An Analysis of ISO 639
Preparing the way for advancements in language identification standards

Peter Constable and Gary Simons,
SIL International

1. Introduction

Globalisation has led to an interest in an increasingly divers variety of languages. Across industry, academic and government sectors, there is a growing consensus that ISO standards for language identification are not meeting current and future needs, and that new work should be done to enhance these standards. In response to these needs, ISO TC 37/SC 2 has resolved to begin a new work initiative to extend the ISO 639 family of standards. Various extensions are being considered, including the following:

- Provide more comprehensive coverage for the world's languages, including the thousands of lesser-known languages that have been attested.

- Provide more comprehensive coverage for language collections, specifically collections based on genetic language relationships.

- Provide systems for extending language identifiers to create identifiers for paralinguistic categories, such as writing system, or identifiers for language varieties based on factors such as style, geographic region, or time period.

In Constable and Simons 2000, we discussed various issues that have not been adequately addressed in existing standards, including ISO 639. The most significant of these issues focuses on matters of definition: stating explicitly the operational definition of "language" being assumed, making clear the type of category (language, language cluster, dialect, etc.) each identifier represents, and documenting explicitly what each identifier denotes.

The importance of these issues of definition would be heightened for any system that attempts to provide comprehensive coverage of the world's languages, or that attempts to provide identifiers denoting several different types of categories. As a result, they are of particular importance for the possible areas of future work on the ISO 639 standards that are being considered.[1]

The starting points for such work—namely, the existing Part 1 and Part 2 standards—are already lacking in these areas of definition. Accordingly, it is an essential first step to any future work to

---

[1] In this paper, we will use "ISO 639" to denote both ISO 639-1 (ISO 639:1988 and successive versions) and ISO 639-2 together.

establish a more explicit set of definitions for the existing standards. Specifically, the following are necessary:

1. Make clear what types of category (language, dialect, collection, etc.) the ISO code elements can refer to and, to whatever extent possible, provide a clear statement of how those category types are defined.

2. Make clear what type of category each individual ISO code element does in fact represent.

3. Provide clear documentation as to what linguistic varieties each ISO code element denotes.

It is with this in mind that we have done a careful analysis of the existing identifiers in ISO 639, and have proposed a mapping of these identifiers to and from the languages listed in SIL's *Ethnologue* (Grimes 2000).

We consider it acceptable to use the *Ethnologue* for this purpose. The *Ethnologue* is not a perfect representation of all the world's languages. Indeed, such a result is impossible in principle.[2] The *Ethnologue* is, nevertheless, among the most complete and generally reliable compilations of information on the world's languages available today. The *Ethnologue* has identified languages with some form of operational definition for *language* in mind, one based on a primary criterion of mutual non-intelligibility, and this definition has been applied with at least some level of consistency across languages.[3] In spite of its limitations, the *Ethnologue* has become a *de facto* standard among many users because of its completeness of coverage, because the complete inventory of languages and the wealth of supporting information are readily accessible on the Web, and because it has been deemed by those users to warrant a sufficient level of their confidence.[4]

The *Ethnologue* assigns a unique three-letter code for each language within its scope. Three features that make it particularly useful are that it is a single source providing comprehensive coverage of all modern, natural languages; that each of its identifiers represents the same type of category (namely, a language, as understood in terms of the operational definition it assumes); and that the denotation of each identifier is well documented and readily accessible on a public Web site. By presenting a thorough and detailed mapping of ISO code elements to languages enumerated in the *Ethnologue*, and by making this an integral part of the *Ethnologue* Web site, we can effectively document the type and denotation of each ISO language identifier to a degree of detail that has been previously lacking.

---

[2] Some of the reasons why this is impossible in principle are considered in Constable and Simons 2000. There are also the very practical considerations of limitations in the ability to conduct the research needed to compile a complete inventory of languages.

[3] We will not pretend to suggest that there is a thoroughly explicit operational definition for language used by the *Ethnologue*—the introduction discusses the factors involved in language identification but falls short of providing a concise definition. Nor are we suggesting that a definition has been applied with complete consistency in every single case. The research behind the *Ethnologue* has, however, been guided by a principle of language identification based on non-intelligibility of one language community in reference to another, and the editor has made every attempt to apply that principle throughout within the limitations of available information. We are not aware of any attempt to compile a comprehensive catalog of the world's languages that has been done with any greater attention to an explicit operational definition applied equally across all languages.

[4] The *Ethnologue*'s inventory and identifiers have been used in a number of research efforts and publications conducted by various agencies. They have also been adopted as the basis for language identification by the Linguist List (http://www.linguistlist.org), the Open Language Archive Community (http://www.language-archives.org) and the Rosetta Project (http://www.rosettaproject.org).

We have presented our proposed mappings in dynamically generated HTML pages that are available online at http://www.ethnologue.com/ iso639/. We acknowledge, though, that definitive mappings can only be specified by the owners of the ISO 639 standards since they are the ones who determine what normative definitions apply to the standards. It is our intent and hope in presenting this paper and the results of our research that the owners of the standards would review these proposed mappings and work with us to revise them as needed so that they may become definitive documentation regarding the denotation of the code elements in the standards.

In section 2 of this paper, we outline the principles by which we determined how to map ISO code elements to languages listed in the *Ethnologue*. In the course of our work, it was necessary to make judgments regarding what the ISO code elements denote, and in so doing we were able to compile in specific detail a number of issues that need to be considered in relation to the ISO standards as they exist at present. These are enumerated in section 3. Regardless of whether or not our mappings are adopted as definitive definitions for what the code elements denote, we believe that the analysis we present will provide important insights into issues in the existing codes that would need to be resolved before new work can advance. Finally, we believe that this analysis will prove helpful in guiding some of the design for new or enhanced standards for language identification. We thus close in section 4 by discussing some implications for such further work.

2. Principles applied in determining the mappings

As with any attempt to establish groupings of languages and to match labels with the range of linguistic varieties they denote, deciding how to map the ISO code elements presents a number of challenges. Sociolinguistic issues related to language variation are often complex within a single local context. In making judgments regarding this collection of language identifiers, one is confronted by sociolinguistic complexities in a number of linguistic and regional contexts.

Overlaying this, there are several issues pertaining to the nature of group identities, which can be especially complex. A given group of people may attribute more than one identity to themselves. Also, identities as perceived within a group may not necessarily match identities as perceived by outsiders. Given two people, the first may perceive the second as belonging to a distinct group, while the second may perceive that both belong to the same group. Identities can be based on ethnicity, location, politics and a number of other factors besides linguistic ones.

Adding another layer to this complexity is the fact that there is often considerable variety in how ethnic or linguistic identities are referred to. Ethnic or language names used by outsiders usually do not match those used by insiders, and even insiders do not always use the same names as one another. Different outsiders, of course, may also use different names for a group than one another, the most obvious instance of this being different language names for a particular language (for example, the English and Spanish names "French" and "francés").

Determining what the ISO code elements denote, therefore, involves assessing the complexities of each sociolinguistic situation, making a determination as to what the relevant linguistic identities are, and matching distinct labels that correspond to each given identity. This would be an enormous task overall. Using the *Ethnologue* allowed considerable simplification, however, since it has done a large amount of the work in assessing what a relevant set of linguistic identities are and in identifying various names that correspond to each. As mentioned above, it is acknowledged that the *Ethnologue* is not perfect in this assessment. Nevertheless, for this exercise, we took the *Ethnologue* at face value and assumed validity in its assessment of linguistic identities.

This constitutes our first principle:

1. We assume the distinctions made in the *Ethnologue* to be valid, and we assume that this exercise does not represent an opportunity to re-examine the *Ethnologue*'s entries in this regard.

This is not to say that there were no cases in which we wondered about the judgments reflected in the *Ethnologue*. It was an acknowledgement, however, that the judgments reflected in the *Ethnologue* are based on years of research and on assessments made by many people with far more expertise on each of the languages involved than we have. We were not, therefore, in a position to propose revisions to the *Ethnologue* at this time.

Another principle that we adopted was the following:

2. We assume complete identity between the two-letter code elements in ISO 639-1 and the corresponding three-letter code elements in ISO 639-2.

This is reflected in tables published by the ISO 639-2 Registration Authority (see http://lcweb.loc.gov/standards/iso639-2/langhome.html).[5] Thus, a two-letter code element never has a different mapping than the corresponding three-letter code elements.

In the case of a number of code elements, the appropriate mapping was clear. For instance, there was no uncertainty that the ISO 639 code elements [af] / [afr] for the language named "Afrikaans" should map to the Ethnologue code [AFK].[6]

Whenever the mapping was less than completely clear, we consulted a variety of sources. The first source for us was, of course, the English language name given in the standards, followed closely by the French language name. In a few situations, it was the French language name that served to clarify what the intended meaning of the code element was. In addition to the English and French names in the standards, we also consulted the MARC Language Codes List (at http://www.loc.gov/marc/languages/). This was done on the assumption that the MARC Language Codes List served as a key source for the codes in ISO 639-2.

In a number of cases, MARC clarified what the correct mapping should be. That was particularly true, for example, in situations in which two unrelated languages share the same name and the ISO standards give only the name, but MARC provides country information or alternate names that give clear indication to one of the possible mappings. MARC was not always helpful: there were situations in which it provided no more information than the standard itself, or worse, situations in which we considered MARC's information to be incorrect.[7] In a few particularly difficult cases, we consulted with other experts who could provide more insight into specific sociolinguistic situations.

In a large number of cases, there were multiple languages in the *Ethnologue* that used the English or French name corresponding to the given code element. Within the Ethnologue entries, this might correspond to the primary language name (either in whole or in part), an alternate name, or a dialect name. For these situations, we adopted the following principle:

---

[5] The discussion of the mappings that follows, therefore, will generally be expressed in terms of mapping from ISO 639-2 only.

[6] In this paper and in our online reports, we adopt the convention of referring to code elements within prose discussions by citing them within square brackets. Ethnologue codes are cited in upper-case letters, while ISO 639 codes are cited in lower-case letters. When code elements from both Part 1 and Part 2 of ISO 639 are cited, they will be separated by a slash; for example, [af] / [afr]. In those cases in which ISO 639-2 contains distinct B and T codes, both will be cited with the T code occurring first; for instance, [sq] / [sqi] / [alb].

[7] Each of these situations is documented by comments added to the report of our mapping provided online.

3. If a code element is classified within the standard as an "individual language code", then the preferred mapping from that code element to the *Ethnologue* is to a single *Ethnologue* language, whenever a clear choice for that language exists.

In these situations, if there was any ambiguity and the MARC documentation did not provide a clear indication of what that one language should be, we employed the following principle:

4. The mapping for an individual language code should be to the unique language (if any) that resembles a major language.

Such assessment of languages was based on a prototype definition using various criteria, including the following:

- The language is spoken by a substantially larger population than the other potential candidates (we looked for a ratio on the order of 10:1 or greater).

- The language has literature, and preferably an established tradition of internal literature development.[8]

- There is evidence that the language is used in mass media.

- The language is an official language in one or more of the countries in which it is spoken.

We will refer to this principle as the *major language variety (MLV) principle*, and to the criteria cited as the *MLV criteria*.

As noted, these criteria are for a prototype definition, one that characterises the most typical instances. It was rare that all of the MLV criteria would apply to one of the potential candidate languages. Thus, we tried to identify a single variety that was a clear best fit for the criteria. If only one of these criteria was met by one of the candidate languages, then that could still be considered sufficient provided none of the other candidates met any of the criteria.

In a limited number of cases, more than one language listed in the *Ethnologue* appeared to satisfy the MLV criteria, suggesting that each would merit its own ISO identifier. In these situations, we adopted the following principle:

5. When there are multiple languages listed in the *Ethnologue* that might possibly be associated with an ISO code element and more than one of them appears to satisfy the MLV criteria, then the code element is mapped indeterminately to those multiple major language varieties as a temporary measure.

Our position is that these cases require further consideration and some resolution. We will discuss this further in the next section.

While it was our preference to map individual-language code elements to single languages listed in the *Ethnologue*, there were many cases in which there were no candidate languages identified by the MLV criteria. For these cases we adopted a further principle:

6. When there are multiple languages listed in the *Ethnologue* that might possibly be associated with an ISO code element but none of those languages is clearly to be preferred by the MLV criteria, and if there is a reasonable basis for associating the name of that code element with a collection of languages, then the code element is taken to represent a collection of languages.

Our mapping makes clear for which code elements this was done, and also identifies what basis for common identity is operating for each collection.

---

[8] That tradition may have been recently introduced, though a long-standing tradition provides better evidence.

As we examined the collections, particularly individual language code elements that we deemed to represent collections, we adopted the following principle limiting the types of collections:

7. With the exception of certain special collections ([art], [mis], [sgn]), the basis for common identity operating within a collection can be one of the following: a genetic classification, location within a common region, or closely related languages with a shared name. Moreover, for every collection, the basis for common identity must be documented.[9]

We will discuss the various types of collections further in the following section.

For the collective language codes based on geographic region (viz. [cau], [cai], [nai], [paa], [phi], [sai]), it was necessary to determine what these included. For this, we adopted the following principle:

8. For collective language codes based on geographic region, the denotation is based on all languages within language families / phyla whose languages are spoken primarily within the specified geographic region. Individual languages that may be spoken outside the region are still part of the collection. When major subgroups of a language family / phylum are distributed between different geographic regions, then the major subgroups of that family / phylum are mapped independently to the respective regions.

Languages covered by more specific code elements are, of course, excluded as called for by the ISO 639-2 standard (see principle 11 below).

The implication of this principle is that a language that is not spoken within a region may be counted among the languages of that region. For example, the *Ethnologue* lists 74 languages in the Arawakan family. One of the languages, Garífuna [CAB] is spoken in Honduras, and another, Island Carib [CAI], is spoken on Dominica. The other 72 languages are spoken on the South American continent. Thus Arawakan is reckoned to be a family of South American Indian languages, and all 74 members of the family are included in the denotation of the collective language code [sai].

A set of significant principles that we employed relate to the terms of usage specified in clauses 4.1.1–4.1.5 of ISO 639-2:

9. We assume that code elements are not used to distinguish between dialects of a language, or between different writing systems of a language.

10. We assume that individual language code elements are not used to denote both a modern language and an ancient language of the same name (even though related).

11. The denotation of a collective language code element does not include languages that are denoted by individual-language code elements or that are included in the denotation of more specific collective language code elements.

For instance, there are ISO 639-2 code elements for Algonquian languages and also for North American Indian languages. Since Algonquian languages are all spoken in North America, it would be appropriate to reckon them as North American Indian languages. The collective "Algonquian" is more specific, however, than "North American Indian", and so mappings that were included for the former were specifically excluded for the latter.

This leads to the final principle we adopted:

---

[9] One benefit of the requirement to document the type of each collection since is that, in doing so, we ensure that each of these categories conforms to an explicit operational definition.

12. In order to provide clear documentation regarding exactly what linguistic varieties a code element denotes, it may be necessary to list varieties that it does *not* denote.

This is particularly important when there are languages that are not covered by a given code element but that use the name that is associated with that code element. So, for example, in presenting the mapping for the code element [aym] "Aymara", we indicate that it maps to Central Aymara [AYM], but also that it specifically does not map (under our proposal) to Southern Aymara [AYC].

One of the difficulties in deciding what range of language varieties is included in a given category is dealing with the tension between grouping and splitting. Some may see two varieties as being variants of the same language and thus group them into a single category, while others may see the two varieties as distinct languages. As discussed in Constable and Simons 2000, such differences amount to differences in operational definitions for *language* that may reflect differing purposes. In our work, this tension was resolved by the MLV principle in combination with our principle of taking the *Ethnologue* as it stands. In a number of situations, this put us in direct conflict with MARC documentation, which often groups multiple varieties under a single code even when a single, major language variety exists. Our decision in dealing with these differences was to remain consistent to the principles we had adopted. The existence of these differences corresponds to a difference in principles. This raises the issue as to what principles are employed in the ISO standards. This is a fundamental issue that requires consideration by ISO. We will discuss this further in section 3.1.

3. Issues arising from the analysis

The full details of our analysis of ISO code elements and their relationship to the languages listed in the *Ethnologue* are given in a Web page at http://www.ethnologue.com/iso639/analysis.asp. The results can be summarized as follows. Of the 389 individual language codes in ISO 639-2, we found that 27 were out of scope for the *Ethnologue* and thus had no equivalent, 320 had a one-to-one mapping, and the remaining 42 mapped to multiple languages in the *Ethnologue*. Of the latter group, 13 were indeterminately mapped to more than one major language variety, 15 were analysed as collections based on a genetic subgroup, 11 as collections based on a shared name, and three as collections based on a geographic region. Among the 55 collective language codes in ISO 639-2, we found that 45 were based on genetic subgroups, six on a geographic region, one on a shared name, and three were based on other criteria.

In the process of performing this analysis we encountered a number of issues that deserve discussion. Many of them, we believe, point to a need for action on the part of ISO in relation to the codes in the current standards. This is particular true for the 42 individual language codes that we were not able to map to a single language. We introduced some of the issues briefly in the previous section; in this section we discuss all the issues at greater length.

3.1 Grouping versus splitting

As mentioned above, there were a number of cases in which we mapped a code element to a single language in the *Ethnologue* while MARC indicated that the same code element is also to be used in reference to other linguistic varieties that are listed in the *Ethnologue* as distinct languages (i.e. not particularly closely related).[10] For example, if MARC usage were followed,

---

[10] These are listed in our analysis report under the heading *Individual language codes / One-to-one mappings with comments*. Note that some of the code elements listed under that heading are not of this sort, however. The comments provided make clear which ones this refers to.

[es] / [spa] would include Aragonese [AXX] and Asturian [AUB], but we excluded these by applying the MLV principle since both of these are official languages of Spain and have language associations promoting their use.

There were many other cases in which the *Ethnologue* lists distinct but closely-related languages with similar names, and we chose to map to one of these.[11] For example, we map the code elements [eu] / [eus] / [baq] to "Basque" [BSQ], but not to two other Basque languages also listed in the *Ethnologue*, Navarro-Labourdin Basque [BQE], and Souletin Basque [BSZ]. Again, this decision was based on the MLV principle. In this particular case, the variety we chose—[BSQ]— has an estimated population of 580,000, there is evidence of literature and efforts toward standardisation, and this variety is identified as an official language of Spain. In contrast, the two other varieties have much smaller populations and no official status.

In situations such as this, we recognised the potential conflict between the linguistic perspective of the *Ethnologue*, which focuses primarily on intelligibility or lack thereof as a basis for making distinctions, and other perspectives that may be based more on perceived identity than on purely linguistic factors. The *Ethnologue* often splits varieties into distinct languages where others may be inclined to group. For instance, in the case of Basque, some might be inclined to say that there is just one language, and that the two varieties corresponding to [BQE] and [BSZ] are simply dialects of that one language.

In all of these situations, ISO must decide whether to define the given code element as denoting only the one variety and to exclude the others, as we have proposed, or to group all of the varieties together. More fundamentally, what is required is a decision about the MLV principle. Should ISO 639 Part 1 and Part 2 code elements identify the single, (relatively) major or standardized variety as per the MLV principle? Or should a different principle be adopted, one that groups all varieties that have a strong shared identity as MARC has done? In light of that decision, all of the mappings should be made consistent with the principle that is adopted.

The implications of the latter alternative should be understood: if some or all of these code elements are mapped to multiple *Ethnologue* codes, then the denotation in every case will need to be reviewed to consider how inclusive it should be. What lies behind these problems is the matter of operational definitions: we will be left with no clear understanding of what operational definition of language is being used. Indeed, it raises the question of whether all of these code elements correspond to a single type of category, or whether there are multiple category types involved, each with its own definition. To pursue this option with any measure of care would really require giving some thought (with documentation) as to what types of categories should be allowed and at least an outline of what the basis for definition of each category might be.

In contrast, by adopting the more restrictive mappings we have proposed, these issues of category types and operational definitions are addressed, at least to some extent: these code elements would all be reckoned to represent a single type of category using the same operational definition reflected in the *Ethnologue*.[12]

---

[11] These are all listed in our analysis report under the heading *Individual language codes / One-to-one mappings* and would be listed as having one or more "false mappings".

[12] The does not eliminate the question of what operation definition is used altogether. As mentioned in footnote 3, the operational definition assumed by the *Ethnologue* has never been stated explicitly, nor is it possible to say how consistently a single operational definition has been applied across all of the languages it lists. Nevertheless, it has been developed with at least some degree of consistency in its definitions, and for the ISO standards to inherit that level of consistency would be a much better situation that having no constraint on operational definitions at all.

A consideration that may provide useful guidance in deciding on the right principle is the needs within the growing use of identifiers in information technology (IT) applications. If different varieties identified by the *Ethnologue* require different linguistic processing (e.g. different spelling checkers), then applications will not be adequately served by identifiers that group those varieties together.

Consideration might also be given to possible implications for additional parts of ISO 639, if created. If there is to be an additional Part 3 with a comprehensive set of distinct codes, and if that standard is adopted for use in IT applications, then any time a Part 1 or Part 2 code element maps to multiple code elements in Part 3, it means that IT applications will need to eschew the Part 1 and 2 codes and only use Part 3. On the other hand, if a Part 1 or 2 code element unambiguously denotes the single variety, then it can continue to be used as a more compact code for major languages, with Part 3 being used for minor relatives.

3.2   Indeterminate mappings of individual-language code elements to equally valid candidate languages

We found 13 cases in which an ISO code element corresponds to two or three linguistic varieties that are listed in the *Ethnologue* as distinct languages and that appear by the MLV criteria to be equally valid candidates for the single language to be mapped to an ISO individual language code.[13] For instance, the code elements [sq] / [sqi] / [alb] correspond to the language name "Albanian". The *Ethnologue* lists four distinct Albanian languages, two of which appear to be equally valid candidates for having ISO identifiers assigned to them: Tosk Albanian is a standardised language spoken by nearly 3 million people and is the official language of Albania. Gheg Albanian also has a standard literature, is spoken by an estimated 1.8 million people, and is an official language of Yugoslavia.

In these situations, we described our mapping as "indeterminate". We have presented the mapping as one-to-many, from the single ISO code element to the two or three *Ethnologue* codes, but we consider these mappings to be temporary only until these situations are resolved. We feel it is necessary for the relevant ISO committees or working groups to consider each of these thirteen cases and determine what the mapping should be. One option is to select one of the candidate languages from the *Ethnologue*. In that event, it would seem natural to consider introducing new code elements for the other languages. A second option is to treat the ISO code element as representing a collection of languages. These collections might include only the two or three candidate languages identified in each situation, though in several cases it would probably make greater sense to include other smaller, related languages as well.[14] A third option would be to make the claim that the varieties that the *Ethnologue* has identified as distinct languages are, in fact, sub-varieties of a single language. If this were done, however, it would need to be accompanied by clear evidence demonstrating that the distinctions made in the *Ethnologue* are invalid, which would then motivate the *Ethnologue* editor to make the change. Otherwise, a mere stipulation that two varieties are one and the same language raises the problems of category types and definitions described in section 3.1.

---

[13] These are all listed in our analysis report under the heading *Individual language codes / One-to-many mappings*.

[14] An issue that would be raised for these collections is what type of collection they represent. The obvious choice would be that they be collections based on shared names, but for that to be the case, many of them would have to include other languages as well. Otherwise, a new type of collection would be involved that would require a separate operational definition.

3.3   Individual-language code elements that represent collections

We found 29 cases in which individual language codes appeared to represent collections of languages rather than single languages. These included collections based on a genetic classification, collections based on shared names, and collections based on geographic region.

In each case, it is necessary to determine whether these code elements do, in fact, represent collections rather than single languages. If a given code element is determined to represent an individual language, then it must be further decided which specific language it does represent. On the other hand, if it is determined to represent a collection, then that should be clearly documented in the standard by adding "languages" to the name, and the complete range of varieties that are (or are not) included in its denotation should be decided.

It should be noted that some of the individual language codes that we have identified as representing collections do have two-letter variants from ISO 639-1. The implication of redefining any one of these codes to represent a collection is that it introduces collective language codes into that part of the standard. This would raise a problem for Part 1 since that part of the standard does not sanction code elements that represent collections of languages. We see no alternative to this result, however. There is no escaping the fact that code elements such as [qu] "Quechua", [bh] "Bihari" and zh "Chinese" represent collections of distinct languages.

3.4   The need to identify collection types

As indicated in section 2, collective language codes (other than [art], [mis] and [sgn]) proved to be of three types: those based on genetic classifications, those based on a geographic region, and those based on a shared name. There is a benefit in documenting the type of collection involved for each of the collective language codes since it provides a means to ensure that all of those codes conform to one of a limited number of operational definitions. It is also of benefit for users since it provides additional information that guides them in knowing what each of the codes represents.

There is also a specific benefit in relation to some of the future work on the ISO standards that is being considered. One proposal has been that there should be a comprehensive set of identifiers for genetic classifications.[15] If such a set is to be developed, it is first necessary to determine exactly what genetic classifications are already covered by the existing code elements.

In the development of a comprehensive set of code elements for genetic classifications, the code elements that are already designated as collective language codes and that we have analysed as being based on geographic region (viz. [cau], [cai], [nai], [paa], [phi], [sai]) would also be of interest since their denotations can be expressed in terms of collections of genetic classifications.

Similarly, the code elements that represent collections based on shared names are of interest when considering genetic classifications because some of them approximate genetic classifications. For instance, we determined that the code element [oji] "Ojibwa" (which we concluded represents a collection rather than an individual language) does not denote all languages of the Ojibwa subgroup of the Algonquian family, but only those that may be referred to using the term "Ojibwa". This encompasses all the members of that subgroup but one: Algonquin [ALG]. We excluded Algonquin on the basis that, in our understanding, it is always referred to using that name and never using the name "Ojibwa". Since such collections can approximate but be

---

[15] If there were a standard set of identifiers covering all languages, it is not entirely clear what purpose that would serve except, perhaps, in subject indexing of information objects.

different from genetic classification, it would be useful to identify which collections are of each of these types as well.

3.5 The problem of individual language codes and collections based on region or ethnicity

Among the code elements currently designated as individual language codes but that we have identified as representing collections, there are three that appear to us to be based on geographic regions: [bih] "Bihari", [him] "Himachali", and [raj] "Rajasthani". There were also two individual language codes—[day] "Dayak" and [kac] "Kachin"—that we mapped to a single language but that appear to be used by MARC to represent culturally-based collections.[16] These code elements are problematic and require more detailed discussion.

The terms "Bihari", "Himachali" and "Rajasthani" correspond to geographic regions within India. Each of these names literally means the speech that is used in that region. It has been documented that there are significant problems involving group identities in India, and these issues are involved here. In the midst of great linguistic diversity, language communities in India do not always report their individual language identity, but sometime choose to use a broader identity. For example, when asked on a census what language they speak, they may not report their individual language but rather may report a broader identity such as a regional identification using a name such as "Bihari". Over time, the preferred identities can change, with the result that censuses taken over a few decades report languages whose number of speakers grow and shrink at impossible rates.

Thus, these terms are actually cover terms for multiple languages and are not the names of specific languages. The problem with their use is that many users are not aware that they represent multiple languages, and so they present a mistaken identity. For the purposes of information technologies, we do not see how identifiers for such regionally-based group identities can be useful.

Within the existing code elements, [bih] "Bihari" presents an even more specific problem: the common meaning of this term explicitly includes the languages Bhojpuri, Magahi and Maithili,[17] yet each of these languages already has its own code element: [bho], [mag] and [mai], respectively. Thus, there is duplicated coverage of these languages within the existing code elements.

The status of [bih], [him] and [raj] needs to be resolved. In the cases of [him] and [raj], it may be possible to specify a single language for each, and to change the name associated with the code to make clear what specific language is referred to. The most obvious candidate languages, however, which are Dogri for [him], and Marwari for [raj], already have their own code elements: [doj] and [mwr]. Thus, it is our judgement that the best resolution of the current confusion is to deprecate the further use of code elements [bih], [him] and [raj].

The code element [day] "Dayak" is also very problematic. The term "Dayak" is a cultural cover term meaning all of the non-Muslim inhabitants of Borneo.[18] There are many distinct languages that are referred to with labels that include "Dayak", and MARC documentation cites several of

---

[16] Note that we did not map these codes to collections since, by our principles, we did not allow for collections based on ethnicity or other cultural factors.

[17] See the *Ethnologue* entry for Bhojpuri.

[18] It appears to be somewhat comparable to the Thai term ชาวเขา ("chaaw kao"), meaning 'hill tribe', or the English term 'tribals', and about as useful for purposes of identifying a particular language variety.

them in relation to [day]. These languages do not correspond to any single genetic classification, however. As it stands, this code element is seriously deficient in its definition.

Among these languages, there are two that are reported in the *Ethnologue* as having significant development and thus may warrant assignment to code elements in ISO 639-2: Iban [IBA] and Ngaju [NIJ]. Iban already has its own ISO 639-2 code element, [iba]. Therefore, one possible resolution for this code element is to specify Ngaju as the specific language that it denotes. Another possibility would be to redefine it as a region-based collective for Borneo languages. In either case, the name should be changed to clearly reflect the meaning. The other possible resolution for this code element is deprecation. It is our recommendation that it be defined to refer to the specific language Ngaju and that the English name be changed to "Ngaju Dayak".

Similar problems exist for the code element [kac] "Kachin". The term "Kachin" is used within northern Myanmar in two ways: it is used as a language name, referring to the language Jingpho [CGP], but it is also used as an ethnonym: in this usage, it is a cover term for a number of people groups that have a collective identity based on a sense of brotherhood derived from a history of cultural association. This usage crosses linguistic boundaries at the highest level within the Tibeto-Burman family. There is no genetic classification that it corresponds to.

The problem for the code element [kac] is the ambiguity between these two uses. MARC documentation clearly indicates that they have applied this identifier in a way that corresponds to the ethnonym. The effect of this is to cause this code element to denote very distinct languages, making it of little use for purposes of language identification within information technologies.[19]

We feel that it is impossible to define this code element as a meaningful and useful collection of languages. On the other hand, it would be quite appropriate to use it for the specific language Jingpho. All that would be required is that the documentation of its denotation make clear that it refers specifically to that language. We recommend, therefore, that the English name associated with this code element be changed to "Jingpho". The only alternative we see would be deprecation.

3.6   Principles for defining regional codes

In section 2, we described the principle that we applied in mapping the code elements for region-based collections, [cau], [cai], [nai], [paa], [phi], [sai]. We repeat that principle here for convenience:

> For collective language codes based on geographic region, the denotation is based on all languages within language families / phyla whose languages are spoken primarily within the specified geographic region. Individual languages that may be spoken outside the region are still part of the collection. When major subgroups of a language family / phylum are distributed between different geographic regions, then the major subgroups of that family / phylum are mapped independently to the respective regions.

MARC appears to use a similar principle for defining this type of category. Their principle is not identical, however. Our assessment of their documentation is that they apply the following, which omits the last clause of our principle:

> For collective language codes based on geographic region, the denotation is based on all languages within language families / phyla whose languages are spoken primarily within

---

[19] The only possible use we can envision would be in subject indexing corresponding to the use of "Kachin" as an ethnonym.

the specified geographic region. Individual languages that may be spoken outside the region are still part of the collection.

The differences can be seen in their handling of the Uto-Aztecan family: they reckon all Uto-Aztecan languages as Central American Indian languages, apparently defining "Central America" in terms of the classical notion of Meso-America (which includes southern Mexico). They specifically list Comanche [COM], a Central Numic language from the Northern Uto-Aztecan branch of Uto-Aztecan; thus they clearly include the Northern Uto-Aztecan subgroup within Central American Indian languages. This implies that they include in that collection a whole subgroup of languages that are spoken entirely in the Western United States as far north as Oregon and Idaho.

If there were a small number of languages from the family that were not spoken in the nominal region and that did not correspond to a major subgroup, we would have included the entire Uto-Aztecan family among Central American Indian languages. To handle all of the Northern Uto-Aztecan subgroup this way, however, seems counter-intuitive. It is certainly not what we believe most users would expect. Instead, we have included the Aztecan branch of Southern Uto-Aztecan in Central American Indian [cai], but have assigned the Sonoran branch of Southern Uto-Aztecan and all of Northern Uto-Aztecan to North American Indian [nai].

Consideration of what users would expect raises a question regarding both our principle and MARC's. Would a user expect a language spoken in a given region to be included in a collection for that region, even though the majority of languages from the same family are spoken in a different region? For example, would users expect Garífuna to be included among Central American Indian languages even though almost every other Arawakan language is spoken in South America? In other words, should the principle for assigning languages to region-based collections be based entirely on the location of the language in question rather than considering the geographic distribution of languages from the same family?

ISO must determine what principle is to be used in assigning languages to region-based collective language codes. Once this determination is made, mapping of these code elements to individual languages should be done accordingly.

3.7   Degenerate collections

Since ISO 639-2 makes clear that the denotation for collective language codes does not include languages that are covered by more specific code elements, it turns out that two code elements— [bat] "Baltic (Other) and [cel] "Celtic (Other)" are somewhat degenerate cases in that they represent only one modern language. Thus, "Baltic (Other)" includes only Prussian [PRG], "Celtic (Other)" includes only Shelta [STH].

As collective language codes, [bat] and [cel] can be used under the terms of ISO 639-2 to refer to ancient languages from those linguistic subgroups. Thus, the set of *all* languages included within their denotation is more than one. We feel, however, that the names presented to users are somewhat misleading: "other" is suggestive of multiple languages, and we believe that users would generally reckon that in terms of modern languages. One is almost inclined to suggest that the single modern languages should be given their own codes, and that these collections should be redefined as "Baltic languages (ancient)" and "Celtic languages (ancient)".

We have no recommendation for action in relation to these codes. We merely highlight them as unusual. We do feel it would be helpful to users, though, to have additional documentation to guide them in the use of these code elements than is currently provided.

We also reckon the code elements [no] and [nor] "Norwegian" as cases of degenerate collections. They are, of course, not designated as collective language codes. Given the introduction of code elements for Nynorsk and Bokmaal, however, it appears clear that "Norwegian" must be intended to refer to at least these two languages, and in practice these code elements have been used ambiguously to represent Nynorsk and Bokmaal. Unless [no] and [nor] are redefined as collective language codes, the type of category they represent is very unclear. Moreover, we are left with a case of duplication. This is always problematic for applications that need to be able to equate and distinguish information objects according to the language in which the information is expressed. Furthermore, this situation is especially problematic because the duplicate code element is ambiguous: given two information objects tagged with identifiers [nor] and [nno] (Nynorsk), there is no basis for saying whether the two objects are in the same or different languages.

It feels, therefore, as though the appropriate thing to do is to redefine [no] / [nor] as a collective language code. Yet, if that is done, then clause 4.1.1 makes clear that it would exclude the more specific identifiers, leaving an identifier with no denotation. "Norwegian" would represent nothing!

Strictly speaking, these code elements are currently deemed to be individual language codes, and there is no formally stated restriction on duplication of individual language codes. Nevertheless, it seems clear to us that the intent of the standards is to prohibit duplications since duplication of code elements creates significant problems for implementations. Moreover, the restriction on duplication with respect to collective language codes in clause 4.1.1 of ISO 639-2 would serve no purpose if that were not the case.

The status of these code elements should be resolved. It does not seem appropriate to us to redefine [no] and [nor] as collective language codes.[20] One alternative is that the normative text of the standard be changed specifically to permit ambiguity in stated cases. As noted above, however, such duplication introduces problems for applications. The only other plausible alternative appears to be deprecation of future use of these code elements.

3.8   Complete versus partial collections

The collective language codes in ISO 639-2 have names that take two forms: "X languages" and "X (Other)" (where X represents the name of a genetic classification or of a region, or is a shared language name). As stated in clause 4.1.1 of ISO 639-2: "A collective language code is not intended to be used when an individual language code or another more specific collective language code is available." Thus, the two forms of name are suggestive of the degree of inclusion: "X languages" suggests that the denotation includes all languages of the category X; "X (Other)" suggests that the denotation includes only some of the languages of the category X, since those that are covered by more specific code elements are excluded.

As we examined the collective language codes, we discovered that the denotation of several collective codes does not match the form used for the name. We found that six codes use the "X languages" form but are only partial in their coverage. For example, not all Algonquian languages are within the denotation of [alg] "Algonquian languages" (for instance, Cree and Ojibwa varieties have more specific code elements and are, therefore, excluded). Conversely, we found three that use the "X (Other)" form but are complete in their coverage. For example, there are no "Papuan" languages that are covered by more specific code elements than [paa] "Papuan (Other)"

---

[20] Note that doing so would introduce collective language codes into Part 1 of the standard.

We recommend that the names of these code elements be changed to accurately reflect the degree of inclusion.[21]

3.9   Many-to-one mappings and duplication

In our analysis, we discovered a number of cases in which there are multiple ISO code elements that map to a single language in the *Ethnologue*. Thus, there appears to be duplication among ISO code elements.

A number of these cases were discussed above: [bho], [mag] and [mai] (Bhojpuri, Magahi and Maithili) overlap with [bh] / [bih] ("Bihari"); [doi] (Dogri) overlaps with [him] ("Himachali"); and [mwr] (Marwari) overlaps with [raj] ("Rajasthani"). (See section 3.5 for further discussion). Also, [no] / [nor] (Norwegian) may be deemed to overlap ambiguously with [nb] / [nob] (Bokmål) and with [nn] / [nno] (Nynorsk), as discussed in section 3.7.

The other cases of apparent duplication are the following:

- [aka] "Akan", [fat] "Fanti" and [tw] / [twi] "Twi" all map to the language Akan [TWS]. Fanti and Twi are listed in the *Ethnologue* as dialects of Akan.

- [bs] / [bos] "Bosnian", [hr] / [hrv] / [scr] "Croatian" and [sr] / [srp] / [scc] "Serbian" all map to the language Serbo-Croatian [SRC].

- [mo] / [mol] "Moldavian" and [ro] / [ron] / [rum] "Romanian" all map to the language Romanian [RUM]. The *Ethnologue* lists Moldavian as an alternate name (and also as a dialect) of Romanian.

- [tur] "Turkish" and [ota] "Turkish, Ottoman (1500-1928)" both map to the language Turkish [TRK].

- [lah] "Lahnda" and [pa] / [pan] "Panjabi" both map to the language Western Panjabi [PNB].

We will briefly discuss each of these cases in a little more detail.

In the case of Akan / Fanti / Twi, MARC documentation specifically mentions this as an instance in which code elements represent dialects rather than languages. Clause 4.1.3 of ISO 639-2 mentions the existence of code elements that represent dialects, though this is considered an exception to the norm.[22] There is nothing in the ISO standards to indicate which code elements represent dialects as opposed to languages, however, which creates confusion for the user. It also turns out to be the case that ISO 639-1 includes a code element that represents a dialect, [tw], while it does not include a code element for the language of which that dialect is a variant.[23]

This situation should be resolved. If the ISO standards are going to allow code elements that represent dialects, then it should be clearly documented which code elements represent that type

---

[21] These are all listed in our analysis report under the headings *Other points of interest / Collectives missing "(Other)"* and *Other points of interest / Collectives with superfluous "(Other)"*.

[22] This is the only case we know of in which ISO code elements represent dialects.

[23] We gather that [tw] must have been introduced to ISO 639-1 on the understanding that it represents a language. In matching up of code elements between parts 1 and parts 2, however, this code element proved to correspond to a code element that was created to represent a dialect.

of category,[24] and what the corresponding language code element is. Another alternative would be to make the claim that these are distinct languages. This would have to be accompanied by sociolinguistic evidence demonstrating that that is a valid assessment.[25] The other possible resolution would be deprecation of the code elements for Fanti and Twi.

In the case of Bosnian / Croatian / Serbian, we recognise that, in the events of recent years, the identities of these varieties are becoming more distinct. The *Ethnologue* editor has not yet seen clear linguistic evidence that these are, in fact, distinct languages. We recognise, however, that this situation is one that needs further consideration in relation to how it is handled in future editions of the *Ethnologue*.[26] We are not recommending any change with regard to ISO code elements.

In the case of Romanian / Moldavian, we are not aware of any basis for claiming a distinction. In the absence of any such information, we suggest that the code element [mol] be deprecated.

The name "Turkish, Ottoman (1500-1928)" suggests to us that the code element [ota] was created to distinguish Turkish literature written in Arabic script from more recent Turkish literature written in Latin script. In other words, the primary distinction is based on script rather than linguistic differences. The alternative is that this represents a linguistic distinction based on time, but there is no other precedent for such distinction in the period since 1500, and a claim of a purely linguistic distinction with such a recent boundary is suspect. Also, the year 1928 corresponds with the year in which orthographic reform for Turkish took place.

Clause 4.1.3 of ISO 639-2 says, "A single language code is normally provided for a language even though the language is written in more than one script." While this wording does not rule out the possibility that different code elements may be assigned on the basis of script, it makes it clear that this is not the norm. Furthermore, the standard goes on to suggest the development of a separate standard for the purpose of designating script. It should be determined whether ISO 639-2 will allow different code elements to represent differences in writing systems rather than distinct languages. We believe that would not be advisable for Part 2. Accordingly, we recommend that the code element [ota] be deprecated and that the distinction for which it was created be handled using a distinct standard for writing systems or scripts.

The case of Western Panjabi is somewhat different. The status of [lah] is tied in with the status of [pa] / [pan], which were among the code elements with as-yet indeterminate mappings considered in section 3.2. It seems clear that [lah] should map to Western Panjabi [PNB]. On the other hand, it is unclear whether [pa] / [pan] "Panjabi" is intended to map to Eastern Panjabi [PNJ], Western Panbaji [PNB], or both. This situation is further complicated by the fact that MARC documentation cites the use of [lah] in reference to distinct languages of the Lahnda subgroup of Indo-Aryan.

We see two possible resolutions:

---

[24] We suggest that assigning code elements for dialects could prove to be problematic: while we can think of ways to formulate operational definitions for *language*, we have not yet seen any way to define an operational definition for *dialect* that would prove useful for obtaining an enumeration of distinct dialects.

[25] See the discussion of Albanian in section 3.2 regarding implications of merely stipulating a particular categorization.

[26] While the *Ethnologue* assumes an operational that is based primarily on mutual non-intelligibility, it also incorporates other factors that aim to identify distinct varieties that may be candidates for separate literature. The increasing cultural divergence among Bosnians, Croats and Serbs may well lead to linguistic divergence in speech as well as divergence in literature. This may well result in a situation in which it is considered appropriate for future editions of the *Ethnologue* to list distinct languages in place of "Serbo-Croatian".

- Redefine [lah] as a collective language code that denotes the languages of the Lahnda subgroup and change the English name to "Lahnda languages".

- Have [lah] denote Western Panjabi and have [pa] / [pan] denote Eastern Panjabi, but change the English name for the latter to "Eastern Panjabi" or "Gurmukhi".

The former alternative avoids conflict with MARC usage of [lah] to refer to several languages, but it still leaves unresolved the indeterminate mapping of [pa] / [pan]. In contrast, the latter alternative resolves the indeterminate mapping of [pa] / [pan], and it provides a solution that adheres to the principles described in section 2. In particular, it upholds principle 3, that individual language codes map to single languages; and it applies principle 4, the MLV principle, by mapping both [lah] and [pa] / [pan] to individual major languages. Thus, we recommend the latter alternative.

3.10 Suggested action items

In summary, from the issues we have raised above, we suggest the following as action items for the relevant ISO committees or working groups.

1. Decide whether the code elements referred to in section 3.1 should be mapped to single *Ethnologue* codes following the MLV principle, or whether a different principle should be used. If the latter alternative is adopted, then additional action items should be added to state the proper principle and review all the mappings in light of that principle.

2. Decide how to resolve the indeterminate mappings discussed in section 3.2.

3. For each of the code elements referred to in section 3.3, decide whether that code element represents a single language or a collection. If the former, decide and document which specific language it denotes. If the latter, then change the name to reflect its status as a collection, document the type of the collection, and document the specific languages that are (or are not) included in its denotation.

4. Assuming the result that some code elements that have been deemed in ISO 639 to be individual language codes have in fact been determined to represent collections, determine how to deal with the implications of this for ISO 639-1.

5. Adopt operational definitions for different types of collective language code elements and document the type for each of the existing collective language code elements. If any code elements that are currently deemed to be individual language codes are re-evaluated and deemed to represent collections, the type collection for these must also be documented.

6. Determine the status of the code elements [bih], [him] and [raj]: assign specific languages and change the names accordingly, or deprecate.

7. Determine the status of the code element [day]: assign to a specific language (such as Ngaju [NIJ]) and change the name accordingly, redefine as a collective (such as "Borneo Languages (Other)") and change the name accordingly, or deprecate.

8. Determine the status of the code element [kac]: assign to the specific language Jingpho [CGP] and change the English name to "Jingpho", or deprecate.

9. Determine the principle to be used in assigning languages to the denotation of collective languages codes based on geographic region.

10. Determine the status of [no] and [nor]: clarify that these are intentionally ambiguous with the more specific code elements, or deprecate.

11. Revise the names for collective language codes to accurately reflect their degree of inclusion.

12. Determine the status of [aka], [fat], [twi] and [tw] (see discussion in section 3.9).

13. Determine the status of [ru] / [ron] / [rum] and [mol] (see discussion in section 3.9).

14. Determine the status of [tur] and [ota]: allow identifiers for writing systems in ISO 639-2 and clearly document them as such, or deprecate [ota].

15. Decide how to resolve the indeterminate mapping of [pa] / [pan] and the potential overlap with [lah] (see discussion in section 3.9).

Once the above action items have been resolved and the mappings to the *Ethnologue* have been revised as necessary, a final action item would be in order:

16. Decide whether or not to adopt a mapping to the *Ethnologue* such as we have proposed (revised as necessary) as normative documentation regarding the denotation of code elements in the ISO 639-1 and 639-2 standards.

4. Possible implications for new or enhanced standards

Momentum is building for an effort to enhance the ISO 639 family of standards toward comprehensive coverage of the world's languages. Specifically, there has been a proposal to develop a "Part 3" of the standard that would offer a set of code elements (possibly using four letters) that encompass all the languages of the world. The lessons learned from the experience of mapping the existing ISO 639 code elements onto the languages identified in the *Ethnologue* provide some insights that could inform that effort. The key insights are:

- An operational definition for every category of language code should be part of the standard.

- The type of category denoted by each code element should be clearly identified.

- The denotation of every language-level code element needs to be defined with much more information than just a name.

- The denotation of every collective code element should be defined by enumerating the more specific collective code elements and the language-level code elements it encompasses.

An operational definition for every category of language code element should be part of the standard. In the current code elements of ISO 639-2, we can distinguish at least three categories of language: ancient languages, modern languages (including recently extinct languages) and constructed (or artificial) languages. Among the collective language codes there are similarly at least four categories: collections corresponding to a single genetic subgroup, collections associated with a particular region, collections based on a shared name, and other collections. There are, of course, other ways to categorize languages and collections and a future standard could end up with more or fewer categories. The main point is that such a standard needs to name the categories for which it provides codes and then provide an operational definition for each category. Such standardised definitions are needed to ensure that all the codes in the new standard consistently meet criteria for what deserves to have a code and for the category of thing a code should represent. In Constable and Simons 2000, we have discussed the need for clarification of categories and operational definitions in more detail.

The type of category denoted by each language code element should be clearly identified. This could be done by means of naming conventions or other signals in the documentation about the code elements, but our proposal is that this be done in the code elements themselves. We recommend that in a four-character code, ranges of letters for the first character could be reserved to denote the category of the code. For instance, the category of ancient languages might be coded using initial characters of *a–c*.[27] The first letter would thus be analogous to a namespace designator.[28] The different namespaces could even be assigned to different registration authorities for the purpose of managing the last three letters of the code. This could be advantageous in that different institutions might prove best qualified to manage the different categories of codes. In Constable and Simons 2000, we have discussed the notion of alternate namespaces for language codes in more detail.

The denotation of every language-level code element needs to be defined with much more information than just a name. Our experience in mapping the current ISO 639 code elements to *Ethnologue* languages demonstrates that a simple name is woefully inadequate for defining the denotation of a language identifier. A basic description such as that given in an *Ethnologue* entry is needed. This implies that the kind of standards document used to define ISO 639 Parts 1 and 2 would not serve to define a comprehensive Part 3. It would not be feasible to produce and maintain a printed volume including the codes and descriptions for 6000+ languages. Rather, a web site operated by the designated registration authority would list the current codes and definitions; the standard would give the operational definitions of categories and describe all other practices and constraints that the registration authority would be required to follow.

The denotation of every collective code element should be defined by enumerating the more specific collective code elements and the language-level code elements it encompasses. If collective code elements represent, by definition, collections of languages, then the web site that documents the code elements should make their denotation explicit by enumerating the code elements found elsewhere in the standard that are part of the collection. These could be language-level code elements, or they could be code elements for more specific collections that are included within the scope of the collection.

5. References

Bright, William, ed. 1992. *International Encyclopedia of linguistics.* Oxford: Oxford University Press.

Constable, Peter, and Gary Simons. 2000. Language identification and IT: Addressing problems of linguistic diversity on a global scale. (SIL Electronic Working Papers, 2000-001.) Dallas: SIL International. Revised version of paper presented at the 17th International Unicode Conference, San Jose, CA. Available online at http://www.sil.org/silewp/2000/001/SILEWP2000-001.html.

Grimes, Barbara F. 2000. *Ethnologue*. 14th edition. 2 volumes. Dallas: SIL International. Web edition available online at http://www.Ethnologue.com.

International Organization for Standardization. 1998. ISO 639:1998(E/F), Code for the representation of names of languages. Geneva: International Organization for Standardization.

---

[27] The letter range of *a–c* is chosen arbitrarily for illustrative purposes only.

[28] In strict terms, it would designate partitions of a single namespace, though the effect is equivalent to designation of independent namespaces that can be used in the same contexts.

International Organization for Standardization. 1998. ISO 639-2:1998(E/F), Codes for the representation of names of languages—part 2: alpha-3 code. Geneva: International Organization for Standardization. Available online at http://lcweb.loc.gov/standards/iso639-2/langhome.html.

Tryon, Darrell, ed. 1995. *Comparative Austronesian dictionary, parts 1–4.* Berlin: Mouton de Gruyter.