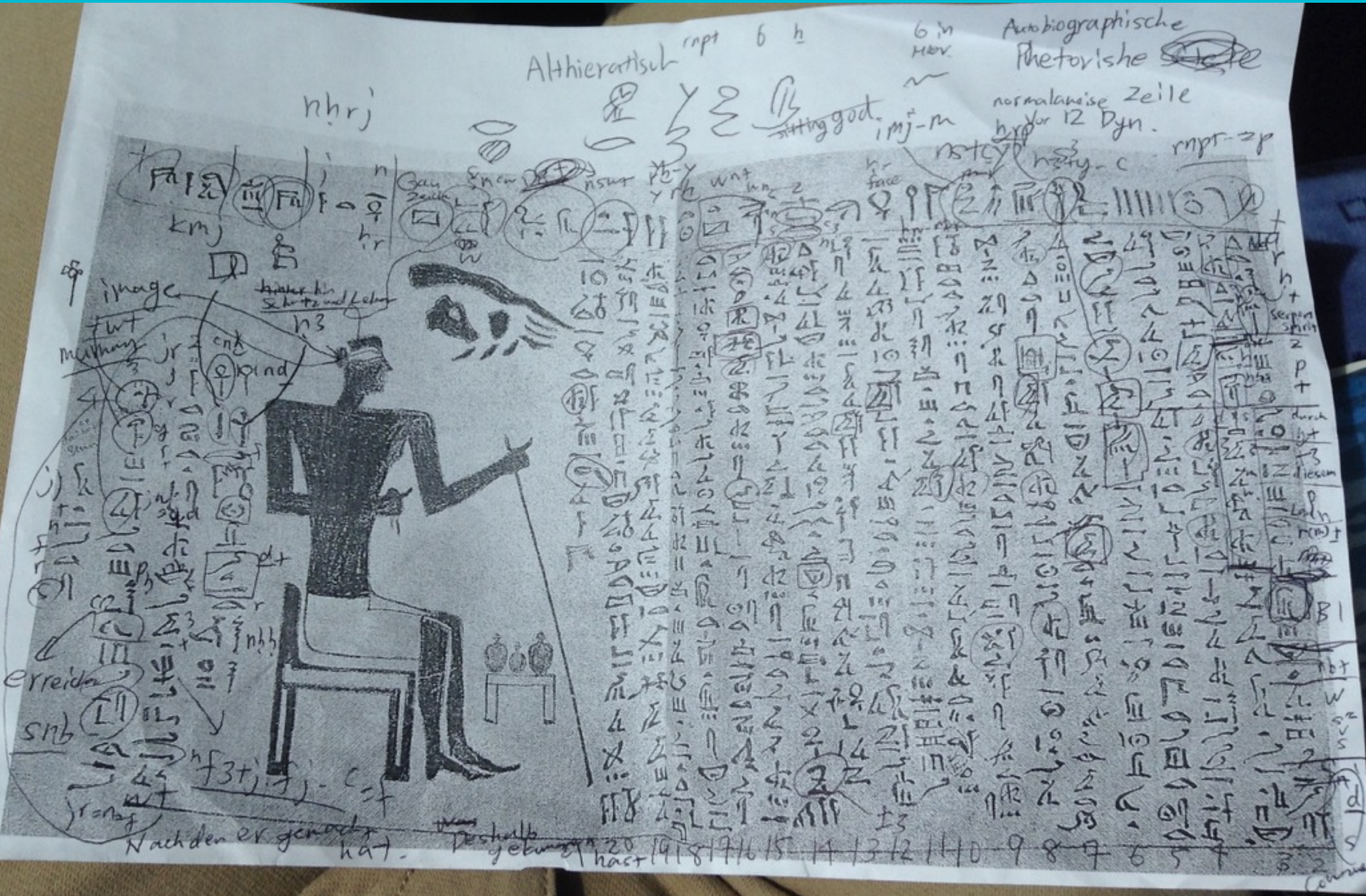
Development in encoding and inputting Ancient Egyptian and Coptic, and their benefits to projects of Egyptological Digital Humanities in the future, in comparison with the case of the East Asian Writing System

So Miyagawa

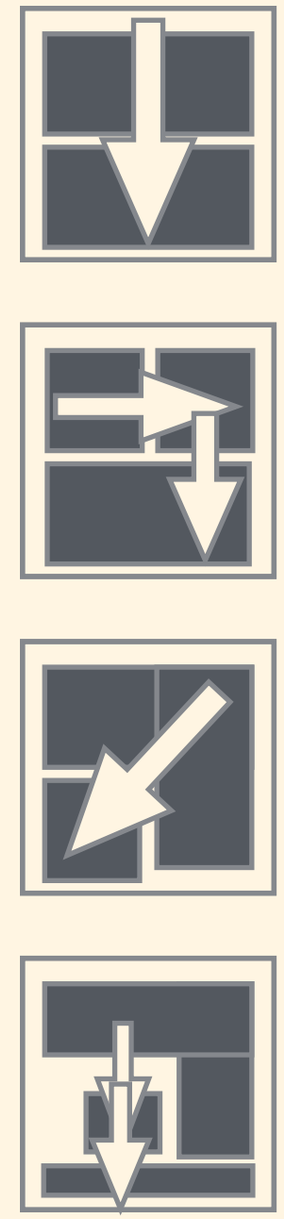
1st Cambridge IT in Egyptology Workshop / Informatique et Egyptologie Workshop 2016, at Fitzwilliam Museum, the University of Cambridge, 11 July 2016

Georg-August-Universität Göttingen (Egyptology & Coptology)
Kyoto University (Linguistics)
research originated in the framework of my work for the SFB 1136

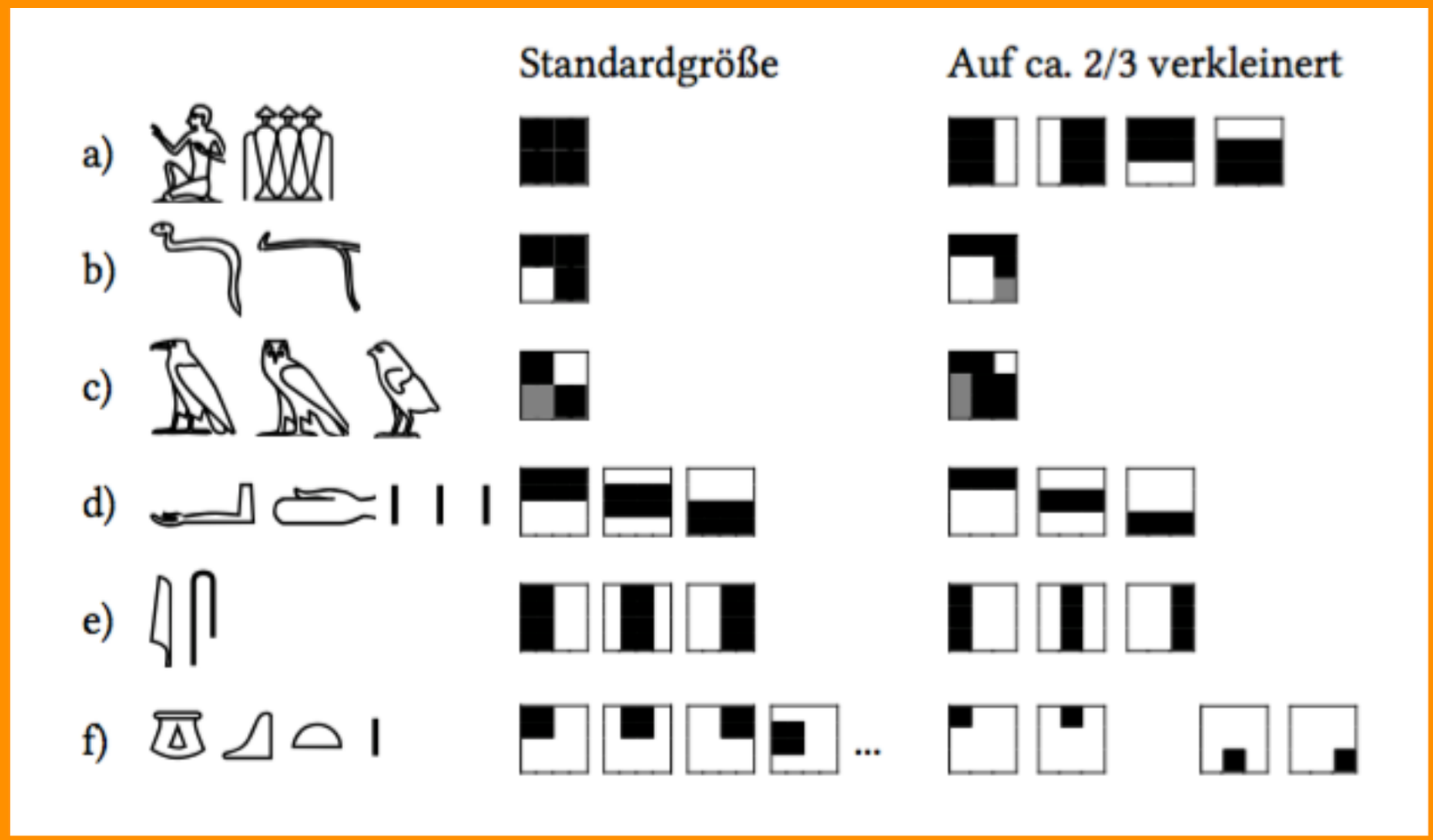
Peculiarity of Egyptian


Autobiographische
Rhetorishe
Althieratisch
rnpt 6 h
6 in Mov.
normalerweise Zeile vor 12 Dyn.
nhrj
sitting god
imj-r
rnpt-zp
kmj
image
h3
mit hand
nsw
face
serpent
durch
diesem
erreichen
snb
Nachdem er gemacht hat.
Deshalb
gekommen hast
20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2

Egyptian

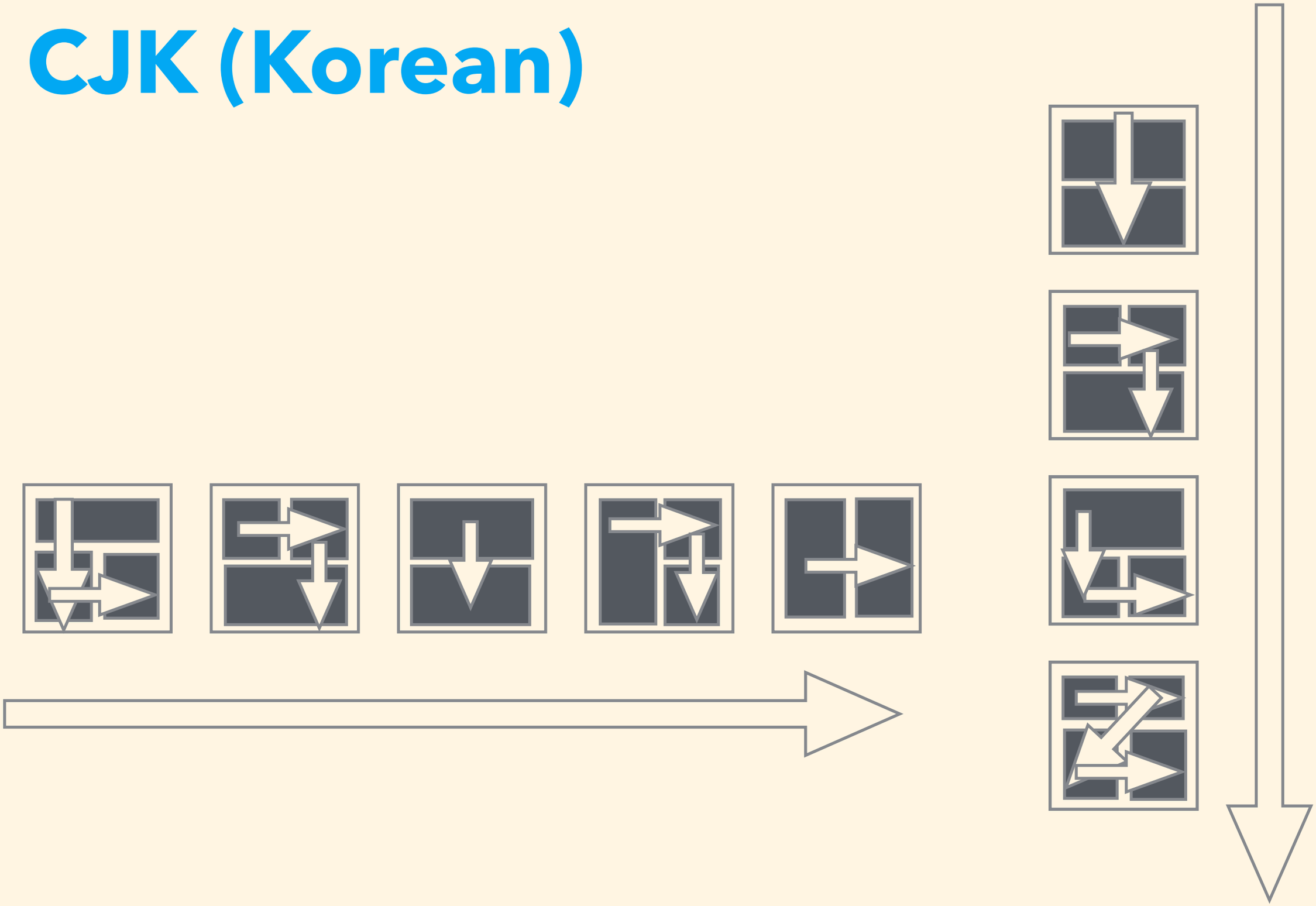


graphemes in a box



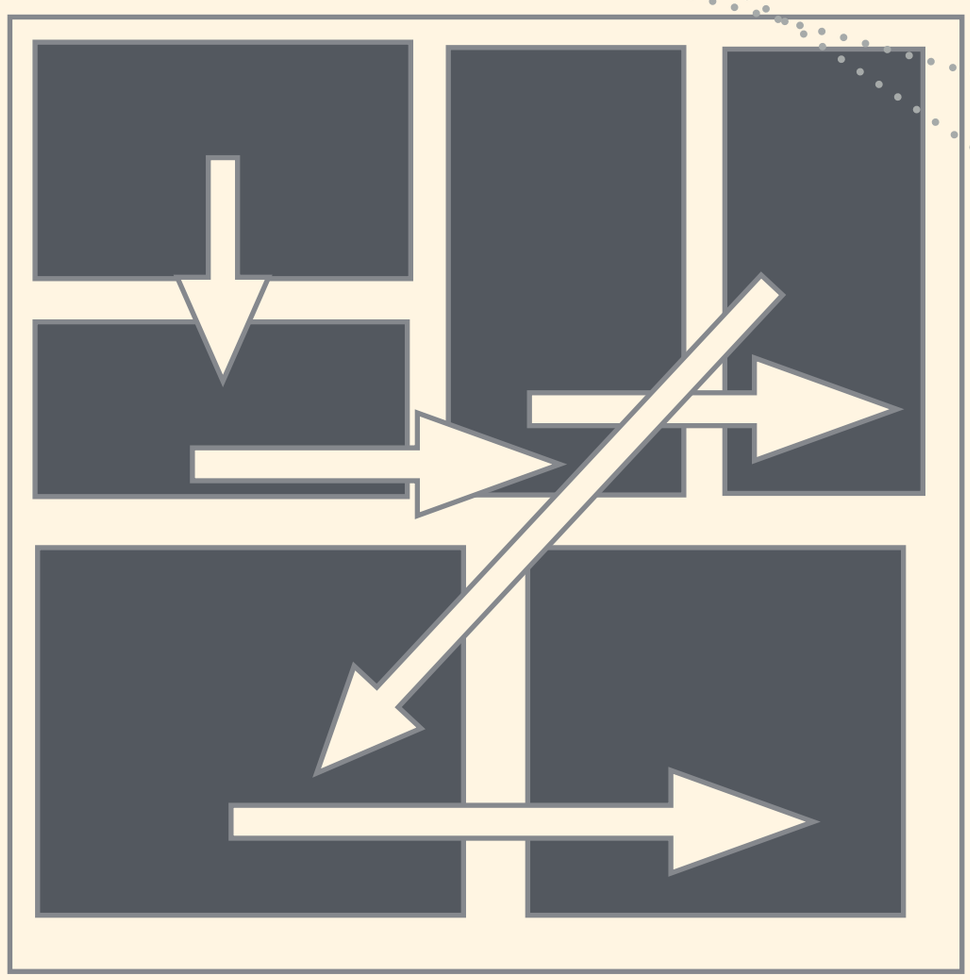
Werning, Daniel (2015), *Propädeutikum*
Hieroglyphisch-ägyptische Schrift und Sprache
Einführung mit Zeichenlernlektionen und Übungshinweisen. p.5.

CJK (Korean)



Korean 뭨 /mwelg/ (/muɔilg/)

ㅁ/m/ ㅜ/u/ ㅓ /ɔ/ ㅣ/i/ ㄹ/l/ ㄱ/g/



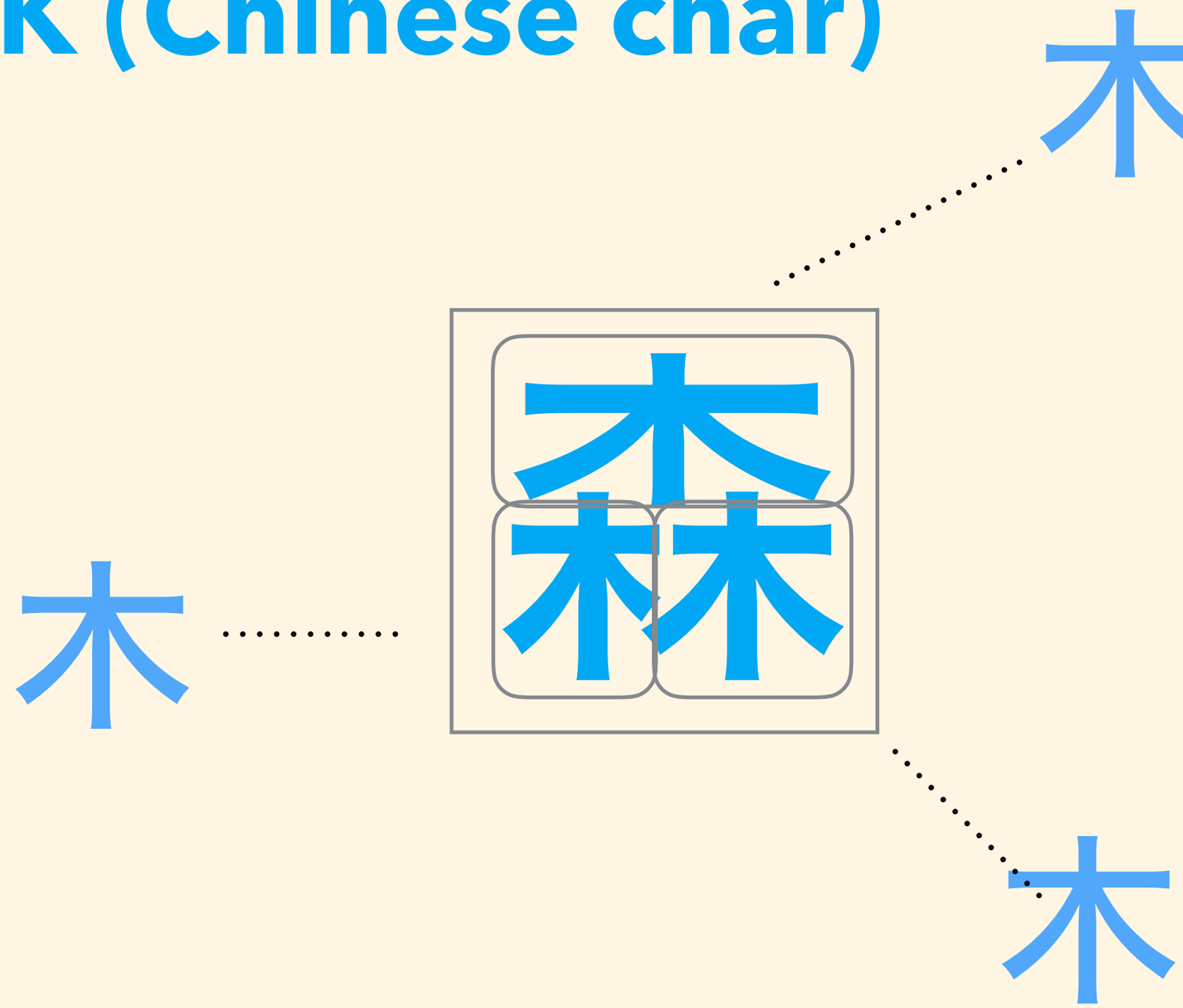
Japanese キロメートル (U+3316)

キ　ロ　メ

キロメートル

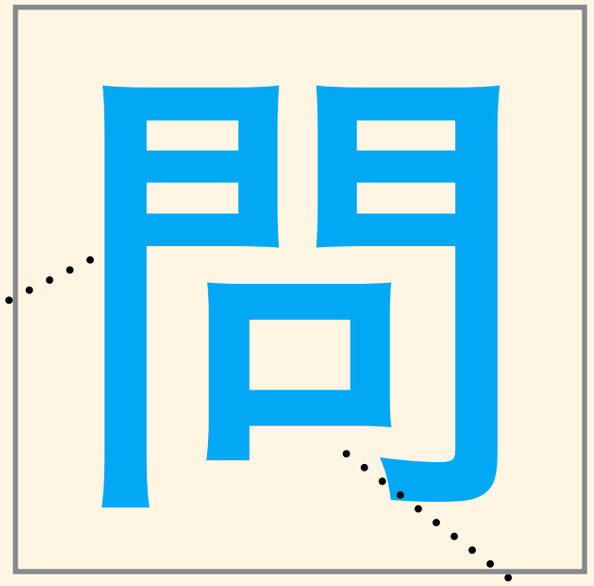
ー　ト　ル

CJK (Chinese char)



CJK (Chinese char)

Sin-Jap. /mon/ “question”



門

1. phonogram
Sin-Jap./mon/
2. ideogram
Sin-Jap./mon/
“gate”

1. determinative
MOUTH
2. ideogram
Sin-Jap./kou/
“mouth”

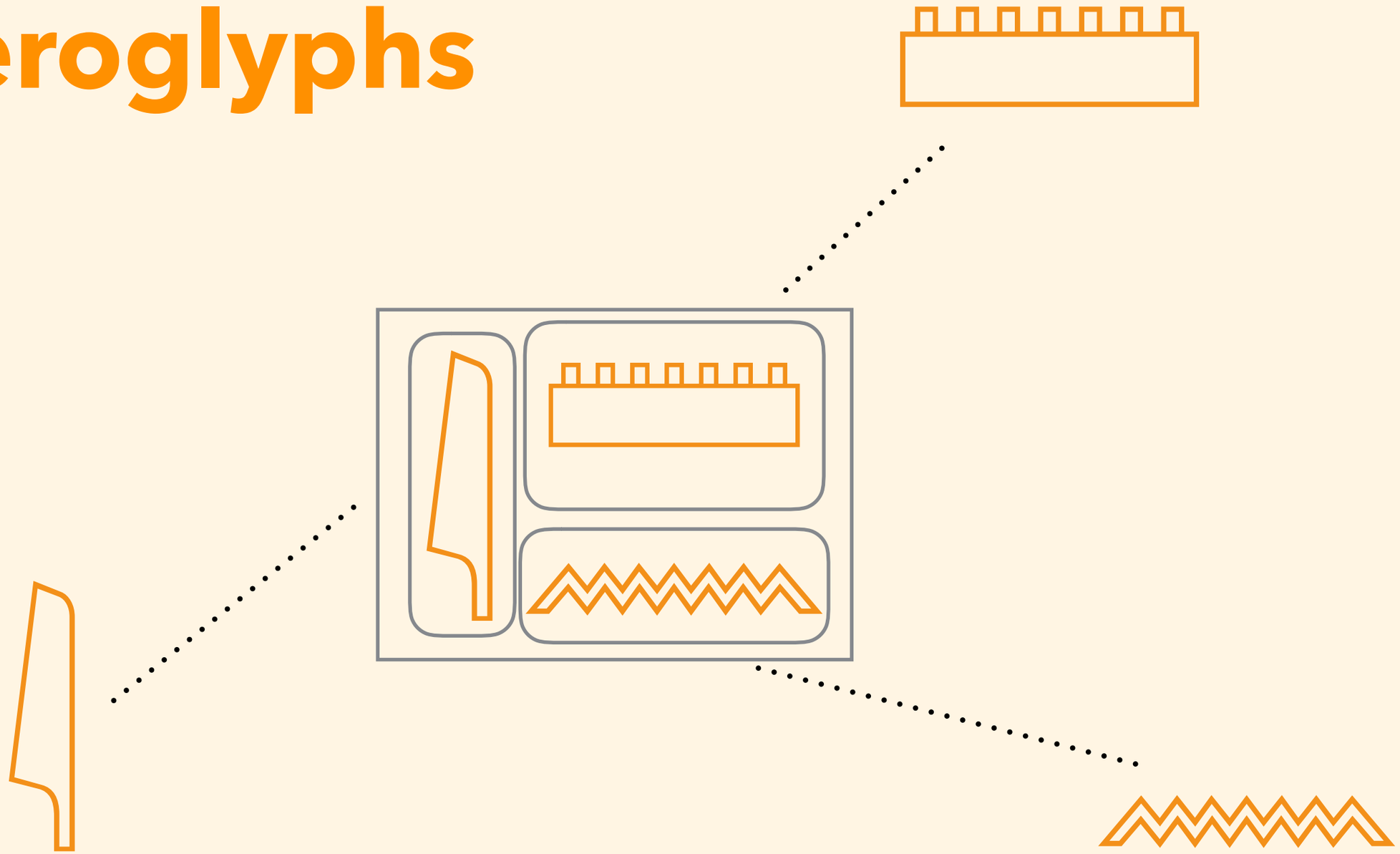
口

CJK

森 U+68EE

木 U+6728

Hieroglyphs



Bob Richmond's EGPZ

U+ec63 Z1&D36

U+ebfa P6&D36&N5&Z1

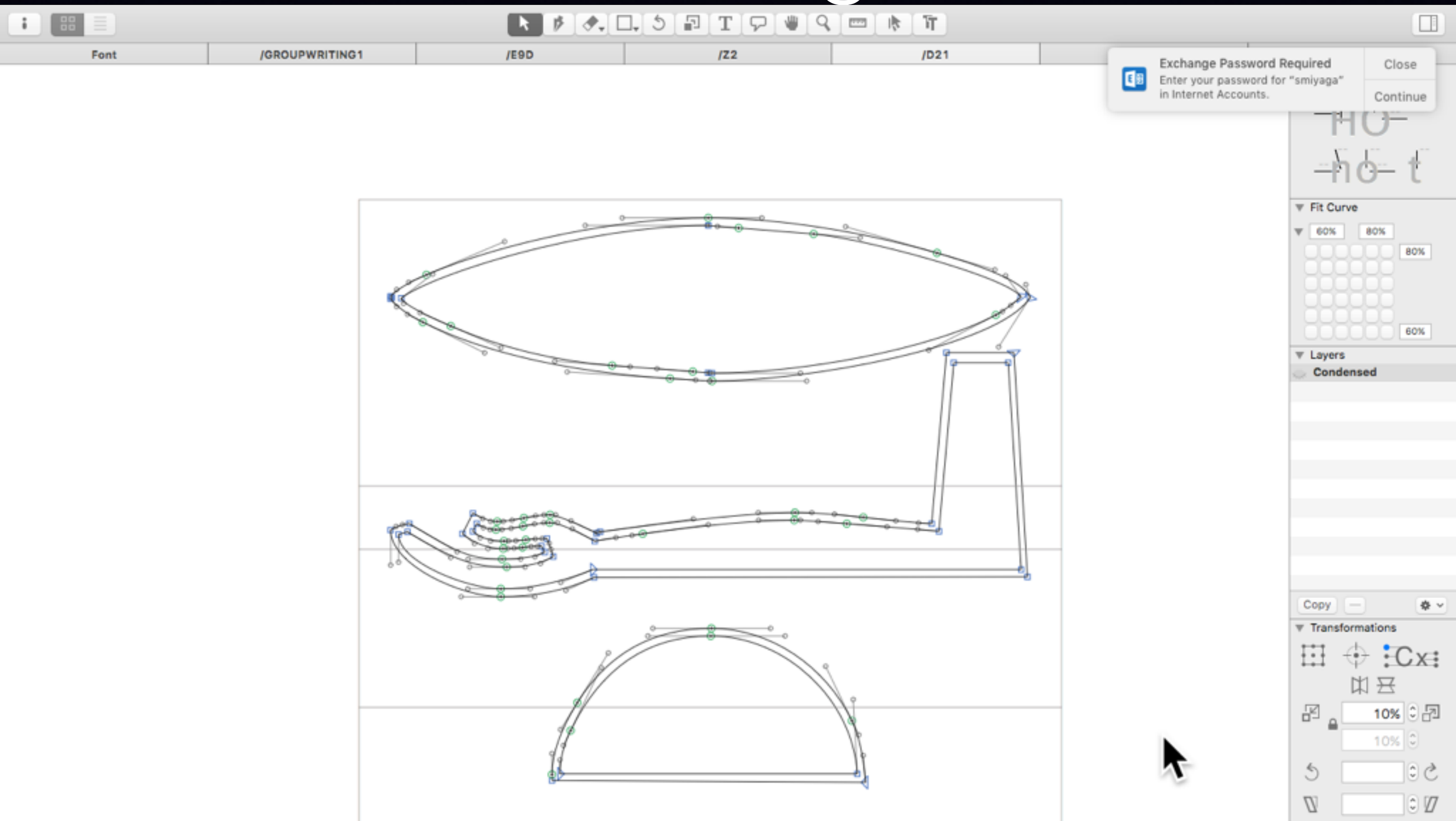
U+ebf9 P5&Z2d&G43

U+ebc0 I10&S43&M17&N35

U+ebbc I10&O34&I9

Photos are from Richmond, Bob. EGPZ 1.0 BETA Specification, 31st August 2007. http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=16FFD7C8DC879E483A2B2A0E5E0E811D?doi=10.1.1.182.4847&rep=rep1&type=pdf, accessed on 2016-01-21.

Making Fonts of Group Writing



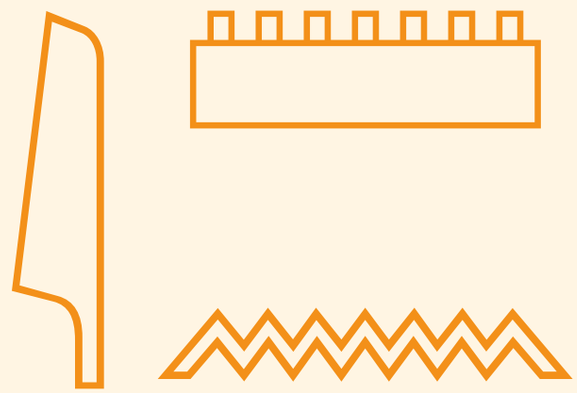
Merit and demerit of CJK way

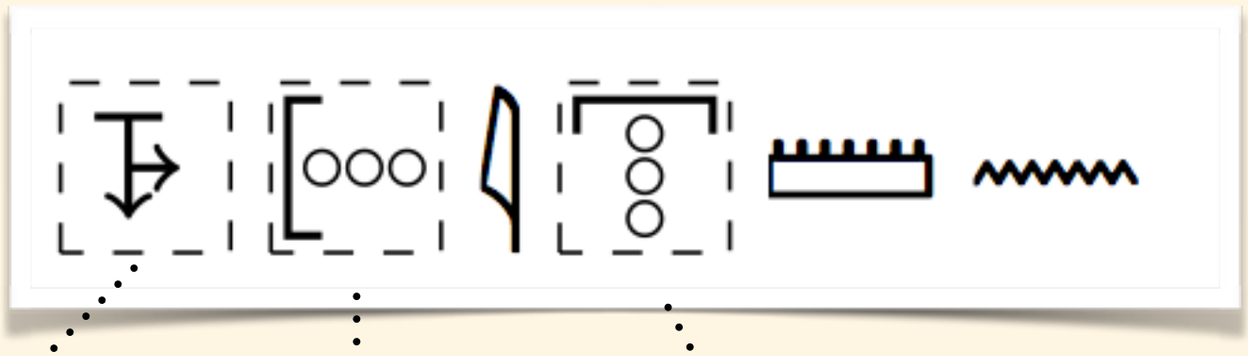
- Make all the combinations of signs as individual characters
- Merit
    - Easy to implement
    - Easy to display
- Demerit
    - Huge number of characters (over 120,000)

Also, you cannot include all

- There is no Unicode for this combination of graphemes

L2-16/177





Using control characters

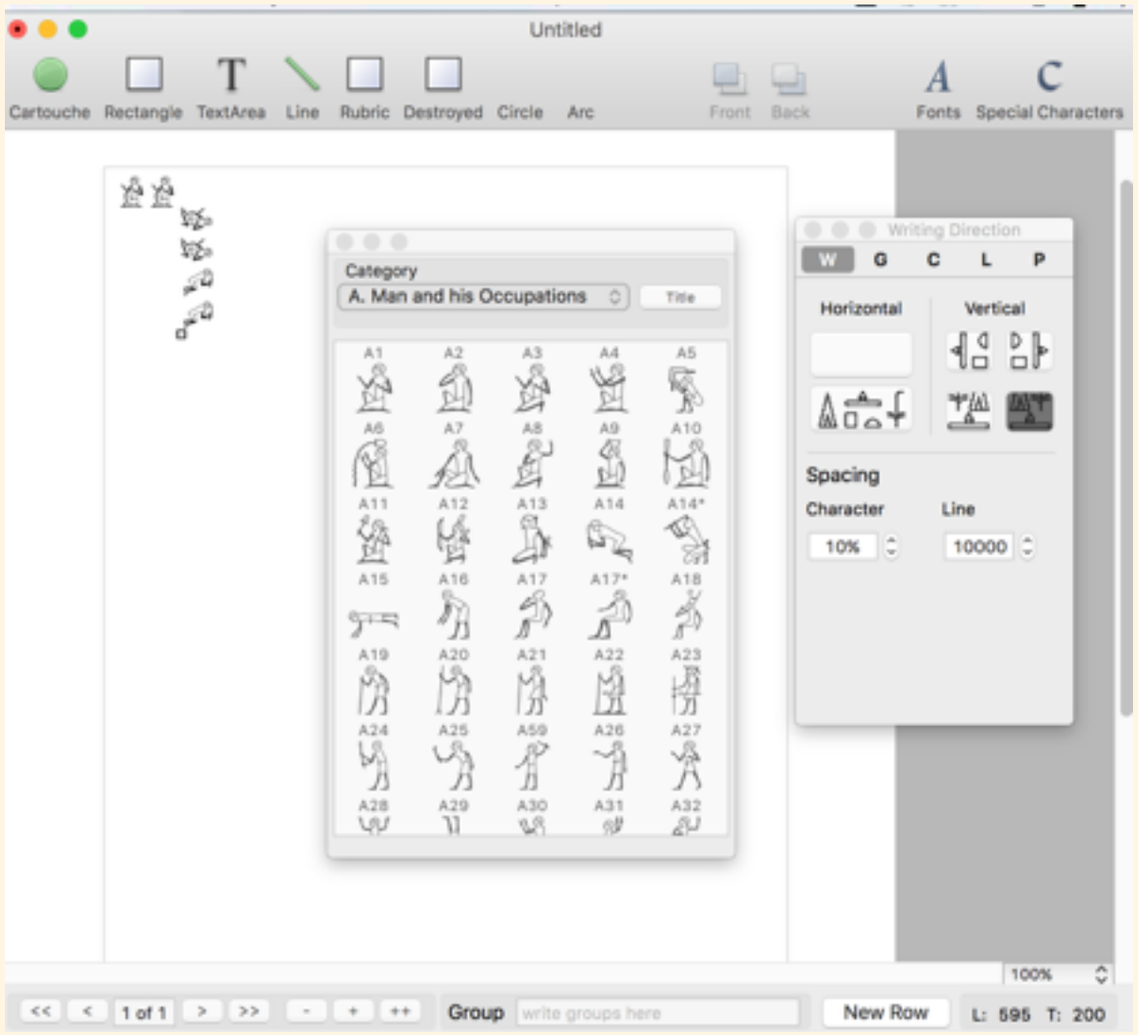
Using Unicode

- The age of Unicode
- flourishing tools based on Unicode
  - Greek: BetaCode —-> Unicode
  - Egyptian: MdC —-> Unicode (?)

Ways to write Hieroglyphs so far

Choose-and-put-the image editors

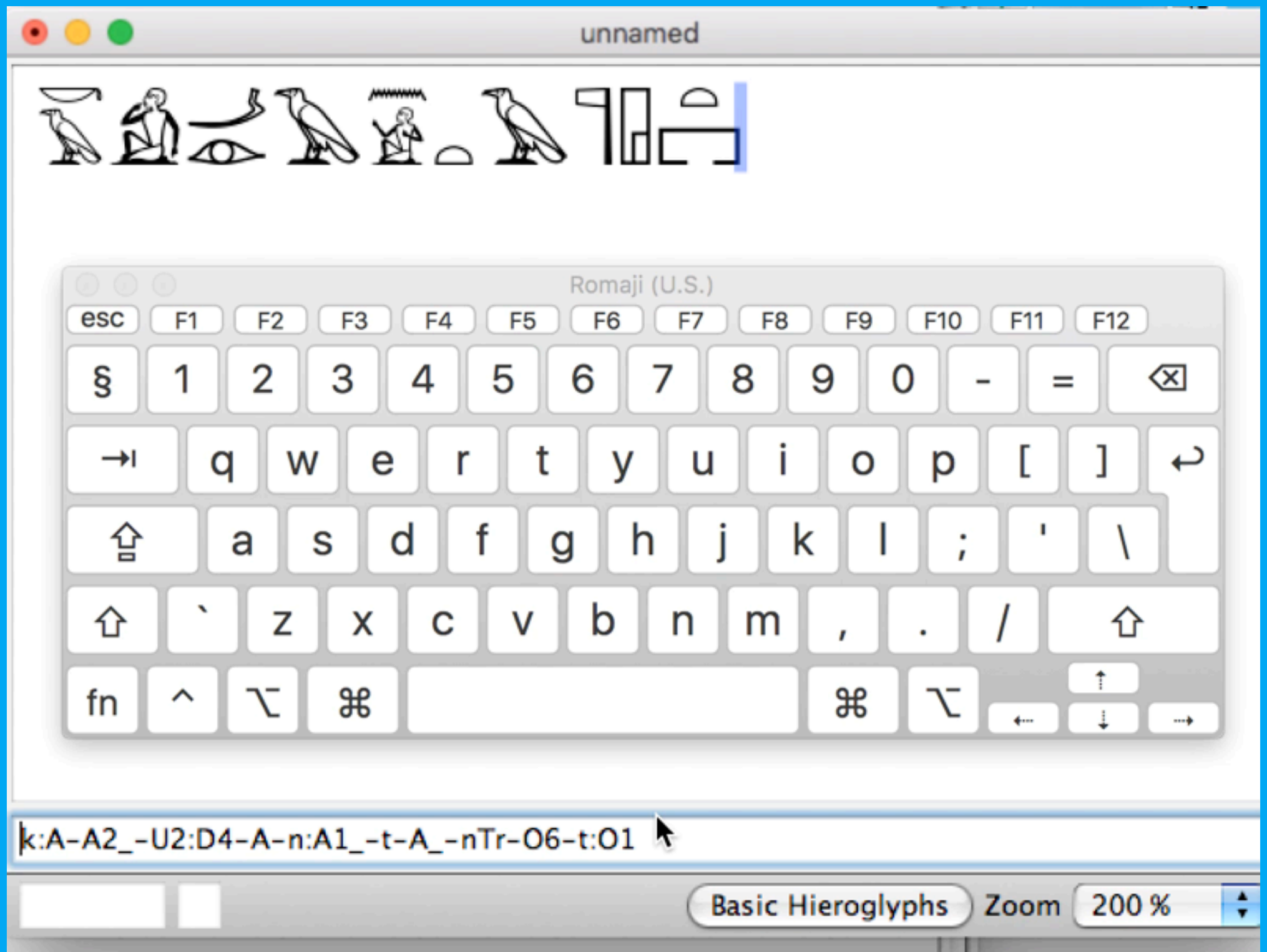
- iGlyph
- GlyphsBasic
- exports as an image file
- and so on



Encoded editors

- JSesh (also used for TLA and Ramses, based on MdC)
- WinGlyph (the math style is intuitive)
- and so on
- exports as an image file

JSesh, making Hieroglyphs



Systems for hi-end users with knowledge

- HieroTeX (Serge Rosmorduc)
- RES (Mark-Jan Nederhof)

hyuuu.htx

Typeset | LaTeX | Macros | Tags | Templates

HieroTeX
before compiling

```
144  \end{itemize}
145  \item 定冠詞の歴史
146  \begin{itemize}
147      \item 中エジプト語指示詞
148      \begin{hieroglyph}pA-A\end{hieroglyph}{\eg pA}
149      \begin{hieroglyph}t:A\end{hieroglyph}{\eg tA}
150      \begin{hieroglyph}n:A\end{hieroglyph}{\eg nA}
151      $\to$
152      コプト語定冠詞 \cop{p t n} \citep[]{allen}
153  \item 現在のDershasheにある第6王朝期の上エジプトの墓に刻まれた碑文が現在確認できるもののうちで初出であるとみられる。\citep{kupreyev}
```

revised.tex

```
697         \glt 「3人の王」
698 \end{exe}
699
700 コプト語では，生産的な名詞の複数語尾屈折がほぼ消失しているため，このような現象は見られない．よって，今回は，数詞と指示詞の数の不一致について言及するにとどめておく．
701
702
703
704 \subsection{「女性単数」的な{\it n-}系列指示詞の用法}
705
706 {\it n-}系列指示詞は，\citet{englund}や\citet{gardiner}や\citet{allen}
    などではpluralとしているが，\citet{allenme}や \citet{ockinga}ではpluralとはせずに，``
    neutral''\citep{allenme}や``common''\citep{ockinga}とだけ表現している．
707 なぜなら，単独で用いられた場合，単数かつ性を指定しないような用法があるからである．
708
709 しかしながら，次のように女性単数として特定できるデータがみられる．
710 \newpage
711 \begin{exe}
712     \ex 『ウェストカー・パピルス』 11,10，紀元前17世紀，P. Berlin inv. 3033 \citep{
        blackman}  \\
713 \begin{hieroglyph}{\leavevmode \loneSign{\Aca GP/38/}\HinterSignsSpace
714 \Cadrat{\CadratLineI{\Aca GD/69/}\CadratLine{\Aca GN/66/}}\HinterSignsSpace
715 \Cadrat{\CadratLineI{\jqABQBDd }\CadratLine{\Aca GN/66/}}\HinterSignsSpace
```

hieroglyph

Find Prev Find All

86 matches

Romaji (U.S.)

LaTeX

HieroTeX -compiled

(42) 『ウェストカー・パピルス』 11,10，紀元前 17 世紀，P. Berlin inv. 3033 (Blackman 1988)

ʿḥʿ-n ḏd-n ꜣs-t n nn nṯr-w ptr [*nꜣ*] [*nt-t*] *n*
そして 言う-PST イシス-F DAT DEM.nn 神-PL 何 nꜣ REL-SG/PL.F 1PL

jj.wjn [*r-s*] *nn jr-t bjꜣy-t n nꜣ n*
来る-1PL.STPro DIR-3SG.F ～ないで 行う-INF 奇跡.SG-F DAT nꜣ GEN

ḫrd-w smj-n n pꜣy-sn jt rḏj
子ども.M-PL 報告する-1PL DAT SG.M-彼らの 父.M.SG させる

jw-t-n
来る-INF-1PL

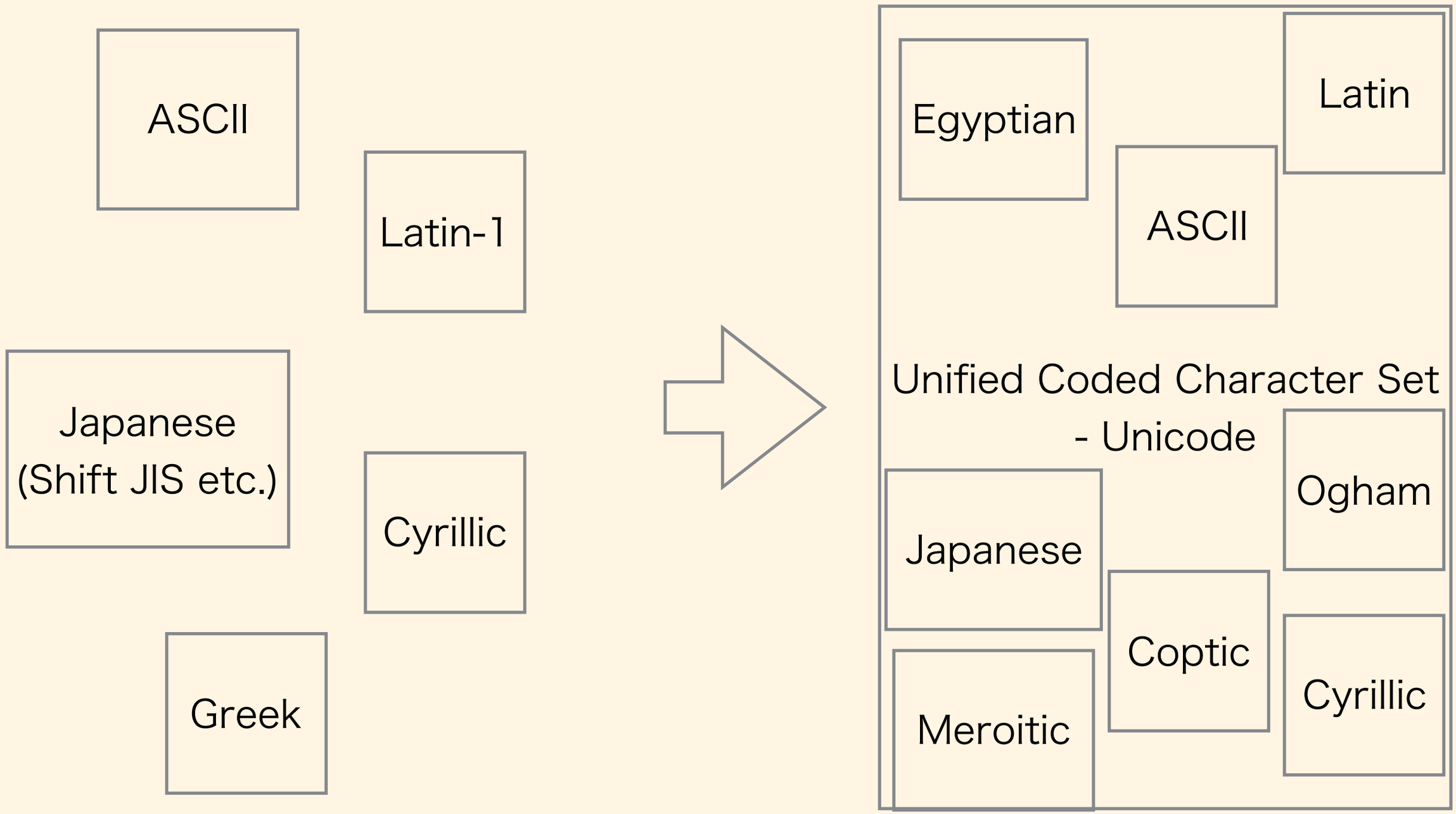
HieroTeX output

「するとイシスが（他の）女神に向かって申しました.「この赤ん坊たちのために，私たちを送りだしたその父君に報告できるような奇跡を行わないで，どうして戻ってきてしまったのでしょう.」(屋形・杉 1978:423)（lit. 戻ってきてしまったというこれ（*nꜣ*; この有様）はなんですか）」

History of coded character set

- ASCII
- ISO/IEC 646 - more variation, switching
- ISO/IEC 2022 - switching, GL, GR, Escape Sequence
- ISO/IEC 8859, Latin-1, Shift JIS and so on
- Unicode, ISO/IEC 10646

Unicode, ISO/IEC 10646



SFB1136 Project -White Monastery Coptic corpus and biblical text reuse

- Now 500,000 words (more technically, tokens)

- Lemmatised, POS-tagged, syntactic information (testing),
- A lot of automatic tools (by Amir Zeldes, Marco Büchler, Krill Bulert and me)
- Automatic OCR (handwritten manuscripts and printed texts)[kirill&me]
- Automatic normaliser[amir]
- Automatic lemmatiser[amir]
- Automatic POS-tagger[amir]
- Automatic syntactic tree generator[amir]
- Automatic Text Reuse detection tool[marco]

=> These all are based on Unicode

output/Luke_71_1L/010001.bin.png

output/Luke_71_1L/010002.bin.png

Promising effects of Unicode on (Pharaonic) Egyptological Projects

- Corpus linguistics based on hieroglyphs, not transcription
- Easy application of OCR software like Ocropy
- Easy application of POS-tagger like TreeTagger
- Easy application of Text Reuse tool like Tracer
- Easy application of corpus linguistic tools like NLTK, VoyantTools
- Easy application of Tree Bank Generator like Arborator
- Easy to search and replace by RegEx

Egyptian Unicode Inputting

- If Unicode is completed, inputting system will be needed.
- I’m making it with Japanese Input System
- Called HieroJIS (Hieroglyphs on Japanese Input System)

HieroJIS

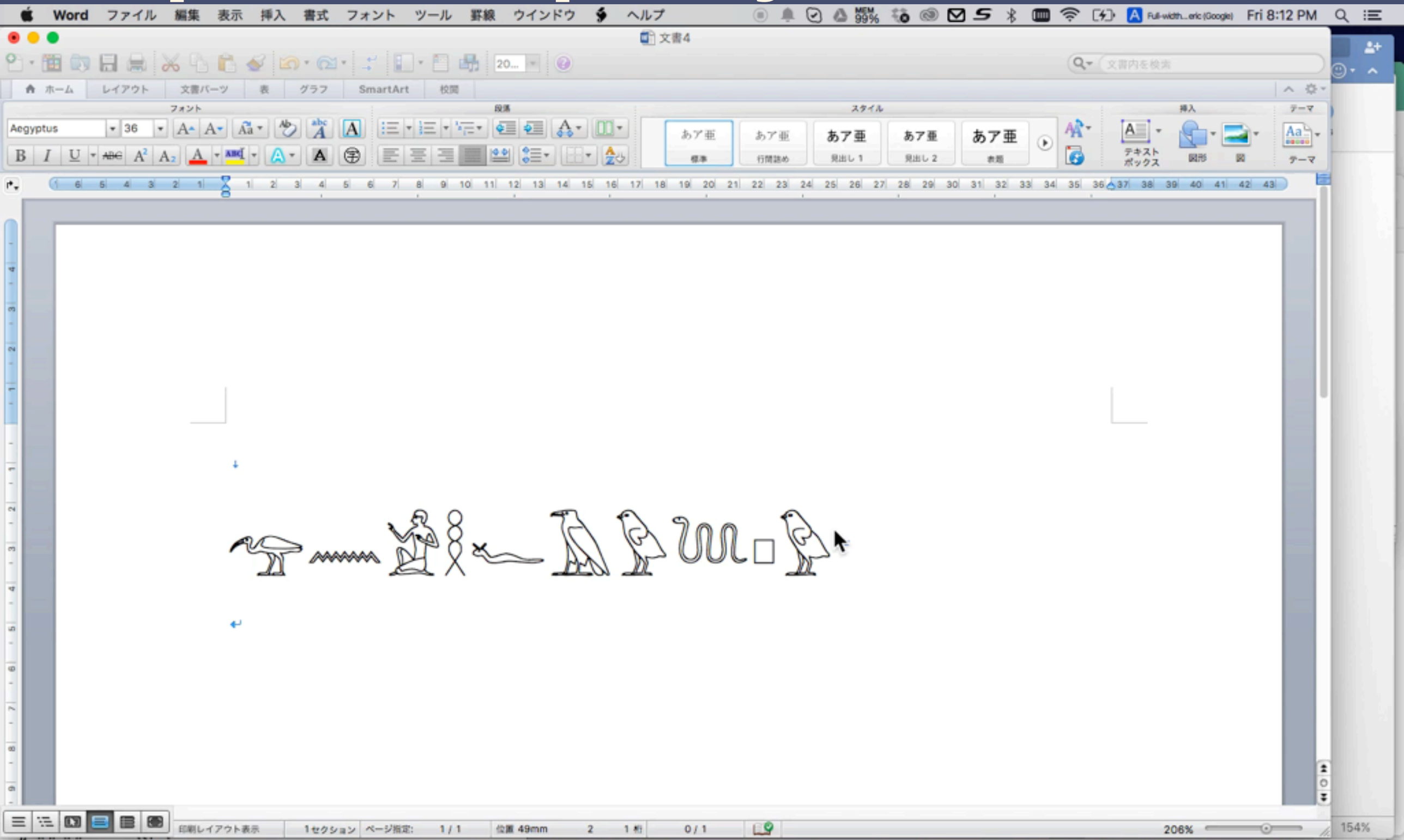
- Four ways of inputting:
  Word-to-Word, Gardiner, Category, Phonetic

- 60,000 entries including TLA lemma (currently)
  candidates on pull-down window

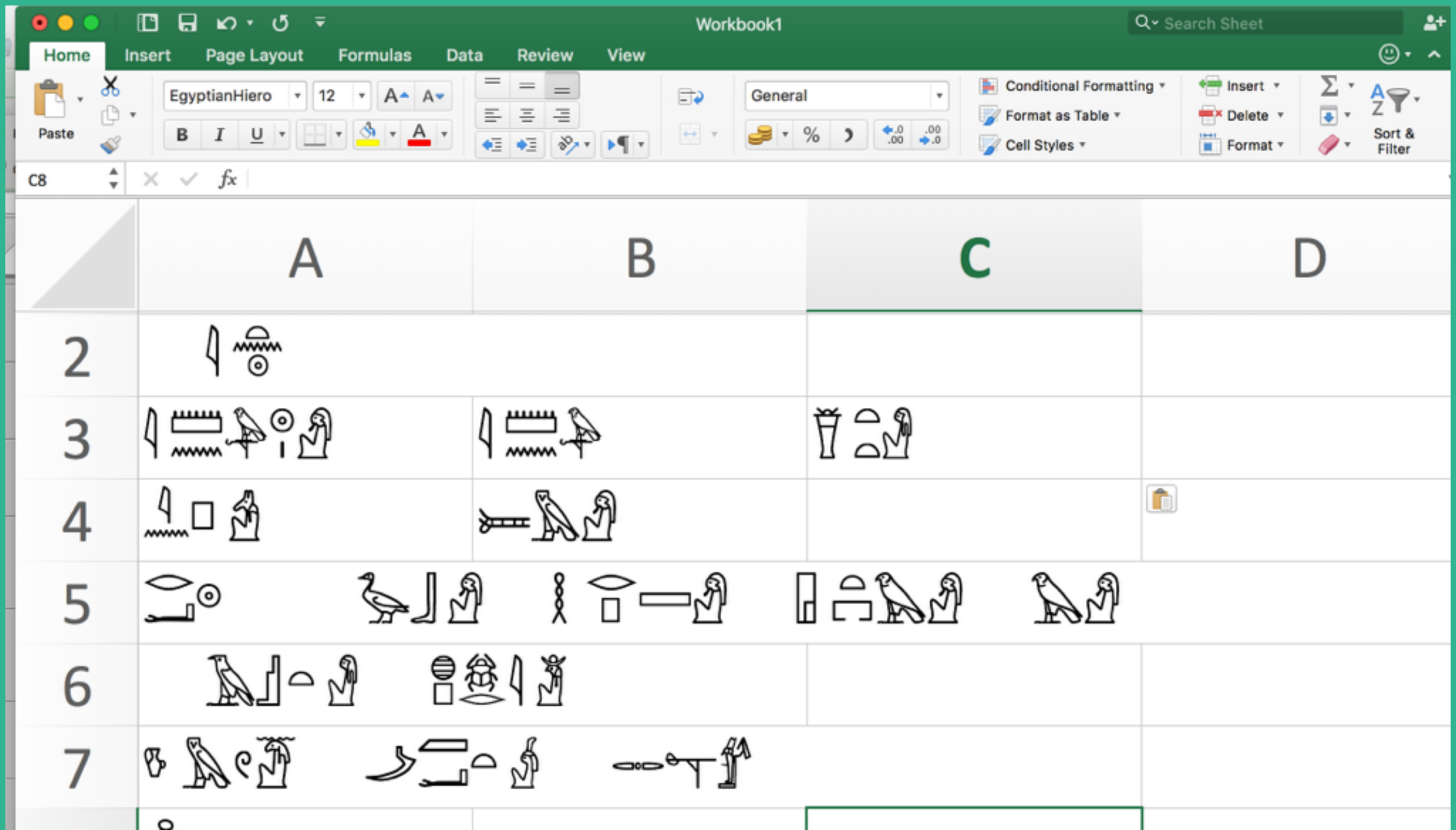
- n-Bunsetsu Longest Concord Method:
  Used in most of the Japanese input systems

- Memorising user history, prediction and optimisation

HieroJIS (Hieroglyphs on Japanese Input System)



SINUHE Project with Marwan Kilani and me



Even on Facebook

Tuesday, July 12, 2016 at 8:23pm

Marwan Kilani 𓋴𓈖𓅱

Unlike · Reply · 1 · 14 hrs

Marwan Kilani replied · 12 Replies

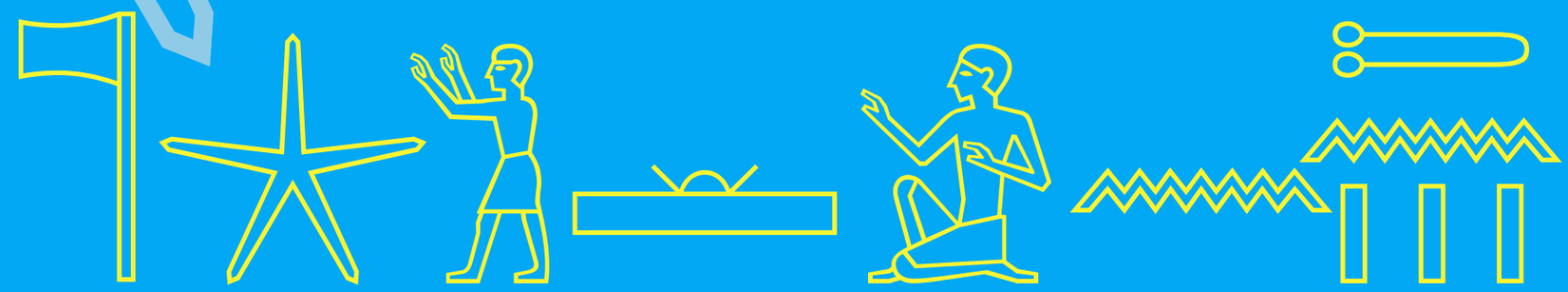
SINUHE = System of INputting Unicode of Hieroglyphic Egyptian

Anyway, it will be marvellous for Egyptology if we have a complete Unicode set and its input system.

And we will discuss mainly Unicode side today.

dw3-ntr=i n=tn

(or *dw3=i n=tn ntr*)



wenemut