Stockholm University

Bridging the Low-Resource Gap in Historical Cryptology

A Multilingual Diachronic Synthetic Dataset for Reproducible Cryptanalysis

Micaella Bruton · Meriem Beloucif · Beáta Megyesi

Stockholms Universitet · Uppsala Universitet

LREC · RESOURCEFUL 2026

What is Historical Cryptology?

Study of **historical[1] codes and ciphers,** and the methods for creating [**cryptography**] and breaking [**cryptanalysis**] them

What?

- Diplomatic correspondence
- Royal and military letters
- Religious / secret society texts
- Personal journals / letters

Where?

- City / State / National archives
- Libraries
- Museums
- Online databases
  - DECODE - 10,106 entries[2]

Why?

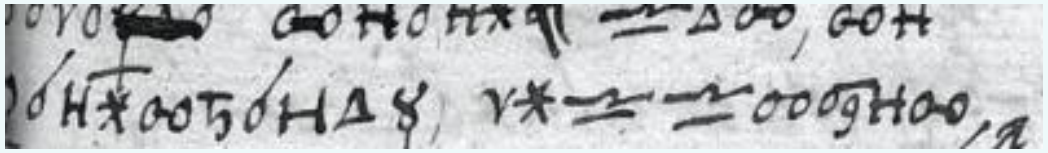
- Protect sensitive information
- Secure private thoughts
- Difficult

Who?

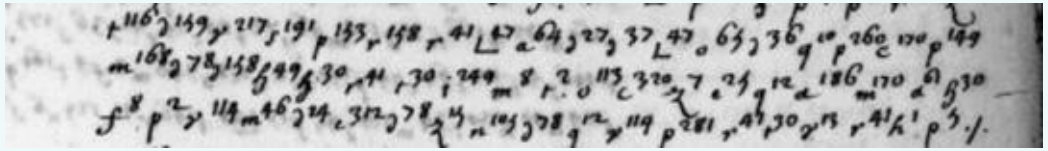
- Historians
- Philologists / Linguists
- Cryptanalysts / Computer scientists

[1]produced no later that the mid-20th century | [2]as of May 2026

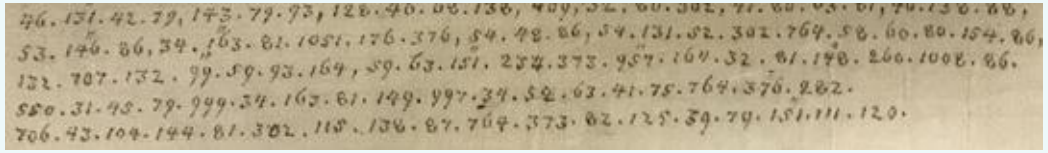
Historical ciphertexts



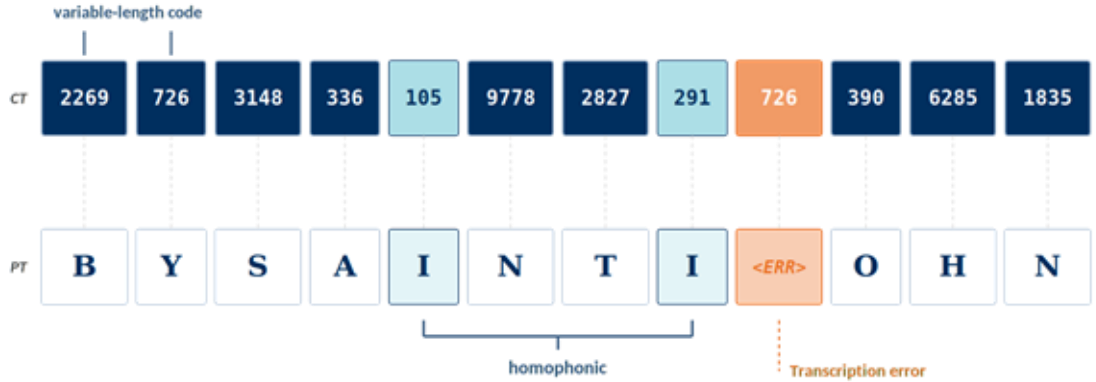
Borg Cipher



BnF Clair. 327, ff.279-280



DECODE #4101



The Need for Data

01

Extreme low-resource

Aligned plaintext–ciphertext pairs are scarce, restricted, and unevenly distributed across languages and centuries.

02

Reproducibility crisis

Prior work generates synthetic data internally and rarely releases it — making meaningful comparison impossible.

03

Synthetic data, validated

We need a principled way to generate AND validate synthetic resources when authentic data is genuinely unavailable.

INTRODUCING

HistCiph at a glance

10	8	4	~4.2M
languages	centuries (1100–1899)	ciphertext variants per text	plaintexts released

Languages covered

Czech · Dutch · English · French · Hungarian · Icelandic · Italian · Polish · Spanish · Swedish

Germanic + Romance + Slavic + Uralic · pre- and post-standardization orthography · diachronically balanced

Open source · Dataset on HuggingFace

Two principles, applied throughout

CONSTRAIN

Generation grounded in domain knowledge

- Real historical plaintext from HistCorp
- Multi-homophone allocation per character
- Variable-length codes (3 / 4 digits)
- Stochastic insertion / deletion noise
- Independent keys per document

VALIDATE

Information-theoretic diagnostics

- Plaintext entropy & redundancy
- Ciphertext entropy & frequency masking
- Homophone allocation patterns
- Unicity distance — recoverability bound
- Compared against theoretical maxima

STEP 1

Plaintext: real historical text, normalized

1

HistCorp

Multilingual historical corpus (Pettersson & Megyesi, 2018)

2

Normalize

Remove whitespace, formatting, punctuation

3

Filter

Remove extended foreign-language passages

4

Balance

Sample evenly across 100-year intervals per language

Why this matters

Plaintext lengths binned at 50 / 100 / 200 / 400 / 600 / 800 / 1000 characters. Diachronic balance preserves authentic orthographic variation across periods — what makes historical decryption hard, and exactly what modern-texts erase.

Homophonic substitution, formally

For each plaintext character $c \in \Sigma$:

$$H_c \subset \Gamma^3 \cup \Gamma^4 \quad \text{with} \quad 1 \leq |H_c| \leq 5$$

$$E(c) \sim \text{Uniform}(H_c)$$

In words:

- Each character maps to up to 5 unique 3- or 4-digit codes
- Each occurrence draws uniformly from that character's set
- Keys are generated independently per document

TRANSCRIPTION NOISE

Per character; p = 0.05

- Deletion: drop the sampled ciphertext token
- Insertion: add a token already present in the ciphertext

Alignment preserved

Affected plaintext positions are marked with the symbol '#', so character-level alignment between plaintext and ciphertext is never lost.

Four ciphertext variants per document

CODE LENGTH

ERRORS	Mixed (3+4 digits)	Fixed (4 digits)
With	**Mixed + Errors** `8148 0511 961 2209 ...` *+ matching plaintext, key, metadata*	**Fixed + Errors** `7533 0280 6091 3906 ...` *+ matching plaintext, key, metadata*
Without	**Mixed, Clean** `5903 1187 840 2910 ...` *+ matching plaintext, key, metadata*	**Fixed, Clean** `7592 3475 6912 6940 ...` *+ matching plaintext, key, metadata*

Each document also carries: year · year_range · text_length · text_id · 4 paired keys

Scale & coverage across languages

Language	Train		Validation		Test	
	Plaintext	Ciphertext	Plaintext	Ciphertext	Plaintext	Ciphertext
Czech	228,235	912,940	28,514	114,056	28,569	114,276
Dutch	689,204	2,756,816	86,135	344,540	86,197	344,788
English	1,204,217	4,816,868	150,505	602,020	150,547	602,188
French	18,002	72,008	2,249	8,996	2,271	9,084
Hungarian	58,285	233,140	7,286	29,144	7,300	29,200
Icelandic	89,679	358,716	11,196	44,784	11,249	44,996
Italian	163,849	655,396	20,472	81,888	20,514	82,056
Polish	261,734	1,046,936	32,792	131,168	32,694	130,776
Spanish	351,566	1,406,264	43,926	175,704	43,966	175,864
Swedish	340,687	1,362,748	42,572	170,288	42,604	170,416

Did our synthetic data behave like real ciphers?

H

Plaintext entropy

How much information is in the source language?

R

Plaintext redundancy

How much structural pattern remains exploitable?

H_CT

Ciphertext entropy

Are plaintext frequencies actually masked?

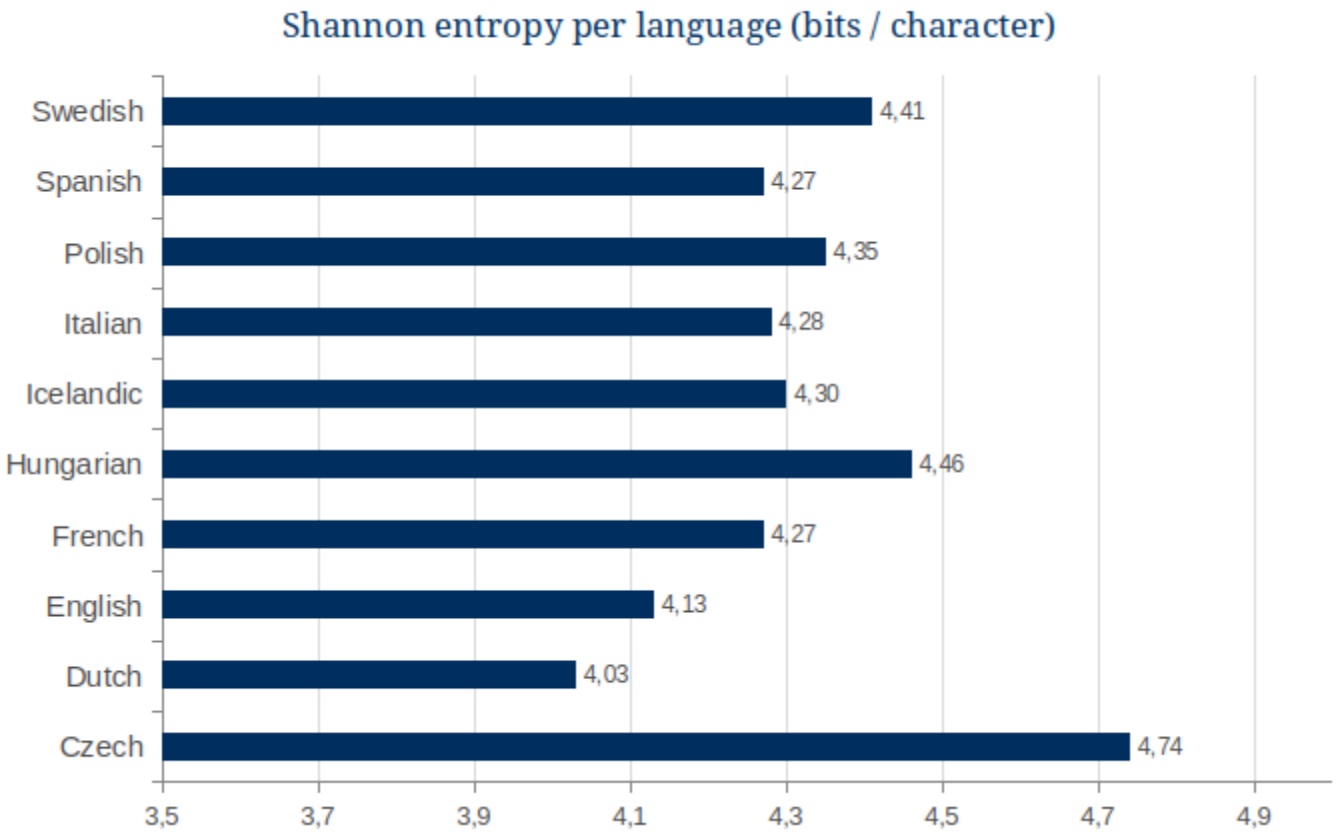
U

Unicity distance

How much ciphertext is needed to uniquely recover the key?

Four lenses on the same question: does this synthetic data behave like the real thing?

Plaintext entropy stays narrow; redundancy varies



4.32

avg. entropy across languages

narrow band, 4.03–4.74 bits/char

1.55

avg. redundancy

spread: 0.65 (Polish) – 2.93 (Spanish)

37 – 138

character inventory size

*long tails reflect **both** historical orthography and noise*

Ciphertext approaches the theoretical maximum

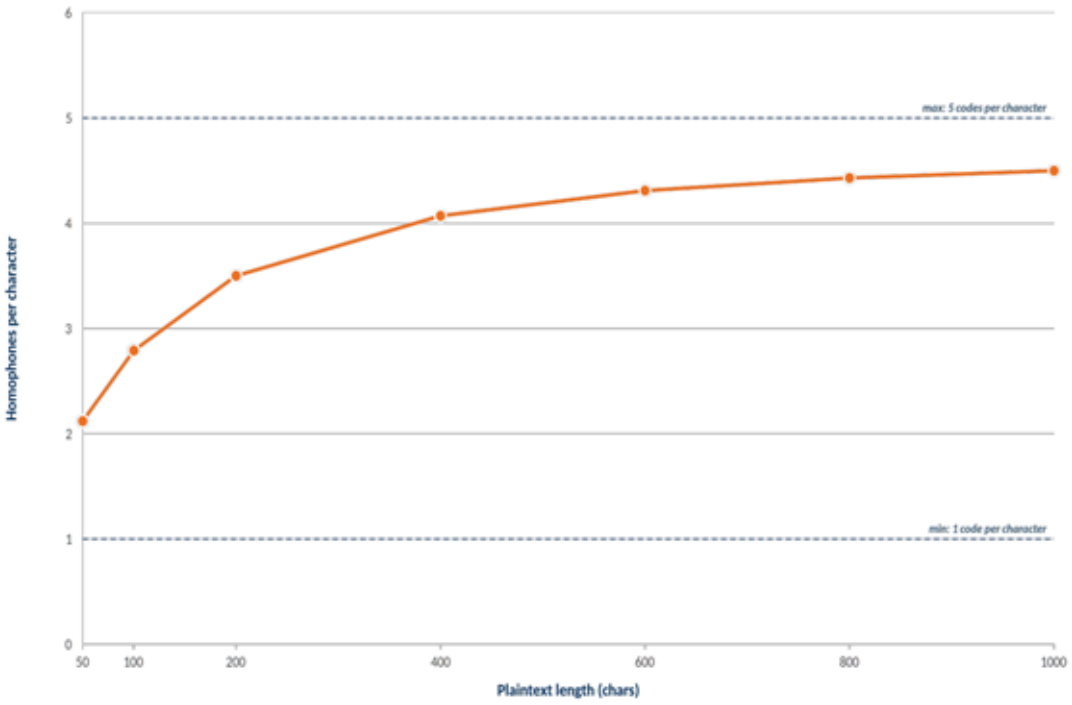
99.46% – 99.80%

of the theoretical entropy bound (13.43 bits / token)

Mean across 10 languages: 13.37 bits

Independent keys per text → entropy estimates are not corpus-wide artifacts.

Homophone allocation grows with text length



Frequent characters reach the cap of 5; rare ones stay at 1

Limitations & honest caveats

Synthetic, not authentic

Real ciphers have key reuse across documents, semantic obfuscation, and codebook entries for sensitive names — none of which we model.

Idealised key design

Authentic historical keys had special-purpose symbols, ad hoc expansions — we use uniform allocation for control.

Fixed noise rate

Transcription error rate is fixed at $p = 0.05$; real archival error rates vary substantially with manuscript condition and editorial conventions.

Long-tail inventories

Large character inventories partly reflect residual loanwords and rare graphemes — small in mass, but inflate theoretical redundancy.

PART II

This isn't really about ciphers.

The methodology is general

1. **Constraint-aware generation**
2. **Quantitative validation against the target**
3. **Open release of data + tooling**

The principle, distilled

STEP 1

Constrain generation

Derive synthetic instances from empirically grounded properties of real-world sources. Encode domain-informed rules — not arbitrary substitutions.

e.g. plaintext from HistCorp + historical key rules

STEP 2

Validate quantitatively

Quantitatively check that the resulting data reproduces key statistical signatures of the target setting. If it doesn't, the generation procedure is wrong.

e.g. ciphertext entropy at 99%+ of theoretical max

Information-theoretic diagnostics generalize as quality control for synthetic data — especially in low-resource contexts

The framework, applied to other domains

Low-resource MT

CONSTRAIN

- Observed alignment patterns
- Sentence-length distributions
- Morphological productivity

VALIDATE

- Source–target mutual information structure
- Conditional entropy symmetry
- Information preservation ratio across languages

Morphological analysis

CONSTRAIN

- Morphological entropy per token
- Redundancy of affix systems

VALIDATE

- Entropy of morphological paradigms
- Minimum description length stability across forms
- Mutual information between stem and inflectional features

Noisy HTR / dialectal text

CONSTRAIN

- Calibrate noise injection to observed error profiles and dialect variation patterns
- Channel noise entropy
- Error distribution entropy

VALIDATE

- Mutual information between clean and corrupted text
- Entropy inflation rate
- KL divergence of character/channel distributions vs real HTR

Domain-informed generation + distributional validation

IN CLOSING

Three takeaways

01

THE RESOURCE

HistCiph: a public collection for historical cryptanalysis

Ten languages, eight centuries, four ciphertext variants per text — open on HuggingFace.

02

THE METHOD

Constrain + validate = empirically grounded synthetic data

Domain-grounded generation rules paired with falsifiable distributional diagnostics — useful far beyond ciphers

03

THE INVITATION

Information-theoretic diagnostics as cross-domain QA

Entropy, redundancy, complexity measures generalize as quality control for any low-resource synthetic resource.

Thank you

Questions, comments, collaborations welcome

RESOURCES

Dataset: huggingface.co/collections/mbruton/HistCiph

Toolkit: github.com/mbruton0426/ChronoFidelius

Contact: micaella.bruton@ling.su.se

Acknowledgements: Riksbankens Jubileumsfond (M24-0028 DESCRYPT) · NAISS (VR 2022-06725) · Nils Kopal