Terminology and language aspects in language coding

Gary Simons
SIL International

TKE 2014 Workshop: Language Codes at the Crossroads
Berlin, Germany, 21 June 2014

The language coding problem

- The international community needs to identify things like:
  - Language of the content in a document or a recording
  - Language of each term in a terminological database
  - Languages supported by a software tool
  - Language proficiencies of people and organizations
- But language name don’t work because:
  - Different languages may have the same name
  - The same language may have different names in different places where it is spoken and in different languages
  - When outsiders don’t know the real name, different people invent different names for the same language.

Enter ISO 639

- The relevant standard is ISO 639:

  Codes for the representation of names of languages

  - *I.e.,* Standardized codes to be used in place of names

- Six parts have been published; three are widely used:
  - Part 1 (1967): About 200 two-letter codes, *e.g.,* en = English
  - Part 2 (1998): Three-letter codes for about 360 individual languages (including all in part 1), *e.g.,* eng = English, and 70 collections, *e.g.,* map = Austronesian languages
  - Part 3 (2007): All individual language codes from ISO 639-2, plus codes for over 7,000 more languages

But there's a terminological problem

- Namely, "What do we mean by *language*?"
- The 3 parts emerged out of different communities
  - Part 1 from the terminology community
  - Part 2 from the library community
  - Part 3 from the linguistics community
- Given that Part 3 includes all the individual languages in Parts 1 and 2, it necessarily lies at the convergence of different notions of what a "language" is

A seminal work on this problem

- Haugen, Einar. 1966. “Dialect, language, nation.” *American Anthropologist* 68:922–35
- The opening sentence:
  - “The taxonomy of linguistic description—that is, the identification and enumeration of languages—is greatly hampered by the ambiguities and obscurities attaching to the terms ‘language’ and ‘dialect.’ ”

Two differing perspectives

- After reviewing how the terms ‘language’ versus ‘dialect’ have been used, he notes there are two fundamentally distinct traditions of use
  - The ***structural*** use
    - — “descriptive of the language itself”
    - — “the overriding consideration is genetic relationships”
  - The ***functional*** use
    - — “descriptive of its social uses in communication”
    - — “the overriding consideration is the uses the speakers make of the codes they master”

The structural view

- The structural view of “language” versus “dialect” is the one most commonly held by linguists.
  - Language is superordinate to dialect.
  - A language is a grouping of related dialects that are intelligible to each other.
  - Standardization does not enter in.
- This is the perspective that was dominant in the code set originally developed for the *Ethnologue,* which is what served as the basis for ISO 639-3.

The functional view

- The functional view of “language” versus “dialect” is the one most commonly held by the public at large.
  - A language has a standardized written form.
  - A dialect is an unstandardized oral variety.
  - A language is thus the medium of communication between speakers of different dialects.
- This is the perspective that was dominant in the formation of ISO 639-1 and 639-2.

Criteria for ISO 639-2

- http://www.loc.gov/standards/iso639-2/criteria2.html
- There should be a sizable and varied literature
  - A request for a new code must cite at least 50 titles
- There should be support by a national or regional language authority or standardizing body
- Evidence of “official” status strengthens the request
- Evidence of extensive use as a medium of instruction in formal education strengthens the request

A third perspective

- A third perspective was evident in the *MARC Code List for Languages* which served as the basis for ISO 639-2.
  - The ***ethnic*** perspecitive
    - — the overriding consideration is the ethnic identity of the users of speech varieties
    - — Logic: “If people have the same ethnic name, then they must have the same language.”
- Examples in Part 2: Cree [cre], Ojibwa [oji], Zapotec [zap]
  - In these cases, there are multiple unintelligible varieties, but no unifying written standard as required by the functional view.
  - The grounds for joining structurally distinct varieties appears to be the shared ethnic name.

Criteria for ISO 639-3

- http://www.sil.org/iso639-3/scope.asp
- Two related varieties are normally considered varieties of the same language if speakers of each variety have ***inherent understanding*** of the other variety.
- Where spoken intelligibility between varieties is marginal, but there is a ***common literature*** or a ***common ethnolinguistic identity*** with a central variety that both understand, they may be varieties of the same language.
- Where there is intelligibility between varieties, but they have well-established distinct ethnolinguistic identities, this can be a strong indicator that they should nevertheless be considered to be different languages.

The easy cases

- The decision for two speech varieties is straight-forward when all three factors align.

Same language	Different languages
Mutually intelligible	Unintelligible
Share a common literature	Use different literatures
Share a common ethnolinguistic identity	Distinct ethnolinguistic identities are encoded in distinct autonyms

The hard cases

- But what about a case in which the factors do not all line up in one column?
  - Depending on your dominant perspective, you'll weight the conclusion to one side or the other.
- When work began on ISO 639-3 in 2002, this created a dilemma for the task of reconciling the Ethnologue codes with the ISO 639-2 codes
- We needed alignment within a single code space:
  - The same thing in both parts must have the same code
  - The same code in both parts must mean the same thing

Irreconcilable differences?

- In many cases *Ethnologue* had multiple languages where ISO 639-2 had only one.
- The case of Arabic
  - The functional view of ISO 639-2 assigned just one code for Arabic [ara] which applied to standard Arabic as well as all spoken varieties.
  - But recognizing that the widely scattered varieties were no longer intelligible after more than a millennium of divergence, the structural view of *Ethnologue* had a code for standard Arabic plus codes for 28 regional varieties

More differences

- There were also cases of the reverse: ISO 639-2 had multiple languages and *Ethnologue* had one.
- The case of Norwegian
  - The functional view of ISO 639-2 assigned codes for Bokmål [nob] and Nynorsk [nno] as distinct languages.
  - The structural view of *Ethnologue* had only one code for Norwegian since it saw these as two ways of writing the same language, as opposed to being distinct languages themselves.

“Macrolanguages” to the rescue

- We reconciled the differences by introducing 55 instances of a new category of codes into ISO639-3:
  - Macrolanguage = “multiple, closely-related individual languages that are deemed in some usage contexts to be a single language”
  - For each macrolanguage that is defined, the standard also lists its member languages
    - Arabic [ara] has 29 member languages
    - Norwegian [nor] has 2 member languages
    - Zapotec [zap] has 47 member languages

A terminological problem

- What really is a macrolanguage?
  - The criterion of "deemed in some usage contexts to be a single language" is rather open ended
  - In the early years of ISO 639-3 we accepted requests to create new macrolanguages and ended up adding some that were based on a "usage context" of shared ethnic identity
- Feedback from Joint Advisory Committee
  - They really should be reserved for alignment between Parts
  - Macrolanguage = "a coded entity that is deemed in some usage contexts to be a single language but which in others corresponds to multiple, closely-related individual languages that also have codes"

Should we tighten even more?

- If this is what “macrolanguage” means, do we really need the category?
  - It is not really a kind of language, but a property of a code
  - We could just use a Linked Data representation (as does Library of Congress at id.loc.gov) to map between Parts and simply infer that a code has the “macro” property
- But there is one current macrolanguage configuration that represents more than just a one-to-many mapping
  - A macrolanguage that represents a diglossic situation has a structure within its relationships and is qualitatively different than a simple grouping of languages

Should we reserve "macrolanguage" as a label just for diglossia?

- *I.e.,* Macrolanguage = "the set formed by a functionally-defined High language and all the structurally-defined Low languages for which it is the unifying standardized form"
- The classic case in the current standard: Arabic [ara] represents Standard Arabic [arb] plus the 29 regional spoken varieties that look to it as their standardized form
- There are known problem cases where Parts 2 and 3 are not fully aligned and the solution will require sorting out a diglossic situation and promotion to macrolanguages:
  - German [deu], Italian [ita], Tibetan [bod]

Improving the standard

- The ISO 639-3 standard provides both:
  - A set of standardized three-letter codes
  - An open process for making changes to the code set
- Thus, fixing the problems in ISO 639-3 depends on participation by the user community
  - Any one who sees something they think is missing or wrong may submit a form to request and justify a change
  - The request is posted on the web for public comment
  - A review panel meets annually to make final decisions
  - Results reviewed by the Joint Advisory Committee

Submitting a change request

- Go to http://www.sil.org/iso639-3/ with links for
  - *Change management* — How it works and annual reports summarizing all change results since 2006
  - *Submitting change requests* — CR form and instructions
  - *Change request index* — Table of all change requests by year, region, family, code, language name with a link for each to a page giving the completed change request form and any other related documents
- In 8 annual cycles (2006 – 2013) we have processed 949 change requests

Some examples

- Mayan languages
  - Nora England submitted 16 merger CRs (2008-048 to 2008-063) to align the standard with the consensus of Mayanists and the Mayan academy. Result: 43 codes were merged into others and retired
- Australian languages
  - Anthony Aristar and Claire Bowern submitted 121 CRs in 2011 and 2012 to clean up the code set for Australia: 4 name changes, 11 splits, and 106 creations of missing languages (mostly extinct)
- Mascoyan languages
  - Hannes Kalisch submitted 4 CRs in 2013 to clean up the Mascoyan family. Result: 2 splits, 2 retired (nonexistent)

Summary

- There is a long tradition of different approaches to understanding “language” versus “dialect”
- Different parts of ISO 639 use different criteria because they embody different perspectives on what constitutes a language
- The macrolanguage concept is used to achieve alignment between Parts 1,2 and Part 3
- Improving the standard should proceed on two fronts
  - Refining the concepts, criteria, and processes it defines
  - Encouraging users to use the open change request system to keep improving the individual codes