The language codes of ISO 639:
A premature, ultimately unobtainable, and possibly damaging standardization

Stephen Morey*, Mark W. Post[+] and Victor A. Friedman[#]
La Trobe University*, University of New England[+] and University of Chicago[#]

1. ISO 639: What is it?

- ambitious attempt to catalogue and standardize various types of references to the languages of the world
- aims to facilitate precise reference to languages, language groups, and language varieties, aiding in cataloguing, archiving, and machine processing (among other potential uses)
- designed to succeed where ordinary language has failed
    - "Adi" and "Abor" ref. to same language (*adi*)
    - "Ama" and "Ama" ref. to different languages (*nyi* and *amm*)
- developed by International Organization for Standardization (ISO) in Geneva, CH
    - provides "requirements, specifications, guidelines or characteristics that can be used consistently to ensure that **materials, products, processes and services** are fit for their purpose" (International Organization for Standardization 2013)
- consists of six parts (Table 1):

Standard	Name (*Codes for the representation of names of languages*)	First edition	Current	No. in list
ISO 639-1	*Part 1: Alpha-2 code*	1967 (as ISO 639)	2002	184
ISO 639-2	*Part 2: Alpha-3 code*	1998	1998	>450
ISO 639-3	*Part 3: Alpha-3 code for comprehensive coverage of languages*	2007	2007	7704 + local range
ISO 639-4	*Part 4: Implementation guidelines and general principles for language coding*	2010-07-16	2010-07-16	(not a list)
ISO 639-5	*Part 5: Alpha-3 code for language families and groups*	2008-05-15	2008-05-15	114
ISO 639-6	*Part 6: Alpha-4 representation for comprehensive coverage of language variants*	2009-11-17	2009-11-17	?

Table 1 – The six parts of the ISO 639 regime (http://en.wikipedia.org/wiki/ISO_639, accessed on 19/07/2013), parts in **bold underline** have already been approved by ISO

- each part has a different "registration authority": International Information Centre for Terminology (Infoterm; ISO 639-1), Library of Congress (ISO 629-2), **SIL International (ISO 639-3)**

2. SIL International[1] and the administration of ISO 639-3

- SIL's *Ethnologue* forms basis for ISO 639-3

- "three-letter codes" (= "language identifiers"), one for every "language" ever spoken

- three-letter codes *in principle* arbitrary, but *in fact* mostly mnemonic (= *meaningful*)
    - ex: *rus* = <u>Rus</u>sian, *spa* = <u>Spa</u>nish, *jpn* = <u>Jap</u>a<u>n</u>ese
    - ex: *czh* = Hui<u>zh</u>ou <u>C</u>hinese, *cjy* = <u>J</u>in<u>y</u>u <u>C</u>hinese, *cdo* = Min <u>Do</u>ng (*c* = <u>C</u>hinese (in English)
    - ex: *nyi* = Ama (reflects alternate name <u>Nyi</u>ma(ng)
    - ex: *clk* = Idu (reflects pejorative exonym <u>Ch</u>u<u>l</u>i<u>k</u>ata)

- submissions, change requests and decisions administered by SIL International
    - *creation* of a new code
        - "languages not previously associated with another language in the code set" (Spanne 2010)
            - e.g. *jkr* "Aka Koro", following Anderson and Murmu (2010)
    - *retirement* of an existing code
        - usually when two codes exist for "the same" language, and one is discarded
            - e.g. *drh* "Darkhat" merged with *khk* "Halh Mongolian", *drh* discarded
    - *update* of an existing code
        - code is re-defined as referencing a language it previously did not, or other facts about the denotatum change

What is *not* possible is:

- code is changed, while code-denotatum relationship remains unchanged
    - e.g. *adl* (to which the Galo community of Arunachal Pradesh, North East India, objected, could not be changed to something preferable or arbitrary, such as *glx* or *zzq*)
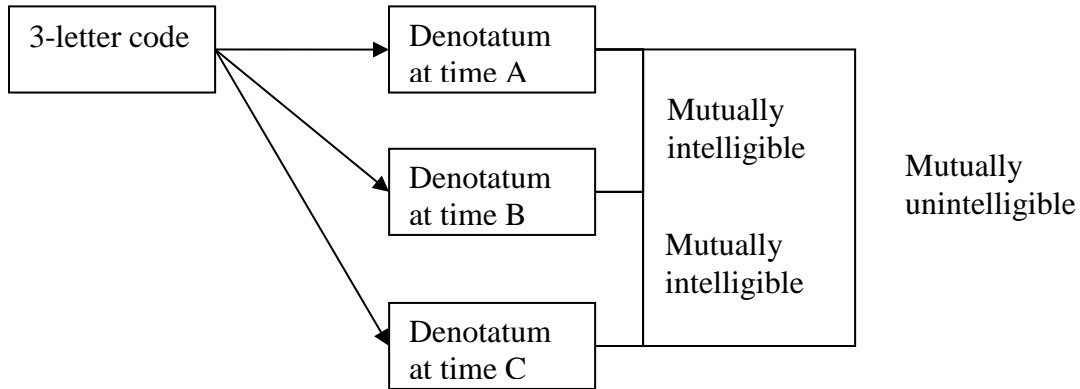
"Since 1989 not one single ISO 639 language identifier has been changed (once they have been assigned). This is a very strong principle, which has been strengthened during the lifetime of the ISO 639 series…Actual changes to the language identifiers can only be done if the scope of the encoded item itself has changed (e.g. if what was seen as one language previously were now to be seen as two or more separate languages)." (Håvard Hjulstad. Secretary of the ISO 639 JAC, email communication to Mark W. Post, 20/10/2009)

---

[1] SIL International is the new name for the erstwhile Summer Institute of Linguistics.

3. What are the problems?

1. Use of three-letter alphabetic codes is problematic

    a. in principle arbitrary, in fact mnemonic (*eng* = English, etc.)
    b. many based on pejorative labels, which "permanently" enshrines offensive designations for (usually minority) speech communities
        i. *clk* "Idu" < Chulikata (pejorative)
        ii. *jnj* "Yemsa" < Janejero (pejorative)
        iii. *dap* "Nyishi" < Dafla (pejorative)[2]

2. Administration of ISO 639-3 by SIL International is problematic

    a. not a disinterested scholarly organization, but in fact a Christian missionary organization whose main aim is translation of the Christian Bible (Dobrin and Good 2009; Olson 2009)
    b. design of Ethnologue and related administration of ISO 639-3
        i. lacks transparency and accountability
        ii. seemingly privileges closed internal resources over open, peer-reviewed scholarly resources
            *   e.g., section on Tani languages continues to ignore Sun (1993), resulting in many incorrect classifications of well-documented languages
            *   new section on Aka Koro based on internal unpublished reports rather than on Anderson and Murmu (2010), despite that the latter attracted worldwide press coverage
        iii. does not cite sources, in general
        iv. does not regularly and proactively seek outside expert commentary
    c. right or wrong, many linguists eschew cooperation with SIL due to its sectarian focus and uneven academic credentials and practices

3. The concept of "permanence" is fundamentally incompatible with the nature of human language

    a. languages are not static entities, but are in fact dynamic processes
    b. languages change over time. Chaucer's English is mostly unintelligible to modern English speakers. Should it receive a distinct ISO 639-3 code? If so, where is the cut-off date?
    c. all languages which are currently referenced by ISO 639-3 *will* change, to the extent that they are eventually unrecognizable. The "permanent" designator is thus certain to – eventually - fail to fulfill its purpose.

---

[2] Between 30/06/13 and 28/11/13 *dap* was changed to *njz*, so perhaps this rule has changed?

```
┌──────────────┐        ┌──────────────┐
│ 3-letter code │ ─────▶ │ Denotatum    │
└──────────────┘        │ at time A    │
        │  ╲            └──────────────┘
        │    ╲                          ┌──────────────┐
        │      ╲        ┌──────────────┐│ Mutually     │
        │        ▶      │ Denotatum    ││ intelligible │       Mutually
        │               │ at time B    │└──────────────┘      unintelligible
        │        ╲      └──────────────┘┌──────────────┐
        │          ╲                    │ Mutually     │
        │            ▶  ┌──────────────┐│ intelligible │
        │               │ Denotatum    │└──────────────┘
                        │ at time C    │
                        └──────────────┘
```

4. The concept of "language" employed by ISO 639 is similar to the concept of "mineral" (static, discrete), and dissimilar to the concept of "biological species" (variable, evolving). But in fact, human language is more like the latter than the former.

   a. there is no commonly-agreed-upon and reliable means of determining the difference between "two languages" and "two varieties of the same language", and dialect chains are common throughout the world. What method will ISO 639 have for drawing boundaries in such cases?
      i. ex. Taupota chain: Ross (1988: 8) has five lgs. while *Ethnologue* has nine[3] (Easton 2007: 54-55)
   b. socio-political concerns regularly enter into the determination of "languagehood" - whose "version" will ISO 639 privilege?
      i. ex. Balkan region: territory from northeastern Italy and southern Austria through former Yugoslavia to northern and eastern Albania, northern Greece, and Bulgaria constitutes a dialect chain in terms of South Slavic (Friedman 1999); speakers call their language "Bosnian", "Croatian", or "Serbian" on the basis of "ethnicity": for Bosnian, Islam, for Croatians, Roman Catholicism, and Eastern Orthodox Christians either Serbian or Montenegrin, depending on "national feeling". An international body that claims final authority in the classification of languages must inevitably support one socio-political agenda over another.
   c. concept of "macro-language" recently introduced by *Ethnologue* and imported into ISO 639 hardly represents a distinct scholarly consensus

5. Some aspects of ISO 639 are at a minimum premature given the present state-of-the-art in linguistics, and may be in principle unattainable

   a. ISO 639-5 aims to catalogue genetic relationships among the world's languages, while…

---

[3] These are the Taupota group within Western Oceanic section of Austronesian. Their codes are *grw, hqw, mum, mvn, tpa, tbo, wag, wed* and *ykk*.

      i.   many languages remain to be described in many parts of the world – so how can they currently be grouped into "families"? (Blench and Post MS-2012)

     ii.   scholars continue to disagree on fundamental issues regarding classification in some areas/language groups

- e.g. for "Sino-Tibetan ~ Tibeto-Burman ~ Trans-Himalayan", c.f. Matisoff (2003), van Driem (2013), Blench and Post (2013)

   iii.   methods for genetically grouping languages seem to currently be evolving

- phylogenetic frameworks

   iv.   studies in language contact and creolization continue to pose problems for the concept of "genetic linguistics" (Burling 2007; McWhorter 2007)

6. ISO 639 has the potential to be misunderstood, misused and abused by governments and other decision-making bodies

    a.   existence or non-existence of ISO codes might support decisions to allocate/deny resources or grant/withdraw recognition

       i.   cf. Indian Government's current requirement of "indigenous script" (Burling 2011)

7. With its veneer of authority, ISO 639 has the potential to mislead scholars, since decisions regarding linguistic statuses and affiliations may be ceded to bureaucrats rather than experts.

4. What is to be done?

At a minimum:

- Administration of ISO 639-3 should be re-assigned to an organization with recognizable scholarly credentials and a proactive commitment to current scholarship, which operates transparently and in partnership with the international community of scholars.
- Labels should be random combinations of letters and numerals, and not mnemonic/meaningful.
- ISO 639-5 and ISO 639-6 must be delayed if not abandoned. There is nothing approaching consensus in these areas.

Ideally:

- The concept of "permanence" should be re-conceptualized. The simple but unworkable concept SYMBOL-DENOTATUM should be re-worked as (something like) a "permanent paper trail". Since we are not computer people, no; we don't know how to do this. But since the current SYMBOL-DENOTATUM

regime is manifestly incompatible with many aspects of the nature of language, it must be transformed somehow.

But ultimately:

- Are these standardization regimes truly necessary? Useful? Productive? Appropriate? Given that human language is not static or unchanging – neither in itself nor in terms of scholars' views about it – is this any more useful or appropriate than developing ISO codes for human religions? Ethnic identities? Cultural practices? Sexualities? Emotions? (for example, to conduct search queries across different literary genres?)

References

Anderson, Gregory D. S. and Ganesh Murmu (2010). 'Preliminary notes on Koro, a "hidden" language of Arunachal Pradesh.' *Journal of Indian Linguistics* 71: 1-32.

Blench, Roger and Mark W. Post (2013). 'Re-thinking Sino-Tibetan phylogeny from the perspective of North East Indian languages.' In Nathan Hill and Tom Owen-Smith, Eds. *Trans-Himalayan Linguistics*. Berlin, Mouton de Gruyter: 71-104.

--- (MS-2012). *Declassifying Arunachali languages: Reconsidering the evidence*. Cambridge, UK and Armidale, AU, Kay Williamson Educational Foundation and University of New England.

Burling, Robbins (2007). 'The lingua franca cycle: Implications for language shift, language change, and language classification.' *Anthropological Linguistics* 49 (3-4): 207-236.

--- (2011). 'Three meanings of "language" and "dialect" in North East India ' In Gwendolyn Hyslop, Stephen Morey and Mark W. Post, Eds. *North East Indian Linguistics Volume 3*. New Delhi, Cambridge University Press India: 35-45.

Dobrin, Lisa M. and Jeff Good (2009). 'Practical language development: Whose mission?' *Language* 15 (3): 619-629.

Easton, Catherine (2007). *Discourses of Orthography Development: Community-Based Practice in Milne Bay (P.N.G.).* . PhD dissertation. Department of Linguistics. Melbourne, La Trobe University.

Friedman, Victor (1999). 'Macedonian Language and Identity: Recent History and Recent Historiography.' In Kramer, Christina and Brian Cook, Eds. *Guard the Word Well Bound: Proceedings of the Third North American-Macedonian Conference on Macedonian Studies, [Indiana Slavic Studies 10]*. Bloomington, IN: Slavica: 71-86.

International Organization for Standardization. (2013). "Standards." Retrieved 19-Jul, 2013, from http://www.iso.org/iso/home/standards.htm.

Matisoff, James A. (2003). *Handbook of Proto-Tibeto-Burman: System and Philosophy of Sino-Tibetan Reconstruction*. Berkeley, University of California Press.

McWhorter, John (2007). *Language Interrupted: Signs of Non-Native Acquisition in Standard Language Grammars*. Oxford, Oxford University Press.

Olson, Kenneth (2009). 'SIL International: An emic view.' *Language* 85 (3): 646-652.

Ross, Malcolm (1988). *Proto-Oceanic and the Austronesian Languages of Western Melanesia*. Canberra, Pacific Linguistics.

Sun, Tian-Shin Jackson (1993). *A Historical-Comparative Study of the Tani (Mirish) Branch of Tibeto-Burman* PhD Dissertation. Department of Linguistics. Berkeley, University of California.

van Driem, George (2013). 'Trans-Himalayan.' In Nathan Hill and Tom Owen-Smith, Eds. *Trans-Himalayan Linguistics*. Berlin, Mouton de Gruyter.