Searching Four-Millennia-Old Digitized Documents: A Text Retrieval System for Egyptologists

Estíbaliz Iglesias-Franjo, Jesús Vilares
Language in the Information Society Group
Universidade da Coruña

www.grupolys.org

ACL LaTeCH 2016: August 11, Berlin
INDEX

- Introduction to Classic Egyptian
- Encoding of hieroglyphic texts
- System architecture
- Conclusions and future work
INDEX

• Introduction to Classic Egyptian
  • Encoding of hieroglyphic texts
  • System architecture
  • Conclusions and future work
Digital Heritage:
- Use of computing and information technologies to study and preserve our cultural legacy

Egyptology:
- Study of Ancient Egypt

Goal: **Text Information Retrieval (TIR) system for hieroglyphic texts**
**HISTORY**

- Longest-attested language:
  - 3300 BC – 17c AD (productive) / today (Coptic Church)
  - Deep changes throughout its evolution

- **Classic (a.k.a Middle) Egyptian:**
  - 2100 BC – 600 BC (spoken) / 5c AD (tradition)
  - Stereotypical image of Ancient Egyptian
HISTORY

• Longest-attested language:
  - 3300 BC – 17c AD (productive) / today (Coptic Church)
  - Deep changes throughout its evolution

• Classic (a.k.a Middle) Egyptian:
  - 2100 BC – 600 BC (spoken) / 5c AD (tradition)
  - Stereotypical image of Ancient Egyptian
CHARACTERISTICS

- Afro-Asiatic language
  - e.g., Arabic, Hebrew, Amharic, etc.
  - Subfamily of its own

- Consonantal
  - Roots formed by consonants
  - Only consonants are written

- Pictographic
  - Symbols portray elements of their world
    → a falcon
• No fixed alphabet
  ▪ Evolved from 800 to 5,000 signs
  ▪ New symbols/variants still appearing
SIGN TYPES

1. **Phonograms:** represent (1-3) sounds
   - (h)
   - (s^3)
   - (hpr)

2. **Logograms/ideograms:** represent the depicted element
   - (eye) (irt)

3. **Determinatives:** not read; denote the semantic group of the word
   - (Man – Human Being)
ARRANGEMENT OF SIGNS

- **Continuous script**: no dividers to separate words or phrases
  
  ![Hieroglyphs](image)
  
  *(iw ȝpdw ḫr nḥt)*
  
  "The birds are on the sycamore"

- **Arranged in non-linear groups**
  
  - No formal rules but principles/heuristics
  - *Horror vacui*
  - Seeking harmony and aesthetics
    
    "sycamore" *(nḥt)* : ![Hieroglyphs](image) → ![Hieroglyphs](image)
DIRECTION OF READING/WRITING

- Very flexible
- Not fixed
- Let’s see some examples!
DIRECTION OF READING/WRITING (cont)
DIRECTION OF READING/WRITING (cont)
in rows
DIRECTION OF READING/WRITING (cont)

in columns
DIRECTION OF READING/Writing (cont)

from-left-to-right
DIRECTION OF READING/WRITING (cont)

from-right-to-left
INDEX

• Introduction to Classic Egyptian
• **Encoding of hieroglyphic texts**
• System architecture
• Conclusions and future work
GARDINER’S LIST

• **Problem:** how to represent hieroglyphs without drawing them

• **Solution:** encoding signs using regular characters
**GARDINER’S LIST (cont)**

- Standard reference (723 signs + 20 var)
- Hieroglyphs encoded as characters:
  
  **Sign code = category letter + seq. number:**

  \[
  \text{ mano } = \text{ B2}
  \]

  - 26 categories according to drawings
  - Numbered sequentially within category

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A. “Man and his occupations”:</strong></td>
<td>![Man 1]</td>
<td>![Man 2]</td>
<td>![Man 3]</td>
<td>...</td>
</tr>
<tr>
<td><strong>B. “Woman and her occupations”:</strong></td>
<td>![Woman 1]</td>
<td>![Woman 2]</td>
<td>![Woman 3]</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
GARDINER’S LIST  (cont)

- Standard reference (723 signs + 20 var)
- Hieroglyphs encoded as characters:
  \[
  \text{Sign code} = \text{category letter} + \text{seq. number:} \quad \text{B}2
  \]

- 26 categories according to drawings
- Numbered sequentially within category

|       | 1 | 2 | 3 | ...
|-------|---|---|---|---
| A. “Man and his occupations”: | ![Man] | ![Man] | ![Man] | ...
| B. “Woman and her occupations”: | ![Woman] | ![Woman] | ![Woman] | ...

...     |  ...  | ...  | ...

• Standard encoding for digitization

• Evolution of Gardiner’s List:
  • Extra codes and rules for accurate representation of features (ASCII only)

• Sign arrangement operators

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Operation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>–</td>
<td>concatenation</td>
<td>Q3-X1-Z4-N1</td>
</tr>
<tr>
<td>:</td>
<td>subordination</td>
<td>X1:Z4:N1</td>
</tr>
<tr>
<td>*</td>
<td>juxtaposition</td>
<td>Q3*X1:Z4</td>
</tr>
<tr>
<td>()</td>
<td>grouping</td>
<td>Q3*(X1:Z4):N1</td>
</tr>
</tbody>
</table>
• **Damaged texts**: special marks (*shades*) attached to sign codes

(b) \(-N5-F12*C10-N36-M17*(Y5:N35)-\>

(c) \(-N5-(F12#13)*C10-N36#13-M17*(Y5:N35)-\>
• Introduction to Classic Egyptian
• Encoding of hieroglyphic texts
• System architecture
• Conclusions and future work
SYSTEM ARCHITECTURE
SYSTEM ARCHITECTURE
TEXT NORMALIZATION

- Regular text:
  - Regular normalization process: standard tokenization, lowercasing, stopwords, etc.

- Hieroglyphic text:
  - No delimiters!
  - Initial approach: **tokenized in sign groups** (delimited by ‘─’ in encoding):
    e.g. “boat” (dpt):
      - D46:Q3*X1–P1
      - D46:Q3*X1
      - P1
SYSTEM ARCHITECTURE
SYSTEM ARCHITECTURE
FRONT-END INTERFACE
FRONT-END INTERFACE
FRONT-END INTERFACE

INPUT QUERY FORMS
SYMBOL PALETTE (JSESH)
FRONT-END INTERFACE

SIGN ARRANGEMENT
SEARCH MODE SWITCH

FRONT-END INTERFACE
FRONT-END INTERFACE

OUTPUT: RELEVANT DOCUMENTS
FRONT-END INTERFACE

ACCESING CONTENT...
Abydos temple of Ramesses II. p. 531-532.

Very long lacuna
FRONT-END INTERFACE

Abydos temple of Ramesses II. p. 531-532.

MATCHINGS
INDEX

• Introduction to Classic Egyptian
• Encoding of hieroglyphic texts
• System architecture

• Conclusions and future work
CONCLUSIONS

• **First TIR system designed to manage Egyptian hieroglyphic texts**
  ▪ Language and writing system
  ▪ Encoding

• **Special care with front-end**
  ▪ Intuitive and easy to use

• **Available at** (free license, open source):
  https://github.com/estibalizifranjo/hieroglyphs
FUTURE WORK

• Study other retrieval solutions:
  ▪ Retrieval models
  ▪ Conflation and matching mechanisms
  ▪ n-Gram based processing

• Take advantage of similarities:
  ▪ Arabic, Hebrew, Japanese, Chinese, etc.

• Create evaluation corpora

• Collaborators?
Thank you very much! Questions?

https://github.com/estibalizifranjo/hieroglyphs

Third-Party Pictures:
• Slides no. 12-16 (cropped pic.): original by Olaf Tausch (Oltau) available at https://commons.wikimedia.org/wiki/File:Edfu_Tempel_42-2.jpg under the Creative Commons Attribution 3.0 Unported license.
• Slide 18 (cropped pic.): example of a handwritten entry, then printed lithographically; obtained from: R.O. Faulkner (2006). "Concise Dictionary of Middle Egyptian". Griffith Institute.
• Slide 22 (cropped pic.): original by Lord-of-the-Light; available at https://commons.wikimedia.org/wiki/File:Medinet_Habou_Cartouche.jpg under the Creative Commons Attribution-Share Alike 3.0 Unported license.
• PPT template: obtained at http://www.free-ppt-templates.com/2013/12/egyptian-ppt-template.html