The DECRYPT Project: Cross-Disciplinary Research for Historical Cryptology

Beáta Megyesi
Department of Linguistics
Stockholm University
´ beata.megyesi@ling.su.se

Abstract

This paper describes the DECRYPT project, which aims to create a research infrastructure for historical cryptology through cross-disciplinary collaboration. The project has curated an extensive collection of encrypted historical manuscripts, including ciphertexts and cipher keys, and has developed a suite of tools for automated cipher analysis. Major components include the cataloging of documents in the DECODE database, the transcription of texts using both manual and AI-supported approaches, as well as cryptanalysis. The project demonstrates the interplay between technological innovation and the critical contribution of subject matter expertise in the humanities, particularly in historical cryptology.

1 Introduction

Historically, scholars and scientists from various fields—such as history, linguistics, philology, computer science, cryptology, and computational linguistics have worked independently and in an uncoordinated manner to crack individual ciphers, each with their own methods and objectives. While some focus on deciphering and interpreting single ciphers, others develop tools to assist in processing historical sources. Regardless of their disciplinary backgrounds, these researchers often face similar challenges when dealing with encrypted documents. By uniting the expertise of these diverse fields to collect and digitize encrypted sources and develop software tools for automatic or semi-automatic decryption, the DECRYPT project aims to establish historical cryptology as a recognized scientific discipline. This effort includes releasing data from encrypted sources and providing public access to transcription and decryption tools.

Funded by the Swedish Research Council from 2018 to 2024 as part of a special initiative to promote high-quality cross-disciplinary research in Sweden, the project received 29.5 million SEK (approximately 3 million Euros) over six years.

This paper gives a brief summary of the DECRYPT project, its goals, and results based on a decade of research. The content of this paper is based on previous publications by members of the project team, see for example (Megyesi et al., 2020), (Héder and Megyesi, 2022) and (Megyesi et al., 2024) and others on the project webpage *de-crypt.org*.

2 The DECRYPT Project

The purpose of the DECRYPT project is to digitize, analyze, and decipher encrypted historical manuscripts, known as ciphers, and to make them accessible through a web service. The project also aims to develop methods and tools for the automatic transcription, analysis, and decryption of various types of historical ciphers.

To achieve the goals, the project brings together expertise from various disciplines, including history, linguistics, philology, cryptography, image processing, and computational linguistics. Since 2019, the project team worked together to answer all the research questions within the cross-disciplinary team and in various subgroups, with project participants, and associated members. At the time of writing, around 20 people are contributing to the project, including professors, researchers, PhD candidates, students, and research assistants.

The project publications, which include over 80 scientific papers published in journals and peer-reviewed conference volumes since 2018, demonstrate the cross-disciplinary nature of the project, with a large number of co-authored articles across subject boundaries.

In the subsequent sections, we describe the main results of the project based on the publications of the team. The interested reader can find more information on the project website.

3 Ciphers and Cipher Keys

Encrypted historical sources in the form of ciphertexts, (i.e. messages written in code) have been widely used throughout history since the invention of writing. These encrypted materials can be found in archives, libraries, and private collections worldwide. In the DECRYPT project, our focus is on encrypted documents from early modern Europe. These texts served various purposes, such as diplomatic and military communications, messages from secret societies, and private correspondence.

Figure 1 shows examples of four ciphertexts along with two cipher key extracts. The cipher keys define the transformation of the plaintext into ciphertext to encrypt the message by replacing the plaintext elements with codes as specified by the key. The examples demonstrate the wide variety of symbols and writing styles used in encryption during the early modern period in Europe.

To learn about the structure and evolution of ciphers, we analyzed over 1,600 cipher keys from 10 European countries between the 15th and 18th centuries (Megyesi et al., 2022). The study revealed that cipher keys became more secure over time, as evidenced by the increasing complexity of symbols, code lengths, and types used, as well as the expanding size and linguistic diversity of nomenclatures, i.e. plaintext elements larger than individual letters (typically syllables and words). Earlier nomenclatures primarily encoded nouns and named entities (personal and place names), while later ones included a broader range of linguistic units such as morphemes. Over time, digit-based codes became more common, replacing earlier symbol sets and metaphorical codes. Additionally, nomenclatures grew in size and required more structured organization. The study highlights regional variations and the growing complexity of cipher keys over time.

To interpret these intriguing and diverse set of documents, we digitize, transcribe, and decrypt the messages. This can be done manually or, for greater efficiency, the process can be partially or fully automated. We have developed a dedicated pipeline for cipher processing, described next.

4 The DECRYPT Pipeline

The pipeline handles a wide variety of historical ciphers stored in the DECODE database, and include a series of tools for transcription and cryptanalysis before historical and/or linguistic contextualization can take place. Figure 2 illustrates the process.

4.1 The Decode Database

Within the project, over 8,000 encrypted sources – ciphertexts and cipher keys – have been collected, and described by a set of metadata adapted to encrypted sources along with their documents of relevance. They are stored in the DECODE database[1] (Héder and Megyesi, 2022), developed to serve as a main resource for research and development in historical cryptology. Users can access the records through the database, and trusted users are able to revise existing ones, or upload new ones. All information is open source with the exception of images with copyright restrictions.

4.2 Historical Language Models

Another important resource for decryption and plaintext language identification is historical texts. We compiled historical language models based on n-grams incl. uni-, bi-, tri-, four-, and five-grams, models that are frequently used in cryptanalysis of historical sources. The models have been generated from HistCorp[2] (Pettersson and Megyesi, 2018), a collection of historical texts in 17 languages from different centuries, as well as from the Gutenberg project[3], an online library of 70,000 ebooks. The models, available on GitHub[4], have been previously applied to the cryptanalysis of English and German ciphertexts (Megyesi et al., 2023), prior to their integration into CrypTool for historical cryptanalysis.

4.3 Transcription

Once the encrypted source has been digitized, preferably as a high-resolution color image (300-400 dpi), the document can be transcribed either manually or (semi-)automatically. Transcription is a challenging task due to the wide variety of symbols arranged in non-sensical sequences for the human eye. During this phase, the various glyphs and the entire symbol set need to be identified, and each symbol transcribed accurately. Manual transcription is labor-intensive and often requires trained expertise in philology. Therefore, tools that streamline and accelerate the process are essential.

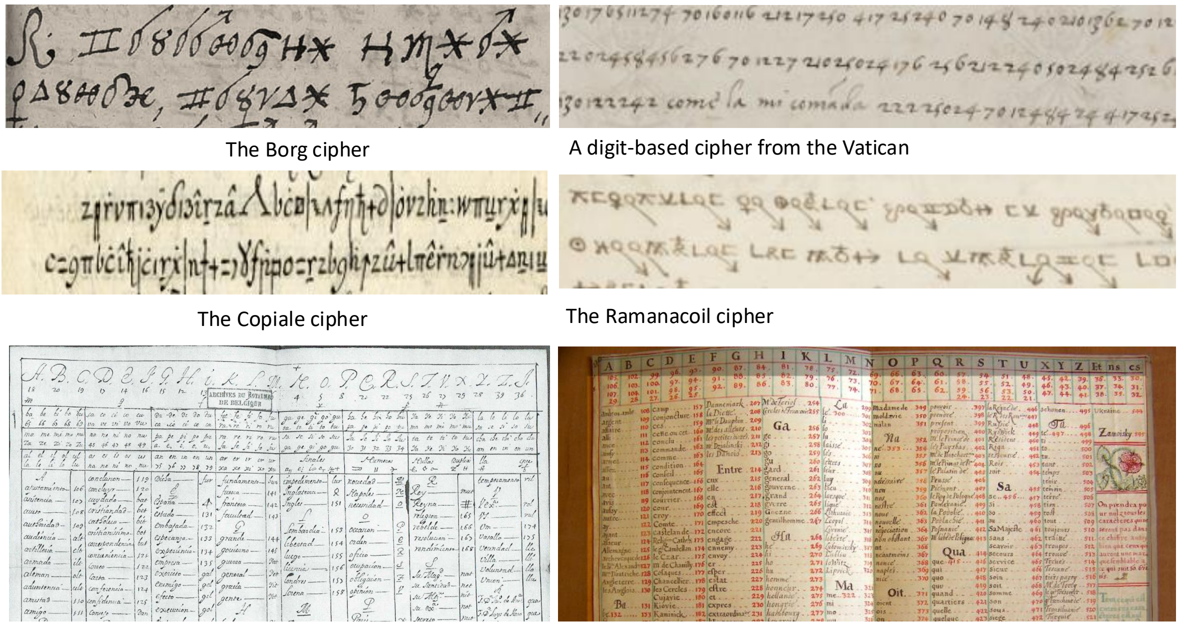
A manual transcription tool, developed by George Lasry, enhances the speed and accuracy of transcription. This tool, called the CrypTool

[1] https://de-crypt.org/decrypt-web/

[2] https://www2.lingfil.uu.se/person/pettersson/histcorp/

[3] https://www.gutenberg.org

[4] https://github.com/CrypToolProject/LanguageStatisticsLibPy



The Borg cipher

A digit-based cipher from the Vatican

The Copiale cipher

The Ramanacoil cipher

Part of a cipher key

Part of a cipher key

Figure 1: Ciphertexts and a cipher key.

Transcriber and Solver (CTTS), allows for precise symbol clustering and labeling. It is available on GitHub[5] and has been integrated into the DECRYPT pipeline to work seamlessly with cryptanalysis. Also, the tool can be used to produce transcribed training data for automatic hand-written text recognition models.

Automatic transcription — whether partial or complete — can be supported by tools like Transkribus[6]. However, results on ciphertexts are often poor due to their unique characteristics, which Transkribus' language models struggle to recognize. Such tools face challenges in dealing with the distinct symbol sets found in ciphers.

To address this, the DECRYPT project developed TranscriptTool, a human-in-the-loop AI system designed to enhance transcription efficiency by learning from user corrections (Szigeti and Héder, 2022). Pre-trained models have been created for the automatic transcription of various symbol systems, ranging from digit-based codes to alphabets with and without eclectic symbols, including Zodiac and alchemical signs derived from individual ciphers (Baró et al. (2019); Souibgui et al. (2021, 2022, 2023)). These models have been integrated into TranscriptTool, which also allows users to train custom models for different symbol sets.

4.4 Cryptanalysis

With a transcription in hand, cryptanalysis can begin. This process involves various techniques, including statistical analysis and plaintext language identification aided by historical language models, and cipher type detection to make educated guesses about the probable cipher type (e.g. simple, homophonic or polyphonic substitution, or transposition).

CrypTool 2 (CT2) and CrypTool-Online (CTO) offer extensive cryptanalysis tools[7], with CT2 providing a desktop version for advanced users and CTO offering an accessible web-based version. The CT2 DECRYPT editor is currently under development to provide an integrated cryptanalysis environment with transcriptions.

We have enhanced CrypTool with additional algorithms to automatically decrypt a wide range of commonly used ciphers (Lasry (2018); Kopal (2018, 2019); Lasry et al. (2021)). Through quantitative studies, we have analyzed historical cipher keys and instructions, successfully decrypting numerous ciphers (Kopal (2018); Lasry et al. (2021); Kopal and Waldispühl (2022)) — including the Mary Stuart letters (Lasry et al. (2023)), which recently attracted significant international media attention.

[5] https://github.com/CrypToolProject/CTTS

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

[7] https://www.cryptool.org

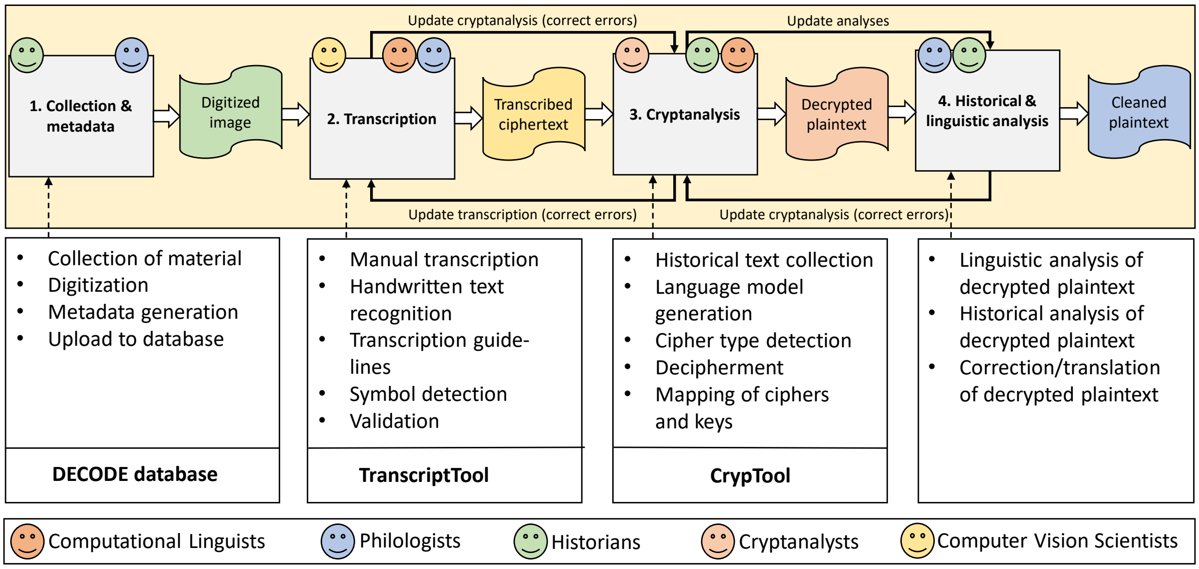

Update cryptanalysis (correct errors)
Update analyses
1. Collection & metadata
Digitized image
2. Transcription
Transcribed ciphertext
3. Cryptanalysis
Decrypted plaintext
4. Historical & linguistic analysis
Cleaned plaintext
Update transcription (correct errors)
Update cryptanalysis (correct errors)
Collection of material
Digitization
Metadata generation
Upload to database
Manual transcription
Handwritten text recognition
Transcription guide-lines
Symbol detection
Validation
Historical text collection
Language model generation
Cipher type detection
Decipherment
Mapping of ciphers and keys
Linguistic analysis of decrypted plaintext
Historical analysis of decrypted plaintext
Correction/translation of decrypted plaintext
DECODE database
TranscriptTool
CrypTool
Computational Linguists
Philologists
Historians
Cryptanalysts
Computer Vision Scientists

Figure 2: The DECRYPT Pipeline

4.5 Historical and Linguistic Analysis

Once decrypted, historians and philologists work to correct and contextualize the text, often translating it into modern languages. This process enriches the Decode database with additional metadata and insights.

5 Challenges and Lessons Learned

While cross-disciplinarity has gained significant promotion and popularity (Láng and Megyesi, 2024), cross-disciplinary teams often face numerous challenges. These include establishing shared goals and research questions, fostering respectful and creative discussions, cultivating an efficient meeting culture, and developing unified publication strategies. To evaluate the management and organization of the DECRYPT project, we conducted a reflexive analysis using a Science and Technology Studies (STS) framework. This approach assessed how effectively AI and the humanities have integrated over the project's span. While the team members were highly satisfied with the project, the STS analysis also provided valuable insights into the difficulties participants encountered, such as issues with terminology, diverse data structures, understanding various methodologies and differing publication practices across disciplines. It also highlighted tensions between academic researchers' focus on publishing and industry partners' need for practical, actionable results, as well as the critical role of expertise and boundary objects in enabling successful collaboration. To navigate these coordination challenges, the team relied on boundary objects—shared concepts or tools that served as common reference points across disciplines. Fostering clear communication, respecting different disciplinary cultures, and setting shared goals from the beginning are vital to ensure smooth project execution in all endeavors.

In the analysis of historical ciphers, while automation tools provide valuable support, human expertise remains indispensable due to the complexity and variability of the ciphers. A significant technological challenge involves the high error rates in automated transcription, particularly in recognizing symbol boundaries, a wide range of symbol types and their variant, and a huge variation of handwriting in historical sources. Even the most sophisticated tools require human expertise to correct errors and adapt models to specific historical scripts .

Manual verification of decipherment results remains also necessary to ensure decipherment accuracy. Difficult tasks include i) the tokenization of the symbol sequences by the identification of code boundaries in script continua, ii) the identification of cipher types iii) and the alignment of ciphertext and plaintext sequences. Automation could assist but is not yet capable of independently solving many of the intricate cryptographic puzzles presented by historical manuscripts.

6 Conclusion

Manuscripts written in rare or unknown scripts are often neglected in historical research due to the complexity involved in analyzing them. These texts are not only linguistically challenging but also hold significant cultural value, containing untapped knowledge and insights. The difficulty lies not only in identifying and transcribing these scripts, but also in interpreting their content in a meaningful way, which demands a wide range of expertise.

Although automation tools like AI can greatly improve efficiency, human expertise is still vital for tasks that require manual verification or nuanced interpretation. It is crucial to develop systems that allow users to make minimal corrections while adapting models to different symbol sets, handwriting styles, and scripts. Moreover, while automation aids cryptanalysis, specialized domain knowledge remains essential, highlighting the steep learning curve for users.

Acknowledgments

The project is financed by the Swedish Research Council, grant 2018-06074, DECRYPT – Decryption of Historical Manuscripts.

References

Arnau Baró, Jialuo Chen, Alicia Fornés, and Beáta Megyesi. 2019. Towards a generic unsupervised method for transcription of encoded manuscripts. In *3rd International Conference on Digital Access to Textual Cultural Heritage (DATECH)*, pages 73–78.

Mihály Héder and Beáta Megyesi. 2022. The DECODE Database of Historical Ciphers and Keys: Version 2. In *Proceedings of the 5th International Conference on Historical Cryptology, HistoCrypt22*.

Nils Kopal. 2018. Solving Classical Ciphers with CrypTool 2. In *Proceedings of the 1st International Conference on Historical Cryptology HistoCrypt 2018*, 149, pages 29–38.

Nils Kopal. 2019. Cryptanalysis of Homophonic Substitution Ciphers using Simulated Annealing with Fixed Temperature. In *Proceedings of the 2nd International Conference on Historical Cryptology, HistoCrypt*, pages 107–16.

Nils Kopal and Michelle Waldispühl. 2022. Deciphering three diplomatic letters sent by maximilian ii in 1575. *Cryptologia*, 46(2):103–127.

George Lasry. 2018. *A Methodology for the Cryptanalysis of Classical Ciphers with Search Metaheuristics*. Kassel university press GmbH.

George Lasry, Norbert Biermann, and Satoshi Tomokiyo. 2023. Deciphering Mary Stuart's lost letters from 1578-1584. *Cryptologia*, 47(2):101–202.

George Lasry, Beáta Megyesi, and Nils Kopal. 2021. Deciphering Papal Ciphers from the 16th to the 18th Century. *Cryptologia*, 45(6):479–540.

Benedek Láng and Beáta Megyesi. 2024. An sts analysis of a digital humanities collaboration: trading zones, boundary objects, and interactional expertise in the decrypt project. *Humanities and Social Sciences Communications*, 11(1):618.

Beáta Megyesi, Bernhard Esslinger, Alicia Fornés, Nils Kopal, Benedek Láng, George Lasry, Karl de Leeuw, Eva Pettersson, Arno Wacker, and Michelle Waldispühl. 2020. Decryption of historical manuscripts: the DECRYPT project. *Cryptologia*, 44(6):545–559.

Beáta Megyesi, Justyna Sikora, Filip Fornmark, Michelle Waldispühl, Nils Kopal, and Vasily Mikhalev. 2023. Historical Language Models in Cryptanalysis: Case Studies on English and German. In *International Conference on Historical Cryptology*, pages 120–129.

Beáta Megyesi, Alicia Fornés, Nils Kopal, Benedek Láng, Michelle Waldispühl, Vasily Mikhalev, and Bernhard Esslinger. 2024. Historical Cryptology. *In Ed: Bernhard Esslinger Learning and Experiencing Cryptography with CrypTool and Sagemath*.

Beáta Megyesi, Crina Tudor, Benedek Láng, Anna Lehofer, Nils Kopal, Karl de Leeuw, and Michelle Waldispühl. 2022. Keys with Nomenclatures in the Early Modern Europe. *Cryptologia*, 0(0):1–43.

Eva Pettersson and Beáta Megyesi. 2018. The HistCorp Collection of Historical Corpora and Resources. In *Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference*, Helsinki, Finland.

Mohamed Ali Souibgui, Alicia Fornés, Yousri Kessentini, and Beáta Megyesi. 2021. Few shots are all you need: A progressive few shot learning approach for low resource handwritten text recognition. *arXiv preprint arXiv:2107.10064*.

Mohamed Ali Souibgui, Alicia Fornés, Yousri Kessentini, and Beáta Megyesi. 2022. Few shots are all you need: A progressive learning approach for low resource handwritten text recognition. *Pattern Recognition Letters*, 160:43–49.

Mohamed Ali Souibgui, Pau Torras, Jialuo Chen, and Alicia Fornés. 2023. An evaluation of handwritten text recognition methods for historical ciphered manuscripts. In *Proceedings of the 7th International Workshop on Historical Document Imaging and Processing*, HIP '23, page 7–12, New York, NY, USA. Association for Computing Machinery.

Ferenc Szigeti and Mihály Héder. 2022. The TRANSCRIPT tool for historical ciphers by the DECRYPT project. In *Proceedings of the 5th International Conference on Historical Cryptology*, pages 208–211.