Decipherment of Historical Manuscripts with Unknown or Rare Writings: The DESCRYPT Project

Beáta Megyesi[1], Alicia Fornés[2], Mihály Héder[3], Raphaela Heil[1], Benedek Láng[4], Nils Kopal[5], Rune Rattenborg[6] and Michelle Waldispühl[7]
[1]Stockholm University, Sweden [2]Universitat Autònoma de Barcelona, Spain
[3]Technical University of Budapest, Hungary [4]Eötvös Loránd University, Hungary
[5]Hochschule Niederrhein, Germany, [6]Lund University, Sweden, [7]University of Oslo, Norway

Abstract

We present a newly funded research program, DESCRYPT, aimed at deciphering and analyzing historical texts with rare or unknown scripts. The project leverages advancements in computational linguistics, artificial intelligence (AI), and image processing, alongside traditional philological methods, to develop innovative tools for transcription, recognition, and interpretation of historical writings with rare/unknown scripts, including ciphertexts. By integrating interdisciplinary expertise, DESCRYPT addresses the challenges posed by complex and undeciphered texts, preserving and unlocking the secrets of our shared cultural heritage.

1 Introduction

Historical texts have long served as windows into the cultural, intellectual, and linguistic landscapes of human civilization. These writings provide unique perspectives on the evolution of languages, belief systems, and societal structures. However, analyzing such texts—particularly those written in rare or undeciphered scripts—poses significant challenges. Traditional philological methods, relying on meticulous manual examination and contextual expertise, are often insufficient when faced with obscure linguistic systems or fragmented historical records.

In recent decades, digital humanities have introduced new opportunities for text analysis through computational tools like 3D modeling, optical character recognition (OCR), hand-written text recognition, topic modeling and other content-based text analytics based on machine learning algorithms. Despite their promise, these technologies struggle with the irregularities that characterize many historical texts, such as degraded materials, non-standard scripts, and irregular and cryptic linguistic structures. Unknown writings share striking similarities with ciphers in that they encode meaning through complex symbols or systems that require decipherment to be understood.

Building on the success of the DECRYPT project (Megyesi et al., 2020), which focused on European early modern cryptographic texts, DESCRYPT expands its scope to embrace a wider range of historical writings. By integrating computational linguistics, AI, cryptanalysis, and traditional linguistic, historical and archeological expertise, DESCRYPT seeks to overcome these challenges and establish innovative methodologies for transcribing, deciphering, and contextualizing rare scripts. This endeavor represents a vital step toward preserving and interpreting the hidden layers of human history.

2 Previous Work

The decipherment of historical writing systems with rare or unknown scripts has evolved significantly, from early breakthroughs to modern computational approaches. Pioneers such as Jean-François Champollion (Champollion, 1822) and Michael Ventris with John Chadwick (Chadwick, 1958) laid the foundation for understanding ancient texts, demonstrating key methodologies that remain influential today.

The field has since integrated computational linguistics, enhancing traditional methods through digital tools. The expansion of digital humanities has led to extensive online repositories cataloging inscriptions across various civilizations, including Egyptian hieroglyphs, Linear B, Runic inscriptions, Ancient South Arabian scripts, Greek and Latin epigraphy, and cuneiform texts. However, challenges persist in data standardization, digitization, and large-scale comparative analysis.

In addition, encrypted manuscripts constitute another category of historical texts that remain

Proceedings of the 8th International Conference on Historical Cryptology, pages 101-105
Poznań, Poland, 16-18 June, 2025

largely undeciphered. Historical cryptology, outlined by David Kahn (Kahn, 1967), has progressed significantly through computational techniques.

Advancements in computational cryptanalysis have enabled breakthroughs in deciphering various encoded texts. Algorithmic approaches have been successfully applied to historical ciphers such as the Zodiac ciphers, the Copiale cipher (Knight et al., 2012), the papal ciphers of the Vatican (Lasry et al., 2020) and the recently deciphered Mary Stuart letters (Lasry et al., 2023). Machine learning has further expanded the potential for decoding unknown scripts, as demonstrated in studies on the Borg cipher (Aldarrab, 2017), anagram-based sources (Hauer and Kondrak, 2016), infilling text in ancient tablets (Papavassileiou et al., 2023) and restoring text along with attributing geographic area and dating of ancient Greek inscriptions (Assael et al., 2022). However, despite progress, computational methods remain limited in fully automating the decipherment of unknown scripts. Instead, interdisciplinary collaboration continues to play a crucial role, as seen in the recent decipherments of Linear Elamite (Desset et al., 2022), the Maya code (Coe, 2011), Amorite vocabulary (George and Krebernik, 2022), and Kushan script (Bonmann et al., 2023).

Automatic script recognition through artificial intelligence has further transformed the field. Techniques such as cluster-based script identification and digital image processing have enhanced manuscript analysis. However, current Handwritten Text Recognition (HTR) tools, such as Transkribus [1] and eScriptorium [2], require extensive labeled training data, rendering them ineffective for rare scripts with limited textual records. Additionally, uncommon scripts often lack Unicode representation, complicating their digital processing.

The importance of interdisciplinary collaboration has gained recognition, particularly in projects such as DECRYPT (Megyesi et al. 2020), which integrates expertise from computational linguists, cryptologists, computer vision specialists, philologists and historians to digitize encrypted documents and develop transcription and cryptanalysis tools. These efforts underscore the necessity of combining traditional scholarship with digital advancements to unravel historical texts.

[1] https://www.transkribus.org

[2] https://escriptorium.inria.fr

Despite persistent challenges, the integration of computational techniques with historical expertise continues to advance the field of decipherment, offering new opportunities for understanding ancient and encoded texts.

3 Objectives

The overarching goal of the DESCRYPT program is to bridge the gap between traditional philological methods and state-of-the-art computational tools. The program is guided by several key objectives:

- **Developing Decipherment Techniques:** Innovate methods for analyzing and interpreting rare scripts, moving beyond manual or semi-automatic approaches.
- **Creating a Digital Corpus:** Assemble a repository of digitized rare scripts, enriched with standardized metadata for preservation and accessibility.
- **Designing Recognition Models:** Develop algorithms capable of transcription and analysis of undeciphered writing systems, ensuring scalability and adaptability.
- **Enhancing User-Centric Research Tools:** Build robust, user-friendly AI tools that integrate feedback from experts, enabling continuous improvement in transcription and analysis accuracy.
- **Fostering Interdisciplinary Collaboration:** Unify expertise from linguistics, cryptanalysis, computer vision, history, and archaeology to create a comprehensive research framework.
- **Preserving Cultural Heritage:** Make rare and historically significant texts accessible and comprehensible, contributing to the understanding of human civilization.

4 Research Questions

The research is framed around questions that address both methodological and technological challenges:

- How can innovative methodologies enhance the analysis and decipherment of rare and unknown scripts?

- What image processing techniques can be adapted to automate the transcription of historical writings across diverse writing systems?
- What strategies can ensure systematic and consistent transcription of symbols from varying notational styles?
- How can interactive platforms incorporate user feedback to refine transcription and decipherment tools?
- What improvements can be made to historical language models to facilitate accurate script identification and interpretation?

These questions guide the program's research design, ensuring that theoretical insights and practical applications are seamlessly integrated.

5 Methodology

The DESCRYPT framework incorporates advanced technologies and traditional expertise to systematically address rare scripts. Its key methodological components include collection, transcription, decipherment and historical and linguistic analysis, as illustrated in Figure 1.

The DESCRYPT program is structured into five work packages (WPs), following a pipeline from the collection of historical sources with rare or unknown scripts to their transcription, decipherment and linguistic and historical analysis. Each component plays an essential role in advancing decipherment techniques through interdisciplinary collaboration.

WP1: Collection and Digitization The first phase focuses on identifying, gathering, and digitizing historical sources from archives, libraries, catalogs, and private collections. High-resolution imaging techniques are used to create digital copies ensuring the best possible preservation and analysis. The sources are then annotated with metadata using the TEI XML standard. To make them accessible for further study, metadata is added in compliance with TEI XML (Consortium, 2023) standards and Linked Open Data (LOD) (Gaitanou et al., 2022), enabling structured documentation and retrieval.

The aim is to create a comprehensive, searchable database integrating rare and undeciphered scripts, including ancient writing systems (e.g., Linear A and B, the Phaistos Disk), historical shorthand systems, early modern artificial language schemes, and encrypted texts. The collection ensures systematic access to digital reproductions of rare scripts while linking to existing scholarly resources.

WP2: Historical Language Models This phase develops historical language models for script recognition, transcription, and interpretation. It involves collecting and standardizing diplomatic transcriptions of historical languages, training models for language identification, and refining them through integration with image processing and cryptanalysis techniques. The corpus includes languages such as Akkadian, Ancient Greek, Classical Arabic, Latin, Old Norse, and Sanskrit. Addressing challenges like non-standard spellings, dialectal variations, and code-switching, the project aims to optimize algorithms for linguistic analysis and transcription correction. One of the key challenges in this process is handling non-standard spellings, dialectal variations, and instances of code-switching, which complicate linguistic modeling and text interpretation.

WP3: Transcription and Image Processing Transcription and image processing play a crucial role in decipherment. Digital images undergo preprocessing techniques to enhance readability, making inscriptions and scripts more legible. Additionally, layout analysis, text segmentation, and symbol recognition algorithms are implemented to automate transcription and streamline the decoding process. To improve readability and transcription accuracy, this phase develops advanced image processing techniques. Methods for preprocessing, document layout analysis, text segmentation, and symbol recognition are implemented to automate transcription for various writing systems. Given the limitations of current handwritten text recognition (HTR) models for rare scripts, a hybrid approach combining computational recognition with user input is developed based on the CTTS tool (Lasry et al., 2023). The outputs include standardized transcription methods, enhanced image analysis tools, and a user-friendly transcription system that can adapt to new sources.

WP4: Interactive Decipherment Platform A critical component of DESCRYPT is the development of an interactive platform that facilitates user engagement in decipherment. The platform integrates AI-driven cryptanalysis, allowing users

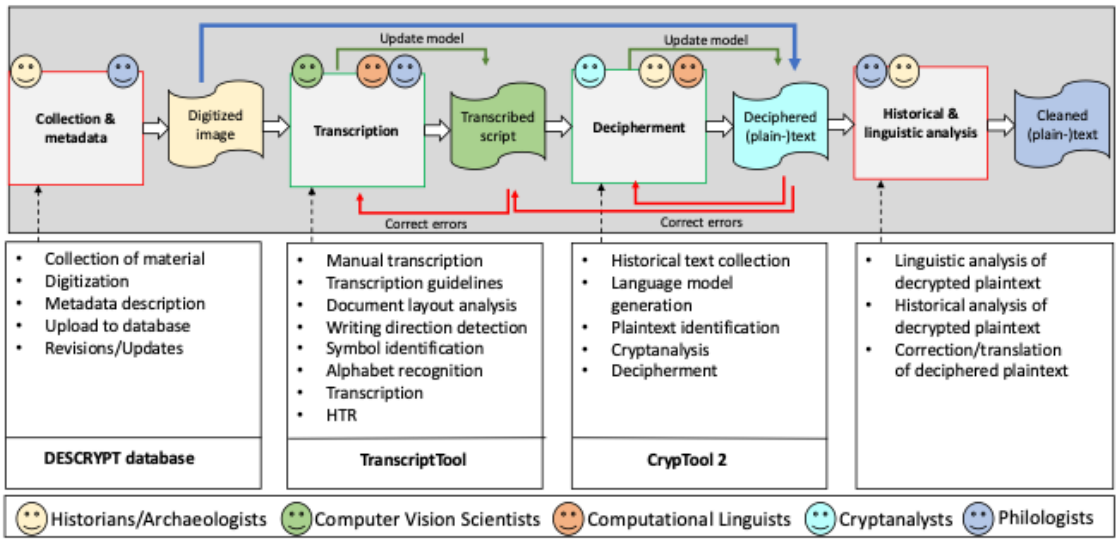

Figure 1: DESCRYPT: methodological overview.

to input corrections, enhancing transcription and decipherment accuracy. Drawing on historical language models and cryptographic algorithms, the platform balances automated analysis with expert insights, allowing users to contribute corrections, thereby improving the accuracy and efficiency of decipherment. The goal is to create a semi-automatic system that combines computational tools with human expertise to decipher rare and encrypted scripts more effectively.

WP5: Testing, Validation, and Refinement The final phase involves testing the developed tools on diverse scripts, evaluating their effectiveness, and refining them based on user feedback. A user-friendly interface and a structured framework ensure accessibility and adaptability. Large-scale experiments validate the reliability of transcription and decipherment models, while continuous improvements enhance the overall system. Large-scale experiments are conducted to assess the scalability and effectiveness of the approaches, ensuring that they can be applied to a wide range of historical texts and encoded scripts. Diverse scripts are used to validate the methodologies, and user feedback is collected to refine processes. Evaluations of interdisciplinary collaboration help identify best practices for integrating computational methods with traditional philological research.

DESCRYPT's success hinges on collaboration across multiple disciplines. Computational linguistics plays a key role in developing historical language models, ensuring accurate representation of rare scripts. Experts in history and philology contribute contextual knowledge and linguistic expertise, aiding in the interpretation of texts. Paleographers provide critical insights into historical handwriting styles, script evolution, and scribal practices, which are essential for the accurate reading and dating of manuscripts. Computer vision specialists enhance image processing and text recognition capabilities, making inscriptions and manuscripts more accessible for study. Cryptologists apply advanced cryptographic techniques to assist in deciphering encoded and unknown scripts. By bringing together leading researchers in archeology, history, linguistics, paleography, computer science, and cryptanalysis, the program fosters a balanced interdisciplinary approach, enabling more effective analysis and interpretation of historical texts to develop a holistic framework for deciphering rare and unknown writing systems.

6 Conclusion

DESCRYPT pioneers a transformative approach to historical text analysis by combining computational methods with traditional expertise. Its interdisciplinary framework aims to contribute to the study of rare and unknown scripts, preserving cultural heritage and enhancing our understanding of human history. By advancing methodologies for transcription, recognition, and interpretation, DESCRYPT paves the way for future innovations in the humanities and beyond. The program's outcomes will benefit multiple disciplines, including

linguistics, history, and computer science, while providing tools for the broader academic community. Ultimately, DESCRYPT demonstrates the power of interdisciplinary collaboration in uncovering the secrets of the past and safeguarding them for future generations.

Acknowledgments

This work has been supported by Riksbankens Jubileumsfond, grant M24-0028: Echoes of History: Analysis and Decipherment of Historical Writings (DESCRYPT).

References

Nada Aldarrab. 2017. Decipherment of historical manuscripts. Master's thesis, University of Southern California.

Yannis Assael, Thea Sommerschield, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Marita Chatzipanagiotou, Ion Androutsopoulos, Jonathan Prag, and Nando de Freitas. 2022. Restoring and attributing ancient texts using deep neural networks. *Nature*, 603:280–283.

Svenja Bonmann, Jakob Halfmann, Natalie Korobzow, and Bobomullo Bobomulloev. 2023. A partial decipherment of the unknown kushan script. *Transactions of the Philological Society*, 121:293–329.

John Chadwick. 1958. *The Decipherment of Linear B*. Cambridge University Press.

Jean-Francois Champollion. 1822. *Lettre à M. Dacier relative à l'alphabet des hiéroglyphes phonétiques*.

Michael D. Coe. 2011. *Breaking the Maya Code*. Thames Hudson Ltd, 3rd edition.

The TEI Consortium, 2023. *Guidelines for Electronic Text Encoding and Interchange P5 Version 4.7.0*. Last updated on 16th November 2023.

Francois Desset, Kambiz Tabibzadeh, Matthieu Kervran, Gian Pietro Basello, and Gianni Marchesi. 2022. The decipherment of linear elamite writing. *Zeitschrift für Assyriologie und vorderasiatische Archäologie*, 112(1):11–60.

Panorea Gaitanou, Ioanna Andreou, Miguel-Ángel Sicilia, and Emmanouel Garoufallou. 2022. Linked data for libraries: Creating a global knowledge space, a systematic literature review. *Journal of Information Science*, 50:204 – 244.

Andrew George and Manfred Krebernik. 2022. Two remarkable vocabularies: Amorite-Akkadian bilinguals! *Revue d'assyriologie et d'archéologie orientale*, 116:113–166.

Bradley Hauer and Grzegorz Kondrak. 2016. Decoding anagrammed texts written in an unknown language and script. In *Transactions of the Association for Computational Linguistics*, volume 4, pages 75–86. MIT Press.

David Kahn. 1967. *The Codebreakers*. New York, 2nd edition.

Kevin Knight, Beáta Megyesi, and Christiane Schaefer. 2012. The copiale cipher. In *ACL Workshop on Building and Using Comparable Corpora (BUCC)*.

George Lasry, Beáta Megyesi, and Nils Kopal. 2020. Deciphering papal ciphers from the 16th to the 18th century. *Cryptologia*, pages 479–540.

George Lasry, Norbert Biermann, and Satoshi Tomokiyo. 2023. Deciphering Mary Stuart's lost letters from 1578–1584. *Cryptologia*.

Beáta Megyesi, Bernhard Esslinger, Alicia Fornés, Nils Kopal, Benedek Láng, Georg Lasry, Karl de Leeuw, Eva Pettersson, Arno Wacker, and Michelle Waldispühl. 2020. Decryption of historical manuscripts: the decrypt project. *Cryptologia*.

Katerina Papavassileiou, Dimitrios Kosmopoulos, and Gareth Owens. 2023. A generative model for the mycenaean linear b script and its application in infilling text from ancient tablets. *Journal on Computing and Cultural Heritage*, 16.