A paper presented at the International Workshop on Resources and Tools in Field Linguistics, LREC 2002 (26-27 May 2002, Las Palmas, Canary Islands).

SIL Three-letter Codes for Identifying Languages: Migrating from in-house standard to community standard

Gary F. Simons

SIL International
7500 W. Camp Wisdom Rd. Dallas, TX 75236, U.S.A.
gary_simons@sil.org

Abstract

A foundational aspect of documenting an endangered language and preserving that documentation for long-term access is identifying the language itself. The web version of the *Ethnologue* has become the de facto standard for identifying the more than 6,800 languages spoken in the world today. The system of three-letter codes that uniquely identify each language has been used within SIL for nearly three decades as an in-house standard, but now there is increasing demand for these codes to be used by other organizations and projects. This paper describes four changes that SIL International is implementing in order to make its set of language identification codes better meet the needs of the wider community. The changes seek to strike a balance between becoming more open while at the same time becoming more disciplined.

The need for language identifiers

A foundational aspect of documenting an endangered language and preserving that documentation for long-term access is identifying the language itself. Effective retrieval of resources depends on the uniform identification of the languages to which they pertain. Simply using language names in metadata is not adequate since the same language is typically known by many names and those names change over time. Furthermore, different languages may be known by the same name. Thus the most effective approach in resource metadata is to use standardized language identifiers. See Simons (2000) and Bird and Simons (2001, section 3.4.1) for a fuller discussion of this point.

The International Organization for Standardization has published a standard for three-letter codes to identify languages (ISO 1998). Known as ISO 639-2, it provides codes for fewer than 400 languages. Thus language documentation efforts (such as ISLE[1], E-MELD[2], OLAC[3], and Rosetta Project[4]) that embrace endangered languages have had to look elsewhere for language identifiers. They have turned to the most widely known and accessed reference work on language identification, the *Ethnologue* (Grimes 2000), now in its 14th edition. With listings for over 7,000 languages, the *Ethnologue* seeks to give a comprehensive accounting of all known living and recently extinct languages in the world. Other languages, such as ancient and constructed languages, are specifically outside the scope of the *Ethnologue*; SIL International is pleased to cooperate with the Linguist List initiative to develop standardized codes for these languages that fall outside the scope of the *Ethnologue* (Aristar 2002).

[1] http://lingue.ilc.pi.cnr.it/EAGLES/isle/ISLE_Home_Page.htm; http://www.mpi.nl/ISLE/
[2] http://saussure.linguistlist.org/cfdocs/emeld/
[3] http://www.language-archives.org/
[4] http://www.rosettaproject.org/

The development of an in-house standard

The system of three-letter language identifiers used in the *Ethnologue* was originally developed as a standard for in-house use almost 30 years ago. The codes were first published with the following explanation in a monograph reporting the results of building a database of languages of the world from the typesetting tapes for the 7th edition (1969) of the *Ethnologue*:

> Each language is given a three-letter code on the order of international airport codes. This aids in equating languages across national boundaries, where the same language may be called by different names, and in distinguishing different languages called by the same name. (Grimes 1974:i)

While the codes were used in the database that generated the 8th and 9th editions of the *Ethnologue*, it was not until the 10th edition that they appeared in the publication itself. The introduction offers this explanation:

> Each language of the world is assigned a unique three-letter code, which is the same in all countries in which that language is spoken. The code helps distinguish the language from other languages with similar names and helps to insure that each language will be counted only once in a world or area statistics. (Grimes 1984:iii)

This system of three-letter codes for identifying languages has been an in-house standard within SIL for nearly three decades. Though the codes were available to the public in print publications, the publications were not widely known. This changed dramatically in 1996 with the publication of a web version of the 13th edition of the *Ethnologue* on the Internet. Before that there were only a few thousand copies of the print publication in circulation. The web edition brought an overnight change with thousands of people consulting the *Ethnologue* every day.

Becoming a community standard

Today with over one million page hits per month on `www.ethnologue.com`, the *Ethnologue* has become a de facto standard for worldwide use. SIL International is responding to this shift from being an in-house standard to becoming a community standard by making changes to the way it manages the three-letter code set.

There are four key changes that are being implemented in order to make the SIL language identification codes better meet the needs of the global community. In sum, the changes seek to strike a balance between becoming more open while at the same time becoming more disciplined. The four changes are:

1. Opening access to the complete code set by making it downloadable
2. Opening the process by which corrections and improvements are made to the *Ethnologue*
3. Tightening the definition of the criteria used for identifying languages
4. Tightening the policies surrounding changes to the code set

The remaining sections describe each of these changes in turn, first discussing the requirements that lie behind the change, and then detailing the solution that is being implemented for it.

Opening access to the complete code set

Heretofore, the three-letter language identifier has been included as part of the *Ethnologue's* description of each language, but the codes have not been published separately as a code set. In particular, organizations and projects that want to use the codes in their own application require the following:

- Users need the complete code set in a form that can be downloaded and in turn imported into a database or other application.

- In addition to the codes themselves, users need to be able to incorporate associated information like countries and alternate names into their applications in order to assist in finding the right code.

SIL International has responded to these requirements by publishing the entire set of language identifiers as a set of tab-delimited tables that can be downloaded for import to a database or other user application (SIL 2002). The structure of the tables (described in SQL statements for creating the database tables) is as follows:

```
CREATE TABLE LanguageCodes (
   LangID      char(3) NOT NULL,        -- Three-letter code
   CountryID   char(2) NOT NULL,        -- Main country where used
   LangStatus  char(1) NOT NULL,        -- L(iving), N(early extinct),
                                        -- E(xtinct)
   Name        varchar(75) NOT NULL)    -- Primary name in that country

CREATE TABLE CountryCodes (
   CountryID   char(2) NOT NULL,        -- Two-letter code from ISO3166
   Name        varchar(75) NOT NULL )   -- Country name

CREATE TABLE LanguageIndex (
   LangID      char(3) NOT NULL,        -- Three-letter code for
language
   CountryID   char(2) NOT NULL,        -- Country where this name is
used
   NameType    char(2) NOT NULL,        -- L(anguage), LA(lternate),
                                        -- D(ialect), DA(lternate)
   Name        varchar(75) NOT NULL )   -- The name
```

The `LanguageCodes` table lists 7,148 distinct language identifiers. Of these, 308 represent extinct languages, 406 are nearly extinct, and the remainder are listed with "living" status. The following shows the entries for the first six languages identifiers in the download table:

```
LangID CountryID LangStatus Name
------ --------- ---------- -------------
AAA    NG        L          Ghotuo
AAB    NG        L          Arum-tesu
AAC    PG        L          Ari
AAD    PG        L          Amal
AAE    IT        L          Albanian, Arbëreshë
AAF    IN        L          Aranadan
```

We see that AAA and AAB denote living languages spoken in Nigeria, AAC and AAD denote living languages spoken in Papua New Guinea, and so on. When a language is actually spoken in more than one country, the `CountryId` gives the country that is considered primary; usually the country of origin or country where most of the speakers are located.

The `CountryCodes` table lists the two-letter identifier and name for 220 countries of the world. The codes are from the international standard known as ISO 3166-1 (ISO 1997). The following shows the entries for the first five codes in the list:

```
CountryID Name
--------- --------------------
AD        Andorra
AE        United Arab Emirates
AF        Afghanistan
AG        Antigua and Barbuda
AI        Anguilla
```

The `LanguageIndex` table documents 37,420 distinct names used for the 7,148 languages. Each entry in this index of names indicate the country in which the name is used. The table thus contains 46,416 records since many of the names are used in more than one country and some are used with more than one language or dialect. The following shows the entries in the name index for the first three language identifiers

```
LangID CountryID NameType Name
------ --------- -------- ------------
AAA    NG        L        Ghotuo
AAA    NG        LA       Otuo
AAA    NG        LA       Otwa
AAB    NG        LA       Alumu
AAB    NG        D        Arum
AAB    NG        LA       Arum-cesu
AAB    NG        LA       Arum-chessu
AAB    NG        L        Arum-tesu
AAB    NG        D        Tesu
AAC    PG        L        Ari
```

We see that AAA has two alternate names in addition to the primary name of Ghotuo. AAB has three alternate names and two dialect names in addition to its primary name. AAC has just one name.

Using this table it is possible to build queries that retrieve sets like all the languages spoken in a particular country, all the countries in which a particular language is spoken, all the languages known by a particular name, or all the names by which a given language is known. When the information provided in these tables is not enough for someone using the codes to be absolutely sure that a proposed code is the right one for a particular language, the user interface for the application can offer a link to the *Ethnologue* web site in order to retrieve a report giving all of the information available on the proposed code. The link is as follows, where `AAA` is the proposed three-letter identifier:

```
http://www.ethnologue.com/show_language.asp?code=AAA
```

Opening the process for changing the *Ethnologue*

The *Ethnologue* is a work in progress; our knowledge of the world's languages is always incomplete and subject to improvement.. Many people who use the *Ethnologue* can give feedback that will make it better and SIL International has always valued this kind of input. Users may have more accurate information on details like locations or names or population figures or language development status. Or they may be able to provide information that would lead to a change to the set of language identifiers. For instance, they may be able to show that what is treated as one language is really two, or vice versa, or that a listed language does not exist or that an existing language is not listed.

It should be easy for any user of the web version of the *Ethnologue* to give feedback that will help to improve the quality of the language descriptions and the set of language identifiers. The *Ethnologue* staff welcomes such input and has an established process for dealing with it. This process involves verifying the information with correspondents in the field, and thus the process is not always a fast one. The current *Ethnologue* web site contains a questionnaire for reporting a language description, but it is too hard to locate and too complex for giving basic feedback. These are some requirements on improving the process for proposing changes to the *Ethnologue*:

- The web page which shows the *Ethnologue* report for a particular language should have a link or button on it that invites the user to give feedback that will help to improve the treatment of that language.
- The user giving feedback should provide basic information such as name, affiliation, and an email address so that the *Ethnologue* staff may contact the contributor for follow-up.
- The user giving feedback should receive an explanation of the process that will be used to process the input.
- The *Ethnologue* staff should notify the contributor of the eventual outcome of the change proposal.

- From the web page for a particular language, any user should be able to see a list of the changes that will be made in the next edition of the *Ethnologue*.

The plan is to incorporate these facilities into the *Ethnologue* web site; at the time of writing they have not yet been implemented.

Tightening the criteria for identifying language

As the process for making changes to the *Ethnologue* is opened up for wider input, it is imperative that the criteria used for identifying languages are clear to all who are giving input since there are so many different notions of what it means to be a different language. The introduction to the *Ethnologue* summarizes the problem like this:

> How many languages are spoken in the world today? No one really knows. What is a language? The term has been used in many different senses. Popular usage often reserves the term 'language' for the major, prestigious speech forms of the world, and uses 'dialect' for everything else. Some people use 'language' to refer to speech forms that share a certain percentage of similar vocabulary, and 'dialect' to refer to speech forms that share higher percentages. Or they may consider varieties to constitute the same language which have similar grammatical and phonological systems. Many people, including some linguists, use the terms 'language' and 'dialect' without always clarifying the sense in which they are being used. (Grimes 2000:vii)

Constable and Simons (2000) discuss the problems involved in developing a standardized code set for identifying all the world's languages. We conclude that many different code sets are possible since different users have different operational definitions of language based on the different purposes they have for identifying languages. Thus one of the most basic challenges involved in developed a set of language identifiers is to establish an operational definition for language and then to assign identifiers consistently on the basis of that definition. A set of language identifiers that merely assigns codes based on the whim of editors and users would not be very useful. The requirements on criteria definition are thus as follows:

- The criteria that will be used by the editors to determine whether two speech varieties should be listed as different languages or as varieties of the same language must be clearly stated.
- When users of the code set propose changes to particular language identifiers, they should do so by demonstrating that the criteria were not applied appropriately in the particular situation. Such changes (once validated) should be made by the editors.
- When users of the code set propose changes because they disagree with the criteria themselves, these changes should not be made (unless there is first a decision to refine the criteria and then to reapply them consistently throughout the entire code set).

This leads us to ask what the *Ethnologue* already says regarding its criteria for identifying languages. While the introduction falls short of giving an operational definition, it does discuss some of the factors. The one given the most weight is intelligibility:

> To those of us who are interested in cross-cultural communication and developing usable literature for speakers of many languages, however, it seems clear that one of the main factors that must be considered in distinguishing 'language' from 'dialect' is how well two linguistically close speech communities understand each other. Marginal intelligibility between two language communities does not allow their speakers to engage in meaningful communication beyond bare essentials. (Grimes 2000:vii)

In addition to identifying the key factor, this statement also identifies the purpose that underlies language identification as carried out by the *Ethnologue*—it is based on a motivation of "developing usable literature for speakers of many languages." Another key statement is the following:

> Variants of the language that are not distinct enough to need separate literature are treated as dialects, and are listed under the language entry and not as separate entries, unless attitudes or other social factors are strong enough that they need to be treated as separate sociolinguistic entities. (Grimes 2000:vii)

From these statements we see that the main criteria are these three: intelligibility, shared literature, and social factors (especially having to do with ethnolinguistic attitudes and identity). A more operational definition of how these factors are used in decision making might be stated as follows:

- If two related varieties share intelligibility and already share a common literature, they are considered the same language.
- Conversely, if two related varieties use literatures that are not intelligible to each other, they are considered to be different languages.
- Where there is no literature, then two related varieties are considered to be the same language if they share intelligibility and they share the same ethnolinguistic identity.
- Where there is intelligibility but such a strong sense of distinct identity as to make the use of shared literature infeasible, the varieties are considered to be different languages.

More work needs to be done by the *Ethnologue* staff to refine these criteria statements. No doubt further points will need to be added in order to handle all known cases. However, the formulation given here should suffice to indicate the basic criteria being used at present and to illustrate the direction we want to go in stating our decision-making criteria more clearly.

Tightening policies for changing the code set

The code set needs to change over time, not only because languages change, but more often because our knowledge about the languages of the world (especially the small and endangered languages) is improving all the time. While the code set was only an in-house standard, changes were made without a lot of attention to the impact of changes in the meaning or a code or the impact of recycling a previously used code.

Now that other organizations are using the code set as a standard in their own applications, SIL International must become more disciplined about how changes to the code set are implemented and documented. These are the main requirements on the process of managing changes to the code set:

- Once a language code has been used in the code set, it may never be reused. In this way, user data that uses the code may become obsolete when the code is retired, but never unexpectedly change in meaning (such as if the code were reused for another purpose).
- Once a language code has been correctly applied to tag an item in user data, it must continue to be the right code to use for that item for as long as the code remains an active member of the code set.
- Users who support applications that use the language codes need to be able to find out how the code set has changed. Furthermore, they need an automated way to find all the data records that may no longer be correctly coded (such as when codes are retired or shift in their range of meaning). For each affected code, there should be an indication of how existing data may need to be changed.

The code management discipline being followed in order to satisfy the first bullet is self evident. The second bullet requires more explanation. When two codes are merged into one, one code is permitted to remain with an extended meaning while the other is retired. This is allowed since all application data previously coded with the code that has been extended in meaning is still correctly coded. However, when the reverse happens (that is, one code needs to be split into two), the original code is retired and two new codes are created. If the original code were retained with a narrower meaning, then some of the existing uses of the code would no longer be valid, thus breaking the second requirement above.

The third requirement is being met by publishing a change history table along with the downloadable code table (SIL 2002). The current plan is to release an updated code table twice per year (on or around January 15 and July 15), With each update a downloadable change history table will also be released. The SQL statement for creating the change history table is as follows:

```
CREATE TABLE ChangeHistory (
   Code         varchar(10) NOT NULL,  -- The affected three-letter code
   Type         char(1) NOT NULL,      -- C(reated), E(xtended),
                                       -- R(etired), (U)pdated
   Date         char(10) NOT NULL,     -- Date of public release
   Description  varchar(255) )         -- Description of change
```

Note that four types of changes are tracked. C is for a code that is newly created. E is for a code that is extended in meaning; the description should tell which other code was merged into it. R is for a code that has been retired; the description should tell which other code or codes replace it. U is for a code for which information in the `LanguageCode` table has been updated; the code and its meaning have not changed, but the name or country or status of the language has been changed. Here are some sample rows from the `ChangeHistory` table:

```
Code Type Date        Description
---- ---- ----------  ---------------------------------------
AOX  C    2002-01-31  Add ATORADA, Guyana, living
APR  E    2002-01-31  Includes [LOA] which was retired
LOA  R    2002-01-31  Merge with [APR]; change all [LOA] to [APR]
AWG  R    2002-01-31  Same as [WMI]; change all [AWG] to [WMI]
CKN  R    2002-01-31  Unable to verify existence; delete from database
AAS  U    2002-01-31  Change from extinct to living
BCJ  U    2002-01-31  Change name from BAADI to BARDI
```

The `ChangeHistory` table holds the cumulative list of all changes that have been made to the code set. Thus it may be queried to learn the complete history of a given code, or to learn all the changes that have been made since a given date. Another key use of the `ChangeHistory` table is in discovering all the codes in an application data set that are now obsolete and thus need to be changed. These will be the codes that are marked as retired in the change history table. Thus a full list of all data records needing to be changed can be found by doing a JOIN on the change history table. For instance, if the column named `Code` in `MyTable` holds a three-letter language code, then the following SQL statement will select all records that have been rendered out-of-date by changes to the code set made since the beginning of 2002:

```
SELECT * FROM MyTable M, ChangeHistory C
WHERE M.code=C.Code AND C.Type='R' AND C.Date >=2002-01-01
```

Note that the `Description` field of the joined result set will describe what needs to be done to bring each offending language code up-to-date.

Conclusion

Language identification is a foundational aspect of documenting an endangered language and preserving that documentation for long-term access. This is because effective retrieval of archived language resources depends on the uniform identification of the languages to which they pertain. The system of three-letter language identification codes used in the *Ethnologue* is proving to be a useful tool for this purpose, and will be even more useful when it can be managed more as a community standard than as an in-house standard. SIL International is therefore endeavoring to implement the changes described in this paper in hopes of better serving the language resources community.

References

Aristar, Anthony. 2002. LINGUIST codes for ancient and constructed languages. Available online at http://www.language-archives.org/wg/language/linguist-20020219.html.

Bird, Steven and Gary F. Simons, editors. 2000. Proceeding of the workshop on web-based language documentation and description, University of Pennsylvania, 12-15 December 2000. Available online at http://www.ldc.upenn.edu/exploration/expl2000/ .

Bird, Steven and Gary F. Simons. 2001. The OLAC metadata set and controlled vocabularies. Proceedings of the ACL/EACL Workshop on Sharing Tools and Resources for Research and Education, Toulouse, July 2001, Association for Computational Linguistics. Available online at http://arxiv.org/abs/cs/0105030.

Constable, Peter, and Gary F. 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 a 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. 1984. Languages of the world: Ethnologue. 10th edition. Dallas: Wycliffe Bible Translators.

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

Grimes, Joseph E. 1974. Word lists and languages. Technical Report No. 2, Department of Modern Languages and Linguistics, Cornell University, Ithaca, NY.

International Organization for Standardization. 1997. ISO 3166-1:1997, Codes for the representation of names of countries and their subdivisions—part 1: country codes. Geneva: International Organization on Standardization. Available online at http://www.din.de/gremien/nas/nabd/iso3166ma/.

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.

SIL International. 2002. SIL three-letter codes for identifying languages. A web page posted at http://www.ethnologue.com/codes.

Simons, Gary F. 2000. Language identification in metadata descriptions of language archive holdings. In Bird and Simons 2000. Available online at http://www.ldc.upenn.edu/exploration/expl2000/papers/simons/simons.htm.