

Artificial linguists

Large language models analyzing language data

Robert Östling

Department of Linguistics
Stockholm University

2024-10-18



Stockholm
University

Traditional computational typology

- We have long used computer processing of parallel texts to investigate linguistic variation
- Language properties leave statistical traces in parallel texts
- Methods are relatively crude



Example

Which language has the most elaborate case system?

Word forms in the Bible occurring where Jerusalem is mentioned

English	Finnish		
Jerusalem	Jerusalem	Jerusalemia	Jerusalemiin
	Jerusalemille	Jerusalemilta	Jerusalemiin
	Jerusalemissa	Jerusalemista	



The next step?

Let's jump on the LLM (Large Language Model – ChatGPT and its kind) bandwagon and see where it leads us!

Why does it seem promising?

- Major LLMs are (probably?) trained on the collected works of linguistics, give or take
- They have showed an amazing ability to solve problems in programming, mathematics, physics, etc.
- Recently also in linguistics (Linguistics Olympiad problems, translation based on grammar + word list)



- Bible translations into some 1500 languages
- Parallel text between familiar languages and under-described ones
- Linguists (including students; Svärd, 2013) can analyze a language based on this material alone
- Can LLMs do the same?



Case study: Kuot

- Kuot is a language isolate spoken by about 1500 people in New Ireland (Lindström, 2002).
- LLMs are unlikely to have seen significant amounts of Kuot text.
 - This can be tested: give them a short exam on Kuot knowledge.
- We use that to test their ability to analyze previously unseen languages.



First attempt, iterate the following on a set of sentences:

- Initial glossing comes from traditional word alignment
- Sample a glossing from GPT-4o for each sentence, conditioned on the current glossing of the other sentences
 - Note the similarity to Gibbs sampling
- One English and one Swedish translation given per sentence

Iteration 0

Urie , lesu la ula bo pugama Oliva .

? , ? ? ? ? hill/mountain ? .

'but Jesus went to the Mount of Olives '

First attempt, iterate the following on a set of sentences:

- Initial glossing comes from traditional word alignment
- Sample a glossing from GPT-4o for each sentence, conditioned on the current glossing of the other sentences
 - Note the similarity to Gibbs sampling
- One English and one Swedish translation given per sentence

Iteration 1

Urie , lesu la ula bo pugama Oliva .

but? , Jesus ? go ? mountain Olives .

'but Jesus went to the Mount of Olives .'

First attempt, iterate the following on a set of sentences:

- Initial glossing comes from traditional word alignment
- Sample a glossing from GPT-4o for each sentence, conditioned on the current glossing of the other sentences
 - Note the similarity to Gibbs sampling
- One English and one Swedish translation given per sentence

Iteration 2

Urie , lesu la ula bo pugama Oliva .

but? , Jesus then? go to? mountain Olives .

'but Jesus went to the Mount of Olives .'

First attempt, iterate the following on a set of sentences:

- Initial glossing comes from traditional word alignment
- Sample a glossing from GPT-4o for each sentence, conditioned on the current glossing of the other sentences
 - Note the similarity to Gibbs sampling
- One English and one Swedish translation given per sentence

Iteration 3

Urie , lesu la ula bo pugama Oliva .

but? , Jesus ? go to mountain Olives .

'but Jesus went to the Mount of Olives .'

How did it go?

Iteration 3

Urie , lesu la ula bo pugama Oliva .

but? , Jesus ? go to mountain Olives .

'but Jesus went to the Mount of Olives .'

Reference

U-rie , lesu la u-la bo pugama Oliva .

3F.SG-DEM , Jesus RELR 3M-go on hill Olives .

'but Jesus went to the Mount of Olives .'



Challenges ahead

- Managing over-confidence and outright hallucinations
- Working with larger amounts of data
- Simulate a human workflow of iterative hypothesis making, data searching, word list construction, morphological analysis, etc.





Echoes of History:

*Analysis and Decipherment of
Historical Writings*

New Approaches to Analyzing Rare and Unknown Scripts

Beáta Megyesi

2024-10-18

Department of Linguistics

Introduction

Motivation

- Importance of historical sources for understanding the past.
- Difficulty in analyzing rare writing systems.
- Individual efforts on a single type of sources.



(Illustration by pbs.org, 2008)



(Illustration by Desset et al., 2022)



(Illustration by Google DeepMind, 2022)

Challenge

- Full analysis require a wide range of expertise.
- Current tools based on AI do not work on small and/or rare datasets.



Linear B.



A Sabaic South Arabian inscription.



Trilingual Cuneiform inscription.



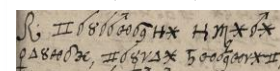
The Phaistos disk.



Codex Runicus.



The Voynich manuscript.



The Borg cipher.



The Ramanacoli cipher.

Purpose of the DESCRYPT Program

Objective

To build a gateway to digital humanities and digital philology, i.e. to develop historical writing research by AI-driven tools for augmented analysis and decipherment.

Goals

- Create a digital corpus of rare/unknown writings.
- Develop recognition models for alphabets and scripts incl. document layout analysis, symbol recognition, and transcription.
- Build an interpretative framework for linguistic and historical analysis by decipherment.



(Illustration by BBC 4, 2024
The Secret History of Writing)



(Illustration by Transcribus, 2023
readcoop.eu)

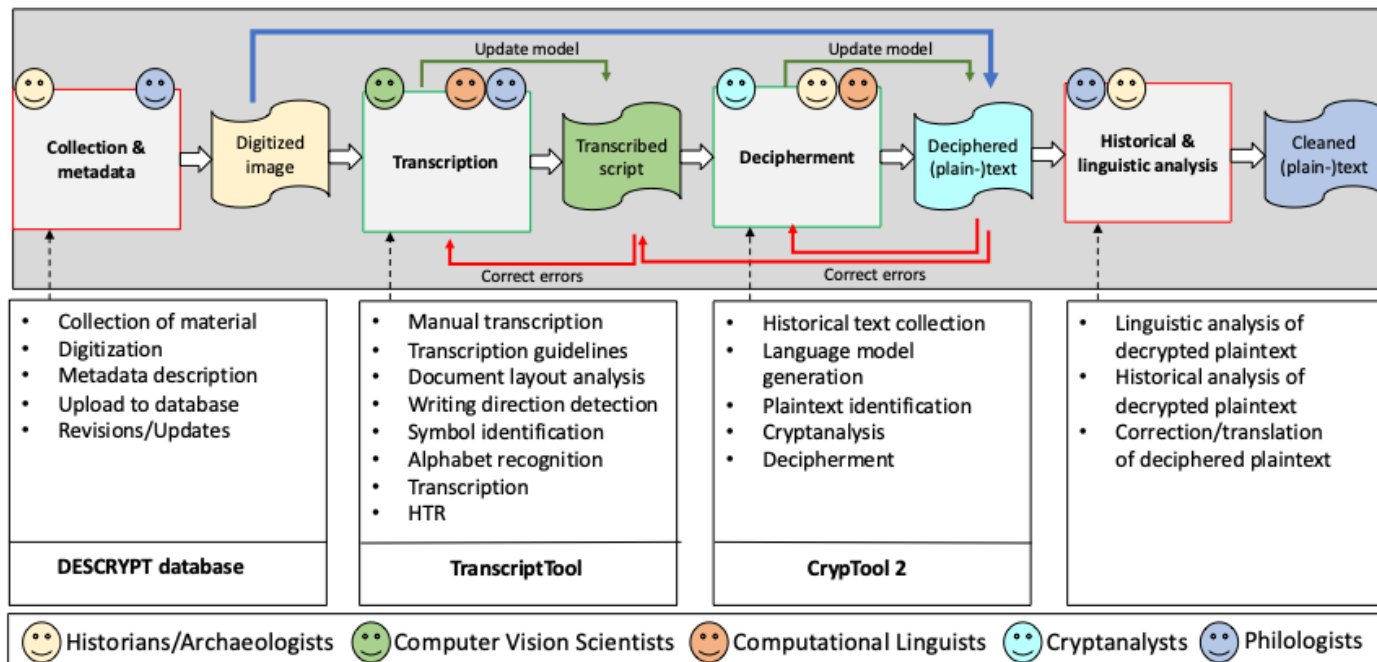
Methodology Overview

Tools

R&D of AI-based transcription and decipherment tools

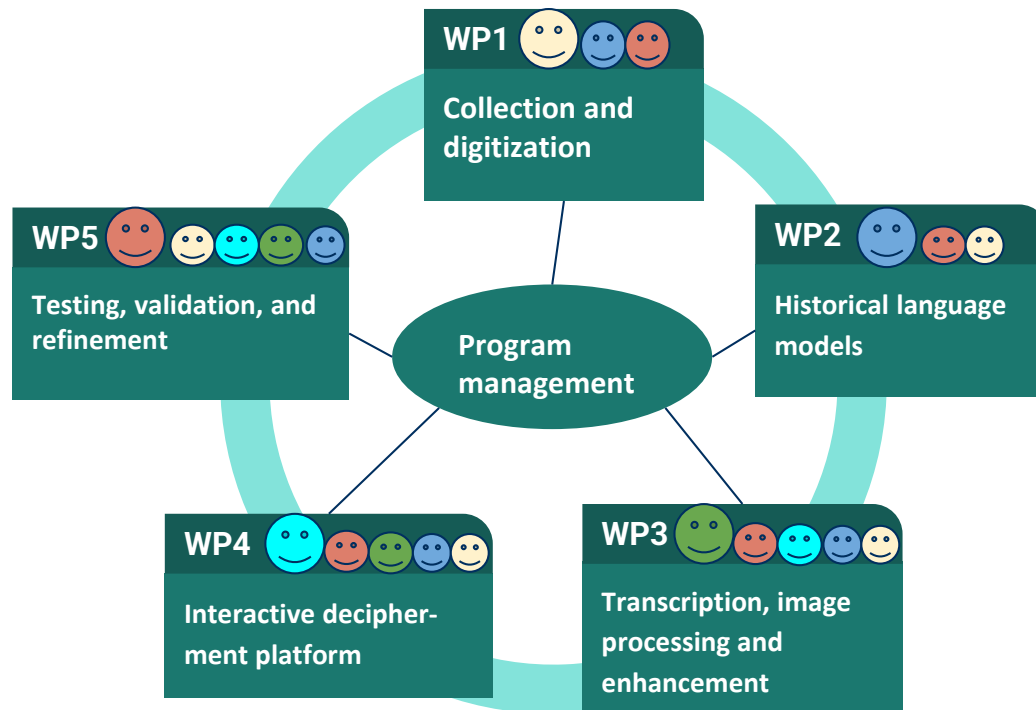
Approach

Interdisciplinary,
collaborative effort



Program Description

5 work packages
running in
parallel



Participants



Benedek Láng
WP1 history



Michelle Waldispühl
co-PI, WP2
historical linguistics



Beáta Megyesi
PI, NLP



Mihály Héder WP5
computer science
system architect



Alicia Fornés WP3
computer vision



Nils Kopál WP4
cryptology



Rune Rattenborg
archeology



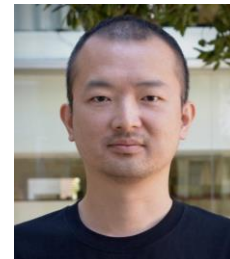
Eva Pettersson
NLP



archeologists,
historians,
linguists,
librarians,
cryptologists, ...



Raphaëla Heil
computer vision
system architect



Lei Kang
computer vision



Vasily Mikhalev
deep learning,
cryptology

Conclusion

The Echoes of History: Analysis and Decipherment of Historical Writings unites cross-disciplinary expertise to interpret historical sources written in rare/unknown scripts.

- **Tools:** Scalable AI-driven framework for historical text analysis and an interactive platform to serve experts
- **Resources:** a digital corpus of historical writings in a standardized format and a historical corpus
- **Knowledge:** New insights into historical languages and AI for humanities for expert interaction
- **Impact:** Contribution to multiple academic fields and preservation efforts of our cultural heritage



(Illustration by <https://artificialpaintings.com/>:
How to Use AI to Explore Historical Data, 2024)



Questions?

Echoes of History:

*Analysis and Decipherment of
Historical Writings*

Research infrastructure for language data in the Swe-CLARIN project

Mats Wirén
Department of Linguistics
Stockholm University

18 Oct. 2024

Research infrastructure

- Typically
 - **Annotated corpora**, whether for text, speech or sign data
 - **Tools** for processing language data
 - Scarce resources!

Case in point: Sign language

- Base-level annotation: ~100 times real time
 - **Glosses** (signs): ~60 times real time
 - **Non-manual signals** (prosody)
 - **Translation** (for example, to Swedish)
- Automatic annotation is not yet mature

Annotated corpora

- **Text:** 10^9 (Korp: ~17 billion word tokens)
 - Lexical and syntactic annotation
- **Sign:** 10^5 (STS: ~200 000 sign tokens)
 - Base-level annotation (glosses, etc.)

*Research in sign language is hampered
by lack of annotated corpora*

What is Swe-CLARIN?

- Part of a national infrastructure consortium for language technology
 - Funded by VR since 2014
 - People at SU: Mats Wirén, Johanna Mesch, Beáta Megyesi, Robert Östling, Gustaf Gren, Moa Gärdenfors

Some past activities

- First part-of-speech annotation
 - Östling, Börstell & Wallin (2015)
- First syntactic annotation (= segmentation of clauses, argument structure)
 - Börstell, Wirén, Mesch & Gärdenfors (2016)
 - Östling, Börstell, Gärdenfors & Wirén (2017)

On-going

- Incorporation of pose information in the STS corpus
 - Gustaf Gren,
Robert Östling
 - Step towards
rationalizing
sign-language
annotation



The good news

- Swe-CLARIN got continued funding for 2025–28
- For the first time, earmarked funding for research infrastructure for sign language
 - Recordings of more informants, annotation of old and new data

So much for sign language...

- We also do text!
- More information on Swe-CLARIN:
<https://www.su.se/english/research/research-projects/swe-clarin>