The language codes of ISO 639:
A premature and possibly unobtainable standardization[1]

Stephen Morey* and Mark Post+
La Trobe University* and Bern University+

<NOTE: This paper is distributed to add to the discussion about ISO 639 at the meeting in Newcastle in February 2013. It is incomplete and not yet ready for citation or further distribution, though we certainly welcome comments on it!>

1. Aims of this paper

The aims of this (draft) paper are:

a) To explain ISO 639 (particularly 639-3) to a scholarly community (academic linguists) that has not yet sufficiently engaged with it,
b) To problematize those aspects of ISO 639 that we have identified as problems,
c) To make suggestions (points 1 to 6 below) of the most important issues to address

ISO 639-3 (*Alpha-3 code for comprehensive coverage of languages*, see below 3) is the best known portion of a 6 part coding system for language set up by the International Organization for Standardization[2] (ISO). Each one of the six parts has, or potentially has, a different registration authority, the registration authority for most of ISO 639-3 being the Summer Institute of Linguistics (a.k.a. SIL International)[3]. The use of these language codes is becoming widespread and in some cases mandatory – they are necessary for archiving, for many grant applications, for setting up wikipedia pages in a language. The actual standard is maintained on the ISO site[4] where language codes are one of the 'popular standards'. The 2007 version is available for CHF 74,00[5], but most people consult the *Ethonologue* on line (http://www.ethnologue.org/ and http://www.ethnologue.com), which is free but has perhaps been updated from the latest available ISO version and also contains considerable information that is assumed to be not part of the existing standard (such as sociolinguistic information, whether there is a Bible translation available &c).

[1] The research work here is funded in part by an Australian Research Council Future Fellowship for the project *A multifaceted study of Tangsa – a network of linguistic varieties in North East India*. We thank colleagues at the Research Centre for Linguistic Typology, La Trobe University, especially Felix Ameka, Tonya Stebbins and Ian Tupper, for their contributions to the discussion of these issues. The views that we express are purely our own.

[2] Readers will notice that the acronym ISO does not derive from the organization's name in English. This is explained on the ISO website (http://www.iso.org/iso/about/discover-iso_isos-name.htm) as follows: "Because "International Organization for Standardization" would have different acronyms in different languages ("IOS" in English, "OIN" in French for *Organization internationale de normalisation*), its founders decided to give it also a short, all-purpose name. They chose **"ISO"**, derived from the Greek *isos*, **meaning "equal"**. Whatever the country, whatever the language, the short form of the organization's name is always ISO.

[3] SIL was invited to do this because they had staffing available to undertake the task, the Ethnologue was plainly the best developed resource of its type and the ISO's existing structures that led to ISO 639-2 and 1 were so patently unsatisfactory and needed to be replaced.

[4] http://www.iso.org/iso/home.html

[5] Swiss Francs. We have not actually purchased the .pdf to look at it. The site for downloading this is http://www.iso.org/iso/catalogue_detail?csnumber=39534

There are 6 main issues that we wish to raise as problems:

1) There are many examples of problems with the existing ISO 639-3 codes, some of which we will illustrate below. It was premature of the ISO to adopt the Ethnologue codes as ISO 639-3 in the way they did.

2) While we can see that there is some purpose and benefit in having codes for purely bibliographical purposes, the nature of language as a fluid entity, combined with our very inadequate knowledge of the linguistic situation in many places and the fact that different scholars, and communities, define language in different ways, means that the use of ISO 639-3 for any policy decisions by Governments, Funding Bodies or other research organizations should be resisted.

At the heart of this issue is the point that while the ISO may wish to regard codes as fixed and permanent (in the way that the standards for a *metre* or a *gram* are), they are not, and we argue, cannot be, because language is constantly changing.

3) The decision making processes within ISO 639-3 need to be overhauled. In particular involvement and input from community members and from linguists working in the field should be sought every time a proposed change to ISO 639-3 is suggested. The final decisions about changes are currently made internally by SIL members.

4) The changing of ISO 639-3 codes **must** be permissible when existing codes are unacceptable for sociolinguistic reasons. In this paper we raise the issue of the Garo, whose language code *adl* encodes within it an incorrect suggestion that Galo is part of the Adi group of languages (section 6). So far the response from the ISO 639 Joint Advisory Committee to requests for change have been dismissive.

5) ISO 639-5 (*Alpha-3 code for language families and groups*) should be abandoned. Any attempt to standardise the grouping of languages is plainly absurd when for many languages in the world their genetic affiliations and family tree structures are either contested or based on research that is far from comprehensive.

There are countless examples of conflicts in the literature to illustrate this. Let us present just this: Suppose the sub-committee charged with the task of organising ISO 639-5 were to accept Greenberg's work on Amerind and/or Starostin's work on Nostratic, both of which are fundamentally at odds with received ideas? Many scholars believe the evidence that Greenberg and Starostin offered, and both were fine scholars, but their theories about language families will always be disputed. This is the key point: language families and language relationships are theoretical, always open to scholarly debate and amendment and fundamentally unstandardisable.

6) ISO 639-6 (*Alpha-4 representation for comprehensive coverage of language variation*) should not be proceeded with. Why on earth does there need to be standardisation right down to the level of the most minute variation? There is no logical end to this. Suppose we encode and standardise in ISO 639-6 the names of different varieties, even that variety may be different in different villages; and different for different speakers within a particular village, due to age, register, or their

own personal experience which may have caused them to mix with other varieties. The logical extension of "comprehensive coverage of language variation" is a code for every individual and every situation, which is ridiculous.

2. Introduction to ISO 639

How many languages are there in the world? Most linguists will be asked this question at one time or another in their careers. The answer will often begin something like, "well, it depends on how you define 'language'!" One may then go on to say something about the notorious difficulties of drawing clear-cut lines around "languages" in situations of dialect continua, multilingualism and language shift. Or, one might start by acknowledging the fact that there are in fact many linguistically under-researched areas of the world, in which a great deal of work must be done before we can provide any truly meaningful statements concerning how many languages are in fact spoken there, what their names are, and what other languages they may be related to.

Or one could instead refer the asker to the latest version of the International ISO 639-3 list of language codes. One current answer[6] by looking at the *Ethnologue* version of ISO 639-3 was 6,909. The ISO 639-3 codes have the imprimatur of a transnational organization, and are increasingly widely accepted in a variety of contexts. To archive language data in most international archives, one must state the ISO 639-3 code for the language in question. When you search international databases like the Virtual Language Observatory (http://catalog.clarin.eu/ds/vlo/), you will search materials connected to ISO 639-3 codes and language names will come up as per the ISO standard. To set up a Wikipedia page in a particular language, the ISO 639-3 code for that language must be enlisted (Dobrin and Good 2009: 626). And while the names given by human beings to their own (or to others') languages are notoriously unstable across time and space, the ISO 639-3 codes are in principle *permanent*; while languages or parts of them may be shifted to new codes, the codes themselves are intended to persist[7] for all time.

While we appreciate the motivation behind ISO 639-3 and sympathize with many of the apparent reasons for its adoption by various organizations and institutions, the purpose of this paper will be to suggest that there are nonetheless several problems with ISO 639-3 in its current form. Some of these problems are institutional, and concern the manner in which the ISO 639-3 regime is administered by the organization invited by the ISO to undertake this task, namely SIL. These problems might in principle be solvable, either by re-organizing SIL's administration of ISO 639-3, or else by developing a different system of administration. But the most serious problem with ISO 639-3 (in its current conception, at least) is that it is unable to capture the true distribution of human languages throughout both time and space.

In what follows, we first present a background overview of the ISO 639 regime, including its motivations, goals, and progressive incarnations. We next discuss various issues surrounding its current administration by SIL International. We follow with two "case studies" from the "Sino-Tibetan" region, which illustrate some of the problems and unintended consequences of the current ISO 639-3 regime, and conclude with a brief summary and some thoughts for further reflection.

[6] http://www.ethnologue.org/ethno_docs/distribution.asp?by=area accessed on 15th January 2013.

[7] They do persist; under the current system they can be retired but never disappear. And the whole aim of standards is to make them universal, presumably in time as well as space.

3. ISO 639-3: What is it and why does it exist?

ISO 639-3 is a coding system for the world's human languages, spoken and signed. It consists of a series of identifiers, or "three-letter codes", each of which is designed to refer always and only to one particular human language, whether that language is used in the world today or may once have been used but is no longer.

What is the justification for creating such codes? According to the Library of Congress website, itself the registration authority for ISO 639-2,

> Using "the" name of a language as the means of language identification in machine applications poses two distinct problems of ambiguity. Firstly, different languages can have identical or very similar names. For example, there are four languages called Lele: Lele [lle] of Papua New Guinea (Austronesian); Lele [lel] of the Democratic Republic of Congo (Niger-Congo, Bantoid); Lele [lln] of Chad (Afro-Asiatic); Lele [llc] of Guinea (Niger-Congo, Mande). Conversely, the same language may be called by multiple different names, for example, one name used by native speakers, another used by speakers of the neighboring language, and yet another used by the national government.
>
> (http://www.loc.gov/standards/ISO 639-2/faq.html):

The Library of Congress makes three important points in this passage, which we highlight here:

(1) different languages may have the same name
(2) the same language may have different names
(3) machines cannot cope with such ambiguities

The ISO 639 regime has been developed as a six-part solution to this problem. Of the six parts envisioned, four (parts 1, 2, 3 and 5) have so far been approved by the International Organization for Standardization, or ISO. These six parts are listed in Table 1.

Table 1: The six parts of the ISO 639 standard (from Wikipedia http://en.wikipedia.org/wiki/ISO_639, consulted on 17/5/2011)

Standard	Name	First Edition	Current	No. Lgs. 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[8]
ISO/DIS 639-4	*Part 4: Implementation guidelines and general principles for language coding*	(As of July 2009 in DIS stage)	-	-
ISO 639-5	*Part 5: Alpha-3 code for language families and groups*	2008-05-15	2008-05-15	114
ISO/FDIS 639-6	*Part 6: Alpha-4 representation for comprehensive coverage of language variation*	(As of July 2009 in FDIS stage)	-	?

The purposes of the first three parts of this standard are as follows, according to the Library of Congress (http://www.loc.gov/standards/ISO 639-2/faq.html):

> *The ISO 639-1 code set was devised for use in terminology, lexicography and linguistics.*
>
> *The ISO 639-2 code set was devised for use by libraries, information services, and publishers to indicate language in the exchange of information, especially in computerized systems. The codes have been widely used in the library community and may also be adopted for any application requiring the expression of language in coded form by terminologists and lexicographers.*
>
> *The ISO 639-3 code set was devised for broad use in a variety of applications where more specific language coding was necessary than the other two standards provided.*

Again according to the Library of Congress, this is how the code lists were developed:

> **ISO 639-1**: *Codes for the representation of names of languages: alpha-2 codes* was developed by the ISO TC37/SC2 in 1988 for use in **terminology, lexicography and linguistics**.
>
> **ISO 639-2**: *Codes for the representation of names of languages: alpha-3 codes* was developed by the ISO TC37/SC2-TC46/SC4 Joint Working Group. Work on the standard was initiated in 1989 because of the inadequacy of the ISO 639-1 two-character code list to represent a sufficient number of languages for **bibliographic and terminology needs**. The list was largely

[8] This number is the one in the Wikipedia site, which differs from the number you gave in paragraph 2. Which says something.

based on the *MARC Code List for Languages* (http://www.loc.gov/marc/languages/), which has been in wide use since 1968.

ISO 639-3: In 2002, ISO TC37/SC2 invited SIL International (www.sil.org) to participate in the development of a new standard based on the language identifiers in the *Ethnologue* that would be a superset of ISO 639-2 and would provide identifiers for all known languages. In 2004 the proposed new standard, ISO/DIS 639-3 was released, incorporating identifiers for living languages from the *Ethnologue* 15th ed. (www.ethnologue.com) and for historical, ancient and constructed languages from the languages database of LinguistList (linguistlist.org), accounting for more than 7000 individual languages. In February 2007, ISO 639-3 was adopted. Elements other than collections listed in ISO 639-2 are a subset of those listed ISO 639-3; every non-collective element in ISO 639-2 is included in ISO 639-3. The denotation represented by alpha-3 identifiers included in both ISO 639-2 and ISO 639-3 is the same in each standard, and the denotation represented by alpha-2 identifiers in ISO 639-1 is the same as that represented by the corresponding alpha-3 identifiers in ISO 639-2 and ISO 639-3.

As we can see from Table 1, only a relatively small number of the world's languages are included in ISO 639-1; mostly, these are "major" national or international languages. More languages were included in ISO 639-2, but still nothing approaching comprehensive coverage was reached. The goal of ISO 639-3 is to achieve comprehensive coverage of all human languages, present, past or (in principle) future. Let's look at a concrete example of how the system has worked in its various incarnations:

In ISO 639-1 and 639-2, Albanian is encoded in the following ways:

(1) ISO 639-1: sq

(2) ISO 639-2: alb (B)
ISO 639-2: sqi (T)

As shown, the two-letter code ISO 639-1 code *sq* was replaced by two three-letter codes in ISO 639-2: a "B" (bibliographic) and a "T" (terminology). There are twenty-one languages with alternative codes in ISO 639-2. In most cases, the "B" form is based on the name of a particular language in English, while the "T" form is based on a romanization of the name of that language in the language itself. Apart from these twenty-one cases, each language enlisted in ISO 639-2 has only one code.

In ISO 639-3, we find that the following set of codes has been assigned to Albanian:

(3) Albanian, Gheg ISO 639-3:aln
Albanian, Arbëreshë ISO 639-3:aae
Albanian, Arvanitika ISO 639-3:aat
Albanian, Tosk ISO 639-3:als

The reason for this proliferation is as follows:

The codes, and the views on human languages that underlie their structure and use, are derived from the *Ethnologue*, a publication of the Summer Institute of

Linguistics, whose staff also design and administer the ISO 639-3 regime (see above). According to the *Ethnologue* (http://www.*ethnologue*.com, consulted 17/5/2011), each of the four varieties of Albanian listed in (3) is spoken in a different country: Serbia, Italy, Greece, and Albania, respectively. There is in fact also a fifth code, ISO 639-3: sqi. This is defined as a "macrolanguage", of which the four varieties listed in (3) are "member languages". Now, the idea of a "macrolanguage", with "member languages" that fall within it hardly represents a distinct consensus in general linguistics. There are other models for understanding the relationships between linguistic varieties, including that of a *dialect chain*. But these issues are not open for debate in the context of ISO 639-3: the existence of "macrolanguages" and "member languages" is one of the many assumptions concerning human language and which are built into the coding system which has been adopted by ISO 639-3. Let us then examine the administration of ISO 639-3 in more detail:

4. SIL, Ethnologue and the administration of ISO 639-3

The first three "parts" of ISO-639 are administered by three different "registration authorities". The registration authority for the ISO 639-1 codes is the International Information Centre for Terminology (Infoterm)[9]; for the ISO 639-2 codes it is the Library of Congress[10] (which also administers ISO 639-5) and for ISO 639-3 codes it is SIL International[11] (http://www.loc.gov/standards/ISO_639-2/faq.html, consulted 14/5/2011).

SIL International became the registration authority by virtue of its maintenance, since the 1950s, of the *Ethnologue*. The *Ethnologue* is a publication both in print and on-line formats (Lewis 2009a and Lewis 2009b), which aims to include information regarding all of the languages that are, or have ever been, used in the world.[12] In addition to the most common colloquial name of a language, *Ethnologue* also provides alternative names, names of dialects or varieties, numbers of speakers and their geographical locations, in addition to other background information regarding apparent religious affiliations of the speakers of this language,[13] and whether and how extensively translations of the Christian Bible into this language are available. All of this information is then linked to a three-letter language code, which is the code that was adopted in 2007as ISO 639-3.

The three-letter language codes used in *Ethnologue* are, as in ISO 639-1 and ISO 639-2, usually an abbreviation (or near-abbreviation) of the "primary" name for the language in question – usually, this is the colloquial name for the language in English. Accordingly, most of the world's "major" languages receive three-letter codes along the lines of *rus* for "Russian", *spa* for "Spanish", *jpn* for "Japanese" and, of course, *eng* for "English". Sometimes, usually in the case of lesser-known languages, it

[9] Email: infopoint@infoterm.org.

[10] Email: ISO 639-2@loc.gov

[11] Email: ISO 639-3@sil.org

[12] The *Ethnologue* website (www.ethnologue.com, accessed 2011-06-16) in fact proclaims itself to be "an encyclopedic reference work cataloging all of the world's 6,909 known living languages." In fact, the *Ethnologue* includes much material concerning languages no longer spoken; the reason for the discrepancy is unclear.

[13] For example, speakers of the language referenced by *spa* (colloquially, Spanish), are declared to be "Christian", something which, despite the best efforts of the Spanish Inquisition and several subsequent regimes, seems to be an overgeneralization; according to Wikipedia, for example (http://en.wikipedia.org/wiki/History_of_the_Jews_in_Spain, accessed 2011-06-17), there are around 50,000 Jews in Spain today, at least some of whom can be presumed to speak their country's language.

happens that the three-letter code which “best fits” as an abbreviation has already been used by a more “major” language; in this case, it becomes necessary to develop an alternative abbreviation, which may be meaningful, or may not be. In many cases, relationships that are believed to exist between languages are referenced by the codes chosen; for example, several languages classified as “Sinitic” include an initial *c*, which presumably derives from English “Chinese”: Huizhou *czh*, Jinyu *cjy*, Min Dong *cdo*, Min Zhong *czo*, Pu-Xian *cpx* and of course Mandarin *cmn*.

In some cases, there is a complete discrepancy between the primary name for a language and the three-letter code assigned to it. Often, such codes are based on language names which have fallen out of use (and often enough, out of repute; see section 6 below). But this may be apparent only to a specialist; to ordinary people, the relationship might appear opaque. Finally, while *Ethnologue* includes much information regarding the classification of languages into groups and families, the classification scheme is not systematically represented in the three-letter codes. This task is apparently being undertaken separately by the Library of Congress under the aegis of ISO 639-5.

As registration authority for ISO 639-3, SIL International entertains requests for changes of certain types to the *Ethnologue*-derived three-letter codes. Many such change requests have been received, from a wide range of individuals and institutions. The process takes around two years: in the “2010” round, change requests were submitted between September 2009 and June 2010. They were then subjected to a decision and finally to a review in March and April 2011. We do not know whether the approvals in 2011 are now included as part of the ISO 639-3 standard or only on *Ethnologue* and waiting to be approved by the ISO. In other words does a change approved by SIL automatically become a change approved by the ISO?

We presume that in both cases the review panel is chosen by SIL International, and primarily includes members of SIL International. Apparently there is no external “expert” review process (that is, a process of contacting people with expertise in the language(s) / language families in question). Outcomes of the 2010 requests were expected to be announced in May 2011. Change requests received after July 1, 2010 will then be deferred until the 2011 series of change requests, the outcomes of which will occur in 2012 and so on. All change requests which have been submitted to them are retained permanently on SIL International’s website, together with the details of any changes approved or denied (http://www.sil.org/ISO 639-3/).

For the 2009 round, 89 requests were considered “recommending 137 explicit changes in the code set. Twelve of the requests are still pending. Of the 77 requests that have been decided, eight have been rejected, five have been partially adopted, and 64 have been fully approved.” (Spanne 2010)

The requests included some *retirements*, which are codes that are seen as no longer needed. In 2010, this included eight previously-coded languages which were merged to other languages, and nine languages which were split into two or more languages, resulting in 20 new language code elements. In addition there were “14 *newly created languages* not previously associated with another language in the code set” (we presume this means 14 newly created language *codes*) and 47 *updates*, including name updates, denotation updates and macrolanguage updates. Table 2, based on Spanne (2010), presents examples of some retirement proposals. As we can see, the codes *drh*, *tnf* and *drw* were “retired” because each of these varieties has been merged with another language. Those codes remain essentially dormant in the system, and will not be reused – unless, perhaps, the situation before their “retirement” is somehow reinstated. In the case of *btb*, SIL International was satisfied that it in fact designated not a language but rather a group of languages, all seven members of which already had codes. Accordingly, the code *btb* could be "retired"[14].

Table 2: Retirements from other than normal split of a language code element

Change Request number	Identifier	Reference Name	Retirement Reason	Retirement Remedy	Outcome
2009-020	drh	Darkhat	Merge	Merge with Halh Mongolian [khk]	Adopted
2009-027	tnf	Tangshewi	Merge	merge with [prs] Dari	Adopted
2009-028	drw	Darwazi	Merge	Merge with [prs] Dari	Adopted
2009-032	btb	Beti (Cameroon)	Split	Beti is a group name, not an individual language name. Member languages are Bebele [beb], Bebil [bxp], Bulu [bum], Eton [eto], Ewondo [ewo], Fang [fan], and Mengisa [mct], all of which already have their own code elements.	Adopted

In this way we can see how the *permanence* of ISO 639-3 is intended to operate. A language may cease to be spoken, but the three-letter code will continue to reference that language in its spoken state. If it is determined that two "languages", each with its own three-letter code, are "in fact" not two languages, but are rather one language (which perhaps happened to receive different designations for whatever historical reasons), then one of the codes can be retired while the other persists. If it is determined that one "language" is "in fact" not one but multiple languages, then the "new" languages are assigned new codes, while the erstwhile code is either retired or, more often, apparently, retained for one of the languages but not for all. These are the types of changes which are permitted. There are also changes which are not permitted. These include cases in which the three-letter code references a pejorative name for a community of language speakers in the code of their language, as well as cases in which the three-letter code references a purported relationship between two or more languages which turns out to be incorrect and possibly offensive to one or more of the referenced groups. We will have more to say about such cases below. But first, in order to understand why the ISO 639-3/*Ethnologue* coding scheme operates the way it does, we will need to understand a little more about the administering organization: SIL International.

SIL International has three main goals, listed by Olson (2009: 648) in this order: (1) Bible translation, (2) Research ("which mostly involves linguistics") and (3) Literacy. As Olson (2009: 650) makes clear, "SIL's articles of incorporation do not mention religious or ecclesiastical activities, but rather the three goals discussed here.

[14] It is possible that btb has already been used in some documentation / archiving / or other place. Presumably someone can look back to the pre 2010 listing to find out what btb was and work out what new code could be better used. But this raises another problem. Suppose a speech variety which had a single ISO code was split in two and the old code was retained for just part of its former use. This would mean that if the code was used prior to the split it would refer to something different from the way the code was used after the split. Maybe if there's a change where there is a split, the old code is always retired to overcome this problem. But what about 'updates' - could these result in changing what the code refers to?

This is because the goals of the organization—including Bible translation—are not religious tasks per se, but rather scholarly ones."

Nevertheless, we would here like to raise the issue of what influence the primacy of the aim of Bible translation might have on the structure of *Ethnologue*. As an explicitly Protestant Christian organization, SIL International members are presumed to have a strong belief in the importance of people being able to read the words of the Bible in their own language. To achieve this result worldwide, the Bible needs to be translated into every language. It is in this context that we can understand the genesis of *Ethnologue*, including its existence, its content and its organization. As can be seen quite clearly from statements made throughout the work, *Ethnologue* is designed to record the extent of Bible translation that has been accomplished or not accomplished for a particular language, and for this to be cross-referenced with information concerning the religious affiliation of the language's speakers. In order to process this information effectively and prioritize research work, translators need to understand how different languages are from one another, how vital languages are, and the sociolinguistic circumstances surrounding the identification and use of languages.

Decisions about changes to the ISO 639-3 standard are therefore presumably affected by the desire to make Bible translations, but, as we shall see below, this may lead to problematic decisions.

One of the issues for academic linguists is that we do not necessarily make or even have the time to critique the details of the *Ethnologue*, and to follow the processes for changing the standards that we have outlined. Many SIL members are making time to do this and as a consequence are getting their ideas recognised as part of the formal standard. As Dobrin and Good (2009) already showed us, as academic linguists we need to engage more with these issues.

But so, perhaps, does the registration authority. Would it be too much to ask that the registration authority has a requirement to keep up with the literature that is published on language classification, and if a paper is written revisiting a classification, rather than requiring the scholar who wrote it to make an application, should there be a process whereby the registration authority initiates that? There are web-based discussion lists that many linguists belong to. As a start, could not the registration authority advertise that when there is a development in our knowledge, everyone on, for example, the Tibeto-Burman list is told that this innovation may be adopted for ISO 639-3? There will need to be a change in the way we conceptualise some of these activities. The present request for change form is based around answering questions whereas our methodology has to be more about asking them.

Another major issue with ISO 639-3 is that the *Ethnologue* on which it is based contains a family tree structure on which it is appears to have been at least partly built. Genetic affiliations are given for each *Ethnologue* coded language. *Ethnologue* is the place where everybody will consult the standard – especially when we consider that it is necessary to pay 74 Swiss francs to download the standard from the ISO website[15]. But when consulting the *Ethnologue* entries, a family tree structure comes with them.

5. Case Study 1: Tangsa

[15] http://www.iso.org/iso/search.htm?qt=iso+639-3&published=on&active_tab=standards accessed 3/6/2011.

We will proceed to discuss the problems of ISO 639-3 in detail, including those that relate to the relationship between ISO 639-3 and the *Ethnologue*, in terms of two language communities that we are familiar with and have researched for some time.

Tangsa is the name given in India to a community of several tens of thousands living on both sides the India-Myanmar border. Under the name Tangsa, they are a scheduled tribe under the Indian Constitution (listed under 'other Naga tribes'). The name was coined in the 1950s. This name was coined in the 1950s by Indian Government Officials. Bipin Borgohain, former Political Officer, Tirap Frontier Division, wrote that

> "the once subjugated but now liberated and resurgent lovable Tangsa (*Tang* = Mountain, *sa* = person), a word which was specially coined by the undersigned and accepted by the tribe and the Government for official use ..." (Foreword written by Bipin Borgohain in Barua 1991: viii).

As best we can tell, the term Tangsa refers to small communities living in what is now Changlang district of Arunachal Pradesh that were not otherwise categorised as belonging to one of the bigger languages like Singpho (ISO 639-3:sgp)[16]. The data collected by Thomas (2009) in India and Statezni and Ahkhi (2011) in Myanmar/Burma, together with our research, show that there are about 70 sub-groups of Tangsa, each speaking a distinct variety, some mutually intelligible and some not.

Prior to the 1950s these groups who were gathered together as Tangsa seem to have been referred to only by their own group name, or sub-tribe name (sub-tribe is the usual English term used in India). A 1927 British map names these as Moklum Naga, Mossang Naga, Jugli Naga, Tikhak Naga and so on, but the term Naga does not appear to have been used in former times as a overarching term for these communities. Around 35 of these groups are found in India. The full list of Tangsa groups is available on the Wikipedia Tangsa site - http://en.wikipedia.org/wiki/Tangsa_people).

Each sub-tribe has it's own endonymn which is not necessarily the same as the 'general name' used by everyone to refer to them. For example the people described as Moklum actually call themselves Muklom, and those who are described as Ponthai call themselves Phong, and those whose general name is Kimsing have the name Chamchang as their ethnonym. If we are going to standardise these sub-tribe names (either as languages under ISO 639-3 or as variants under ISO 639-6), should we use the endonym (i.e. Chamchang) or 'general name' (i.e. Kimsing)? There are arguments for both – if we use something like kim for *Kimsing* other people will know what is being referred to, but if we used maybe chg for *Chamchang* we are honouring their own endonym. What need is there to standardise either of these?

The situation of Tangsa is further complicated by the fact that at least some of the peoples now called Tangsa have an indigenous term, *Hawa* or *Hewe* to refer to the larger group to which they all belong, but this does not cover all. In Burma, since 2003, the term Tangshang has been used to cover the groups called Tangsa in India (and perhaps some of those called Nocte too). The name Tangshang is derived from *Tang Nyuwang* and *Shang Nyuwang*, two siblings in the oral history. It is not cognate with Tangsa.

[16] If such a group was within what is now Tirap district, it got classified as Nocte (ISO 639-3:njb). Groups that are found in both places, like the Ponthai (Phong) thus get called both Tangsa and Nocte, depending on where they live.

We will not discuss the linguistic diversity of Tangsa in detail here, but it is certainly the case that while some of the language varieties subsumed under it are mutually intelligible, others are not.

By way of demonstrating the issues associated with the original adoption of the *Ethnologue* codes for ISO 639-3, Table 3 presents the entry for the Tangsa in the on-line version of *Ethnologue*, accessed in August 2006, at the time the *Ethnologue* codes were under consideration for adoption. Here the language is termed *Naga, Tase*. The immediate problem presented by this entry is in the name *Naga, Tase* and its designation as 'a language of India'. *Tase* is a cognate form to *Tangsa*, in the Chamchang (Kimsing) variety, incidentally the variety chosen by the Bible Society of India for a Bible Translation. So the principle name of the language adopted by ISO 639-3 is in a form of the name used by just one of the subgroups, privileged largely because of the Bible translation.

In both the name adopted by the Ethnologue and ISO and also implied by the code *nst*, the term *Tase* is used in combination with an overarching cultural term 'Naga'. This would be something like calling French 'European, French', but somewhat more problematic. While many Tangsas are happy to be associated as Nagas, a significant proportion do not want to be included under the heading 'Naga'. In part this is because of the association of Naga ethnicity with militant Christianity as witnessed by the motto 'Nagaland for Christ'[17]. Within India at least, we believe that a majority of the non-Pangwa Tangsas are not Christian, with most in the Tikhak group being Buddhist or followers of Rangfra, and most Moklums being followers of Rangfra, which is a kind of codification of traditional animist practices. Within the Pangwa groups, however, most are Christians.

[17] When searching for reference to this on-line, most of the websites that come up are those maintained by groups who are opposed to Christian conversion. Certainly there is evidence of forced conversion by some Naga insurgents of some Tangsa people to Christianity. As far as we can tell, the vast majority of Tangsas who have converted, however, have done so from their own free will. Even this much discussion, however, will be enough to convince our readers that the use of the term Naga is not without difficulty.

Table 3: Sample *Ethnologue* entry for ‘Naga, Tase’

NAGA, TASE: a language of India

The following is the entry for this language as it appeared in the 14th edition (2000).
It has been superseded by the corresponding entry in the 15th edition (2005).

SIL code: NST

ISO 639-2: sit

Population	17,000 in India (1997 IMA). Population total both countries 17,000 or more.
Region	Southeastern Arunachal Pradesh, Changlang District. Also spoken in Myanmar.
Alternate names	TASEY, TANGSA, RANGPAN, CHAM CHANG
Dialects	LONGPHI, YOGLI, HAVE, KHEMSING, LUNGCHANG, LUNGRI, MOKLUM, PONTHAI, RONGRANG, RONRANG, TAIPI, TIKHAK, SANKE (SHANGGE), SANGCHE.
Classification	Sino-Tibetan, Tibeto-Burman, Jingpho-Konyak-Bodo, Konyak-Bodo-Garo, Konyak.
Comments	Some dialects are widely divergent. Close to Nocte Naga. ‘Tase’ is the name of the language; ‘Tangsa’ of the people. ‘Tangsa’ means ‘hill people’. A Scheduled Tribe in India. SOV. Literacy rate in second language: 20%. Traditional religion, Christian (some). NT 1992.

Also spoken in:

Myanmar	*Language name*	NAGA, TASE
	Alternate names	TASE, TASEY, CHAM CHANG, TANGSA, RANGPAN
	Dialects	GASHAN, HKALUK, SANGCHE, SAUKRANG, LANGSHIN, MAWRANG, MYIMU, SANGTAI, TULIM, LONGRI.
	Comments	Some dialects are widely divergent. SOV. NT 1992. See main entry under India

Amongst the ‘Alternate names’ in Table 3, we see (1) names for the whole group, Tangsa and other alternate forms Tase, Tasey; (2) forms that appear to relate only to some groups within Tangsa – Rangpan, which appears to be an alternative term for Pangwa a term for some but not all Tangsa sub-tribes; and (3) Cham Chang which is the name of a single subgroup. The entry lists the following subgroups of Tangsa in India: “Longphi, Yogli, Have, Khemsing, Lungchang, Lungri, Moklum, Ponthai, Rongrang, Ronrang, Taipi, Tikhak, Sanke (Shangge), Sangche.” Khemsing represents the same subgroup as Chamchang. This list of fourteen represents less than half of the known Tangsa groups in India. This is not intended as a criticism of *Ethnologue*, the research necessary to discuss the Tangsa varieties even in the way we did above had not been done in 2007 and a better entry would not have been possible then. What we do question is that this was adopted by the ISO for a permanent standard when so little was known about the Tangsa group of languages.

Table 3 also includes the ISO 639-2 code, *sit*, which refers to 'Sino-Tibetan Languages', a group of several hundred languages in the ISO 639-3 code (based on *Ethnologue*). Within ISO 639-2 and 639-1, only the three national languages within Sino-Tibetan, Chinese, Burmese and Tibetan are represented in both, and within ISO 639-2 the Karen languages are added. This is shown in Table 4:

Table 4: ISO 639-2 and 639-1 entries for Sino-Tibetan languages

Language names	ISO 639-2 (B)	ISO 639-2 (T)	ISO 639-1
Chinese	chi	zho	zh
Burmese	bur	mya	my
Karen Languages	kar		
Tibetan	tib	bod	bo

(http://www.loc.gov/standards/ISO 639-2/php/code_list.php accessed 17/5/2011)

Clearly ISO 639-2 is completely inadequate as regards Sino-Tibetan languages, and the motivation for adoption ISO 639-3 is clear.

As already mentioned above, it appears that the name *Naga, Tase* was privileged because the Chamchang (Kimsing) variety was the one chosen by the *Bible Society of India* for the Tangsa Bible.

The SIL methodology is to undertake surveys of languages to establish the extent of intelligibility and come up with recommendations for setting up Bible translation projects and their concomitant literacy programs. (For an example of such a survey, see Kondakov forthcoming for the Koch languages in North East India).

On the basis of such survey studies, which are ongoing on the Burma side (Statezni and Ahkhi 2011), we can perhaps expect a proposal to split ISO 639-3: nst (Naga Tase) into smaller units on the basis of mutual intelligibility, but in such a way that it will be useful for the purposes of Bible translation and for literacy programs among other things. We do not think that this is necessarily a proper basis for a scientific classification of languages.

6. Case Study 2: "Sino-Tibetan" languages of the Eastern Himalaya

The Eastern Himalaya, roughly including the area between Bhutan, Assam, Tibet and Burma (mostly falling within the modern-day Indian state of Arunachal Pradesh), is one of the most under-researched linguistic regions in all of Asia. The majority of languages which are believed to be spoken in this region are so sparsely described that, for most purposes, they should be considered undescribed and unclassified. The area remains in principle off-limits to foreign researchers, although very tenacious individuals have been known to obtain the entire suite of national and state-level permissions that are required. The area is explicitly off-limits to foreign religious missionaries under any circumstances, meaning that linguistic research by SIL staff members is only indirectly possible, through local proxies.

There is a small literature on Eastern Himalayan languages, which for some languages extends back nearly a hundred and fifty years. The older literature, most of it written by British military officers or (more commonly) civil administrators posted in the region, makes numerous references to various tribes, and sometimes includes wordlists, grammatical sketches, and statements concerning the names and linguistic affiliations of various groups.[18]

Due to difficulty of access to most areas, much of the information published in such works was obtained second-hand; an administrator would encounter an individual from one region, ask him or her who lived in the next region and what language they spoke, and so forth and so on. Unsurprisingly, much of the information recorded was of questionable reliability. In many cases, "exonyms" (names used by outsiders) were recorded rather than "autonyms" (names used by a community to reference itself); for reasons we need not go into, many of these exonyms are nowadays considered to be pejorative.

Many of these are exonyms are now "permanently" etched into the ISO 639-3 code. Some examples are as follows:

Idu is a language spoken in the upper reaches of the Dibang river valley. It has traditionally been considered as part of a (speculatively Tibeto-Burman) "Mishmi" tribal cluster including the mutually-unintelligible languages Miju and Digaru (Taraon). *Ethnologue* has assigned a three-letter code to each of these three languages. The codes assigned to Miju and Digaru are *mxj* and *mhu*, respectively, the logic behind which is as yet unclear to us (the *m* seems to reference "Mishmi", whereas the second two letters may reference names with which we are unfamiliar). The code assigned to Idu, however, is *clk*. *Clk* derives from the Assamese word *chulikata* [sulikata], which literally means 'the one(s) with the cut hair'. The term references a characterization of the hairstyle traditionally worn by Idu that was presumably proffered by an Assamese-speaking informant to an author or administrator who was unable to visit the (highly remote) Idu-speaking area. It is a linguistically meaningless designation, and not one particularly appreciated by Idu speakers themselves.

Nyishi is a Tibeto-Burman language of the Tani subgroup (Sun 1993) spoken in west-central Arunachal Pradesh, with a large number of speakers with considerable regional prestige. In earlier times, Nyishi-speaking tribespeople appear to have been referred-to as "Dafla". Although it is unclear where this name derives from, it has long been disapproved-of by Nyishi tribespeople, who are liable to take offence if the name is used publicly. And it is used publicly, both in the *Ethnologue* and every time the ISO 639-3 code is published: the three-letter code *dap* approximates the Assamese pronunciation of *Dafla* (*f* and *ph* are in free variation in Upper Assamese).

Mising is another Tibeto-Burman language of the Tani subgroup, whose speakers primarily reside in the plains of Assam. Both the name "Mising" and an at one time more common name "Miri" are seemingly indigenous; however, only the name "Mising" is in use as a modern-day autonym; "Miri" is considered to be insulting to most Mising speakers. However, this name is referenced in the *Ethnologue*/ISO 639-3 code *mrg* (the *g* is of uncertain function, but seems arbitrary; *mir* and *mri* were already taken, by Mixe and Maori respectively).

Finally, Galo is also a Tibeto-Burman language of the Tani subgroup, which was classified as a Western Tani language by Sun (1993). Linguistically, Galo is opposed to the "Adi" cluster of Eastern Tani languages, with which it is nevertheless in sustained contact. For reasons discussed in Post (in press), there has for some time been a colloquial *tribal* alignment of Galo with "Adi" groups, a relationship which broke down some years ago when Galo leaders "officially" declared their group's non-inclusion in the Adi cluster on both cultural and linguistic grounds. Apparently taking vernacular tribal alignments to count as linguistic alignment, the *Ethnologue*

[18] A few representative examples include Robinson (1849), Needham (1886), and Dunbar (1915).

set up Galo as a *variety* of Adi which is simply "sociolinguistically distinct" (http://www.ethnologue.com/show_language.asp?code=adl, accessed 2011-06-17). The *Ethnologue*'s incorrect assessment[19] of the linguistic relationship between Galo and Adi is encoded in the three-letter codes: all Eastern Tani languages are referenced by the code *adi*, while Galo is referenced by the code *adl* – a clear reference to its being seen by *Ethnologue* staff as "a variety" of Adi.

In 2008, during a period of vigorous linguistic engagement by the Galo community (who were at the time lobbying for recognition by the State of Arunachal Pradesh as an "official third language"), Galo leaders became aware of the existence of the ISO 639-3 code referring to their language. The reaction was at first muted; numerous archaic designations are on the books in India, and the representative body of a given group simply goes through the process of applying to the Indian Government to have the designation changed. The Indian Government generally yields to community wishes in such instances, and changes pejorative designations to more acceptable names; Galo leaders therefore assumed that the ISO would be no less accommodating. They were mistaken. A request to change the designator *adl* to something more palatable (or at least arbitrary), was twice submitted to the SIL/*Ethnologue* administrators, once in 2008 and again in 2009, by the community-authorized Galo Language Development Committee, supported by the second author of this paper. Both requests were rejected. The reason given was that three-letter codes *cannot* change; they can only be created or retired. Galo community leaders were incredulous upon learning this. "What right does SIL/*Ethnologue* have to impose an incorrect name on our people," I was asked. An appeal was submitted, which was referred to ISO 639's Joint Advisory Committee. They had the following to say in reply:

> 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. The ISO 639 Joint Advisory Committee is committed to maintain the stability of the standard. Actual changes to the language <u>identifiers</u> can only be done if the scope of the encoded item itself has changed (e.g. if what was seen as <u>one</u> language previously were now to be seen as two or more separate languages).
> This does not seem to be the case when it comes to Galo. (email correspondence from Håvard Hjulstad on behalf of the ISO 639-3 Joint Advisory Committee to Mark W. Post, 2009-17-09)

In other words, the wishes of a community not to be insulted every time their language is publicly referenced does *not* fall within the purview of ISO 639's concerns. Such wishes are, however, consistent with the United Nations Declaration on the Rights of Indigenous Peoples (UNDRIP) Article 13, Section 1, which states:

> **Indigenous peoples have the right to** revitalize, use, develop and transmit to future generations their histories, languages, oral traditions, philosophies, writing systems and literatures, and to **designate and retain their own names** for communities, places and persons. (UNDRIP 13:1, available online: http://www.un.org/esa/socdev/unpfii/en/drip.html emphasis by the authors)

[19] It is unclear what sources were consulted by SIL/*Ethnologue* when conducting their assessment, as these are not cited (the word "reportedly" is, however, used). What *is* clear is that the most authoritative information source available on the topic, the well-known 1993 University of California at Berkeley PhD thesis of Tian-Shin Jackson Sun, was not consulted. If it had been, the incorrectness of SIL's assessment would have been obvious.

The UNDRIP is primarily aimed at states, and may not in principle apply to transnational organizations such as the ISO, which are not signatories to the UNDRIP. But it is the sense of both authors that most linguists who interact with indigenous peoples feel quite strongly that most principles stated in the UNDRIP should be adhered-to as a matter of universal practice (indeed, most of these requirements are made of linguists nowadays, whether explicitly or implicitly, when we sign off on the ethics guidelines and requirements that are imposed on us by our local universities). We believe it important to point out, therefore, that UNDRIP principles are not, in this type of case, being adhered-to by ISO 639-3/*Ethnologue*.

7. Fundamental problems with ISO 639-3

In criticising the adoption of the *Ethnologue* as the permanent standard for the classification of languages, we do not intend this as a criticism of *Ethnologue* or of SIL International which maintains *Ethnologue*. They should indeed be congratulated for many aspects of their work, but as we have already mentioned, the aim of Bible translation places a particular direction on approach to the issues of language classification and has affected the document that has emerged.

Rather, our criticism is aimed at the International Organization for Standardization for adopting the *Ethnologue* with little questioning, and at the many international organizations that now require the use of the ISO 639-3 codes for many different purposes. For example, when archiving Tangsa texts at DoBeS (http://www.mpi.nl/DoBeS), we have no choice but to put in ISO 639-3 codes for every language item that we archive, and given the likelihood already mentioned that ISO 639-3:nst will be changed at some point, those labels will quickly become redundant and misleading.

Fundamentally, though, we question the need for standardization at all. There are certainly arguments for favouring some language varieties for the development of literacy – perhaps to preserve one Tangsa variety, or three, by means of literacy / Bible translation is better than losing all of them. We do not propose to discuss these arguments here, but rather to point out that this kind of consequence will come as a result of this type of standardization. Standardization, with all its concomitant loss of diversity, is after all the aim of the ISO, and we shouldn't expect that a standardization of this kind should be any different.

We are concerned that governments will adopt the ISO 639-3 standard to make decisions about languages. Consider the situation of the Gbe group of languages in Ghana, Togo and Benin, the best known of which is Ewe. According to *Ethnologue*, consulted on 3/6/2011, there are 21 Gbe languages. One, Ewe[20], is spoken in Ghana; four are in Togo and 16 are in Benin. But this is not an accurate depiction of the linguistic variety within Gbe – there is not 16 times more variation in the Gbe languages of Benin than in the Gbe languages of Ghana. This apparent distortion appears to have arisen because (1) Ewe in Ghana is a standard language and has a Bible translation, and (2) SIL survey work in Benin has established the presence of 16 varieties[21]. Not only is there more variety in Ewe than the current *Ethnologue* picture suggests (and consequently the current ISO 639-3), but the family tree arrangement there is also questionable.

[20] Written Éwé in *Ethonologue*, though we are assured by Felix Ameka, pc, that the word does not have high tones and shouldn't be written this way.

[21] We thank Felix Ameka for sharing his knowledge of Ewe and the other Gbe varieties.

If the Government of Ghana bases its literacy policies on this, then it will produce materials only in standard Ewe, leading to dialect leveling and standardization, possibly without intention. If the Government of Benin basis its policies on the same information, it may need to produce 16 literacy programs and that may not be necessary if the variation is not as great as the codes make it appear.

Another problem is that while we are required to use these codes for archiving, there is a good chance that some will change. We have already listed ISO 639-3:nst in our archives for Tangsa at the DoBeS archive, Max Plank Institute, Nijmegen. There is a good chance that these will change in time, so if we use them, should we perhaps use the following formula 'ISO 639-3: nst as at 24/5/2011'? Perhaps they need to be treated more like webpage references?

We raise the following other issues for discussion

1) Linguistic community and language codes:
    a. codes shouldn't be pejorative, and if they are they should be changed,
    b. language communities should have the right to choose or at least approve their own codes,
2) Classification / grouping of codes
    a. classification of codes – is it reasonable that some codes are 'variants of others', in other words a iconic way of suggesting groupings that may not be valid,
    b. would entirely arbitrary codes be better than ones that are partially based on language names?

Ultimately, however, whatever changes are made to improve ISO 639, we think it is doomed to fail because of the nature of human languages: in principle, all languages which are currently spoken today will change. Accordingly, there will be a time when use of a code assigned in 2011 becomes inappropriate. And because of the graduality of language change, there is in principle no means of determining when that time will be, when it arrives.

While understanding the impulse toward standardization, we need to take account of the fact that the world just doesn't cooperate with our desires to make it behave the way we want it to. Linguistic Diversity is probably just not able to be standardized, and attempts to do so simply mean a lot of people end up wasting a lot of time and energy in the process.

8. Coda: Some useful references in the literature parts of which may be incorporated in a final version of this paper.

<This section is incomplete, but there are some studies which have referred to these issues and related ones that could be taken into account>

Dobrin and Good (2009:624)

> Yet the discipline of linguistics has, largely through its own inertia, allowed SIL to take over leadership roles that we argue would be more appropriately held by the academic community. The tensions here are felt most acutely within documentary linguistics, the core of the endangered-language research paradigm, since it benefits more directly than any other subfield from SIL infrastructure. Documentary linguists were made keenly aware of this situation with the 2007 adoption of the three-letter *Ethnologue* codes as a central component of the first comprehensive ISO standard for language identification, ISO 639-3, and the

> concomitant establishment of SIL as the registration authority overseeing the standard's update. ISO granted SIL, a missionary organization, authority over this international linguistic standard because the academic community was unable to offer an alternative. The *Ethnologue* is simply the closest thing that exists to a comprehensive listing of the languages of the world. Even before their adoption by ISO, the *Ethnologue* codes had already become the de facto standard, not only for individual linguists, but also for major digital language archives and funding programs.

(2009:625)

> So officializing the codes has had little effect on academic practice. It has, however, made it clear that academic linguistics has virtually nothing to say about an aspect of its object of study that is of intense and legitimate interest outside the discipline: when asked what the languages of the world are, it is only SIL that is ready to answer. One academic organization has officially questioned this new arrangement: the Society for the Study of the Indigenous Languages of the Americas (SSILA). SSILA neither supports nor condemns SIL's missionary activities. It does, however, acknowledge that a standard is not secure until it enjoys support from the full range of its users, something the *Ethnologue*-derived ISO 639-3 codes do not have (Epps et al. 2006, SSILA 2006).

(2009:626)

> For example, before a Wikipedia can be created in a new language, it is a requirement that the language be assigned a valid ISO code. The Wikimedia Foundation defers to ISO on this matter because it lacks the expertise to determine whether a speech variety should be deemed a language, and as we know, the stakes of such decisions are high. In this new era, what counts as a language as opposed to a dialect is no longer so much about who has an army and a navy; it is about who has a three-letter code.

The next two pages, here highlighted in yellow, are taken from section of a paper by Tonya Stebbins, *On becoming an object of study: legitimisation in the discipline of Linguistics*, which contains a detailed description based on Easton 2007 of a situationi in PNG:

Easton (2007:85-117) provides an account of the creation of language boundaries in a small area of Papua New Guinea, on the north coast of Milne Bay Province.

In the early days of contact with outsiders, geography functioned as a powerful filter that had a great deal to do with the identification of a language:

> 'Languages' were discovered and named when missionaries, government officials and plantation owners came into contact with the people living near deep harbours and land considered suitable for plantations, mission stations and government stations (Easton 2007:86).

The establishment of mission stations consolidated the status of some speech varieties into languages and expanded their influence far beyond traditional domains. In the present case, the first mission was established at Wedau with a second following at Taupota. In conjunction, grammars and dictionaries were produced for both varieties (Easton 2007:87). Ultimately only Wedau succeeded in achieving the status of a mission lingua franca and only the grammar and dictionary for Wedau were ever published.

From the time of the first settlement of the missions, there was a perception that a dialect chain existed between Wedau and Taupota. This perception was possible because of the lenses through which mission staff viewed the languages concerned. When Wedau assumed the status of mission lingua franca, one of the effects was to de-rank Taupota and obstruct, by negligence, its development into a written language (cf. Easton 2007:89-90). At the same time, the basic division of the region into these two main 'languages' persisted. It is found in the work of Ray (1907, 1938), Capell (1943, 1954, 1962, 1969), Lithgow (1976) and Ross (1988). Although these authors used a range of techniques in coming to their classifications, including comparison of lexical and grammatical similarities, typological features, and lexicostatistics, the basic distinction between Wedau and Taupota was taken as a given (see Easton 90-100). This division became part of the filtering process through which these languages were viewed.

Mirroring the practice of naming the language group after the most prominent variety, Ray classified Galavi, Wedau, Taupota and Awayama together with Kehelala, Tawala, Maiwara, Wagawaga and Bohilai into the 'northern or Wedau group' (Easton 2007:92), whereas Ross (1988:195) names the corresponding (though not identical group of named languages) after Taupota.

Easton contrasts the official linguistic classification of these languages with the ethnoclassification of the communities concerned. This research highlights the fact that communities also view their language through filters and lenses. Easton invited representatives from each of the villages in the region between Wedau village to the west and Galuwahi to the east to report on the speech varieties in the area (Easton 2007:208-219). One of the most intriguing findings of her study is that the ethnoclassifications of the groups involved do not match each other to such an extent that tensions are created across different communities when these issues are raised (as they are in the context of orthography development and language planning).

As a result of this variation, it was also not possible for Easton to produce a map summarizing the ethnoclassifications of the region since to do so would inevitably have privilege some language groups and their perspectives over others (Easton 2007:220). Irvine and Gal (2000:50-51) describe how similar considerations could also have challenged the makers of language maps in Africa in the 1880s if they had not been filtered out of consideration.

In summarising the discourses that communities used in constructing their own and other people's speech varieties as emblematic of particular social groupings, Easton identifies each of the semiotic processes (filters) identified by Irvine and Gal. She notes, for example, how people in each village could describe and imitate the 'tune' of the language in neighbouring villages. This is an example of iconization since the tune becomes a marker of other types of cultural and social boundaries (Easton 2007:231). She continues,

> 'In Topura, Aigura, Yapoa, and Awauya, where the original language has been lost, phonological and lexical features of their variety of Wedau have become the most important factor in creating boundaries between themselves and Wedau, not only linguistically, but also culturally (Easton 2007:231).'

Perhaps the strongest driver for differences in the allocation of language boundaries was the need for each community to distinguish itself from surrounding groups. In this context, fractal recursion plays an important role.

> Cultural oppositions based on land, or history can be reflected recursively in the creation of the 'other' as they are reflected onto language. Differences in the realisation of phonemes

> become the salient expression of otherness which extends from a much deeper difference. ‘Imagined others’ are created (Easton 2007:231-2).

This process can also work the other way, so that ‘salient linguistic differences become objects creating cultural and social boundaries (Easton 2007:232).’

Finally, erasure can be used to remove awareness or representation of both difference and similarity. Easton demonstrates the use of erasure to overlook difference with reference to the orthographic choices made by the Wedau and Wamira communities who chose the same orthographic symbol to represent phonetically distinct (though phonemically equivalent) sounds apparently to reinforce their shared identity (Easton 2007:244-5). She also notes (2007:232) that similarity can be erased, for example in the refusal of one community to recognise that the same speech variety is used in a neighbouring community, resulting in the less powerful community being excluded from orthographic decision making for this variety.

<extract from Stebbin’s summary finishes here>

Easton (2007: 54-55) points that for the Taupota chain, as defined by Ross (1988: 8), “somewhat arbitrary lines have sometimes been drawn to differentiate ‘languages’. This, she maintains, is the reason why Ross (1988) has 5 Taupota chain languages, but *Ethnologue*, and hence ISO 639-3, has nine[22]. These differences, Easton points out, are due to difference in methodology. We would claim that these differences are still subject to academic debate which is fair enough, but one of them, that of *Ethnologue*, is privileged because it has been accepted by the ISO as a permanent classification.

Easton (2007: 101-2) demonstrates how the historical accidents such as the locations of missions or the field sites of linguists led to the privileging of two varieties within a chain as ‘separate languages’ and some intermediate varieties were treated as deviations or dialects of these languages. Some varieties didn’t even get this much recognition, as she says, “the remaining intermediate speech varieties came to be ‘erased’ as variants of no significance.”

We are concerned that the whole ISO 639-3 process has the potential to do this worldwide.

9. References

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

Dunbar, G. D.-S. (1915). "Abors and Gallongs Part I: notes on certain hill tribes of the Indo-Tibetan border." *Memoirs of the Asiatic Society of Bengal* 5(Extra number): 1-86.

Easton, Catherine. 2007. *Discourses of orthography development: community-based practice in Milne Bay (P.N.G.)*. PhD dissertation. Melbourne: La Trobe University.

Kondakov, Alexander. Forthcoming. ‘Koch Dialects of Meghalaya and Assam: A Sociolinguistic Survey’. In G. Hyslop, S. Morey and M. Post (Eds.). *North East*

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

Indian Linguistics, Volume 5. To be published by Cambridge University Press, India.

Lewis, M. Paul (ed.), 2009a. *Ethnologue*: Languages of the World, Sixteenth edition. Dallas, Tex.: SIL International.

Lewis, M. Paul (ed.), 2009b. *Ethnologue*: Languages of the World, Sixteenth edition. Dallas, Tex.: SIL International. Online

Needham, J. F. (1886). *Outline grammar of the Shaiyang Miri language: As spoken by the Miris of that clan residing in the neighborhood of Sadiya, with illustrative sentences, phrase-book and vocabulary*. Shillong, Assam Secretariat Press.

Olson, Kenneth S. 2009. 'SIL International: An emic view' *Language* 85.3: 646-653.

Post, M. W. (in press). "The Siyom River Valley: An essay on intra-subgroup convergence in Tibeto-Burman". In G. Hyslop, S. Morey and M. W. Post, Eds., *North East Indian Linguistics Volume 5*. New Delhi, Cambridge University Press India.

Robinson, W. (1849). "Notes on the Languages spoken by the various Tribes inhabiting the Valley of Assam and its mountain confines." *Journal of the Royal Asiatic Society of Bengal* 18(1): 311-318, 342-349.

Saul, Jamie. 2005. *The Naga of Burma*. Bangkok: Orchid Press.

Spanne, Joan. 2010. ISO 639-3 Change Requests Series 2009: Summary of Outcomes. SIL International document, downloaded from XXX on 15/5/2011.

Statezni, Nathan and Ahkhi. 2011. So near and yet so far: dialect variation and contact among the Tangshang Naga in Myanmar. Presentation given at NWAV Asia-Pacific I conference, Delhi, India, 23-26 February 2011.

Stebbins, Tonya. *On becoming an object of study: legitimisation in the discipline of Linguistics.*

Sun, T.-S. J. (1993). *A Historical-Comparative Study of the Tani (Mirish) Branch of Tibeto-Burman* PhD Dissertation, University of California.

Thomas, Mathew. 2009. *A sociolinguistic study of linguistic varieties in Changlang District of Arunachal Pradesh.* PhD thesis submitted to the Centre for Advanced Study in Linguistics, Annamalai University, Tamil Nadu, India.