Using rule-based computational linguistics for Australian languages: Electronic resources for Murrinh-Patha

Melanie Seiss & Rachel Nordlinger

University of Konstanz & Melbourne University

7th European Australianists Workshop 2012

03./04.04.2012
How computational linguistics and Australian languages can profit from each other

**Profits for Australian languages**

- another way of describing and conserving a language
- applications useful for, e.g.:
  - promoting literacy among language speakers (e.g. Arrernte Footy, Lareau et al. 2011)
  - promoting language skills for language learners

**Profits for computational linguistics**

- test existing tools and methods on different language types
- test tools and methods on languages without other resources (corpora etc.)
- develop new tools and methods
Computational tools and methods

- CL strives to develop language-independent tools and methods
- problem: even tools/methods which claim to be language-independent often assume a specific language structure (Bender 2009)
- reality: main focus on English (fixed word order, few verb forms, little inflection etc.)
- even German problematic for some (statistical) methods (Rehbein and van Genabith 2009)

But:

- rule-based tools and methods can handle complex linguistic data, e.g.
  - Grammar Implementation Plattform XLE (based on LFG)
  - XFST Morphology
A Brief Introduction to Murrinh-Patha

Murrinh-Patha

- polysynthetic
- non-Pama-Nyungan
- Southern Daly subgroup, together with Ngan’gityemerri (Green 2003)
- spoken in and around Wadeye, NT by approx. 2500 speakers
- lingua franca of region (still acquired by children, still spoken in everyday life)
What makes Murrinh-Patha so difficult?

- bipartite verbal structure
- complicated verbal template
- complex number system
- morphophonemics
Murrinh-Patha bipartite verbs

- MP verbs (mostly) consist of a classifier stem and a lexical stem

(1) *manganta*
    
mangan - rta
    3sgS.SNATCH(9).nFut - hug
    ‘He/she hugged him/her.’

- classifier stem: inflected for subject person, number, tense
  (portmanteau forms, over 50 forms per paradigm); ‘rather general’ meaning

- lexical stem: uninflecting, ‘more specific’ meaning
The Murrinh-Patha verbal template & dependencies

<table>
<thead>
<tr>
<th>Class.</th>
<th>RR</th>
<th>SubjN/Obj</th>
<th>RR</th>
<th>IBP</th>
<th>Lex</th>
<th>TNS</th>
<th>Adv/Prt</th>
<th>SubjN/ObjN</th>
<th>Adv/Prt</th>
</tr>
</thead>
</table>

Class: classifier stem, marked for tense, aspect & subject number
SubjN: subject number markers for dual & paucal subject
Obj: object agreement marker
ObjN: object number marker for dual & paucal
RR: reflexive / reciprocal marker
IBP: incorporated body part
Lex: lexical stem
TNS: tense marker
Adv: Adverbial
Prt: Particle

(adapted from Blythe 2009)
Murrinh-Patha Number System

Complex number system for subject and object:

- singular (sg), dual (du), paucal (pauc), plural (plural)
- sibling vs. non-sibling (in dual and paucal only)
- gender: female (fem) vs. male (in dual and paucal only)
marked by a combination of the classifier stem and separate morphemes (Nordlinger 2010a)

**Example:** ‘They saw it.’

(2a) $\text{Bam-nginxha-ngkardyu}$ ‘They 2 fem non-sib’
(2b) $\text{Bam-ninthha-ngkardyu}$ ‘They 2 male non-sib’
(2c) $\text{Pubamka-ngkardyu}$ ‘They 2 sibling’
(2d) $\text{Pubamka-ngkardyu-ngime}$ ‘They paucal fem non-sib’
(2e) $\text{Pubamka-ngkardyu-neme}$ ‘They paucal male non-sib’
(2f) $\text{Pubamkardu (Pubam-ngkardyu)}$ ‘They plural, they paucal sib’
Object Marking

- Direct and indirect object marking on the verb
- same categories as for subject marking
- discontinuous object markers for non-sibling categories

(3a) *Bam-ngi-ngkardu* ‘He/she saw me.’
(3b) *Bam-nganku-ngkardu-ngintha* ‘He/she saw us (2 fem non-sib).’
(3c) *Bam-nganku-ngkardu* ‘He/she saw us (2 sibling).’
(3d) *Bam-nganku-ngkardu-ngime* ‘He/she saw us (paucal fem non-sib).’
(3e) *Bam-pun-ngkardu* ‘He/she saw us (plural/paucal sib).’
Competing subject and object markers:

- Subject number and object marker compete for same slots
- Object marker has priority

(5a)  
\[ \text{bam} -\text{ngintha} -\text{ngkardu} \]
3sgS.SEE(13).nFut -du.f -see
‘They two (non-sibl) saw him/her.

(5b)  
\[ \text{bam} -\text{ngi} -\text{ngkardu} -\text{ngintha} \]
3sgS.SEE(13).nFut -1sgDO -see -du.f
‘They two (non-sibl, female) saw me.

- Many more interdependencies between morphemes, see Nordlinger (2010b) for an overview.
Morphophonemtics

- surface form is often different from the component parts:

(5a) mam-watha $\rightarrow$ mampatha
(5b) bam-ngkardu $\rightarrow$ bamkardu
(5c) mangan-rta $\rightarrow$ manganta
Electronic Resources for Murrinh-Patha

- **Electronic Dictionary:**
  - can process Murrinh-Patha words and phrases
  - decomposes the input for the user and looks up the meaning parts

- **Translation System:**
  - takes English input and generates Murrinh-Patha output
  - can be used to translate simple sentences
  - is especially intended to learn about the Murrinh-Patha verb form and its number system
Components of the Implementation
Resources used for the lexicon

- Lexical items
  - automatically extracted from Joe Blythe’s Toolbox dictionary and Street (1989)
  - Additional vocabulary from Walsh (1987) and fieldnotes from Rachel Nordlinger
Morphology

- Finite State morphology built with XFST (Beesley & Karrtunen 2003)
- 2 level morphology:
  bam+class13+3P+sg+3sgDO+ngkardu+LS : bamngkardu
- Inbuilt mechanisms to model the long distance dependencies between morphemes (e.g., subject number markers)
- Allows for modeling of morphophonemic processes, e.g.
  \[ n g k \rightarrow k || m_-, n_- \]
  → bam+class13+3P+sg+3sgDO+ngkardu+LS : bamkardu
XLE Grammars

- XLE Parser developed at PARC (Crouch et al. 2011, Butt et al. 1999)
- implementation based on Lexical-Functional Grammar formalism
- used by the ParGram-Group for large-scale grammar implementation: English, German, French, Norwegian, Japanese, Urdu, Hungarian, Indonesian ...
XLE Grammar Output

"bamkardu kardu kigay"

```
CS 2:   ROOT
     
      S
       
       VP  NP
          |
          |
          V   NC   N

bamkardu  kardu  kigay
```

- c(onstituent)-structure
- f(unctional)-structure
XFR Rewrite Rules for Translation

\[ \text{PRED}(%V, \text{see}) \Rightarrow \text{PRED}(%V, \text{ngkardu}). \]

\[ \text{PRED}(%V, \text{boy}) \Rightarrow \text{PRED}(%V, \text{kigay}), \text{NOMCLASS}(%V, \text{kardu}). \]
Architecture of Translation system

Disambiguation Module:
- checks if the f-structure of the English input has a plural subject or object
- If no plural is present, redirects to the transfer rules
- If a plural is present, prompts the user to give more information
Disambiguation Module (simplified)
Translation Example

Which English sentence do you want to translate?

They saw him.

Is the subject a group of 2, a small group (ca. 3-10) or a bigger group (over ca. 10)?

◇ a group of two ◇ a small group (ca. 3-10) ◇ a big group (over ca. 10)

Are they siblings?

◇ yes ◇ no

What are siblings?

Are they male or female or mixed?

◇ male ◇ female / mixed

Translation Result:

English: They saw him.
Murrinh-Patha: Bamninthangkardu.
Translation Example 2

Which English sentence do you want to translate?

They saw me.

Translate

Is the subject a group of 2, a small group (ca. 3-10) or a bigger group (over ca. 10)?

- a group of two
- a small group (ca. 3-10)
- a big group (over ca. 10)

Are they siblings?

- yes
- no

What are siblings?

Are they male or female or mixed?

- male
- female / mixed

Translation Result:

English: They saw me.
Murrinh-Patha: Bamningkardunintha.
Example:
Input: *bamkardu kardu kigay*

Tries parsing:
1) NP: kardu kigay bamkardu
2) kardu kigay bamkardu

Only 2) gives grammatical output

script extracts information

Lookup in dictionary:
*ngkardu* + classifier 13: ‘to see’
*kigay* + NC *kardu*: ‘teenage boy’
Dictionary Example

Type in a Murrinh-Patha word or phrase:

**bamkardu**

Result for 'bamkardu':

Verb: 'ngkardu' + Classifier Stem 13:

'to see, to look'

Examples:

bangkardu mani 'I'll have a look' pipi nginarr nhinhi damkardu thurran? 'do you look at your wife's mother?'

More Info
Dictionary Example 2

Type in a Murrinh-Patha word or phrase:

```
kardu kigay bamkardu
```

Search

Result for ‘kardu kigay bamkardu’:

Verb: ‘ngkardu’ + Classifier Stem 13:

‘to see, to look’

Examples:

bangkardu mani ‘i’ll have a look’ pipi nginarr nhinhi damkardu thurran? ‘do you look at your wife’s mother?’

Noun: ‘kigay’ + Noun Classifier ‘kardu’:

‘teenage boy’
Both, the Translation System and the Dictionary offer more information to the user after the initial output:

- morphological analysis
- show form with different tense, subject and object number information
- show various paradigms (keeping other information stable):
  - show form in all tenses
  - show form with all subject numbers
  - show form with all object numbers

→ Can be used to study structure, detect patterns, etc.
More Information Example

More information for 'bamngingkardunintha':

Get dictionary entry and examples

Verb: 'ngkardu' + Classifier Stem 13:
'to see, to look'

Examples:

bangkardu mani 'i'll have a look' pipi ngingarr nhinhi damkardu
thurran? 'do you look at your wife's mother?'

Get morphological information

Morphology for bamngingkardunintha:

<table>
<thead>
<tr>
<th>bam</th>
<th>Classifier 13, Subject Information: 1. or 3. Person sg (I or he/she/it), Tense: non-Future</th>
</tr>
</thead>
<tbody>
<tr>
<td>ngi</td>
<td>Direct Object Information: 1. Person singular (me)</td>
</tr>
<tr>
<td>ngkardu</td>
<td>Lexical stem</td>
</tr>
<tr>
<td>nintha</td>
<td>Subject: two male non-Siblings (overwrites info from classifier)</td>
</tr>
</tbody>
</table>

Show form with all possible tenses

Show form with all possible subject numbers
Future Work

- build web-based application (so far perl tk interface)
- broaden coverage
- more fine-grained feedback
- add sound files (difficult)
- build applications for Murrinh-Patha speakers learning English using the same underlying Murrinh-Patha grammar
References


