Automatic alignment of hieroglyphic and transliteration

Mark-Jan Nederhof

2008-07-09

Context of this work: AELalign

- Originally motivated by learning and teaching.
- In general: distributed effort.
- Automatically combine resources from different creators.
- Resources are:
  - hieroglyphic,
  - transliteration,
  - translation,
  - lexical annotation.

Software and corpus

In operation since 2000:

`http://www.cs.st-andrews.ac.uk/~mjn/egyptian/texts/`

- XML format 'AELalign' presented at **I&E 2002** (Pisa).
- Abandoned MdC, and introduced RES encoding of hieroglyphic (also presented at **I&E 2002**).
- Progress reported at **I&E 2006** (Oxford).

Progress since 2006

- Visual editor for hieroglyphic (RES).
- Leverhulme grant; generalise beyond Ancient Egyptian.
- **This talk:** initial experiments with automatic alignment of hieroglyphic and transliteration.

Comparisons

Relative to some other annotated corpora:

- corpus of project contains few texts as yet;
- lexical annotation is very limited.

Present project:

- Considered resources are means, not end.
- Main research objective: investigate distributed creation of resources.

Alignment

Consider an ancient text.

- One resource is sequence of hieroglyphs.
- Another resource is list of words of transliteration.
- We want to display them together on the screen (interlinear representation).

Central question

How easy is it to match each transliteration word to corresponding sequence of hieroglyphs?

Problems to overcome

- Signs can have several meanings.
- E.g. match two consecutive signs to one word, or match them to two consecutive words?
- Wrong decision for one word causes problems for what follows.
- Ideally, try all solutions and pick best one.

Cf. automatic transliteration (work by Rosmorduc):

- Alignment is easier (transliteration already given).
- Alignment here is more difficult (no dictionary or grammatical knowledge is assumed).

Requirement for practical use

Robustness

- Idealised input: accurate result.
- Less than ideal input: still 'reasonable' result.

Is essential for this task, due to:

- imperfections in sign list,
- errors by scribe, or by modern scholar,
- idiosyncratic writings.

Simple model of orthography

- Distinguish phonograms and determinatives.
- (Ideograms treated as phonograms.)
- Word is written as list of phonograms together covering all letters from transliteration, followed by determinatives.
- Penalties if such a solution cannot be found.
- Minor issues: numbers, dual, plural.

Method

- Extracted sign list with meanings from *Grosses Handwörterbuch Ägyptisch-Deutsch*.
- Produced hieroglyphic and transliteration for *Shipwrecked Sailor*.
- Manual marking of first hieroglyph for each word of transliteration.
- Automatic marking of same.
- Check agreement.

Results

Among 1014 words, only 16 errors:

- 9 due to readings missing from sign list
- 5 due to honorific transposition
- 2 due to other inadequacies of orthographic model

Very robust: recovers quickly after stumbling over problematic words.

Conclusions

Automatic alignment of hieroglyphic and transliteration:

- Even with simple model: high accuracy.
- Room for improvement: sign list, honorific transposition.
- Obviates need for manual alignment.

Other issues:

- Quantitative study of orthography.
  E.g. how often is certain sign used as phonogram?
- Use for learning and teaching.