification for the entry. Presumably, this is the reason why there are almost as many spurious languages in E16/E17/E18 as there are missing living languages.

E16/E17/E18 is not alone in not citing the individual justification for language listings. Nearly all modern language listings for continent-sized areas produced by linguists have the same policy of not citing explicit sources (or are derivative of the *Ethnologue*), for example, Dixon 2002 for Australia, Tryon 2006 for the Pacific, Masica 1993 for the Indo-Aryan languages of South Asia, Maho 2003 for the Bantu languages, Bradley 2007 for Southeast Asia, and so on. In fact, the only contemporary language listings produced by linguists that do provide individual justifications are Goddard 1996 and Mithun 1999 for North America, Adelaar & Muysken 2004 for the Andes region of South America, and van Driem 2001 for the Himalayan region. In particular, LINGUIST List,[11] which is in charge of listing extinct languages for ISO 639-3, has followed the practice of not tying entries to sources. As a standard of comparison, this listing contains more errors of all kinds mentioned in this review, on a far simpler task.

7. Conclusion. From a scientific perspective, there is really only one serious fault with E16/E17/E18, namely, that the source for the information presented is not systematically indicated. Furthermore, the introduction contains a number of items where the description of the principles behind E16/E17/E18 is questionable. Nevertheless, *Ethnologue* is an impressively comprehensive catalogue of world languages, and it is far superior to anything else produced prior to 2009. In particular, it is superior by virtue of being explicit. Most works with an overlapping goal produced by linguists contain extraordinary amounts of vagueness in language definition, borders, justification, and scope. I have listed upward of five hundred missing extinct and living languages and several hundred spurious languages, so the number of errors that could have been prevented with more work is far from negligible. The remaining entries, as far as I have been able to tell, match one-to-one with a speech community recognizable in the literature. A redivision of those speech communities along the lines of mutual intelligibility

[11] Under http://multitree.org/codes/, accessed 20 January 2012.

would recognize fewer languages (about 85%) than E16 (likely also for E17/E18). The number 85% can be ascertained with confidence intervals, so there are limits to the eagerness to split. Many languages are known only through SIL surveys, and the language inventory as a whole is reasonably well informed. There is a rapid stream of change requests submitted to ISO 639-3 on behalf of the *Ethnologue* editor covering many of the languages highlighted in the present review. Therefore, I look forward to an even sharper 19th edition.

REFERENCES

Adelaar, Willem F. H., and Pieter C. Muysken. 2004. *The languages of the Andes*. (Cambridge language surveys.) Cambridge: Cambridge University Press.

Bakker, Peter, and Mikael Parkvall. 2010. Catalogue of pidgin languages. Paper presented at the second Atlas of Pidgin and Creole Language Structures (ApiCS) Conference, 11–14 November 2010.

Bartlett, P. O. 2006. Artificial languages. *Encyclopedia of language and linguistics*, 2nd edn., ed. by Keith Brown, vol. 1, 488–90. Amsterdam: Elsevier.

Blust, Robert. 2008. Is there a Bima-Sumba subgroup? *Oceanic Linguistics* 47.45–113.

Bradley, David. 2007. East and southeast Asia. *Encyclopedia of the world's endangered languages*, ed. by Christopher Moseley, 349–424. London: Routledge.

Cage, Ken. 2003. *Gayle—the language of kinks & queens: A history and dictionary of gay language in South Africa*. Johannesburg: Jacana Media.

Campbell, Lyle. 1997. *American Indian languages: The historical linguistics of Native America*. (Oxford studies in anthropological linguistics.) Oxford: Oxford University Press.

Cochran, William G. 1963. *Sampling techniques*. 2nd edn. New York: Wiley.

Day, William. 1985. Optimal algorithms for comparing trees with labeled leaves. *Journal of Classification* 2.7–28.

Dieu, Michel, and Patrick Renaud. 1983. *Situation linguistique en afrique centrale—inventaire préliminaire: Le cameroun*. (Atlas linguistique de l'Afrique centrale.) Paris and Yaoundé: Agence de Coopération Culturelle et Technique (ACCT); Centre Régional de Recherche et de Documentation sur les Traditions Orales et pour le Développement des Langues Africaines (CERDOTOLA); Direction Générale de la Recherche Scientifique et Technique (DGRST), Institut des Sciences Humaines. [Carries the date 1983 but did not come out of the presses until 1985.]

Dixon, R. M. W. 2002. *Australian languages: Their nature and development*. (Cambridge language surveys.) Cambridge: Cambridge University Press.

Dixon, R. M. W. 2012. How many languages? *Basic linguistic theory, vol. 3: Further grammatical topics*, 463–64. Oxford: Oxford University Press.

Frawley, William J. (ed.) 2003. *International encyclopedia of linguistics*. 2nd edn. Oxford: Oxford University Press.

Gippert, Jost. 2012. Language-specific encoding in endangered language corpora. *Potentials of language documentation: Methods, analyses, and utilization* (Language Documentation & Conservation special publication 3), ed. by Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann, Dagmar Jung, Anna Margetts, and Paul Trilsbeek, 17–24. Honolulu: University of Hawai'i Press.

Goddard, Ives (ed.) 1996. *Handbook of North American Indians, vol. 17: Languages*. Washington, DC: Smithsonian Institution.

Gray, Russell D.; Alexei J. Drummond; and Simon J. Greenhill. 2009. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* 323.479–83.

Grimes, Barbara F. (ed.) 1988. *Ethnologue: Languages of the world*. 11th edn. Dallas: SIL International.

Grimes, Barbara F.; Joseph E. Grimes; Malcolm Ross; Charles E. Grimes; and Darrell Tryon. 1995. Listing of Austronesian languages. In Tryon 1995, 121–80.

Hammarström, Harald. 2005. Review of *Ethnologue*, 15th edn, ed. by Raymond G. Gordon, Jr. *LINGUIST List* 16.2637. Online: http://linguistlist.org/issues/16/16-2637.html.

Hammarström, Harald. 2008. Counting languages in dialect continua using the criterion of mutual intelligibility. *Journal of Quantitative Linguistics* 15.34–45.

HAMMARSTRÖM, HARALD. 2010. Defining intelligibility on formal languages. Paper presented at the conference of the Centre for Language Technology, Gothenburg, 9 November 2010.
HAMMARSTRÖM, HARALD; ROBERT FORKEL; MARTIN HASPELMATH; and SEBASTIAN NORDHOFF. 2014. Glottolog 2.3. Leipzig: Max Planck Institute for Evolutionary Anthropology. Online: http://glottolog.org. Accessed on July 16, 2014. Database available online: http://dx.doi.org/10.5281/zenodo.10899.
KIEßLING, ROLAND, and MAARTEN MOUS. 2004. Urban youth languages in Africa. *Anthropological Linguistics* 46.303–41.
LASTRA, YOLANDA. 1990. El náhuatl del sur de Puebla. *Anales de Antropología* 27.383–90.
LASTRA DE SUÁREZ, YOLANDA. 1986. *Las áreas dialectales del náhuatl moderno*. México: Universidad Nacional Autónoma de México.
MAHO, JOUNI FILIP. 2003. A classification of the Bantu languages: An update of Guthrie's referential system. *The Bantu languages* (Routledge language family series), ed. by Derek Nurse and Gérard Philippson, 639–51. London: Routledge.
MAHO, JOUNI FILIP. 2009. Nugl online: The online version of the new updated Guthrie list, a referential classification of the Bantu languages. Gothenburg: University of Gothenburg, Department of Oriental and African Languages. Online: http://goto.glocalnet.net/mahopapers/nuglonline.pdf.
MASICA, COLIN P. 1993. *The Indo-Aryan languages*. (Cambridge language surveys.) Cambridge: Cambridge University Press.
MITHUN, MARIANNE. 1999. *The languages of native North America*. (Cambridge language surveys.) Cambridge University Press.
MOÑINO, YVES. 1977. Conception du monde et langue d'initiation la'bi des gbaya-kara. *Langage et cultures africaines: Essais d'ethnolinguistique*, ed. by Geneviève Calame-Griaule, 115–47. Paris: François Maspéro.
MUYSKEN, PIETER. 2009. Kallawaya. *Lenguas de Bolivia, vol. 1: Ambito andino*, ed. by Mily Crevels and Pieter Muysken, 147–67. La Paz: Plural Editores.
POMPEI, SIMONE; VITTORIO LORETO; and FRANCESCA TRIA. 2011. On the accuracy of language trees. *PloS One* 6.6.e20109. Online: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0020109.
TRYON, DARRELL T. (ed.) 1995. *Comparative Austronesian dictionary: An introduction to Austronesian studies*. (Trends in linguistics: Documentation 10.) Berlin: Mouton de Gruyter. 4 vols.
TRYON, DARRELL T. 2006. Australasia and the Pacific. *Atlas of the world's languages*, 2nd edn., ed. by R. E. Asher and Christopher Moseley, 97–126. London: Routledge.
VAN DRIEM, GEORGE. 2001. *Languages of the Himalayas*. (Handbuch der Orientalistik 2:10.) Leiden: E. J. Brill. 2 vols.
WHALEN, DOUG H., and GARY F. SIMONS. 2012. Endangered language families. *Language* 88.155–73.

[harald.hammarstroem@mpi.nl]

[Received 23 February 2013;
revision accepted 29 June 2015]