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

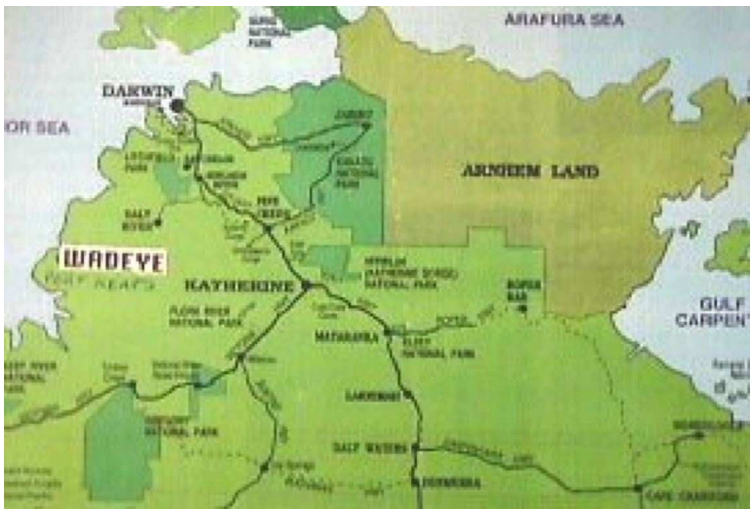
Australian languages

- 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

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

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 every day 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; 'rather general' meaning
- Lexical stem: uninflecting, 'more specific' meaning

Inflection on classifier stems

Encoded in portmanteau forms:

- subject person/number marked on classifier stem
- 4-way number contrast: singular, 1. inclusive, dual, plural
- 3-way person contrast
- 5-way tense/aspect contrast: non-Future (nFut), Past Imperfective (PImpf), Future (Fut), Future Irrealis (FutIrr), Past Irrealis (PstIrr)

$\Rightarrow$ more than 50 forms per paradigm

The Murrinh-Patha verbal template & dependencies

1	2	3	4	5	6	7	8	9	10
Class.	RR	SubjN/ Obj	RR	IBP	Lex	TNS	Adv/Prt	SubjN/ ObjN	Adv/Prt

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)

The Murrinh-Patha verbal template & dependencies

1	2	3	4	5	6	7	8	9	10
Class.	RR	SubjN/ Obj	RR	IBP	Lex	TNS	Adv/Prt	SubjN/ ObjN	Adv/Prt

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)

Subject Number

- marked by a combination of the classifier stem and separate morphemes (Nordlinger 2010a)

Example: *'They saw it.'*

(2a) *Bam-ngintha-ngkardu* 'They 2 fem non-sib'

(2b) *Bam-nintha-ngkardu* 'They 2 male non-sib'

(2c) *Pubamka-ngkardu* 'They 2 sibling'

(2d) *Pubamka-ngkardu-ngime* 'They paucal fem non-sib'

(2e) *Pubamka-ngkardu-neme* 'They paucal male non-sib'

(2f) *Pubamkardu (Pubam-ngkardu)* 'They plural, they paucal sib'

Object Marking

- Direct and indirect object marking on the verb
- same categories as for subject marking
- discontinuousobject 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) *bam* *-ngintha* *-ngkardu*
3sgS.SEE(13).nFut -du.f -see
'They two (non-sibl) saw him/her.

(5b) *bam* *-ngi* *-ngkardu* *-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.

Morphophonemics

- Surface form is often different from the component parts:

(5a) mam-watha → mampatha

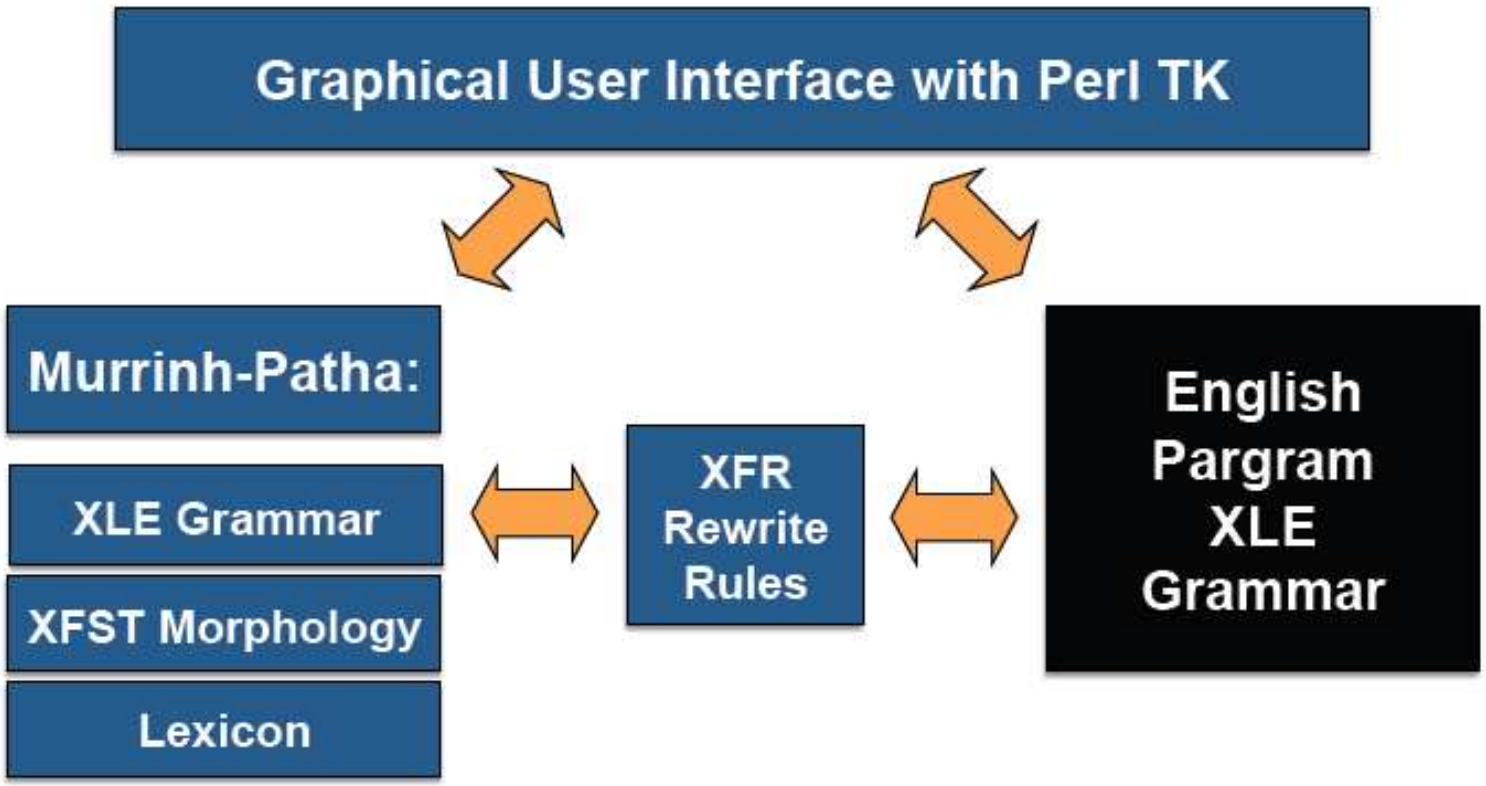
(5b) bam-ngkardu → bamkardu

(5c) mangan-rta → manganta

Electronic Resources for Murrinh-Patha

- Electronic Dictionary:
  - The electronic dictionary can process Murrinh-Patha words and phrases.
  - It decomposes the input for the user and looks up the meaning parts.
- Translation System:
  - The translation system takes English input and generates Murrinh-Patha output.
  - It can be used to translate simple sentences.
  - It is especially intended to learn about the Murrinh-Patha verb form and its number system.

Components of the Implementation



Resources used for the lexicon

- entries automatically extracted from Street (1989)
- Additional vocabulary from Walsh (1987), fieldnotes from Joe Blythe and Rachel Nordlinger

- Used as entries in
  - XFST Morphology
  - XLE Grammar (verbs only)
  - XFR Rewrite Rules (translation)
  - Dictionary entry (definitions & examples)

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., discontinuous object markers)
- allows for modeling of morphophonemic processes, e.g.

  [n g k $\rightarrow$ k || m _ , n _]

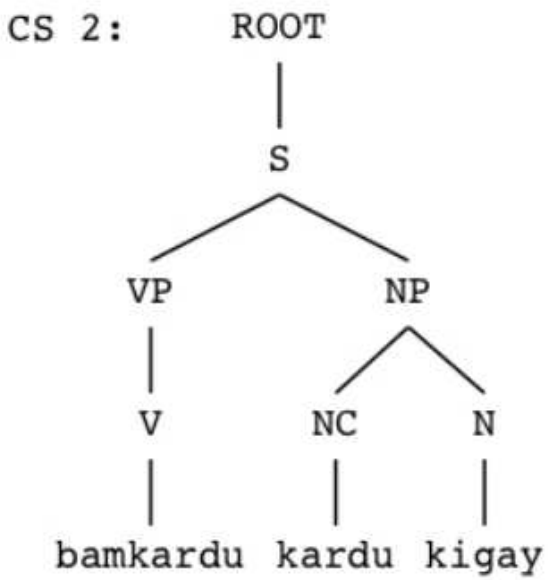
  $\rightarrow$ bam+class13+3P+sg+3sgDO+ngkardu+LS : bamkardu

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, ...

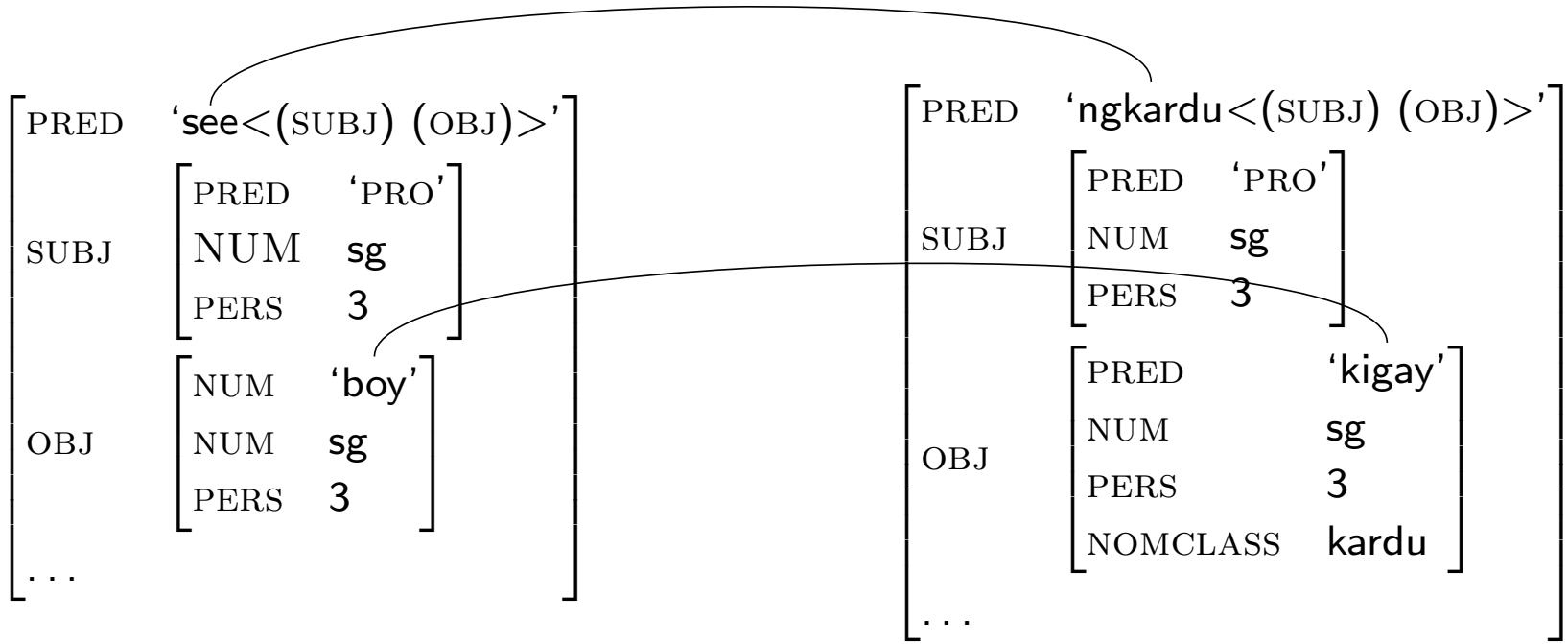
XLE Grammar Output

"bamkardu kardu kigay"




```
[PRED   'ngkardu<[1-SUBJ:PRO], [35:kigay]>'
SUBJ    [PRED 'PRO'
         NUM sg, PERS 3]
OBJ  35 [PRED 'kigay'
         NUM sg, NomClass kardu, PERS 3]
CHECK   [_CS 13, _DO pres]
TAM     [TENSE non-fut]
1 VTYPE complex_pred
```

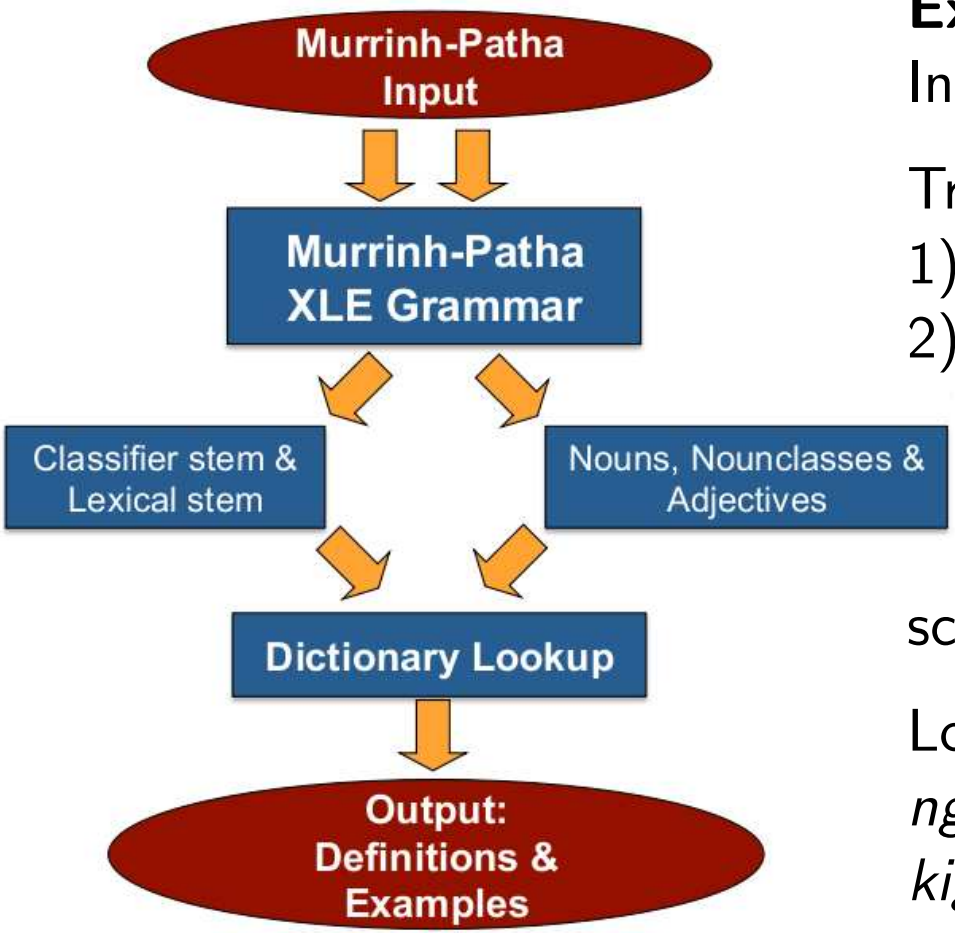
XFR Rewrite Rules for Translation



PRED(%V, see)
⇒
PRED(%V, ngkardu).

PRED(%V, boy)
⇒
PRED(%V, kigay),
NOMCLASS(%V, kardu).

Electronic Dictionary



Example:

Input: *bamkardu kardu kigay*

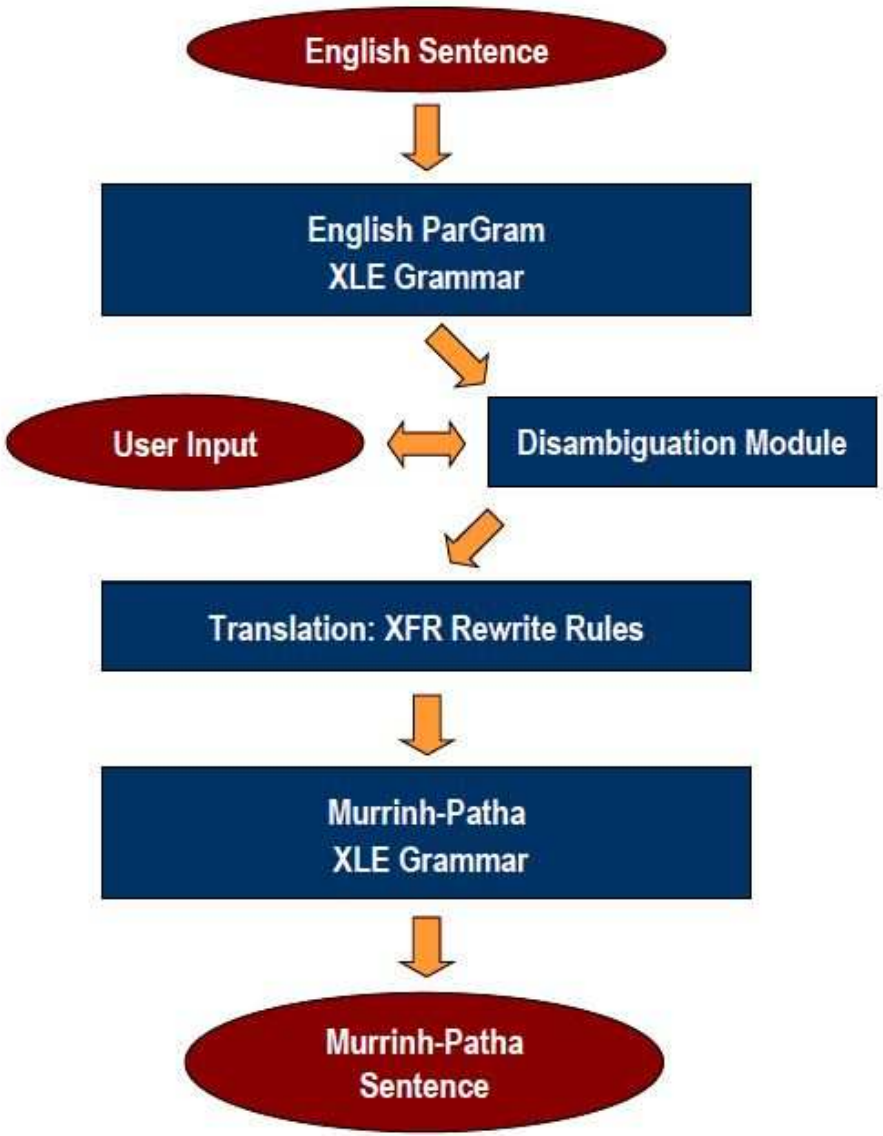
Tries parsing:
1) NP: kardu kigay bamkardu
2) kardu kigay bamkardu

Only 2) gives grammatical output

script extracts information

Lookup in dictionary:
ngkardu+classifier 13: 'to see'
kigay + nounclass *kardu*: 'teenage boy'

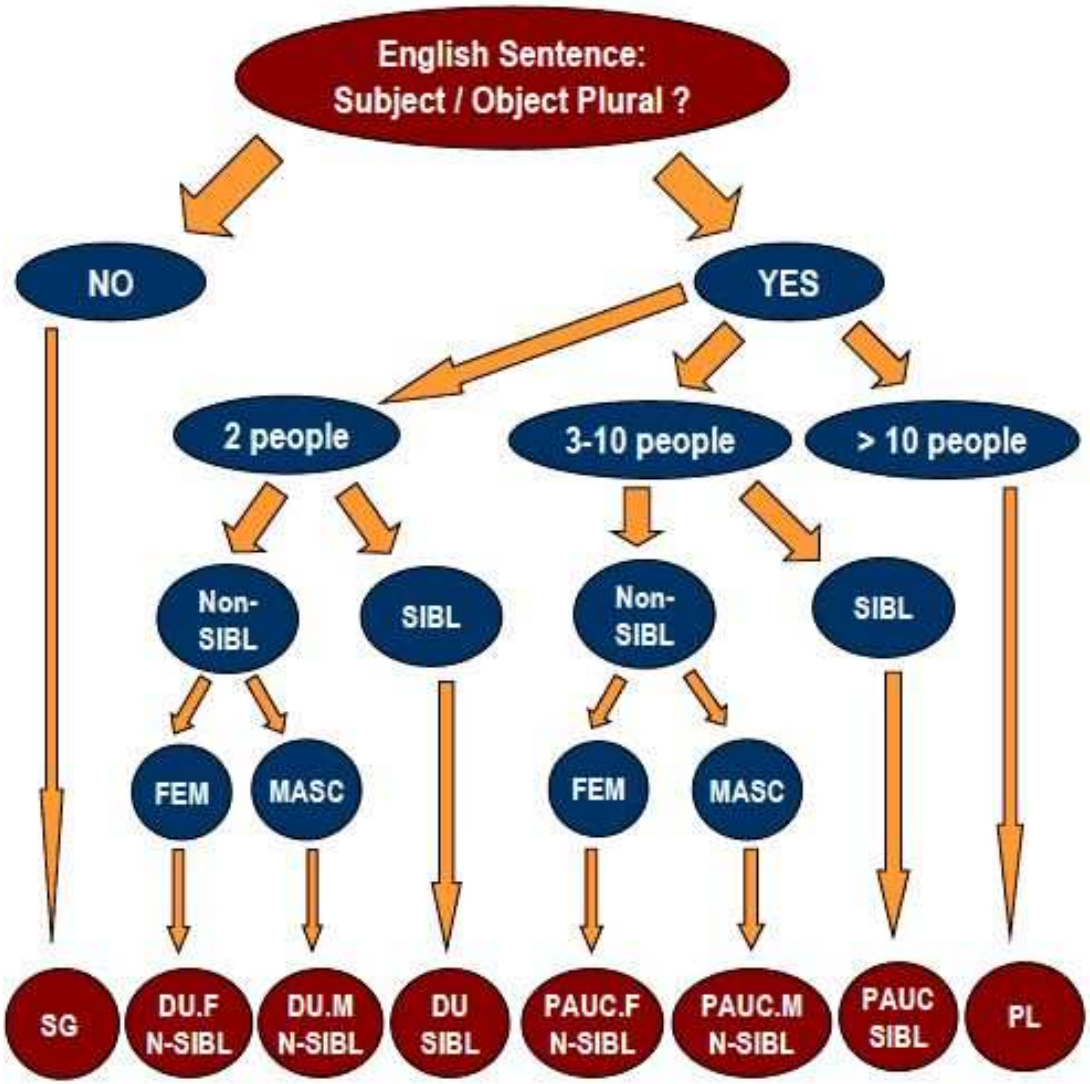
Architecture of Translation system



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, promts the user to give more information

Disambiguation Module



More Information

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

$\longrightarrow$ Can be used to study structure, detect patterns, etc.

Future Work

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

References

Beesley, Kenneth R. and Karttunen, Lauri. 2003. *Finite State Morphology*. Stanford: CSLI Publications.

Blythe, Joe. 2009. Doing referring in Murriny Patha conversation. PhD thesis. University of Sydney.

Crouch, Dick, Mary Dalrymple, Ronald Kaplan, Tracy Holloway King, John T. Maxwell III and Paula Newman. 2011. XLE Documentation. Palo Alto Research Center.

Green, Ian. 2003. The genetic status of Murrinh-patha. In N. Evans, ed., *The Non-Pama- Nyungan Languages of Northern Australia*, page 125158. Canberra: Pacific Linguistics.

Lareau, François, Marc Dras, Benjamin Börschinger and Robert Dale. 2011. Collocations in Multilingual Natural Language Generation: Lexical Functions meet Lexical Functional Grammar. *Proceedings of the Australasian Language Technology Association Workshop (ALTA 2011)*, 95-104. Canberra, Australia.

Nordlinger, Rachel. 2010a. Agreement Mismatches in Murrinh-Patha Serial Verbs. In Yvonne Treis and Rik De Busser (eds). *Selected Papers from the 2009 Conference of the Australian Linguistic Society*.

Nordlinger, Rachel. 2010b. Verbal morphology in Murrinh-Patha: evidence for templates. *Morphology* 20(2).

Street, Chester. 1989. Murrinh-Patha vocabulary. Electronic MSWord file.