Workshop on Language Identifying Codes

Newcastle, February 9 2013-01-28

Programme

9:20 – 9:30	Opening
9:30 – 10:30	**Invited presentation**: ISO639: Where we are and how we got there Gary Simons (SIL International)
10:30 – 11:00	Code splitting in ISO 639-3 Brenda Boerger (SIL International)
11:00 – 11:30	**Coffee break**
11:30 – 12:00	**Invited presentation**: Towards a new improved setting for the ISO639 standard: the FROLIC approach Sebastian Drude (The Language Archive, MPI Nijmegen)
12:00 – 12:30	PARADISEC and ISO-639-3 Nick Thieberger (PARADISEC / University of Melbourne)
12:30 – 2:00	**Lunch**
2:00 – 2:30	Language Identifying Codes in the AUSTKIN Project Tom Honeyman (Australian National University)
2:30 – 3:00	Some issues from the field: who makes which distinctions on the basis of what evidence? Simon Musgrave (Monash University)
3:00 – 3:30	Capturing the neoglots: mixed languages, reviving languages and other hybrids Michael Walsh (AIATSIS)
3:30 – 4:00	**Coffee Break**
4:00 – 5:00	Discussion, future plans

The workshop will take place at the Newcastle Museum:
http://www.newcastlemuseum.com.au/

This workshop is funded by the Australian National Data Service.

Towards a new improved setting for the ISO639 standard: the FROLIC approach

Sebastian Drude (The Language Archive, MPI Nijmegen)

In this talk I plan to address some of the issues around language identifying codes that have been seen as problematic in the current technical and organizational setting. These issues have been raised by several linguists and by other people involved with the standardization efforts.

I plan to show a possible strategy for improving the current technical and administrative setting so that as much expertise as possible goes into the additions and changes to the different parts of the standard. Although currently substandard 639-3 is certainly the mostly used and most often discussed substandard, special attention should also be given to sub-standards 639-4, 639-5 and 639-6.

A group of linguists in Europe and the US interested in standardization, which already has successfully worked together in the RELISH project, has submitted "FROLIC", a project proposal for a new cooperative US-EU project that addresses this topic. Even if FROLIC is not granted, we all should unite and synchronize our efforts to improve the setting for maintaining the different parts of ISO639 now, taking advantage of the upcoming international congress on language documentation ICLDC.

Code splitting in ISO 639-3

Brenda H. Boerger

SIL International

This paper illustrates the open process for making changes to the ISO 639-3 code set, by discussing a code split I initiated. It delineates the steps and review procedures by which Nalögo [nlz] and Natügu [ntu] came to be recognized as distinct languages each with its own code. These two varieties represent opposite ends of a dialect chain on Santa Cruz Island in the Solomon Islands, which was originally represented by a single code [stc]. I summarize the lexical, textual, and sociolinguistic evidence used to address the three criteria for categorizing separate languages: lack of mutual intelligibility, lack of a common literature, and separate ethno-linguistic identities.

Natügu has already received considerable language development, primarily through work by the Natügu Language Program, for which I was advisor for over twenty years. But now, by having its own international language code, Nalögo is positioned to receive further attention from two directions. First, speakers can appeal to the Solomon Islands government for language development activities as it moves toward vernacular education for all language groups in the country. And second, Nalögo will have greater visibility internationally, perhaps leading to documentary linguistics work, which could in turn give input into the vernacular education materials requested by the community. Therefore, the code split shows one way the ISO 639-3 processes already in place can contribute toward a community achieving its goals.

PARADISEC and ISO-639-3

Nick Thieberger (PARADISEC / University of Melbourne)

As recently argued by Kamusella (2012), the regime of language codes is problematic for a number of reasons ranging from the very definition of 'language' to the contingent nature of language naming. While recognising that there are many difficulties associated with establishing standard codes, I suggest that we still need some mechanism for identifying languages, esepcially based on my experience with the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) which has used ISO-639-3 codes since its catalog began in 2003. In the main, these codes have served a useful function and are uncontroversial in identifying the language variety recorded in the collection item. We have always had an additional field for 'Language as given' and have had instances where we have used a code for a nearby language when there was none for the named variety included in the collection. We also provide a geographic reference for the language, either as a bounding box drawn by the depositor or else derived from the ISO-639-3 code (using centroid data bought from GMI[1]). Each of these metadata elements facilitates locating the item in the collection and that is its main function. In this talk I will summarise some of Kamusella's arguments with reference to the experience of building the PARADISEC collection of linguistic material. I will explore ways of using standards that do not necessarily reify language varieties.

Kamusella, T. 2012. The global regime of language recognition. *International Journal of the Sociology of Language* 218:59-86.

Language Identifying Codes in the AUSTKIN Project

Tom Honeyman (Australian National University)

The AUSTKIN project (Dousset 2010) centres around the compilation of a detailed database of the kinship terminology of Australian languages. The second phase of this project is focusing on social category systems (moieties, sections, subsections and others). In this talk I will discuss the problems revolving around the linking of language identifiers to clusters of terms that can apply across or within languages, especially those cases where there is insufficient data to identify the language itself.

All data in the collection is linked to an AIATSIS language code, and also to the terms in which it was originally described in source materials. This dual strategy has so far enabled us to sometimes unify seemingly disparate sources under the one language code, or to identify and flag insufficiently precise language codes as they have arisen during data collection.

Modern language identifiers often reflect the current understanding of the bounds of languages and language groupings. I discuss the need to be sensitive to the preservation of information on the groupings and other distinctions that factor into the understanding of the analyses held by the researchers at the time.

Linking data points to language identifiers enables the data collected to be inter-operable with other research. This has also allowed us to group together similar sources and give them an approximate centroid location, linked to the language identifier. As an authoritative source of codes that seeks to be more inclusive, I will discuss both the negative and positive aspects of using the AIATSIS codes for identifying languages in Australia.

Finally I argue for the coexistence (and not compromise) of language codes with more detailed, specific and project internal identification of languages.

References:

DOUSSET L. et al. 2010. Developing a database for Australian Indigenous kinship terminology: The AustKin project, Australian aboriginal Studies2010 (1): 42-56.

Some issues from the field: who makes which distinctions on the basis of what evidence?

Simon Musgrave (Monash University)

This paper examines some language varieties in the northern part of Ambon Island in the Maluku province of East Indonesia, considering the status accorded them by ISO-639 codes and by my own research, and assessing these statuses in light of sociolinguistic factors.

According to Collins (1982:90):

> the language spoken along the north coast [of Ambon Island – SM] from Seit to Tial and in Laha on Ambon Bay is called Hitu after its most prestigious village. There are three main dialects: Hitu-Tulehu, Seit-Kaitetu, and Laha.

In another publication, Collins (1983:100) treats the languages of Seit, Kaitetu, Laha, Hitu and Tulehu as distinct. *Ethnologue* and ISO-639 assign separate codes to the three dialects of Collins 1982: Hitu is [htu], Tulehu is [tlu] and Laha is [lhh]. My own research includes only data from the Hitu-Tulehu area, and it suggests that at the level at which Collins 1982 makes distinctions (i.e. that of dialect), there are further distinctions which could be made. The question then arises as to whether it would be useful to make such distinctions within a system of identifying codes, and if so at what level.

The situation is further complicated by the sociolinguistic situation in this region. There are at least two relevant factors here. Firstly, the speakers of these languages do not make clear distinctions between these varieties. The normal way of referring to any of these indigenous languages is by the Malay term *bahasa tanah* 'language of the land' and this applies equally to any of these varieties (and many others also). If more specificity is sought, a variety may be referred to by associating it with a village, but this does not necessarily correspond to the geographic range of the variety. These issues can be exemplified by the problems encountered in finding a way of referring to the variety from Tulehu which was the focus of my research. The second sociolinguistic issue is that these varieties are seriously endangered; very few people under the age of 30 have good command of them, and even older speakers use many Malay words when speaking their indigenous variety. This means that it is very difficult, if not impossible, to gather data which would reliably differentiate varieties and also that the distinctions which comparative linguists make might be different depending on the point at which their data is collected.

Collins, James T. 1982. Linguistic research in Maluku: a report of recent fieldwork. *Oceanic Linguistics* 21:73-146

Collins, James T. 1983. *The Historical Relationships of the Languages of Central Maluku, Indonesia*. Canberra: Pacific Linguistics (D-47)

Capturing the neoglots: mixed languages, reviving languages and other hybrids

Michael Walsh
AIATSIS

When I got into Australian languages around 40 years ago, a number of terms were yet to appear. These include: Neo-Nyungar; Light Warlpiri; Gurindji Kriol – so-called 'mixed languages'. With the upsurge of language revitalization over the last 30 years or so there have been many instances of 'new languages' based on language documentation and targeted linguistic engineering. This latter category has not typically taken on new labels that reflect their recent and evolving status. So, it is not usual to encounter terms like neo-Kaurna (Adelaide area, South Australia), neo-Wiradjuri (central New South Wales) or even neo-Arrernte (central Australia). However, these forms of speech need to be captured within the ISO-framework.

This presentation will focus on examples from Australian Languages but the issues raised have relevance for the numerous instances of linguistic hybridity across the world. ISO has more comfortably embraced the traditional linguistic classification for which the touchstone is hierarchy. The not infrequent occurrence of hybrid varieties presents a challenge for typical hierarchical models of linguistic connectedness. Such connectedness does not involve a 'daughter' of a 'single parent' but a descendant of two or more predecessors. The capture of such neoglots is a significant issue for the modification of the ISO classification.