Hunting the Snark

The problem posed for MT by non-concatentative morphologies

They sought it with thimbles,
They sought it with care,
They pursued it with forks and hope,
The threatened its life with a railway share,
They cajoled it with smiles and soap.
A common question

● Why not use Google Translate?
  ● 65 languages
  ● saves time & money
  ● etc...
Genesis 1.1

- Ancient Greek - LXX
  - ἐν ἀρχῇ ἐποίησεν ὁ θεὸς τὸν οὐρανὸν καὶ τὴν γῆν. ἡ δὲ γῆ ἦν ἀόρατος καὶ ἀκατασκεύαστος καὶ σκότος ἐπάνω τῆς ἀβύσσου καὶ πνεῦμα θεοῦ ἐπεφέρετο ἐπάνω τοῦ ὕδατος.

- English...
  - in the beginning epoίisen God heavens and the earth. And the earth unseen and HN akataskeúastos and Scott epáno avússou and Spirit of God epeféreto epáno waters.
One or two problems

- **Fixable**
  - better Gk lexicon
  - and grammar
  - i.e. put in more knowledge

- **Harder to fix**
  - Target language
    - any of 7,000 +
    - no lexicon
    - no grammar
    - different language tomorrow...
ParaTExt

- A translator’s workbench (language independent)
- Editing environment, Base texts, Model texts
Checking & Review

- Key term list
  - Automatic analysis of text for consistency
- Semi-automatic morphology analysis
  - Spelling checks
Checking & Review

- Automatic interlinear back-translation
Spelling review

- Word list tool
  - parsed by
    - morphology
    - syllable
Beyond translation

- Literacy - Dictionaries & Concordances
Glossing Technologies

• Provide
  • Language independent
  • Lemmatisation
  • Morphology analysis
    – driven by glossing

• Problems
  • Orthography
    – spaces...
  • Complex morphologies
Complex Morphologies 1

- Concatenative
  - >75% of languages
  - e.g. Bantu languages
    - Swahili
    - verb -pend-

- Word Form Template:
  - [Pre]Stem[Suff]

- akipenda, anakupenda, atanipenda, mlipenda, mpende, nakupenda, nawapenda, nilipenda, ninakupenda, [-]pendana, [-]pendea, [-]pendwa, sikupendi, tulipenda, tutapenda, ulipenda, ungependa, utapenda, walipenda, wanaupenda, watapenda
Complex Morphologies 2

• Non-concatenative
  • <25%
  • e.g. Semitic languages
    – Amharic, Arabic, Hebrew, Syriac

• Template:
  – \([m][m][m][m]\)
    • $ = \text{stem}
    • m = \text{morphs}
Finding morpheme structures

• Premise:
  • Valid morpheme structures will occur in a text with statistically significant frequency

• Three ways to find morphemes
  • Statistically
  • Minimum Description Length
  • Paradigm analysis
    – i.e. by stem association
Paradigm Analysis

- Concatenative
  - Find possible morphs
    - examine initial and final n-grams
  - Validate
    - build inflection paradigms

- Non-concatenative
  - Find possible morphs
    - ?
  - Validate
    - build inflection paradigms
Finding morph templates

• Pre-requisites
  • a lexicon of surface forms in the target language

• Method
  • compare each form in the lexicon with every other form and note common sequences.
Hebrew - Rendering

- UTF-8
  - Difficult to render zero width glyphs
- Michigan-Claremont Encoding
  - 7-bit ASCII
  - Easier to render morph templates
    - Remove cantillation

Example MC Encoding:

UTF-8           MC
With Cantillation & Vowels:
ברלמסיה        ב.:/R")$I73YT
Without Cantillation:
ברלמסיה        ב.:/R")$IYT
Consonants alone:
ברלמסיה        BR)$YT

fig. 5.
Example 1 - QF+AL / MFLA+

- Four matched characters: MF+AL
- Rule:
  - successors must follow below and to the right
- Three solutions:
  - F+, FA, FL

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Q</td>
<td>F</td>
<td>+</td>
<td>A</td>
<td>L</td>
</tr>
<tr>
<td>1</td>
<td>M</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>F</td>
<td>F</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>L</td>
<td></td>
<td></td>
<td>L</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>A</td>
<td></td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>+</td>
<td></td>
<td>+</td>
<td></td>
</tr>
</tbody>
</table>

fig. 1 match matrix for QF+AL / MFLA+
Assessing the solutions

- Match coordinates:
  - $F(2,2)$
  - $+(3,5)$
  - $A(4,4)$
  - $L(5,3)$

- **S1**
  - \{ $F(2,2)$, $+(3,5)$ \}

- **S2**
  - \{ $F(2,2)$, $A(4,4)$ \}

- **S3**
  - \{ $F(2,2)$, $L(5,3)$ \}

---

```
<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\varnothing$</td>
<td>$Q$</td>
<td>$F$</td>
<td>$+$</td>
<td>$A$</td>
</tr>
<tr>
<td>2</td>
<td>$M$</td>
<td>$F$</td>
<td>$F$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>$L$</td>
<td></td>
<td></td>
<td>$L$</td>
</tr>
<tr>
<td>4</td>
<td>$A$</td>
<td></td>
<td></td>
<td></td>
<td>$A$</td>
</tr>
<tr>
<td>5</td>
<td>$+$</td>
<td></td>
<td></td>
<td>$+$</td>
<td></td>
</tr>
</tbody>
</table>
```

fig. 1 match matrix for QF+AL / MFLA+
Solution Value (V)

- \[ V = 1 + \left(1 - \frac{d}{f}\right) \]
- where
  - \( d \) = distance between the two x or y coordinates, whichever is the greater.
  - \( f = 10 \) (distance beyond which it is unlikely the two items are related)

- S1: F(2,2), +(3,5)
  - \( V = 1.7 \)

- S2: F(2,2), A(4,4)
  - \( V = 1.8 \)

- S3: F(2,2), L(5,3)
  - \( V = 1.7 \)
Example 2 - YIM:LO+ / YIQ:+OL

- Matched items:
  - Y (1,1)
  - I (2,2)
  - : (4,4)
  - L (5,7)
  - O (6,6)
  - + (7,5)

fig. 2 Match matrix for yiq:tol / yim:lot
Example 2 - Solutions

- S1 \{ Y(1,1), I(2,2), :,(4,4), +(7,5) \}  
  \[ 1.9, 1.8, 1.7 = 5.814 \]
- S2 \{ Y(1,1), I(2,2), :,(4,4), O(6,6) \}  
  \[ 1.9, 1.8, 1.8 = 6.156 \]
- S3 \{ Y(1,1), I(2,2), :,(4,4), L(5,7) \}  
  \[ 1.9, 1.8, 1.7 = 5.814 \]
Example 3 - YIQ:+:LW. / YIM:L:+W.

- Y(1,1)
- I (2.2)
- : (4,4)
- : (4,6)
- L(5,7)
- : (6,4)
- : (6,6)
- +(7,5)
- W.(8,8)

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ø</td>
<td></td>
<td></td>
<td></td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td>Y</td>
</tr>
<tr>
<td>Y</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td></td>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>:</td>
<td></td>
<td></td>
<td>:</td>
<td></td>
<td></td>
<td>:</td>
<td></td>
</tr>
<tr>
<td>:</td>
<td></td>
<td></td>
<td>:</td>
<td></td>
<td></td>
<td>:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>:</td>
<td></td>
<td>:</td>
<td>:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>:</td>
<td></td>
<td>:</td>
<td>:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>L</td>
</tr>
<tr>
<td>L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>L</td>
<td></td>
</tr>
<tr>
<td>W.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>W.</td>
</tr>
</tbody>
</table>

*fig 3. Match matrix for YIQ:+LW / YIM:L:+W*
E.g. 3 Solutions

- **S1**\{Y(1,1), I(2,2), :(6,4), L(7,5), w.(8,8)}
  - $1.9 \times 1.6 \times 1.9 \times 1.7 = 09.8192$

- **S2**\{Y(1,1), I(2,2), :(4,4), L(7,5), w.(8,8)}
  - $1.9 \times 1.8 \times 1.7 \times 1.7 = 09.8838$

- **S3**\{Y(1,1), I(2,2), :(4,4), :(6,6), w.(8,8)}
  - $1.9 \times 1.8 \times 1.8 \times 1.8 = 11.0808$

- **S4**\{Y(1,1), I(2,2), :(4,4), +(5,7), w.(8,8)}
  - $1.9 \times 1.8 \times 1.7 \times 1.7 = 09.8838$

- **S5**\{Y(1,1), I(2,2), :(4,6), +(5,7), w.(8,8)}
  - $1.9 \times 1.6 \times 1.9 \times 1.7 = 09.8192$
Solution complements

- **Eg. 1** QF+AL / MFLA+
  - S1{F(2,2), +(3,5)} Q--AL, M_LA--
  - S2{F(2,2), A(4,4)} Q_+L, M_L+_
  - S3{F(2,2), L(5,3)} Q+_A, M_L__

- **Eg. 2** YIQ:+OL / YIM:LO+
  - S1 { Y(1,1), I(2,2), :(4,4), +(7,5) }
    __Q__OL, __M_LO_`
  - S2 { Y(1,1), I(2,2), :(4,4), O(6,6) }
    __Q_+L, __M_L_+
  - S3 { Y(1,1), I(2,2), :(4,4), L(5,7) }
    __Q_+O, __M__O+

- **Eg. 3** YIQ:+LW. / YIM:L:+W.
  - S1{Y(1,1), I(2,2), :(6,4), L(7,5), W.(8,8)}
    __Q_:__, __M__:+
  - S2{Y(1,1), I(2,2), :(4,4), L(7,5), W.(8,8)}
    __Q_:__, __M__:+
  - S3{Y(1,1), I(2,2), :(4,4), :(6,6), W.(8,8)}
    __Q_:__, __M__:+
  - S4{Y(1,1), I(2,2), :(4,4), +(5,7), W.(8,8)}
    __Q_:L, __M:L:+
  - S5{Y(1,1), I(2,2), :(4,6), +(5,7), W.(8,8)}
    __Q_:L, __M:L___
Hebrew Results

- Lexicon
  - Hebrew forms in Genesis – 4,431

- Templates Generated
  - 52,357

- Best 1% of templates
  - Complements
    - 50 stems,
    - 42 valid

- Build inflection paradigms...

work in progress...
Jon Riding

Linguistic Computing at British & Foreign Bible Society

jon.riding@biblesociety.org.uk
http://lc.bfbs.org.uk