DE CRYPT
VETENSKAPSRÅDET
THE SWEDISH RESEARCH COUNCIL
RIKSBANKENS JUBILEUMSFOND
FRÄMJAR HUMANIORA OCH SAMHÄLLSVETENSKAP
Stockholms universitet

Unlocking Hidden Histories:

AI and Expert Collaboration in Deciphering Rare Scripts

Beáta Megyesi

RESOURCEFUL-2025

DECRYPT

Decryption of Historical Manuscripts

Vetenskapsrådet 2018-2024

DESCRYPT

Echoes of History:

Analysis and Decipherment of Historical Writings

Riksbankens Jubileumsfond 2025-2032

New Approaches to Analyzing Rare and Unknown Scripts

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 Kopal WP4
cryptology

Rune Rattenborg
archeology

Eva Pettersson
NLP

archeologists,
historians,
linguists,
librarians,
cryptologists, …

Raphaela Heil
computer vision
system architect

Lei Kang
computer vision

Vasily Mikhalev
deep learning,
cryptology

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 adapted to small and rare datasets.

Linear B.

A Sabaic South Arabian inscription.

Codex Runicus.

The Voynich manuscript.

The Borg cipher.

Trilingual Cuneiform inscription.

The Phaistos disk.

The Ramanacoil cipher.

Purpose

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.

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

Goals

- Create a digital annotated 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 Transcribus, 2023
readcoop.eu)

VETENSKAPSRÅDET
THE SWEDISH RESEARCH COUNCIL
RIKSBANKENS JUBILEUMSFOND
FRÄMJAR HUMANIORA OCH SAMHÄLLSVETENSKAP
Stockholms universitet

Challenges

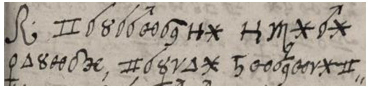
Linguistic challenges

- Undeciphered or poorly understood language.
- No living speakers make annotation speculative.
- Lack of standardized writing system. Variations in symbols, ligatures, and diacritics.

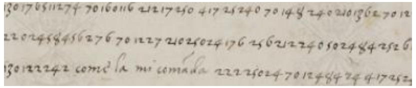
Data scarcity

- Lack of systematic collections and annotated texts
- Remote locations, private collections, restricted archives, political and bureaucratic barriers conflicts and policies
- Ethical concerns in digitization and access.

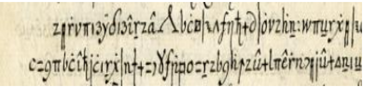
The Voynich manuscript, 15th century
The Beinecke rare book and manuscript library, Yale University



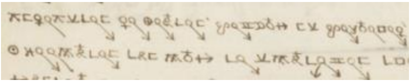
The Borg cipher



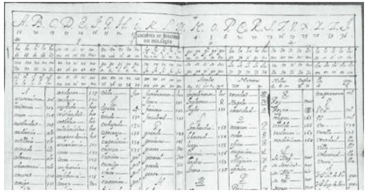
A digit-based cipher from the Vatican



The Copiale cipher



The Ramanacoil cipher



Part of a cipher key

Part of a cipher key

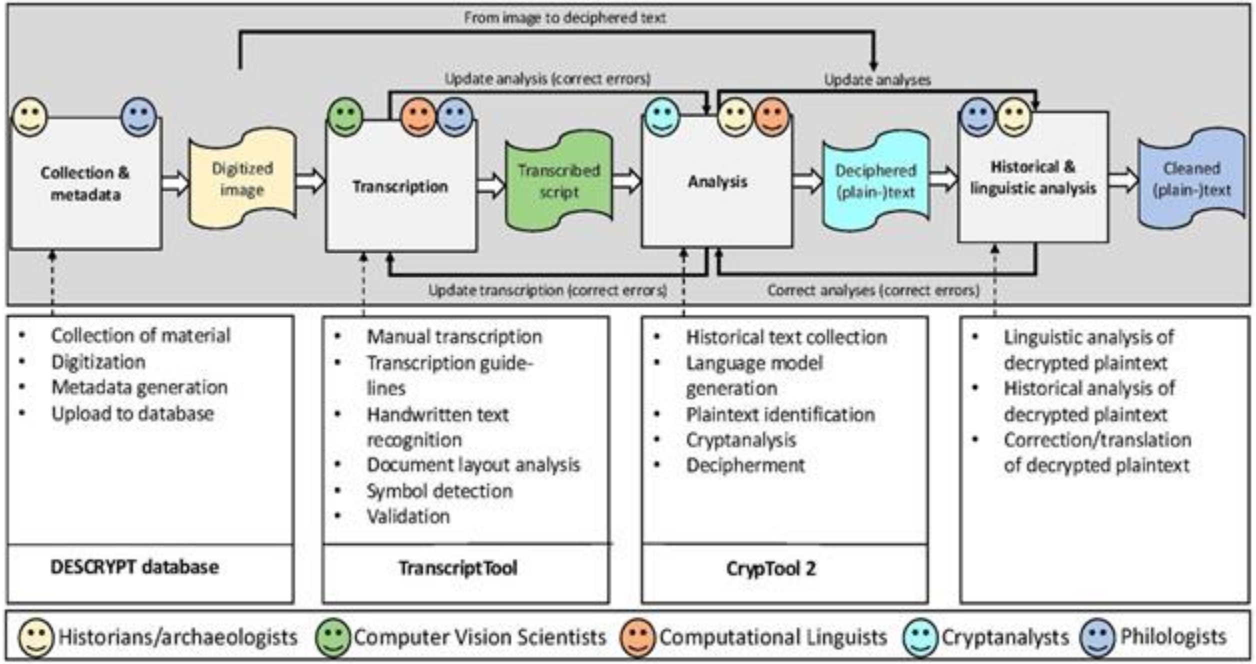
Methodology Overview

Tools

R&D of AI-based transcription and decipherment tools

Approach

Interdisciplinary, collaborative effort



From image to deciphered text

Update analysis (correct errors)

Update analyses

Collection & metadata → Digitized image → Transcription → Transcribed script → Analysis → Deciphered (plain-)text → Historical & linguistic analysis → Cleaned (plain-)text

Update transcription (correct errors)

Correct analyses (correct errors)

- Collection of material
- Digitization
- Metadata generation
- Upload to database

DESCRYPT database

- Manual transcription
- Transcription guidelines
- Handwritten text recognition
- Document layout analysis
- Symbol detection
- Validation

TranscriptTool

- Historical text collection
- Language model generation
- Plaintext identification
- Cryptanalysis
- Decipherment

CrypTool 2

- Linguistic analysis of decrypted plaintext
- Historical analysis of decrypted plaintext
- Correction/translation of decrypted plaintext

Historians/archaeologists Computer Vision Scientists Computational Linguists Cryptanalysts Philologists

Data Collection

Global collaboration

- with local experts and communities. Personal contact in archives.
- Establish research networks.

Visibility and meeting platforms

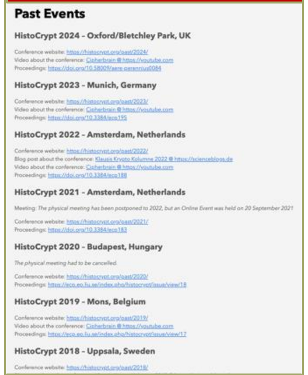
- for experts and the public

Research infrastructure

- Encourage open-access digitization.
- Establish user-friendly research infrastructure for sharing.
- Share **not too early, not too late**…

archeologists, historians, linguists, librarians, cryptologists, …

Conferences



Past Events

HistoCrypt 2024 – Oxford/Bletchley Park, UK

Conference website: https://histocrypt.org/past/2024/
Video about the conference: Cipherbrain @ https://youtube.com
Proceedings: https://doi.org/10.58009/aere-perennius0084

HistoCrypt 2023 – Munich, Germany

Conference website: https://histocrypt.org/past/2023/
Video about the conference: Cipherbrain @ https://youtube.com
Proceedings: https://doi.org/10.3384/ecp195

HistoCrypt 2022 - Amsterdam, Netherlands

Conference website: https://histocrypt.org/past/2022/
Blog post about the conference: Klausis Krypto Kolumne 2022 @ https://scienceblogs.de
Video about the conference: Cipherbrain @ https://youtube.com
Proceedings: https://doi.org/10.3384/ecp188

HistoCrypt 2021 - Amsterdam, Netherlands

Meeting: *The physical meeting has been postponed to 2022, but an Online Event was held on 20 September 2021*

Conference website: https://histocrypt.org/past/2021/
Proceedings: https://doi.org/10.3384/ecp183

HistoCrypt 2020 - Budapest, Hungary

The physical meeting had to be cancelled.

Conference website: https://histocrypt.org/past/2020/
Proceedings: https://ecp.ep.liu.se/index.php/histocrypt/issue/view/18

HistoCrypt 2019 - Mons, Belgium

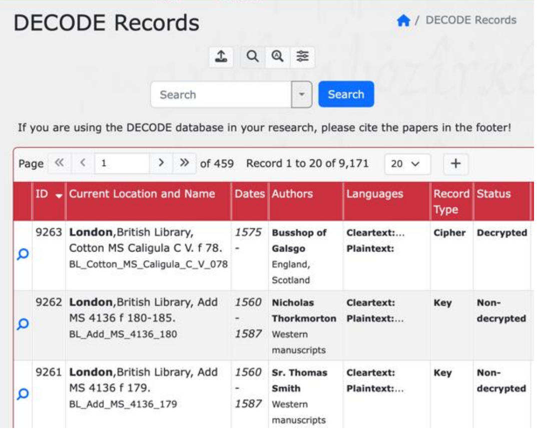
Conference website: https://histocrypt.org/past/2019/
Video about the conference: Cipherbrain @ https://youtube.com
Proceedings: https://ecp.ep.liu.se/index.php/histocrypt/issue/view/17

HistoCrypt 2018 - Uppsala, Sweden

Conference website: https://histocrypt.org/past/2018/

HistoCrypt.org

Platform for sharing



DECODE Records

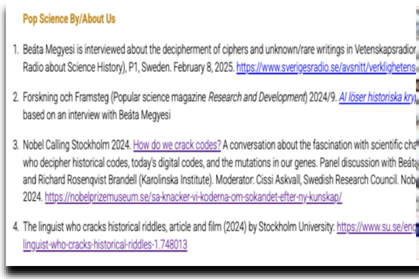
If you are using the DECODE database in your research, please cite the papers in the footer!

Page 1 of 459 Record 1 to 20 of 9,171 20

ID	Current Location and Name	Dates	Authors	Languages	Record Type	Status
9263	**London**,British Library, Cotton MS Caligula C V. f 78. BL_Cotton_MS_Caligula_C_V_078	*1575 -*	**Busshop of Galsgo** England, Scotland	**Cleartext:...** **Plaintext:**	**Cipher**	**Decrypted**
9262	**London**,British Library, Add MS 4136 f 180-185. BL_Add_MS_4136_180	*1560 - 1587*	**Nicholas Thorkmorton** Western manuscripts	**Cleartext:** **Plaintext:...**	**Key**	**Non-decrypted**
9261	**London**,British Library, Add MS 4136 f 179. BL_Add_MS_4136_179	*1560 - 1587*	**Sr. Thomas Smith** Western manuscripts	**Cleartext:** **Plaintext:...**	**Key**	**Non-decrypted**

In the press

MysteryTwister
THE CRYPTO CHALLENGE CONTEST



Pop Science By/About Us

1. Beáta Megyesi is interviewed about the decipherment of ciphers and unknown/rare writings in Vetenskapsradior Radio about Science History), P1, Sweden. February 8, 2025. https://www.sverigesradio.se/avsnitt/verklighetens
2. Forskning och Framsteg (Popular science magazine *Research and Development*) 2024/9. *AI löser historiska kry* based on an interview with Beáta Megyesi
3. Nobel Calling Stockholm 2024. How do we crack codes? A conversation about the fascination with scientific cha who decipher historical codes, today's digital codes, and the mutations in our genes. Panel discussion with Beáta and Richard Rosenqvist Brandell (Karolinska Institute). Moderator: Cissi Askvall, Swedish Research Council. Nob 2024. https://nobelprizemuseum.se/sa-knacker-vi-koderna-om-sokandet-efter-ny-kunskap/
4. The linguist who cracks historical riddles, article and film (2024) by Stockholm University: https://www.su.se/eng linguist-who-cracks-historical-riddles-1.748013

Annotation

Lack of standardization

- Encoding issues: No universal transcription. No Unicode. No standard tools.
- Missing metadata. No annotation guidelines. No consistent annotation.

Lack of annotated data/corpora

- Standard HTR tools fail non-Latin scripts.
- Crowdsourcing annotation is difficult due to limited expertise in these scripts.

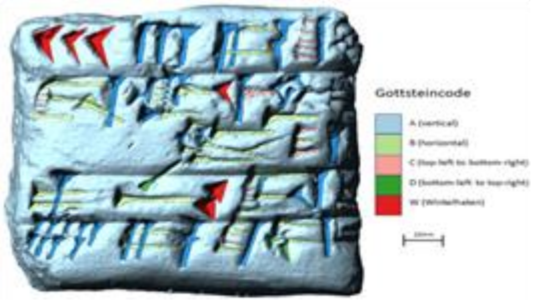
DECODE Metadata

Current location: archive

Provenance: country, date of use, writer/sender

Type: Key/Cipher

Content: Number of pages, Status, Cipher type, Symbol set, Plaintext lang(s), Cleartext lang(s)



(Illustration by Homburg et al. 2022)

AI & Manuscript Studies

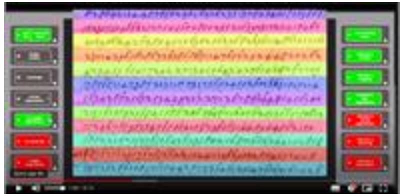
VETENSKAPSRÅDET
THE SWEDISH RESEARCH COUNCIL
RIKSBANKENS JUBILEUMSFOND
FRÄMJAR HUMANIORA OCH SAMHÄLLSVETENSKAP
Stockholms universitet

Current AI tools can assist in the identification, transcription, and classification of scripts.

Challenges

- Incompatibility with existing NLP and OCR Models
- AI requires **large datasets**, which are not available.
- AI models lack **cultural-contextual understanding.**
- **Tasks:** inventories of signs, document layout analysis, identification of writing direction, positional frequency and co-occurrences of signs, grammatical patterns, archaeological and historical contextualization.

(Illustration by DALL-E and M.Héder)



(Transcript Tool by Szigeti & Héder)

Development of Tools

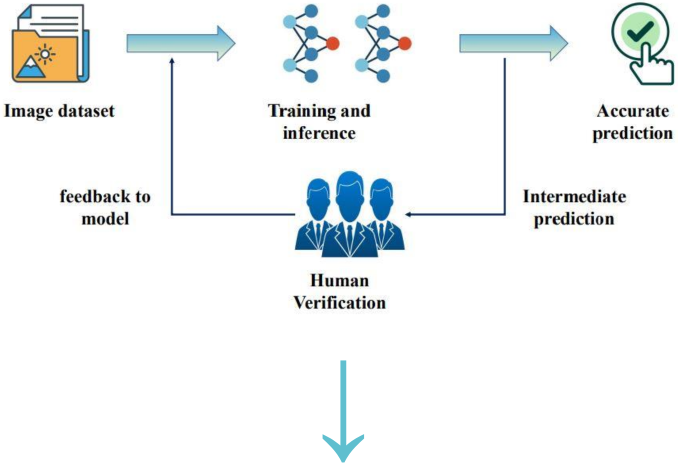
VETENSKAPSRÅDET

RIKSBANKENS JUBILEUMSFOND
FRÄMJAR HUMANIORA OCH SAMHÄLLSVETENSKAP

Stockholms universitet

Building adaptive models with experts

- to annotation of alphabets, layouts, scripts, languages;
- to overcome the challenge of sparse data by data generation;
- by reinforcement, semi-supervised, continual and few-shot learning.
- to learn from the expert input by a few corrections through Human-in-the-loop AI UX research, active learning and iterative refinement

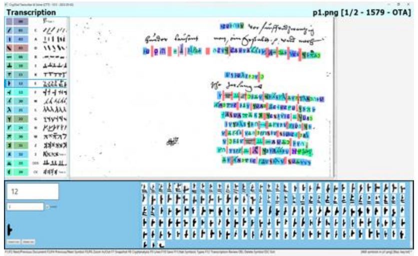


robust tool ecosystem for transcription and decipherment

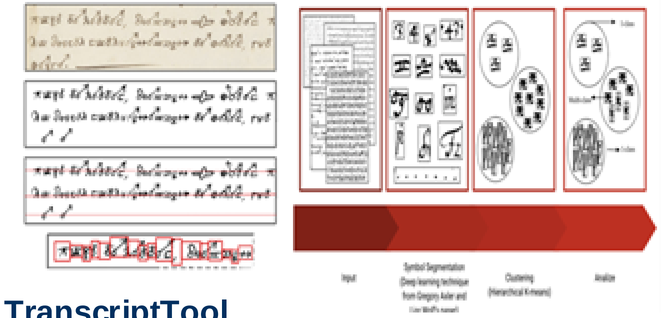
Transcription

- **Recognition Models** for alphabets and scripts
- **Transcription Tools:**
  - ➢ **Manual:** CrypTool Transcriber and Solver (CTTS)
  - ➢ **Automatic**: The TranscriptTool Integration with image processing for enhanced accuracy
- **Outcome:** Manual and semi-automatic tools for transcription

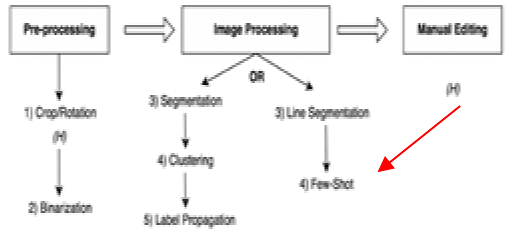
CTTS



Clustering


Input
Symbol Segmentation (Deep learning technique from Gregory Axler and Lior Wolf's model)
Clustering (Hierarchical K-means)
Analize

TranscriptTool


Pre-processing → Image Processing → Manual Editing

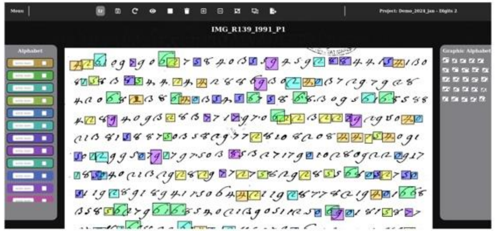
1) Crop/Rotation (H) → 2) Binarization

OR

3) Segmentation → 4) Clustering → 5) Label Propagation

3) Line Segmentation → 4) Few-Shot

(H)



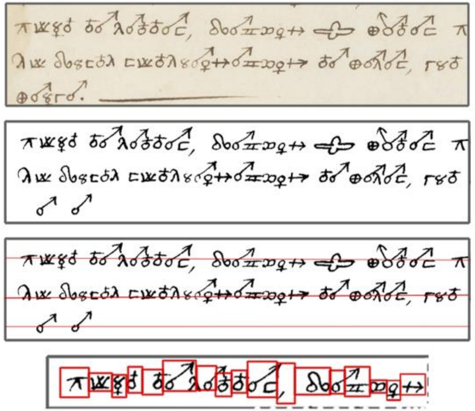
HTR: Unsupervised

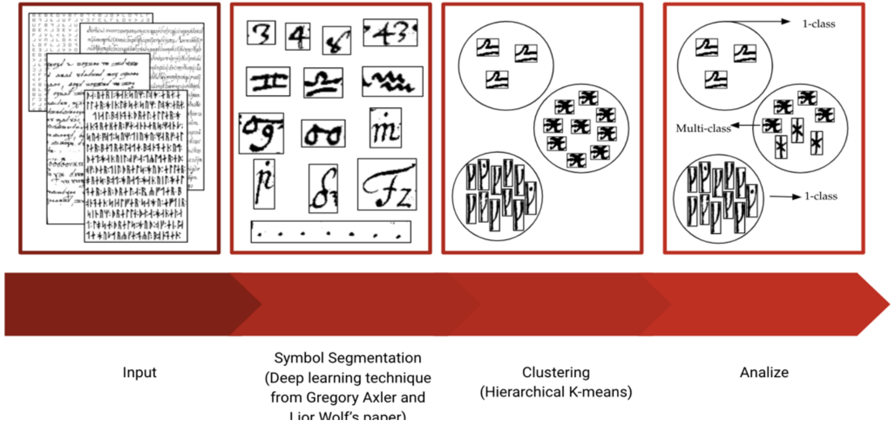
Aims:

- Discover the alphabet
- Find ciphers with the same cipher alphabet
- Transcribe ciphers with various alphabets
- Clustering

User effort:

- Choice of settings for binarization, line and character segmentation, label propagation, output generation
- Cleaning the clusters





Input

Symbol Segmentation (Deep learning technique from Gregory Axler and Lior Wolf's paper)

Clustering (Hierarchical K-means)

Analize

Chen, j., Souibgui, M.A., Fornés, A., and Megyesi, B. 2021. **Unsupervised Alphabet Matching in Historical Encrypted Manuscript Images**. In *Proceedings of the 4th International Conference on Historical Cryptology*. HistoCrypt 2021.

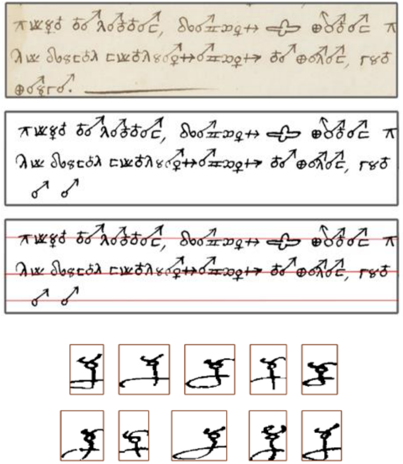
HTR: Supervised few-shot

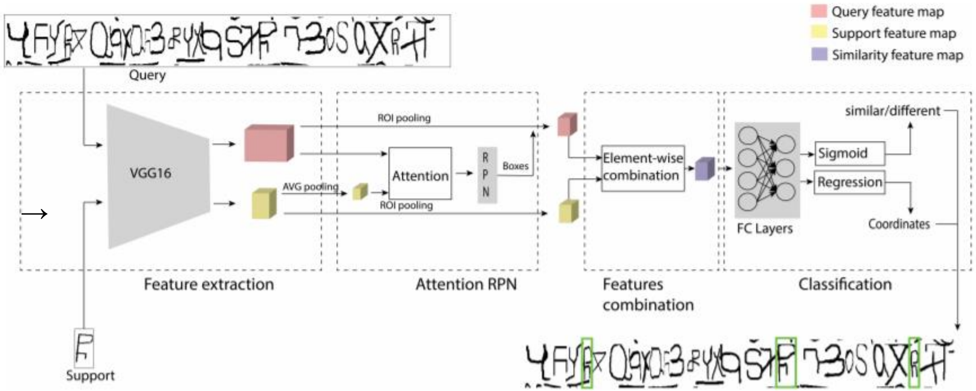
Aims:

- Provide transcription for ciphers with various symbol sets
- High recall and precision
- Few-shot model architecture

User effort:

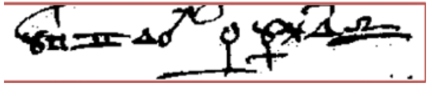
- Preprocessing: Line segmentation
- Create supporting alphabet – 10 examples for each symbol type
- Output validation and correction

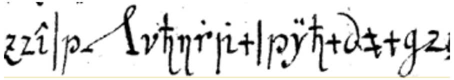


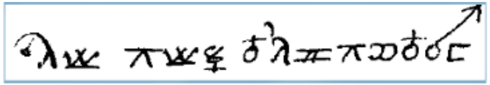


Souibgui, M. A. (2020) and Fornés, A and Kessentini, Y and Tudor, C. "(2020)
A Few-shot Learning Approach for Historical Ciphered Manuscript Recognition"
25th International Conference on Pattern Recognition ****Best Student Paper Award****

HTR: Evaluation







Model	Borg (I-D)		Copiale (I-D)		Ramanacoil (O-D)	
	Precision	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**
Clustering	57.63%	74%	89.61%	73%	**93.71%**	33%
Few-Shot	**96.6%**	**85%**	**96.62%**	**79%**	59.47%	**93%**

- Supervised few-shot wins for in-domain data
- Clustering is preferred for symbol recognition in out-domain data
- Few shot is better in coverage, can be improved by given examples of all (less frequent) symbols

🌸 Thanks to Giacomo Magnifico! (2021)
Lost in Transcription: Evaluating Clustering and Few-Shot learningfor transcription of historical ciphers, Master’s thesis in language technology, Uppsala University

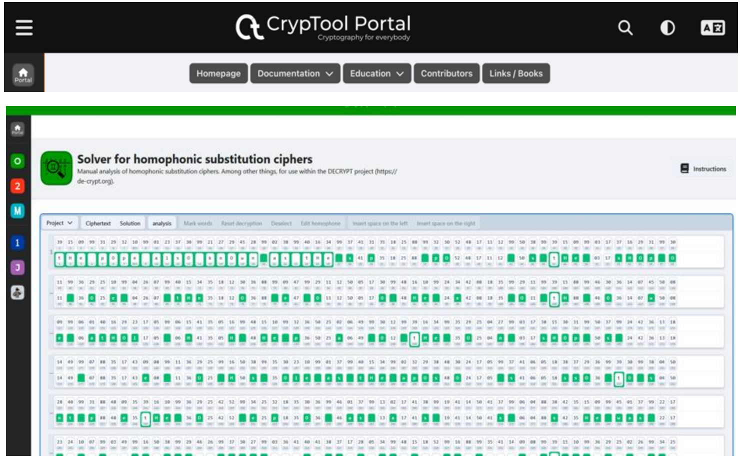
VETENSKAPSRÅDET
THE SWEDISH RESEARCH COUNCIL

RIKSBANKENS JUBILEUMSFOND
FRAMJAR HUMANIORA OCH SAMHÄLLSVETENSKAP

Stockholms universitet

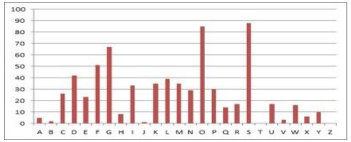
Decipherment

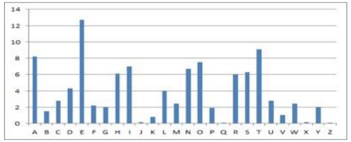
- **Assumptions**: Languages, Dating, Code structure, Space, nullities, cancellation
- **Encoding types**: simple substitution, homophonic substitution, codebooks
- **Language models**: Character- and word-based n-grams for languages and time periods
- **Attacks**: One or several documents: frequencies, clustering, hill climbing, …

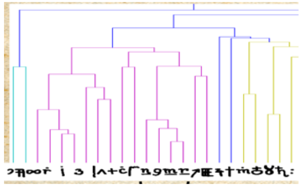


Solver for homophonic substitution ciphers

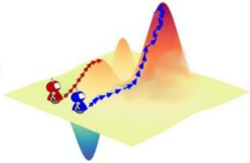
Manual analysis of homophonic substitution ciphers. Among other things, for use within the DECRYPT project (https://de-crypt.org).







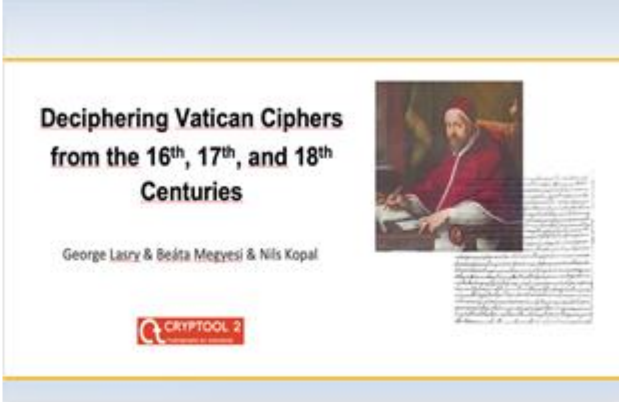
- Hill climbing
  - Start with empty or random plugboard
  - Test various plugboard connections
    - If score improves, keep change
    - Otherwise discard change
  - Restart process if stuck



VETENSKAPSRÅDET
THE SWEDISH RESEARCH COUNCIL
RIKSBANKENS JUBILEUMSFOND
FRÄMJAR HUMANIORA OCH SAMHÄLLSVETENSKAP
Stockholms universitet

Conclusion

- **Tools:** Scalable AI-driven framework for historical text analysis with minimal corrective inputs from experts and an interactive platform to serve experts
- **Resources:** a digital corpus of historical writings in a standardized format and a historical corpus
- **de-crypt.org**, https://github.com/decrypt-project/
- **Publications:** ca 100 scientific papers
- **Conference:** histocrypt.org
- **Outreach**: university courses, exhibitions, museums, the press, YouTube, MysteryTwister, tutorial



Deciphering Vatican Ciphers from the 16th, 17th, and 18th Centuries

George Lasry & Beáta Megyesi & Nils Kopal

CRYPTOOL 2

Decoding Mary, Queen of Scots (1542-1587)
George Lasry, Satoshi Tomokiyo, Norbert Biermann

Future

Progress depends on:

- collaboration across disciplines to bridge the gap between technology and humanities,
- open, annotated and standardized data across scripts and languages, and
- hybrid approaches and adaptive AI models that require minimal data input by experts.

Thank you!

Questions?

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