A Morphological Guesser for a Morphologically Rich Language
Melanie Seiss, Universität Konstanz

Introduction
- Murrinh-Patha (MP) is a polysynthetic non-Pama-Nyungan language spoken by approx. 3000 people in Wadeye, Northern Territory, Australia.
- Overall aim: High coverage morphological analyzer for MP
- Aim here: a morphological guesser to detect new lexical items
- Challenge: MP verbal template very complex:
  - 9 different morpheme slots with complex interdependencies
  - High degree of syncretism for morphemes
  - Constraints needed to restrict combinatorial possibilities
- Idea: morphological analyzer + guesser which relaxes constraints stepwise

Challenges posed by MP verbs
- Verbal Template:
  - Phonological rules apply on morpheme borders:
    1. *bam – ngkardu → bamkardu
  - Syncretism: mainly for classifier stem (CS) forms, e.g.
    2. *bam
    1sgS.CS(13).nFut / 3sgS.CS(13).nFut
    1sgS.CS(18).nFut / 3sgS.CS(18).nFut
  - Classifier stems and lexical stems (LS) determine semantic meaning of verb together, but not every lexical stem can combine with every classifier stem and vice versa:
    3. *bam-ngkardu
    1sgS.CS(13).nFut-see / 3sgS.CS(13).nFut-see
    *1sgS.CS(18).nFut-see / *3sgS.CS(18).nFut-see
  - Interdependency for subject & object markers (1./2. person):
    4. *bam-ngkardu
    1sgS.CS(13).nFut-see / 3sgS.CS(13).nFut-see
    *1sgS.CS(18).nFut-see / *3sgS.CS(18).nFut-see
  - Interdependency for classifier stems and tense markers
    5. a. be-ngkardu-dha
    1sgS.CS(13).PImpf-see-PImpf
    b. ba-ngkardu-nu
    1sgS.CS(13).Fut-see-Fut
  - For even more complex dependencies see Nordlinger (2010), for their computational implementation Seiss (2011)

Morphological Guesser
- Morphological guesser proposed by Beesley & Karttunen (2003) for open class items to make morphological analysis more robust and to detect new lexical items
- Not possible for MP nouns as they are usually uninflected
- Morphological guesser built for MP lexical stems
- Problem of guesser for lexical stem:
  - High number of guesses due to phonological rules
  - If dependencies are not modeled in guesser, there are also too many guesses for each lexical item.
  - If dependencies are modeled, it might not find many nonstandard uses.
  - Refined lookup strategy needed

Strategy for Morphological Analyzer
- All forms of 38 CS
- 898 LS (1732 LS + CS comb.)
- 1140 Nouns
- 72 Borrow
- 175 Adj
- 41 Adv
- 55 Interj...

1. Step: Full Constraints
2. Step: relaxed constraint on CS & LS
3. Step: Guesser based on relaxed constraints on CS & LS
4. Step: relaxed constraint on tense markers
5. Step: Guesser based on relaxed constraints on tense markers

Evaluation
- tested on a corpus of bible translations
- Development corpus: 10 032 words / 1516 types
- Test corpus: 12 316 words / 1871 types
- Negative Testing:

<table>
<thead>
<tr>
<th>Types – development</th>
<th>1.</th>
<th>2.</th>
<th>3.</th>
<th>4.</th>
<th>n.f.</th>
</tr>
</thead>
<tbody>
<tr>
<td>97.8</td>
<td>71.6</td>
<td>3.2</td>
<td>20.3</td>
<td>0.7</td>
<td>1.9</td>
</tr>
<tr>
<td>98.4</td>
<td>87.9</td>
<td>1.0</td>
<td>5.6</td>
<td>1.8</td>
<td>2.0</td>
</tr>
<tr>
<td>94.1</td>
<td>58.4</td>
<td>6.1</td>
<td>24.5</td>
<td>2.0</td>
<td>3.0</td>
</tr>
<tr>
<td>97.9</td>
<td>85.9</td>
<td>1.2</td>
<td>6.3</td>
<td>2.1</td>
<td>2.4</td>
</tr>
</tbody>
</table>

- Newly found candidates for lexical + classifier stem combinations: 113 (109 from Strategy 2)
- Still high number of possible analyses per type (in average 10.3 analyses per type)
- Syncretisms make positive testing difficult

Conclusion
- Strategy allows for reliable detection of new lexical stems and new combinations of classifier and lexical stems
- Heuristic to extract most probable guess from possibilities needed
- Stepwise guesser also built in electronic dictionary for more useful feedback

Acknowledgements & References
Many thanks go to Rachel Nordlinger and Joe Blythe who provided me with data and information about the language. For the lexicon of the morphological analyzer, Joe Blythe’s toolbox dictionary has been automatically extracted.