TIR over Egyptian Hieroglyphs
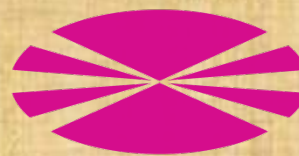
Estíbaliz Iglesias-Franjo, Jesús Vilares

Language in the Information Society Group
Universidade da Coruña
www.grupolys.org

LYS

INDEX

INDEX

CONTEXT

- 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**

CHARACTERISTICS

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

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

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

CHARACTERISTICS

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

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

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

CHARACTERISTICS (cont)

- **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

    (*ḥ*)   (*s3*)   (*ḫpr*)

2. **Logograms/ideograms:** represent the depicted element

    : *"eye"* (*ỉrt*)

3. **Determinatives:** not read; denote the semantic group of the word

    : [MAN – HUMAN BEING]
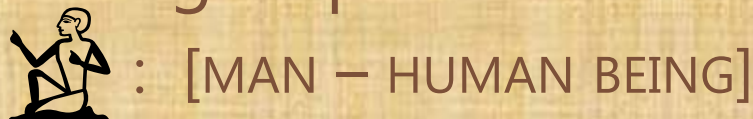
ARRANGEMENT OF SIGNS

- **Continuous script:** no dividers to separate words or phrases

  𓇋𓅱 𓄿𓊪�𓅭𓃾𓏏𓐍𓎛𓏏 (*iw ꜣpdw ḥr nht*)

  "The birds are on the sycamore"

- **Arranged in non-linear** *groups*
  - No formal rules but principles/heuristics
  - *Horror vacui*
  - Seeking harmony and aesthetics

    "sycamore" (*nht*) : 𓈖𓏏𓆭 → 𓈖𓏏𓆭

DIRECTION OF READING/WRITING

- Very flexible: four possible ways

INDEX

PROBLEM

"I'm glad they set that 140-character limit..."

• Solution: **encoding signs using characters**

GARDINER'S LIST

- Standard reference (723 signs + 20 var)

- Hieroglyphs encoded as characters:

Sign code = category letter + seq. number

 = **B2**

- – 26 categories according to drawings
- – Numbered sequentially within category

	1	2	3	...
A. *"Man and his occupations"*:				...
B. *"Woman and her occupations"*:				...
...	...	...	...	

GARDINER'S LIST (cont)

- Standard reference (723 signs + 20 var)

- Hieroglyphs encoded as characters:

  Sign code = category letter + seq. number

  = **B2**

  – 26 categories according to drawings
  – Numbered sequentially within category

	1	2	3	...
A. *"Man and his occupations"*:				...
B. *"Woman and her occupations"*:				...
...	...	...	...	

MANUEL DE CODAGE

- **Standard encoding for digitization**
- Evolution of Gardiner's List:
  - Extra codes and rules for accurate representation of features (ASCII only)
- **Sign arrangement operators**

Symbol	Operation	Example	
–	concatenation	`Q3-X1-Z4-N1`	
:	subordination	`X1:Z4:N1`	
*	juxtaposition	`Q3*X1:Z4`	
( )	grouping	`Q3*(X1:Z4):N1`	

- **Damaged texts**: special marks (*shades*) attached to sign codes

(a)   (b)   (c)

(a)   `<-N5-F12*C10-N36-M17*(Y5:N35)->`

(c)   `<-N5-(F12#13)*C10-N36#13-M17*(Y5:N35)->`

INDEX

SYSTEM ARCHITECTURE
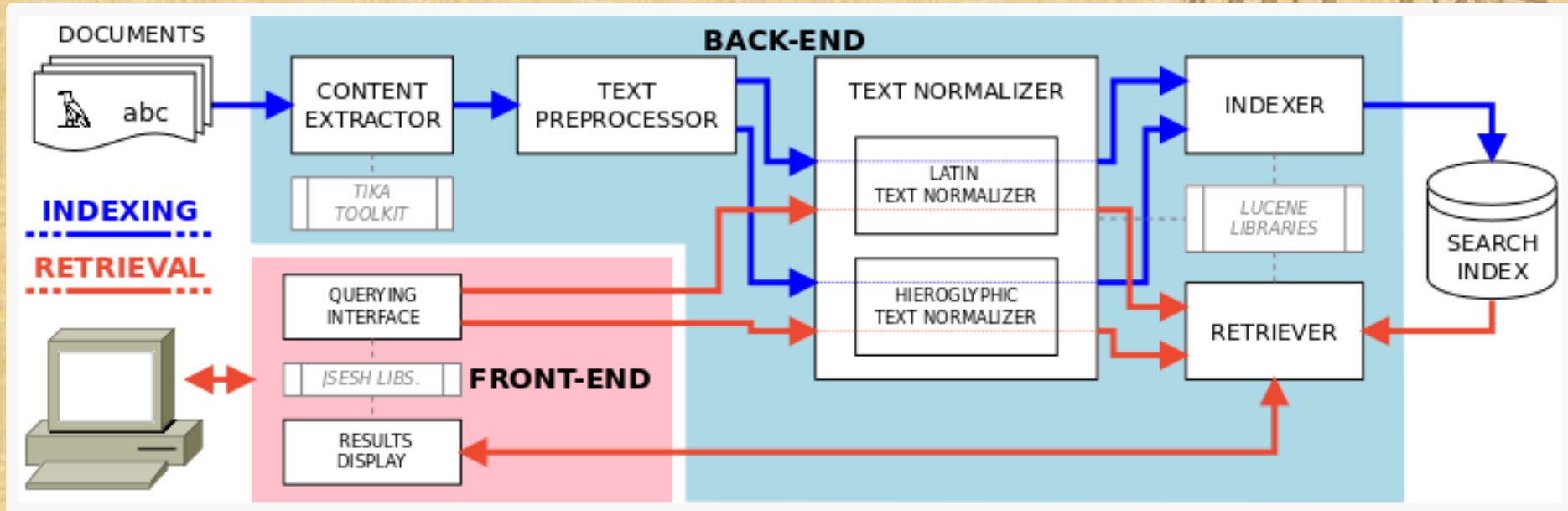
SYSTEM ARCHITECTURE
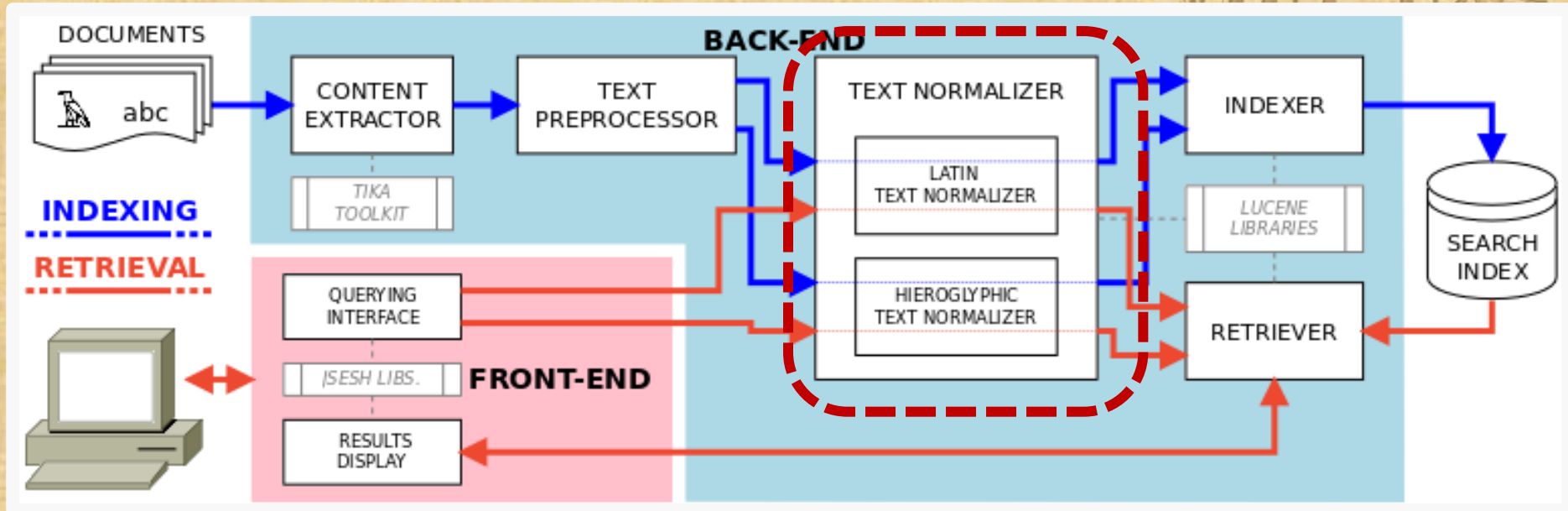
TEXT NORMALIZATION

- **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
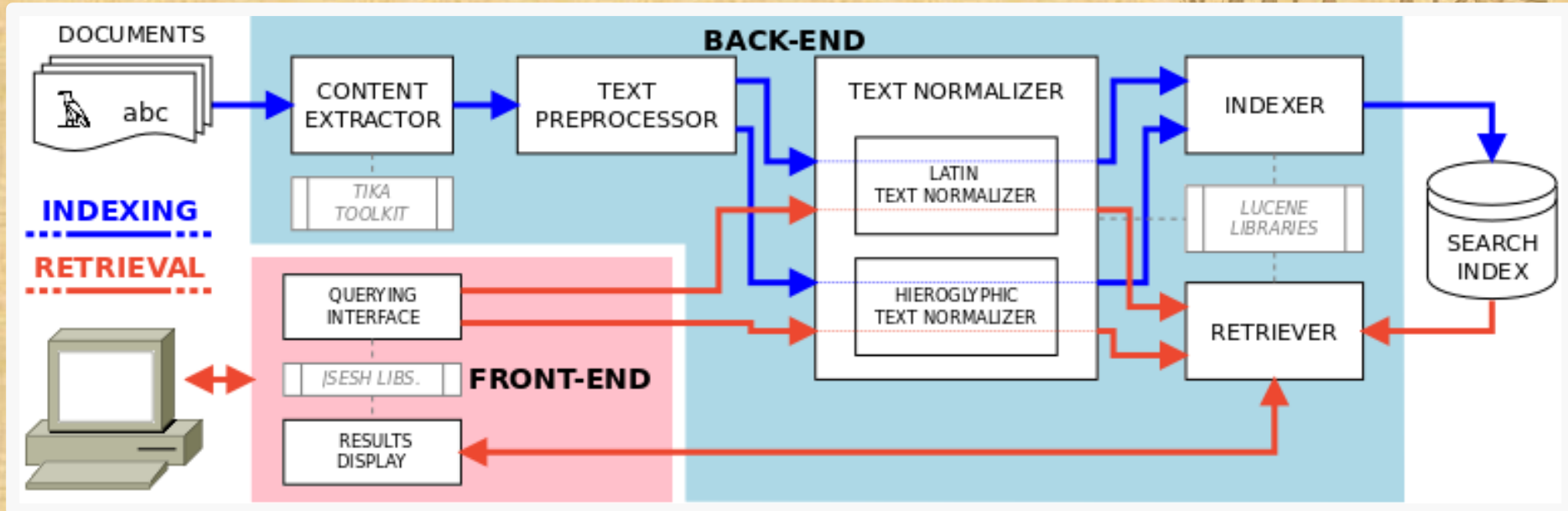
SYSTEM ARCHITECTURE

FRONT-END INTERFACE

FRONT-END INTERFACE

Family: Y. Writings, games, music

Sub-Family: all

Sub-Sub-Family:

☑ Show all   filter signs containing   Search

values
    dmD, mDAt

Y1

☐ user Pal.

File   View   Help

temple ramesses

t:Y1:3-D37:k-wn:n:z-mi:t:t-nTr*Hwt:k

Group vertically   Group Horizontal

Cartouches   Shades

Search

☐ Approximate latin search

☐ Approximate hiero search

Results

Luxor
Ankhnesneferibre.odt
King of Upper Egypt (Part 1).txt
Ramses
Temple of Philae.pdf
Rosseta Stelae
Templo de Filae
Osiris sarcofagus.hie

INPUT AREA

FRONT-END INTERFACE

INPUT QUERY FORMS

Family: Y. Writings, games, music
Sub-Family: all
Sub-Sub-Family:
☑ Show all   filter signs containing   Search

values
    dmD, mDAt

Y1

☐ user Pal.

File   View   Help

temple ramesses

t:Y1:3-D37:k-wn:n:z-mi:t:t-nTr*Hwt:k

Search

☐ Approximate latin search

☐ Approximate hiero search

Results

Luxor
Ankhnesneferibre.odt
King of Upper Egypt (Part 1).txt
Ramses
Temple of Philae.pdf
Rosseta Stelae
Templo de Filae
Osiris sarcofagus.hie

FRONT-END INTERFACE

SYMBOL PALETTE (JSESH)

FRONT-END INTERFACE

FRONT-END INTERFACE

SEARCH MODE SWITCH

FRONT-END INTERFACE

OUTPUT: RELEVANT DOCUMENTS

FRONT-END INTERFACE

FRONT-END INTERFACE

File   View

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

Very long lacuna
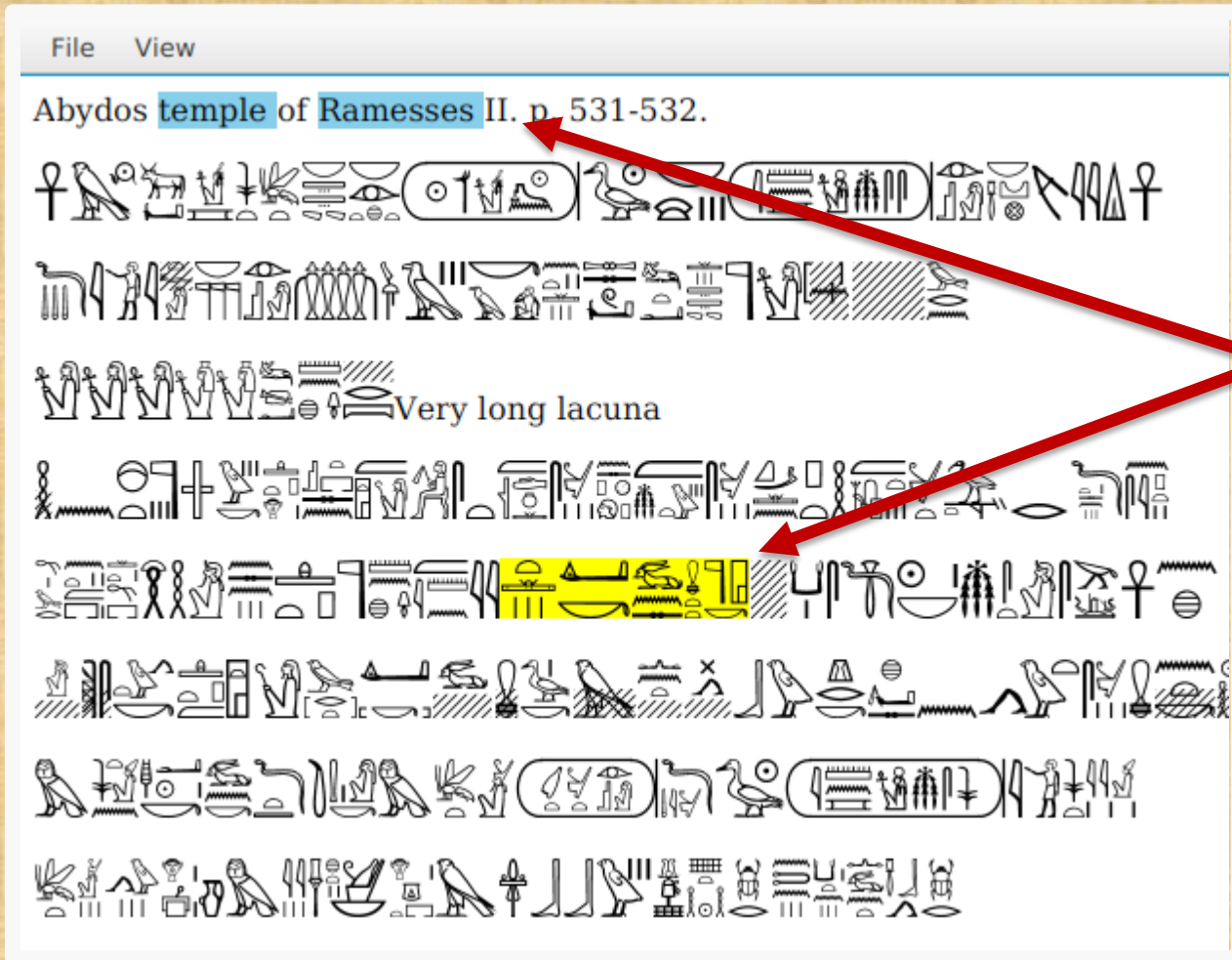
MATCHINGS

INDEX

CONCLUSIONS

- **<u>First</u> 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 an evaluation corpora**

Thank you very much! Questions?

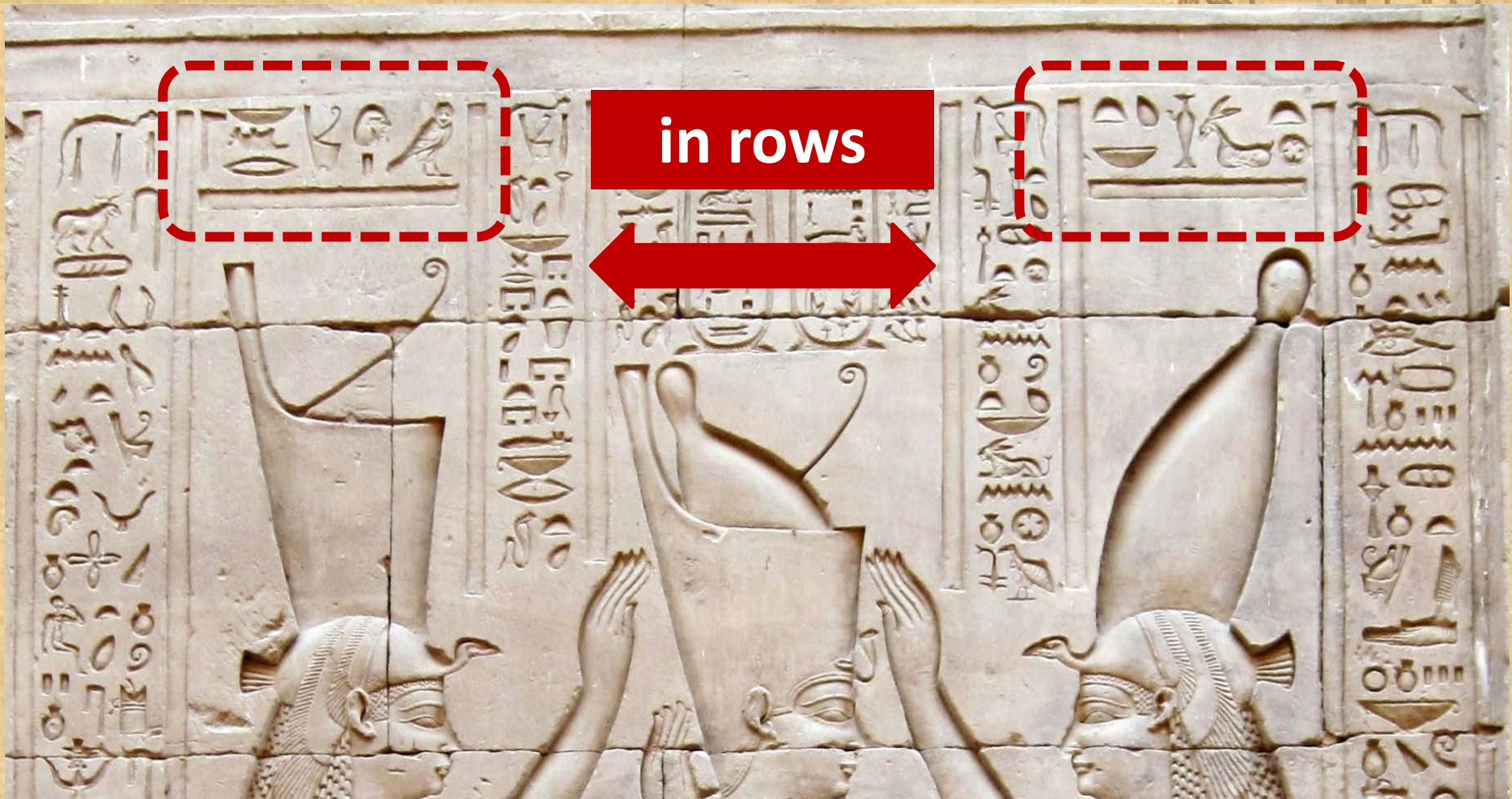https://github.com/estibalizifranjo/hieroglyphs

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

RELATED WORK

- Closely linked to the development of **classic-style text editors**:

  - GLYPH (1986): MS DOS, Windows, Mac
    - Laid the foundations of future editors

  - JSesh (2014): open source, Java
    - Currently the most popular

RELATED WORK (cont)

- **NLP/Text Mining**:

  - (Very) Initial stages: lack of corpora

  - Automatic transliteration (Barthélemy and Rosmorduc, 2011)

  - Language modelling (Nederhof and Rahman, 2015a)

  - Text categorization (Gohy et al., 2013)