LANGUAGE IDENTIFYING CODES: REMAINING ISSUES, FUTURE PROSPECTS

Simon Musgrave (Monash University, Australian National Corpus)
Linda Barwick (PARADISEC, University of Sydney)
Michael Walsh (AIATSIS, University of Sydney)
Andrew Treloar (Australian National Data Service)
Overview

• The importance of accurately identifying languages
  • Within research communities
  • In a wider context
• ISO639-3 – working with what we have
  • Improving linguistic input – Australian examples
  • Improving processes – registration authorities
• Looking to the future
  • Current developments in ISO639/TC37
  • Ways of influencing outcomes:
    • ARGILaRe
    • RDA Working Group
Language codes in disciplines outside linguistics

• Language coding could improve discovery and accessibility of web and library resources in other disciplinary areas e.g.:
  • Song
  • oral history
  • (ethno)biology
  • ...
• Accurate identification of languages can facilitate reintegration of cultural knowledge across domains
• This is a benefit for public and researchers
Implementation example

• Particularly useful for organizing and management of multilingual collections
• E.g. Western Arnhem Land song project
  • ISO639-3 (language identifiers) already adopted
  • potential use for 639-5 (language families and groups)
  • decisions on 639-6 (language variants) may be impossible without linguistic advice
• Only if appropriate infrastructure available:
  • standards agreed
  • tools for coding available
  • platforms for aggregation created, etc
• Need for dialogue with the library/cataloguing community
42 Australian languages in Western Arnhem Land Song corpus
Language coding for song

- Special song languages/registers: coding 'spirit language' used in song; coding songs entirely in vocables
  - Could use ISO639-3 mis ‘uncoded languages’
- Granularity: code switching within a single song item (e.g. 4 languages in single Malgarrin text)

<table>
<thead>
<tr>
<th>mulurn</th>
<th>kanarra</th>
<th>puratj</th>
<th>parraya</th>
<th>wantinya</th>
</tr>
</thead>
<tbody>
<tr>
<td>shade/leaves</td>
<td>leaves</td>
<td>brush</td>
<td>you go</td>
<td>[unknown]</td>
</tr>
</tbody>
</table>

Murrinh-patha mwf, Gija (Kitja) gia, English eng, Gija gia, Djaru (Jaru) ddj

- Indeterminacy: language-distinguishing grammatical markers may be absent, though possible to determine higher order language grouping (Arandic, Bininj Kunwok)
Wider communities

• Mac Developer Library: For language designations, you can use either the ISO 639-1 or ISO 639-2 conventions.

• Osborn, D. 2010. African Languages in a Digital Age. Cape Town: HSRC Press. This set of standards [ISO639] serves several purposes, including the identification of the languages of web content and the selection of appropriate locale information. (p73)

• W3C Internationalization. 2009. Language tags in HTML and XML
  All language tags must begin with a primary language subtag…….These codes come from, and are kept up to date with, ISO 639 language codes.
  (http://www.w3.org/International/articles/language-tags/)
**Yolngu:** [http://www.ethnologue.com/language/duj](http://www.ethnologue.com/language/duj)

<table>
<thead>
<tr>
<th>ISO 639-3</th>
<th>duj</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Alternate Names</strong></td>
<td>Dual, Duala, Wulamba</td>
</tr>
<tr>
<td><strong>Location</strong></td>
<td>Northern Territory, Arnhem Land, Roper river.</td>
</tr>
<tr>
<td><strong>Language Maps</strong></td>
<td>Northern Australia</td>
</tr>
<tr>
<td><strong>Language Status</strong></td>
<td>5 (Developing).</td>
</tr>
<tr>
<td><strong>Classification</strong></td>
<td>Australian, Pama-Nyungan, Yuulngu, Dhuwal</td>
</tr>
<tr>
<td><strong>Dialects</strong></td>
<td>Datiwuy (Daatiwuy), Dhuwal, Dhuwaya, Djapu, Liyagalawumirr, Liyagawumirr, Marrakulu, Marrangu. Similar to Djangu [dhg].</td>
</tr>
<tr>
<td><strong>Language Use</strong></td>
<td>Vigorous. Also use Djambarrpuynu [dir].</td>
</tr>
</tbody>
</table>
http://www.ethnologue.com/language/djr

<table>
<thead>
<tr>
<th>ISO 639-3</th>
<th>djr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alternate Names</td>
<td>Djambarbwingu, Jambapuing, Jambapuingo</td>
</tr>
<tr>
<td>Location</td>
<td>Northern Territory, Elcho island.</td>
</tr>
<tr>
<td>Language Maps</td>
<td>Northern Australia</td>
</tr>
<tr>
<td>Language Status</td>
<td>5 (Developing).</td>
</tr>
<tr>
<td>Classification</td>
<td>Australian, Pama-Nyungan, Yuulngu, Dhuwal</td>
</tr>
<tr>
<td>Language Use</td>
<td>Lingua franca for 2,000 (1990 UBS).</td>
</tr>
</tbody>
</table>
**Djangu**

A language of **Australia**

<table>
<thead>
<tr>
<th>ISO 639-3</th>
<th>dhg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alternate Names</td>
<td>Budalpudal, Burada, Buralbural, Buratha, Dangu, Dhangu, Dhangu’mi, Warameri, Waramiri, Warramiri, War-ramirri, Warumeri</td>
</tr>
<tr>
<td>Location</td>
<td>Northern Territory, Arnhem Land, Elcho island.</td>
</tr>
<tr>
<td>Language Maps</td>
<td>Northern Australia</td>
</tr>
<tr>
<td>Language Status</td>
<td>5 (Developing).</td>
</tr>
<tr>
<td>Classification</td>
<td>Australian, Pama-Nyungan, Yuulngu, Dhangu</td>
</tr>
<tr>
<td>Dialects</td>
<td>Dhangu-Djangu, Gaalpu (Kalbu), Golumala, Ngaymil, Rirratjingu, Wangurri.</td>
</tr>
</tbody>
</table>
Gumatj

A language of **Australia**

<table>
<thead>
<tr>
<th>ISO 639-3</th>
<th>gnn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alternate Names</td>
<td>Gomadj, Gumait, Gumaj</td>
</tr>
<tr>
<td>Population</td>
<td>240 (2006 census)</td>
</tr>
<tr>
<td>Location</td>
<td>Northern Territory, Yirrkala</td>
</tr>
<tr>
<td>Language Maps</td>
<td>Northern Australia</td>
</tr>
<tr>
<td>Language Status</td>
<td>5 (Developing)</td>
</tr>
<tr>
<td>Classification</td>
<td>Australian, Pama-Nyungan, Yuulngu, Dhuwal</td>
</tr>
<tr>
<td>Dialects</td>
<td>Mangalili</td>
</tr>
<tr>
<td>Typology</td>
<td>SOV</td>
</tr>
<tr>
<td>Language Use</td>
<td>Many also use Gupapuyngu [guf] or English [eng]</td>
</tr>
</tbody>
</table>
**Gupapuyngu**

**A language of Australia**

- **ISO 639-3**: guf
- **Alternate Names**: Gobabingo, Gubabwingu
- **Location**: Northern Territory, Arnhem Land, Milingimbi, Elcho Islands.
- **Language Maps**: Northern Australia
- **Language Status**: 5 (Developing).
- **Classification**: Australian, Pama-Nyungan, Yuulngu, Dhuwal
- **Dialects**: Dhuwala, Gupapuyngu, Madarra, Manggalili, Munyuku, Walangu, Wubulkarra. About 45 related dialects. Similar to Gumatj [gnn].

http://www.ethnologue.com/language/guf
Yolngu according to Wikipedia

Varieties

Yolŋu Matha consists of about six mutually intelligible languages divided into about thirty clan varieties and perhaps twelve different dialects, each with its own Yolŋu name. Put together, there are about 4600 speakers of Yolŋu Matha. While there is extensive variation between these dialects, there is generally common mutual intelligibility, hence the umbrella group of Yolngu Matha. The linguistic situation is very complicated, since each of the 30 or so clans also has a named language variety. Dixon (2002) distinguishes the following:[1]

- Dhangu
  - Dhanğu
  - Nhançu/Jarnango (Golpa)
  - Djangu
- Dhuwal
  - Dhuwal (Dhay'yi, Gupapuyngu)
  - Ritharrŋu
- Djinang
  - Djinann
  - Djinba

Bowern (2011) adds the varieties in parentheses as distinct languages.

<table>
<thead>
<tr>
<th>Language</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dhuwaya</td>
<td>100.00% Baby Gumatj (Courtenay &amp; Alpher 1975, Morpphy 1983, Walker 1984, Ganambarr &amp; Sommer 1978), Baby <strong>yolngu</strong> (Walker 1984, Ganambarr &amp; Sommer 1978) [Amery MS 2823:20] [Other names]</td>
</tr>
<tr>
<td>Murngin</td>
<td>100.00% <strong>yolngu</strong>+languages+(NT+SD53) [Thesaurus]</td>
</tr>
<tr>
<td>Ritharrngu</td>
<td>100.00% <strong>yolngu</strong> (Ritharrngu) [Horton name]</td>
</tr>
<tr>
<td>Dhay'yi</td>
<td>100.00% <strong>yolngu</strong> (Daii) [Horton name]</td>
</tr>
<tr>
<td>Dhiyakuy</td>
<td>100.00% <strong>yolngu</strong> (Diakui) [Horton name]</td>
</tr>
<tr>
<td>Djinang</td>
<td>100.00% <strong>yolngu</strong> (Djinang) [Horton name]</td>
</tr>
<tr>
<td>Djinba</td>
<td>100.00% <strong>yolngu</strong> (Djinba) [Horton name]</td>
</tr>
</tbody>
</table>
Yolngu flop!???
Improving processes

• Different parts of ISO639 currently have different registration authorities
  • Part 1 – International Information Centre for Terminology
  • Part 2 – Library of Congress
  • Part 3 – SIL International
  • Part 5 – Library of Congress
  • Part 6 - Geolang

• Moving to a single registration authority would improve consistency of processes
Current developments

- ISO639-5 – language families and groups (2008), currently 114 codes
- ISO639-6 – comprehensive coverage of language variants (2009), four letter codes, number of codes assigned is not clear
- ISO639-4 - Implementation guidelines and general principles for language coding (most recent version 2010)
Current developments

• There are clear problems with parts 5 and 6
• Development of part 6 may have stalled
• Expert input is important for all parts
• But getting part 4 (general principles for language coding) as good as possible is very important
Influencing outcomes

• Representation on ISO Technical Committees is by national standards bodies
• Standards Australia has observer status only
• Discussion of issues happens at level of Working Groups
• Processes are opaque – difficult even to track membership of working groups
ARGILaRe

- Australian Reference Group for Interoperability of Language Resources
- Formed in February 2013
- Any interested people can join
Research Data Alliance

- **RDA:**
  The purpose of the Research Data Alliance is to accelerate international data-driven innovation and discovery by facilitating research data sharing and exchange, use and re-use, standards harmonization, and discoverability.
Research Data Alliance

- RDA (rd-alliance.org):
  The purpose of the Research Data Alliance is to accelerate international data-driven innovation and discovery by facilitating research data sharing and exchange, use and re-use, standards harmonization, and discoverability.

- Improvements in identifying the language of resources is certainly relevant

- A Working Group within RDA is addressing problems for Standardisation of Categories and Codes
Formed in 2012 by research funders from
- EU (500M)
- US (300M)
- AU (23M)

Members in every continent except Antarctica
- But strongly biased towards US, Europe

Plenary 2 took place in Washington DC in September 2013
- [https://www.rd-alliance.org/future-events](https://www.rd-alliance.org/future-events)

Plenary 3 taking place in Dublin in March 2014
- [https://rd-alliance.org/rda-third-plenary-meeting.html](https://rd-alliance.org/rda-third-plenary-meeting.html)
Working Groups and Interest Groups

- **Interest Groups**
  - people concerned with a particular class of problems (by discipline or by kind)
  - may have ongoing existence
  - will probably spin off series of working groups

- **Working Groups**
  - focussed on a particular problem
  - will run for 12-18 months
  - will produce a piece of infrastructure (broadly interpreted) for deployment
  - should lead to more data being exchanged
Mirror committee

- Standards Australia allows for participation: “in the work of international Technical Committees via a national mirror committee”
- The RDA Working Group is exploring the possibility of an Australian Mirror Committee for ISO TC 37
- A case for Net Benefit has to be made
  - This will require collaboration outside of research communities
Conclusion

• ISO639 has flaws but it is not going away
  • Would there ever be a right time to standardise?
• Efforts should be devoted to incremental improvement
  • Submit change requests
  • Join in efforts of groups like ARGILaRe and RDA
  • Find ways of collaborating with:
    • Various research communities
    • Interested parties outside academia