ISO 639-3

Where are we and how did we get here?

Gary Simons
SIL International

Workshop on Identifying Codes for Languages
Newcastle, Australia, 9 February 2013

Overview

- Where are we?
  - The ISO 639 family of standards
- How did we get here?
  - Philosophy of language codes
  - Historical timeline of ISO 639
  - Two conflicting views of what “language” means
  - A mechanism for resolving the conflict
- Where from here?
  - Changing the code set
  - Changing the infrastructure

The problem

- What problem is ISO 639 trying to solve?
- The international community needs to identify the:
  - Language a document is written in
  - Language spoken or signed in a recording
  - Language documented in a dictionary or grammar
  - Language of each term in a terminological database
  - Source language of a translated document
  - Languages supported by a software tool
  - Languages covered by translation services
  - Language proficiencies of people and organizations

Language names don't work

- Different languages (in different parts of the world) may have the same name.
- The same language may have different names in various places where it is spoken.
- The same language may have different names in various other languages.
- In the absence of a standard name, different people refer to the same language by different names.
- The preferred name for a language may change over time.

Enter ISO 639

- ISO: International Organization for Standardization
- TC37/SC2/WG1: Technical Committee 37, Sub-Committee 2, Working Group 1: “Language Coding”
  - 29 countries participating , 10 observing
- 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

ISO 639-1

- Part 1: Alpha-2 code
  - About 200 two-letter codes, *e.g.,* en = English
  - First published 1967
  - Registration Authority: Infoterm, Austria
    - http://www.infoterm.info/standardization/iso_639_1_2002.php

ISO 639-2

- Part 2: Alpha-3 code
  - Three-letter codes for about 360 individual languages and 70 collections of languages, *e.g.,* eng = English, map = Austronesian languages
  - First published 1998
  - Registration Authority: Library of Congress, USA
    - http://www.loc.gov/standards/iso639-2/

ISO 639-3

- Part 3: Alpha-3 code for comprehensive coverage of languages
  - All individual language codes from ISO 639-2, plus codes for over 7,000 more languages
  - First published 2007
  - Confirmed in 2010 review
  - Registration Authority: SIL International, USA
    - http://www.sil.org/iso639-3/

Side bar: Codes are *not* abbreviations

- With around 400 codes in Part 2, most possible three-letter combinations were available so most codes look like abbreviations, but they aren't.
  - [rom] Romany vs. [roa] Romance languages, [roh] Romansh, [ron] Romanian
- With Part 3, almost half of the 17,576 possible combination are now taken
  - Codes are arbitrary; mnemonic match is not possible
  - The letters chosen have no significance or structure
  - Best thought of as three-digit base 26 numbers

ISO 639-4

- Part 4: General principles of coding of the representation of names of languages and related entities, and application guidelines
  - No language codes
  - First published 2010
  - Yet to be confirmed

ISO 639-5

- Part 5: Alpha-3 code for language families and groups
  - All collective codes from ISO 639-2, plus codes for about 50 more groups of languages, *e.g.,*
    pqe = Eastern Malayo-Polynesian languages
  - First published 2008
  - Confirmed in 2011 review
  - Registration Authority: Library of Congress, USA
    - http://www.loc.gov/standards/iso639-5/

ISO 639-6

- Part 6: Alpha-4 code for comprehensive coverage of language variants
  - Thousands of four-letter codes for
    - Language variants and language groupings all the way up to: wrld = World
    - Arranged in a hierarchy including codes from Parts 2, 3, 5
  - First published 2009; yet to be confirmed
  - Registration Authority: GeoLang Ltd., UK
    - http://www.geolang.com/Iso639-6/
    - Browse or search by alpha-4 code, parent code, name
    - No descriptions, no download tables, no mechanism for change requests

Uptake of ISO 639-3

- Used to catalog 190,000 language resources by 44 archives participating in OLAC (Open Language Archives Community), *e.g.,* ASEDA, PARADISEC
- A recognized encoding scheme in the Dublin Core standard: DCMI Metadata Terms
- Recognized as a source of language codes by the Library Congress for use in MARC and MODS cataloging
- Included within Best Current Practices 47, "Tags for Identifying Languages," of IETF (Internet Engineering Task Force)
- "Wikimedia does not decide for itself what is a language and what is a dialect. We follow the ISO 639 standard. Every Wikimedia language edition is required to have a valid ISO 639-1 or ISO 639-3 code."

Overview

- Where are we?
  - The ISO 639 family of standards
- How did we get here?
  - Philosophy of language codes
  - Historical timeline of ISO 639
  - Two conflicting views of what “language” means
  - A mechanism for resolving the conflict
- Where from here?
  - Changing the code set
  - Changing the infrastructure

What is motivating this?

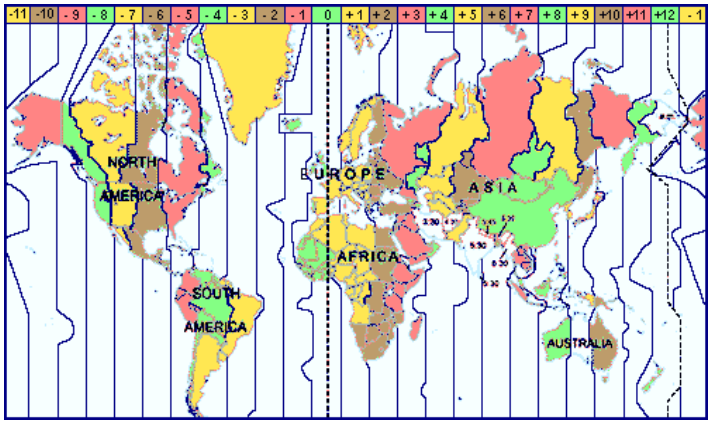
- "How many languages are there in the world?"
  - As linguists, we know that a precise answer is impossible
  - The majority of languages are not adequately described
  - There are not clear cut boundaries
  - It depends on how you define language
- Many linguists thus regard the idea of standardized language codes with suspicion
  - "if you agree that the above problems are true, why do you pursue this anyway?"

Life without standards

- Standards are motivated by the common good
  - They let us “interoperate”
- Without standards:
  - We would be cheated in the marketplace
  - There would be no interchangeable spare parts
  - We could not plug in our devices anywhere
  - There would be no real-time long distance communication
  - We would have train wrecks all the time

The standardization of time

- Two approaches to reckoning time
  - Solar noon vs. Standard time



- **Source:**
  - Blaise, Clark. 2000. *Time lord: Sir Sandford Fleming and the creation of standard time.* New York: Pantheon Books
  - Simons, Gary F. 2009. Linguistics as a community activity: The paradox of freedom through standards. In Will Lewis et al (eds.), *Time and Again: Theoretical Perspectives on Formal Linguistics. In honor of D. Terence Langendoen.* pages 235–250. Amsterdam: John Benjamins.

An apt analogy

- The perils of train travel without standard time:
  - Passengers miss the train when they miscalculate the time
  - Trains wreck when wrong train on track at the wrong time
- The perils of cybersearch without standard codes:
  - Users miss relevant material when their queries use a name that differs from what the material uses (=low recall)
  - Users experience an information wreck when their queries using names retrieve mostly irrelevant things, either wrong language or not language-related at all (=low precision)

Two approaches to local time

	Solar Noon	Time Zones
How it works	Local time is a continuous function; there is an unlimited number of noons	Local time is a step function; there are exactly 24 noons
How it interoperates	Convert by a number of minutes that must be looked up for any east-west travel	Convert by whole hours only over long distances
Optimized for	Convergence with physical reality (= truth)	Interoperation and the common good

Codes as points versus zones

- Which approach best serves linguistics community?
  - Codes as solar noons that reify single varieties, or
  - Codes as time zones that cover a range of varieties?
- Linguists as descriptivists may prefer
  - Point approach because of greater accuracy
- But linguists are also promoters and researchers
  - Zone approach means more people will find their work
  - Zone approach means they will find relevant materials in closely related varieties when they search
- Each ISO 639-3 code is a zone over all varieties of a language; together, all the zones cover all the earth.

History: Roots of ISO 639-1

- In the 1960s, the terminology subcommittee (TC 37) of ISO developed two-letter codes for languages dealt with by its user community.
- Published in 1967 as ISO/R 639:1967, *Symbols for languages, countries and authorities*
- Replaced in 1988 by ISO 639:1988, *Codes for the representation of names of languages*

History: Roots of ISO 639-2

- During the 1960s, the US Library of Congress developed the MARC (Machine-Readable Cataloging) standards to facilitate sharing of catalog records between libraries.
- They developed three-letter codes for use in header field 008/12-14 (Language) to identify the language in which the work is written, with more detail including other languages specified in field 041 Language Codes. Still published today as the *MARC Code List for Languages*.
- In 1979, work began on making that an American National Standard which culminated in the 1987 publication of ANSI Z39.53, *Codes for the Representation of Languages for Information Interchange*

History: Roots of ISO 639-3

- In 1974, Joseph Grimes wrote about creating the *Ethnologue* database of all known living languages:
  - "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."
  - The codes were behind the scenes in the database that generated the 8th (1974) and 9th (1978) editions
  - Beginning with the 10th edition (1984) they appeared in the publication itself

History: ISO 639-2 emerges

- During 1980s, the international standard for library cataloging was not using the international standard for language coding.
- In 1989, the library community (represented by ISO TC 46) created a Joint Working Group with TC 37 to work out a three-letter code standard that would work for both.
- In 1998, this resulted in the publication of ISO 639-2 and establishment of the Joint Advisory Committee
- In 2002, original ISO 639 was republished as ISO 639-1

History: ISO 639-3 emerges

- In 2000, OLAC was launched. Following the extension rules for IETF language tags, it used Ethnologue codes (as `x-sil-abc`) for living languages and Linguist List codes (as `x-ll-xyz`) for ancient languages.
- ISO TC 37/SC 2 was under pressure from its user base to provide codes for all languages. They approached SIL International in 2001; a formal work item ensued in 2002.
- In 2005, after over 600 code changes to align with ISO 639-2, the codes of the Draft International Standard appeared in the 15th edition of Ethnologue.
- Full adoption of ISO 639-3 in 2007

Conflicting views of “language”

- ISO 639-3 lies at the convergence of two very different notions of what a “language” is
- Einar Haugen (1966, “Dialect, language, nation,” *American Anthropologist* 68:922–35) describes these perspectives and labels them as:
  - Structural — the overriding consideration is the genetic relationship among varieties
  - Functional — the overriding consideration is how the varieties are used in communication

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 lies behind ISO 639-1 and ISO 639-2
  - A prerequisite for getting a code is that there be at least 50 books published in the language.

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 lies behind the code set originally developed for the *Ethnologue,* in which most languages were unwritten.

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 variants.
  - Recognizing that the widely scattered variants 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 variants

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 languages themselves.

“Macrolanguages” to the rescue

- We reconciled the differences by introducing 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
- By introducing 55 macrolanguages to ISO 639-3 we were able to reconcile the differences.

Overview

- Where are we?
  - The ISO 639 family of standards
- How did we get here?
  - Philosophy of language codes
  - Historical timeline of ISO 639
  - Two conflicting views of what “language” means
  - A mechanism for resolving the conflict
- Where from here?
  - Changing the code set
  - Changing the infrastructure

Change management process

- The standard provides both:
  - A set of standard three-letter codes
  - An open process for making changes to the code set
- Thus, where ISO 639-3 goes from here depends on 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 2006–2012 summarizing all change results
  - *Submitting change requests* — 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 summarizing the changes requested along with the completed change request form, any supporting documents, and the Registration Authority's rationale in the case of a rejected change

How's business?

- In 7 annual cycles (2006–2012) we have processed 918 change requests with these results:

Outcome	Change requests	Per cent
Adopted in full	820	89.3%
Adopted in part	9	1.0%
Rejected	72	7.8%
Pending better documentation	13	1.4%
Stuck because 639-2 is affected	4	0.4%

Example: A name change

- Change request: CR 2008-003
- Affected code: adl
- Request:
  - Change reference name from “Adi, Galo” to “Galo”
- Rationale:
  - “Galo do not identify as a ‘subtype’ of Adi, nor is it linguistically sound to insist that they are (since there is in fact no clearly-defined language ‘Adi’). ‘Adi’ means ‘mountain/hill (people)’ in most Tani languages, and is not a linguistic label at all.”
- Outcome: Adopted

More examples

- Mayan languages
  - Nora England submitted 16 merger CRs (2008-048 to 2008-063) to align the standard with the consen-sus 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 during 2011 and 2012 to clean up the code set for Australia: 4 name changes, 11 splits, and 106 creations of missing languages (mostly extinct)

Total codes changed: 2006-2012

Action	Changes	Reason	Changes
Retired	223	Merged	115
		Split	79
		Non-existent	21
		Duplicate	8
Created	506	New	302
		From split	204
Updated	480	Name change	400
		Widened via merger	62
		Macro scope	18
Total	*1209*		*1209*

Possible infrastructure changes

- Status quo: Parts are under 4 Registration Authorities
  - TC 37 / SC 2 wants to unify it under a single RA
- Key features of proposed “next generation” ISO 639
  - ISO 639 Management Board — One representative per TC 37 national member body
  - ISO 639 Secretariat — Handles CR work flow
  - Scientific Advisory Board — Reviews the CRs
- Progress
  - Planned change announced in 2009, but revision of standard has not moved beyond Working Draft

Conclusion

- ISO 639-3 language codes are arbitrary, but internationally standardized, labels for a comprehensive system of “language zones” that cover the earth.
- That system is being used widely to enable tasks like resource discovery, information sharing, and implementation of language software tools.
- An open process of change management is working to allow the various user communities to continually improve the set of codes.

The criteria in ISO 639-3

- 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.