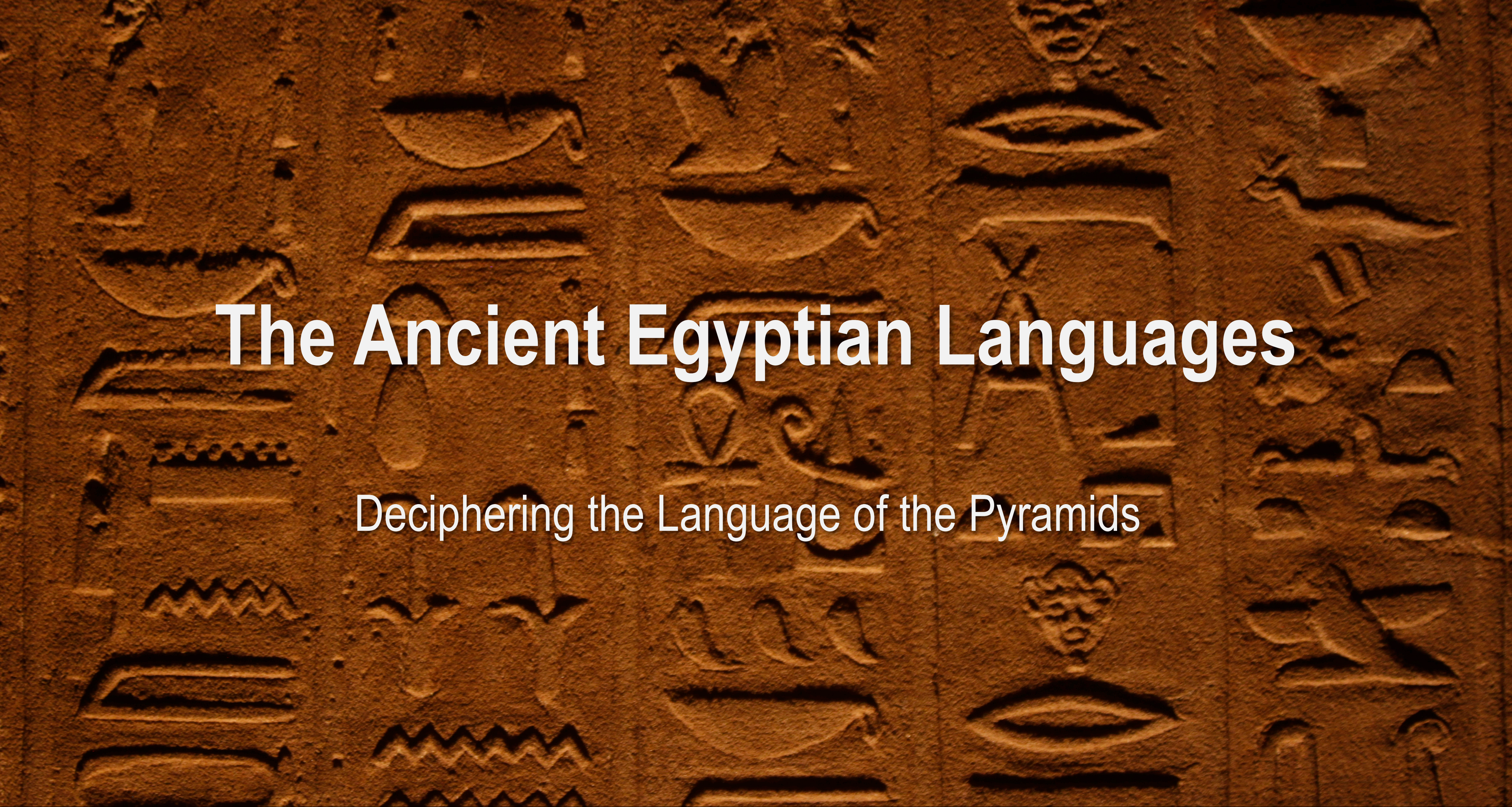
A Linguistic Journey through Ancient Egypt Using NLP

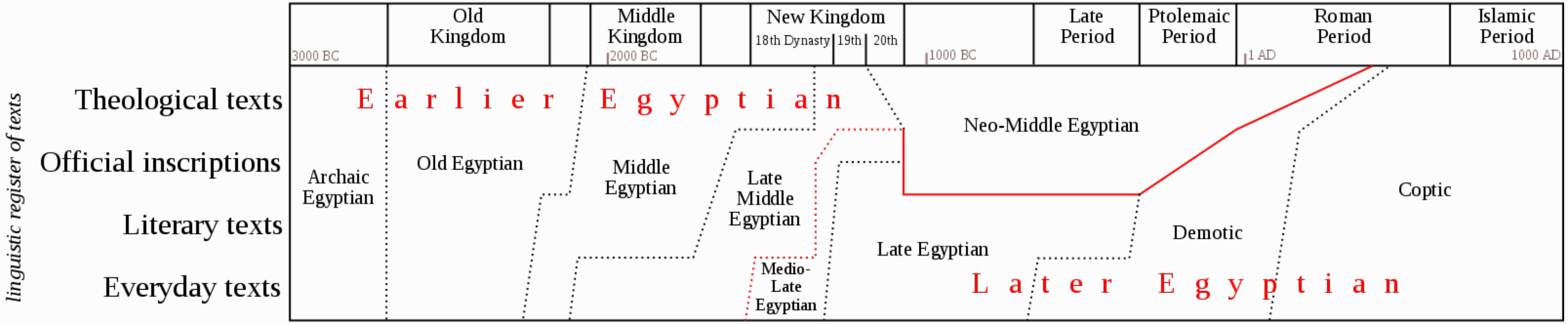
Ricardo Muñoz Sánchez



The Ancient Egyptian Languages

Deciphering the Language of the Pyramids

Ancient Egyptian Through Time



Taken from Wikimedia

Ancient Egyptian Languages

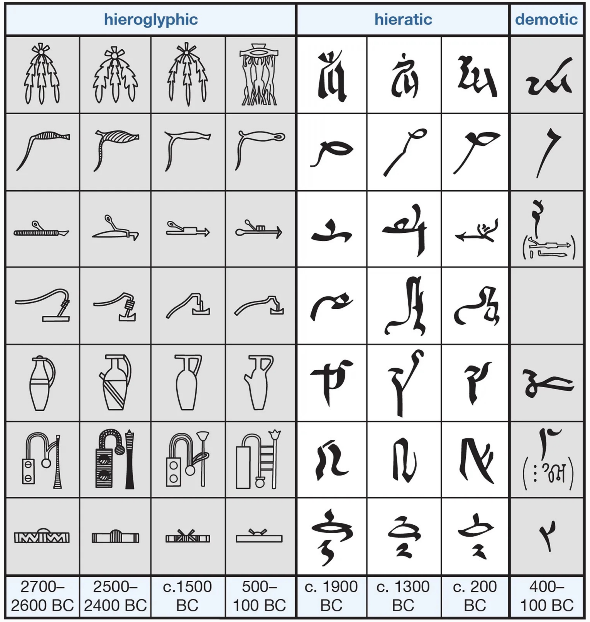
- They are Afro-Asiatic languages
  - Word roots are important
  - There are two grammatical genders (masculine and feminine)
  - Vowels are rarely written down

- They are all extinct languages
  - Coptic is still in liturgical use

Writing Systems

- Hieroglyphic (from c. 3200 BC to c. 400 AD)
- Hieratic, a cursive version of hieroglyphs (from c. 3200 BC to 3rd century AD)
- Demotic became the official script during the 26th Dynasty (664–525 BC)
- Coptic script, based on the uncial Greek script with letters borrowed from Demotic (from 2nd century AD)

Evolution of the Ancient Egyptian Scripts



hieroglyphic				hieratic			demotic
2700–2600 BC	2500–2400 BC	c.1500 BC	500–100 BC	c. 1900 BC	c. 1300 BC	c. 200 BC	400–100 BC

Taken from Encyclopaedia Britannica, based on the same table by Möller (1919, p. 78)

Hieroglyphs – The Symbols

- **Phonograms** represent sounds
  - **Uniliteral signs** represent a single sound:
    𓄿 can be transliterated as *ꜣ*
  - **Biliteral signs** represent two sounds:
    𓃹 can be transliterated as *wn*
  - **Triliteral signs** represent three sounds:
    𓆣 can be transliterated as ḫpr

Hieroglyphs – The Symbols

- **Phonograms** represent sounds (e.g. 𓄿, 𓃹, and 𓆣)

- **Logograms** represent whole words or ideas

  𓀗 represents *the king of Lower Egypt*

  𓁦 represents *Maat*, goddess and personification of order

  𓆾 represents “emmer”, a kind of wheat

Hieroglyphs – The Symbols

- **Phonograms** represent sounds (e.g. 𓄿, 𓃹, and 𓆣)
- **Logograms** represent whole words or ideas (e.g. 𓀾, 𓁦, and 𓇉)
- **Determinatives** clarify the meaning of the word
    - 𓇳 denotes words related to the sun, day, and time
    - 𓊛 denotes words related to boats

Hieroglyphs – The Symbols

- **Phonograms** represent sounds (e.g. 𓄿, 𓃹, and 𓆣)
- **Logograms** represent whole words or ideas (e.g. 𓀵, 𓁬, and 𓆭)
- **Determinatives** clarify the meaning of the word (e.g. 𓇳 and 𓊛)
- **Typographical signs** can either give semantic meaning or serve as fillers
    - 𓏤 can be used to fill-in space within stacks of symbols

Hieroglyphs – Some Issues

- A symbol might belong to more than one category

  As we mentioned, ☉ can be a determinative

  But it can also be a logogram for *the sun*

  Moreover, it can be a phonogram for *n*

Hieroglyphs – Some Issues

- A symbol might belong to more than one category

- Phonemic values might be repeated within a word

  𓏠 has phonetic value *mn*

  𓈖 has phonetic value *n*

  𓏠𓈖 means *to endure* and has phonetic value *mn*

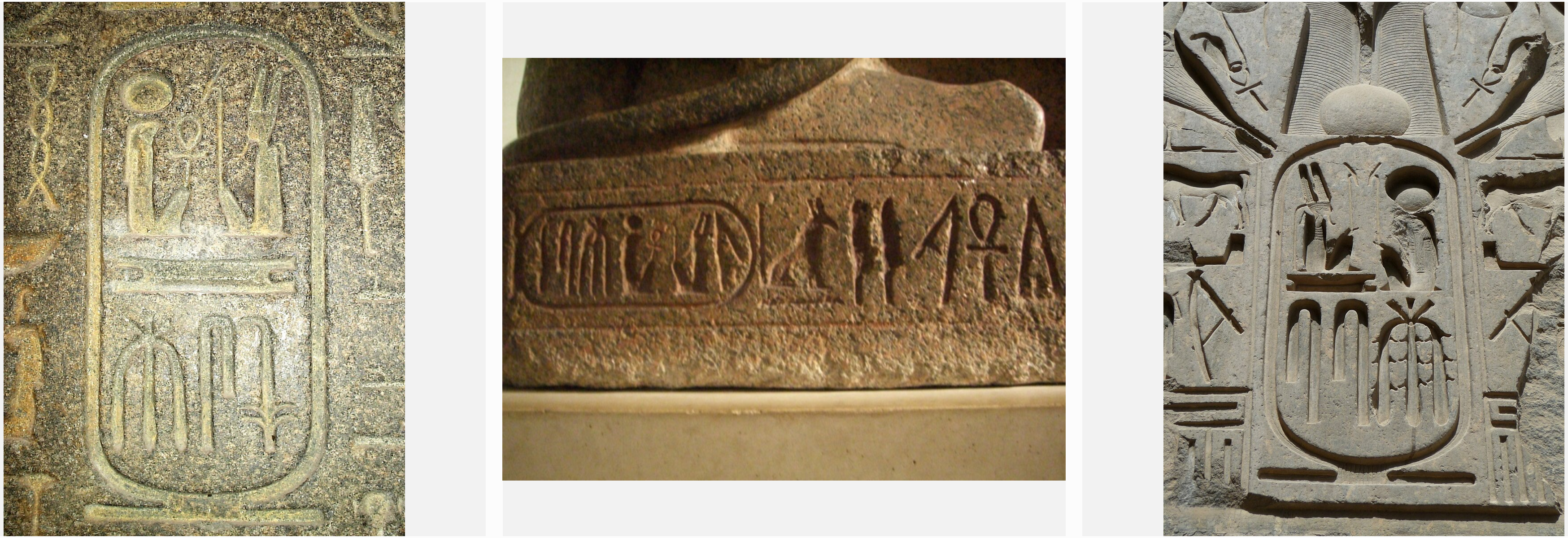
Hieroglyphs – Some Issues

- A symbol might belong to more than one category
- Phonemic values might be repeated within a word
- There were no end-of-word or end-of-sentence markers

Hieroglyphs – Some Issues

- A symbol might belong to more than one category
- Phonemic values might be repeated within a word
- There were no end-of-word or end-of-sentence markers
- Text was written in any direction

An Example of Text Directionality – Cartouches



Cartouche of Ramses II, taken from Wikimedia: [1], [2], and [3]

Hieroglyphs – Some Issues

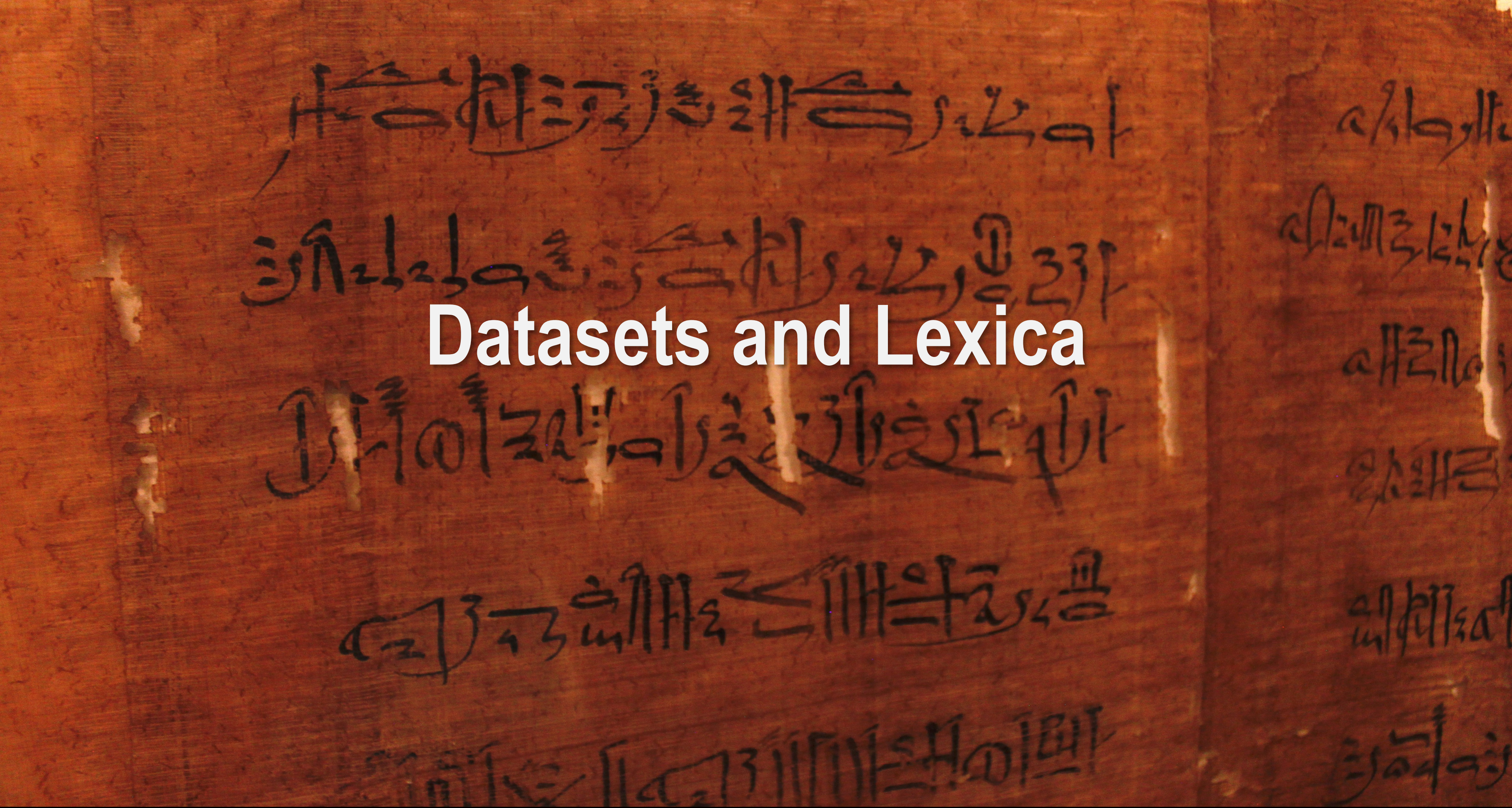
- A symbol might belong to more than one category
- Phonemic values might be repeated within a word
- There were no end-of-word or end-of-sentence markers
- Text was written in any direction
- Scribes were artists and valued the aesthetics of their work

Deir al-Medina – Tomb of an Artisan

But What About Coptic?

- We have been putting Coptic to the side
- It underwent several big changes from Middle and Late Egyptian
  - Uses an alphabet (i.e. no ideograms)
  - Has no repetition of phonemes
  - Has a lot of loanwords from Ancient Greek
  - Has morphological and grammatical variation
- Different fields tend to focus on it (e.g. theology)

Datasets and Lexica



Smaller Corpora & Lexica

- Nederhof and Rahman (2015)
  - Middle Egyptian
  - Hieratic transliteration
  - Contains two texts, with the script being linearized

- The Chicago Demotic Dictionary
  - A lexicon maintained and updated from 1972 to 2012
  - Contains both words and scans of the documents where they originally came from

Multilingual Corpora

- OPUS Collection
  - Contains parallel data for translation
  - One of the languages included is Coptic
- Nordhoff and Krämer (2022)
  - Dataset with morpheme annotation for several low-resource languages
  - Includes examples in Middle and Late Egyptian, as well as Coptic
  - They do not mention the number of examples for the different languages

The Ramses Project

- Began in 2008, last updated in 2016
- It aims to collect all available texts in Late Egyptian (c. 1350 – 700 BC)
- They started with the most relevant texts for studying the language and culture
- Contains a bevy of annotations

The Ramses Project

- Annotations
  - Inflections
  - Lemmata
  - Spellings
- Metadata
  - Where the text comes from
  - State of preservation of the documents
  - Alterations or editings of the texts
- Other
  - Comments or criticisms on the annotator’s choices

Thesaurus Linguae Aegyptiae (TLA)

- Released in 2004 and updated until 2012
- Texts ranging from the Old Kingdom to Roman times
  - Contains text in Old, Middle, and Late Egyptian, as well as Demotic and Coptic
  - It includes the oldest pyramid texts
- Only has lemmatization and morpho-syntactic annotation

Thot Sign List (TSL)

- A collection of graphemes that have been attested either in hieroglyphic or hieratic text
- Its first release contains 1,203 signs, 4,842 functions, and 21,834 tokens
- All of the (Unicode) hieroglyphic symbols in this presentation come from there!

Coptic Scriptorium

- Designed to be used to study a wide variety of subjects
- At the time of its release it contained less than 60 thousand manually annotated words
- It currently contains 850 thousand annotated words
- It was created in 2013 and released in 2016

Other Resources for Coptic

- Database and Dictionary of Greek Loanwords in Coptic
  - Contains Coptic lemmas coming from Ancient Greek

- The Marcion Project
  - A lexicon that contains over 11 thousand head words and over 87 thousand items

- Coptic WordNet
  - Created using both of these dictionaries and the Coptic section of the TLA

Coptic Scriptorium

- All three of the resources in the previous slide were incorporated into the Coptic Scriptorium

- It was also used to create a treebank for the Universal Dependencies project

- Its latest version was released in October 2023

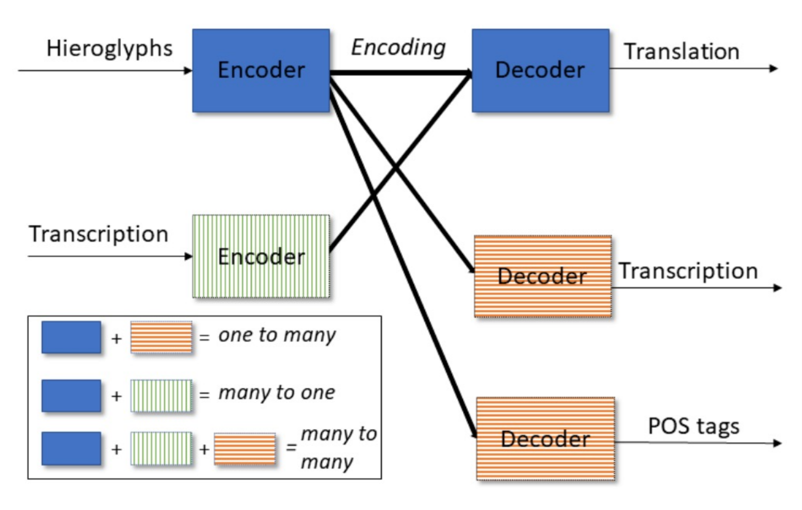
NLP for Middle and Late Egyptian

Transliteration

- The writing system is complex and lacks standardization
- Human transcription is time consuming
- Hieroglyphic encoding still requires more characters
- Thus, transliteration is still an open problem

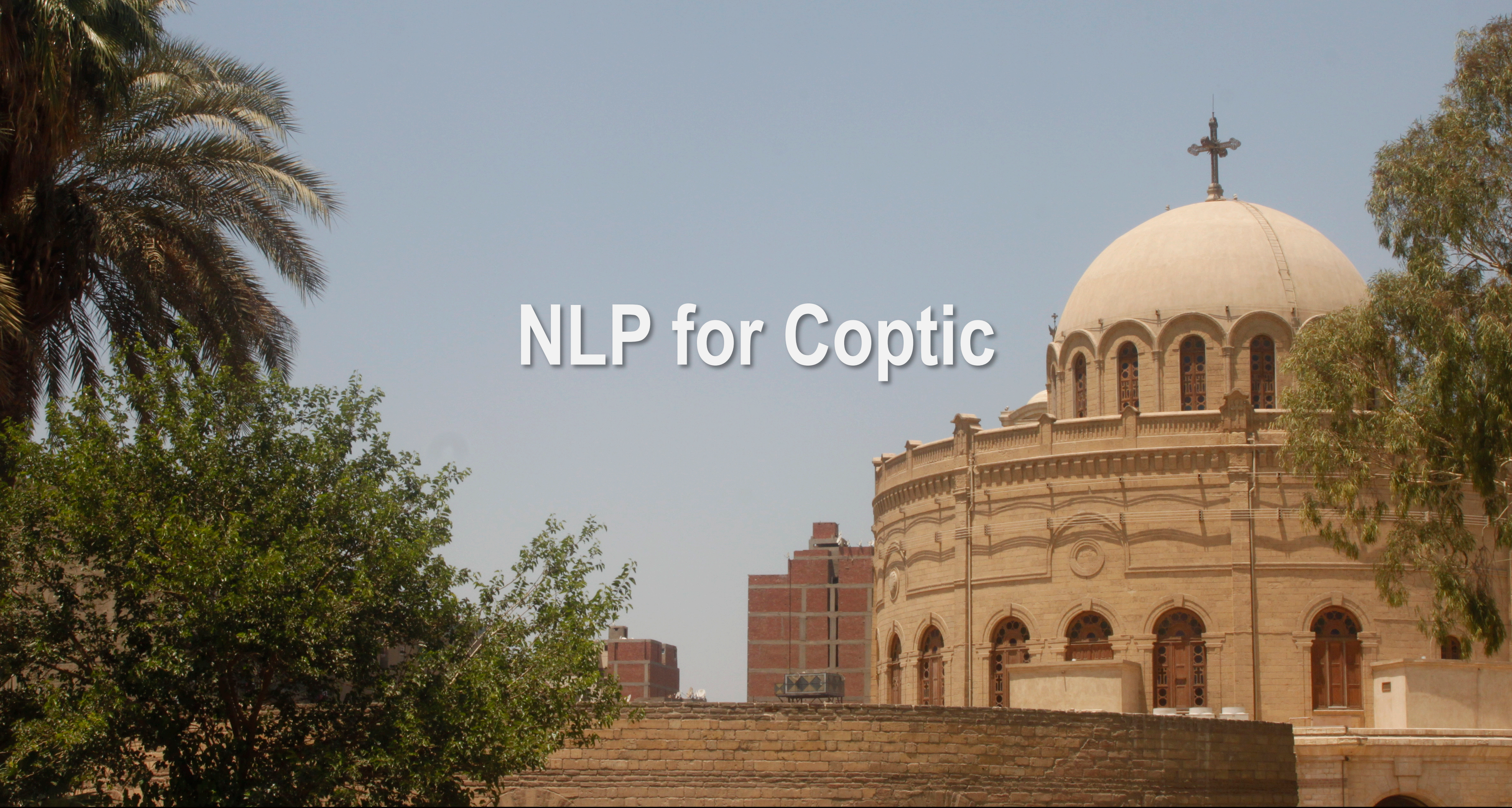
Translation and POS Tagging

- Only one paper deals with both of these tasks and it does them in tandem (Wiesenbach et al., 2019)
- They compare three approaches
- The many to one system was the best performing



Other Tasks

- Text Classification
  - Helps with document organization and management
  - Allows us to give us insights into the different registers used
- Text Retrieval
  - Allows for searching and querying
  - Diminishes manipulation of the original documents
- Semantic Representations
  - Semantic maps allow us to study variation through time

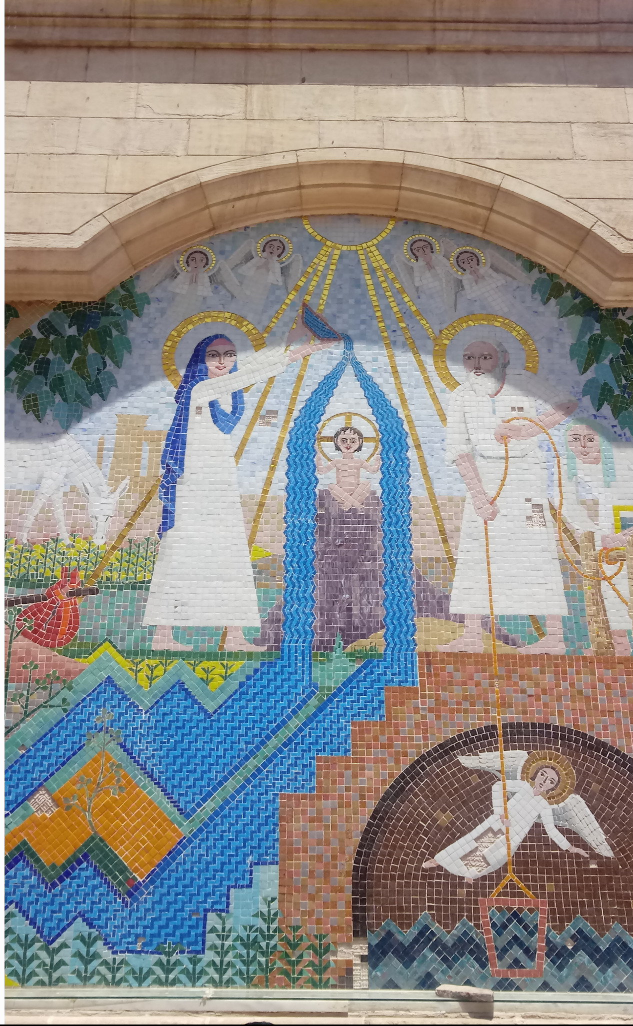


NLP for Coptic

NLP for Coptic

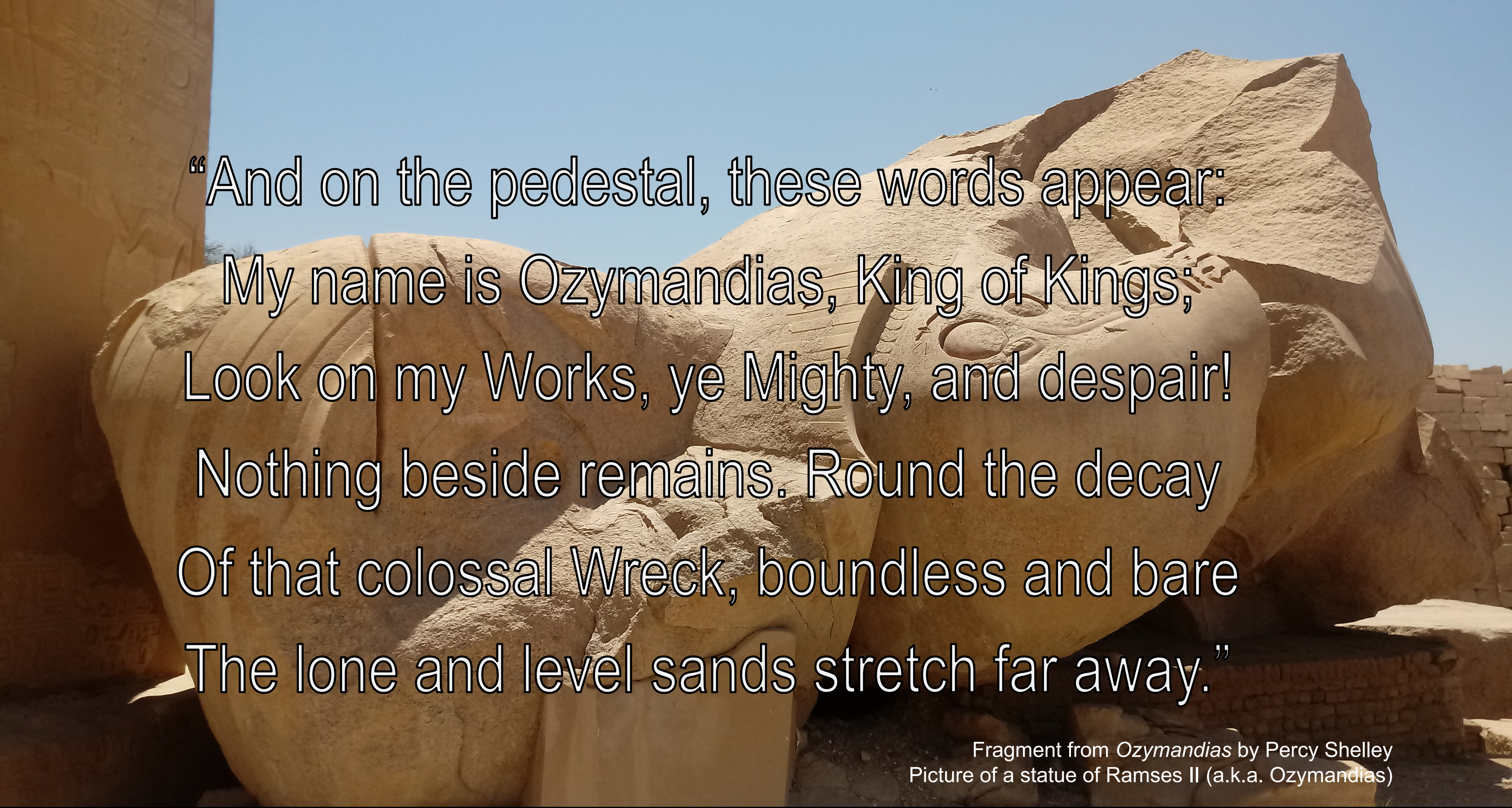
- Morphological Analysis
  - Most of the work has focused on Sahidic Coptic
  - The idea is to focus either on the didactical part or on more theoretical models

- Named Entity Recognition
  - The idea is to identify named entities in Coptic versions of the Bible



UD for Coptic

- There is an UD treebank for Coptic
  - The treebank scores well in most quality metrics for treebanks (Kulmizev and Nivre, 2023)
  - However, it ranks low for variability
- It was used to develop a pipeline for NLP analysis
  - This pipeline is available as part of the Coptic Scriptorium project
- It has also been used as part of other works



“And on the pedestal, these words appear:
My name is Ozymandias, King of Kings;
Look on my Works, ye Mighty, and despair!
Nothing beside remains. Round the decay
Of that colossal Wreck, boundless and bare
The lone and level sands stretch far away.”

Fragment from *Ozymandias* by Percy Shelley
Picture of a statue of Ramses II (a.k.a. Ozymandias)

GÖTEBORGS
UNIVERSITET

SPRÅKBANKENTEXT

Ricardo Muñoz Sánchez
ricardo.munoz.sanchez@svenska.gu.se
rimusa.github.io