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, Stockholms Universitet
{micaella.bruton, beata.megyesi}@ling.su.se, meriem.beloucif@lingfil.uu.se

Abstract

Many NLP tasks suffer from limited aligned supervision in the target domain. Historical cipher decryption represents an extreme case: aligned *plaintext–ciphertext* pairs are scarce, access to decrypted archives is restricted, and prior work often relies on synthetic data that is neither released nor evaluated for realism. This limits reproducibility and obscures whether models trained on synthetic benchmarks transfer to archival conditions. We introduce **HistCiph**, the first publicly available multilingual collection of historically grounded plaintext–ciphertext datasets for classical ciphers. Spanning ten languages (Czech, Dutch, English, French, Hungarian, Icelandic, Italian, Polish, Spanish, Swedish) and multiple centuries, the collection combines diachronically balanced historical plaintext with independently generated homophonic substitution keys and controlled transcription noise. Synthetic generation is explicitly constrained by documented properties of historical ciphers, including multi-homophone allocation and variable-length codes. We validate the datasets using information-theoretic diagnostics—entropy, redundancy, frequency masking, and unicity distance—showing that ciphertext distributions approach theoretical bounds while preserving cross-linguistic variation. HistCiph provides a reproducible benchmark for neural decryption and alignment, and illustrates a principled framework for empirically grounded synthetic data generation in low-resource NLP.

Keywords: synthetic data, low-resource NLP, historical text, substitution ciphers, entropy, unicity distance

1. Introduction

Large collections of historically encrypted documents remain preserved in archives across the world. These materials often span diplomatic correspondence, military communication, political intelligence, and private letters which would provide historians and other scholars with additional insights into important events throughout our history (Megyesi et al., 2019; Yin et al., 2019; Kopal and Waldispühl, 2022; Antal et al., 2023; Bruton and Megyesi, 2025). Although the cipher systems used in many of these documents are now considered cryptographically weak, their decryption remains challenging.

Messages are often short, contain transcription errors, and employ variable-length homophonic codes—substitution systems in which a single plaintext (the original, readable text) character can be represented by multiple alternative numeric codes of different lengths—which makes them time- and labour-intensive to manually decipher (Megyesi et al., 2024a; Kambhatla, 2024). Moreover, the underlying plaintexts reflect historical spelling conventions and orthographic instability that differ substantially from modern language norms; this complicates automated approaches that rely on distributional statistics learned from contemporary language data (Megyesi et al., 2023; Desenclos and Lasry, 2024). Together, these factors make historical cipher decryption a difficult computational problem.

Despite growing interest in automatic cryptanalysis—including computational methods for recovering encryption keys, plaintext messages, and detecting cipher types (e.g., simple, homophonic or polyphonic substitution, transposition)—historical ciphertext decryption remains an extremely low-resource problem. Few publicly available plaintext–ciphertext pairs (aligned examples of original text and its encrypted version) exist, as decrypted archival materials are often restricted or inconsistently digitized (Antal et al., 2023). As a result, recent work typically encrypts its own synthetic data and reports experimental settings without releasing the generated ciphertexts, limiting reproducibility and comparative evaluation (Leierzopf et al., 2021b; Fürthauer et al., 2022; Ahmadzadeh et al., 2022; Kimura et al., 2022; Bruton and Megyesi, 2025; Simhadri et al., 2025).

Existing publicly available cipher datasets are generally small, rely on modern language data rather than historical texts, embed ciphertexts within generative language model prompts instead of providing aligned sequence pairs, and often include only a single clean ciphertext variant per plaintext. To our knowledge, there is currently no publicly available dataset of historically grounded plaintext–ciphertext pairs explicitly designed for computational sequence modelling and reproducible cryptanalysis research.

To address this gap, we introduce **HistCiph**, the first publicly available multilingual collection of historical plaintext–ciphertext pairs spanning ten lan-

Proceedings of The Fourth Workshop on the Role of Resources in the Age of Large Language Models (RESOURCEFUL 2026), pages 13–24 May 11, 2026.

B	**Y**	S	**A**	**I**	N	T	**I**	**O**	H	N	T	H	**E**	F	**I**	R	S	T
2269	**726**	3148	**336**	**105**	9778	2827	**291**	**390**	6285	1835	9976	9673	**834**	2759	**644**	9387	1127	2827
T	H	E	P	E	R	S	O	N	S	O	F	T	H	E	P	L	A	Y
1801	6496	8294	9976	1817	0560	8272	3148	9387	8272	1127	1037	9670	5597	3899	0675	3238	7585	4667

Table 1: Excerpts showcasing aligned English plaintext–ciphertext pairs. Top example shows a plaintext from the year range 1500-1599 with its ciphertext variant employing variable length codes, and the bottom shows a plaintext from the year range 1800-1899 with its ciphertext variant containing fixed length codes. Bolded characters showcase variable length code, and coloured groupings showcase homophonicity where the plaintext character (here *E*, *T*, and *I*) has several codes.

guages and multiple centuries (1100–1899). As decrypted archival plaintext–ciphertext pairs are scarce and often inaccessible, HistCiph operationalizes this critically low-resource setting by deriving synthetic training data from real-world historical sources in a principled way.

Plaintexts are drawn from diachronic corpora included in HistCorp and normalized (e.g., whitespace and punctuation removal, foreign-language filtering) to reflect historical cipher writing practices (Pettersson and Megyesi, 2018). Each text is then encrypted in four different ways while explicitly constraining the encryption process to reflect attested properties of historical homophonic practice (e.g., multi-symbol allocations, variable-length numeric codes, and transcriptional irregularities). Crucially, we do not treat the resulting data as synthetic "by fiat": we evaluate whether the generated ciphertexts exhibit representative cryptographic behaviour by quantifying their statistical and information-theoretic characteristics. In particular, we measure plaintext entropy and redundancy, ciphertext entropy and frequency masking, homophone allocation patterns, and unicity distance, to validate that the synthetic data preserves realistic difficulty regimes and supports meaningful benchmarking under archival conditions.

HistCiph is designed for low-resource historical cryptanalysis and related sequence modelling tasks. By combining diachronic orthographic variation with frequency-masking homophonic encryption, the dataset provides a controlled but realistic benchmark for character-level modelling, cross-lingual transfer, and decryption under limited supervision. The dataset contains balanced plaintext lengths ranging from 50–1000 characters, explicit train/validation/test splits, and rich metadata including origin year, century range, and text length. Our contributions are as follows:

- release the first publicly available multilingual datasets of historical plaintext–ciphertext pairs;
- provide balanced splits across ten languages and multiple centuries, with explicit metadata supporting diachronic analysis;
- implement a controlled homophonic encryption procedure with optional transcription noise to simulate archival conditions, and;
- provide detailed statistical analyses of plaintext and ciphertext properties, including entropy, homophone distribution, and unicity distance.

2. Related Work

Automatic cryptanalysis and classical cipher detection have received increasing attention in recent years. However, most existing work relies on internally generated synthetic data rather than publicly released plaintext–ciphertext corpora.

Several studies on cipher classification and decryption generate their own encrypted datasets using modern language corpora. For example, Ahmadzadeh et al. (2021) and Ahmadzadeh et al. (2022) evaluate neural approaches on ciphertexts derived from modern plaintexts but do not release the encrypted datasets. Similarly, Gopinathan et al. (2006) and Leierzopf et al. (2021a) perform decryption experiments on self-encrypted data, with Leierzopf et al. (2021a) even utilizing historical data, but the encrypted data used for experimentation is not publicly released. While these studies demonstrate methodological advances, the absence of publicly available plaintext–ciphertext pairs limits reproducibility and downstream benchmarking.

Some small-scale classical cipher datasets are available online. For example, the Machine-Learning-Based Classical Cipher Classifier repository[1] provides encrypted examples for cipher-type classification, but the plaintext inputs consist primarily of extremely short word-level or artificial text segments and are not designed for diachronic analysis or decryption. Additionally, 65 datasets are currently hosted on HuggingFace[2] that appear when searching *cipher*, but they use modern plaintext data and provide limited or no metadata. In many cases, they utilize modern encryption systems and ciphertexts are embedded within prompt templates for generative language models and are not intended as standalone cryptanalysis resources.

[1] github/Manbendra2014/Machine-Learning-Based-Classical-Cipher-Classifier

[2] huggingface/cipher-datasets

Several do not contain ciphers or encrypted text at all. To our knowledge, there is currently no publicly available dataset of historical plaintext–ciphertext pairs spanning multiple centuries, and current datasets do not provide explicit metadata supporting diachronic and low-resource modelling. HistCiph is designed to fill this gap.

Synthetic data generation has become a common strategy in NLP for mitigating data scarcity, especially in low-resource and domain-shifted settings. Prior work has used synthetic examples to support machine translation, information extraction, and robustness evaluation, for instance via back-translation and self-training, rule-based or templated generation, and more recently through large language models (LLMs) that produce task-specific text paired with labels (Chimalamarri and Sitaram, 2021; Abudouwaili et al., 2023; Meyer and Buys, 2024; Evuru et al., 2024; de Gibert et al., 2025; Nadăş et al., 2025). Synthetic data can substantially improve coverage and controllability, enabling targeted perturbations (e.g., noise injection, style shifts, or constrained vocabularies) and systematic evaluation under specific conditions that are difficult to obtain at scale in naturally occurring corpora (Evuru et al., 2024; Nadăş et al., 2025).

At the same time, a recurring challenge is ensuring that synthetic datasets are representative of the phenomena encountered at test time. Synthetic corpora may inadvertently simplify the task, erase long-tail variation, or introduce artifacts that models exploit, leading to inflated performance that does not transfer to real-world data (Meyer and Buys, 2024; Nadăş et al., 2025). Recent NLP work therefore emphasizes grounding generation procedures in empirical properties of the target domain and validating synthetic outputs through structural and distributional diagnostics (Chimalamarri and Sitaram, 2021; Abudouwaili et al., 2023; Evuru et al., 2024).

For example, augmentation strategies for Turkic languages preserve vowel–consonant class and syllabic structure when hallucinating new word forms, explicitly validating that generated data respects phonological constraints (Abudouwaili et al., 2023). Constraint-based generation frameworks extract lexical or syntactic patterns from gold data and enforce them during LLM prompting to ensure structural faithfulness (Evuru et al., 2024). In low-resource machine translation, large-scale synthetic parallel corpora generated via forward translation have been evaluated using neural quality estimation metrics, demonstrating that even noisy synthetic data can yield substantial gains when authentic supervision is scarce (de Gibert et al., 2025). Work on Dravidian languages further highlights the importance of linguistically motivated segmentation and native-speaker validation when constructing synthetic or morphologically derived resources (Chimalamarri and Sitaram, 2021).

These studies underscore the importance of constraint-aware generation and post-hoc validation principles that also guide our design of homophonic encryption and entropy-based diagnostic evaluation. While prior work primarily evaluates synthetic data through structural faithfulness (e.g., syllable preservation or constraint satisfaction) or downstream task performance, comparatively less attention has been paid to corpus-level distributional diagnostics that quantify how closely generated data matches the statistical properties of real-world sources. In contrast, our approach incorporates information-theoretic measures such as entropy, redundancy, and unicity distance to assess whether synthetic data generation preserves global complexity characteristics.

Historical cipher systems themselves have been extensively studied across a wide range of periods and geographic regions (Meister, 1902, 1906; Kahn, 1996; Megyesi et al., 2024b). Surviving codebooks, keys, and archival analyses provide detailed insight into how classical encryption systems—particularly homophonic substitution ciphers—were constructed and used in practice (Kopal and Waldispühl, 2022; Lasry et al., 2023; Desenclos and Lasry, 2024). We therefore possess substantial knowledge about historical key design, symbol allocation strategies, frequency masking techniques, and common transcription practices. This knowledge enables our application of information-theoretic diagnostics that complement structural and task-based evaluation in historical cryptography. More broadly, such diagnostics provide an additional layer of quality control for synthetic data in low-resource settings.

3. Dataset Construction

In the present study, we draw on historical cryptographic scholarship to generate ciphertexts that are structurally grounded in documented encryption practices. Rather than relying on arbitrary synthetic substitutions, our encryption procedure is designed to approximate historically attested simple and homophonic substitution systems. In particular, we model features such as variable-length numeric codes with and without whitespaces and nullities, transcription errors, various ciphertext lengths and languages, thereby producing data that is both computationally controlled and historically plausible.

The HistCiph collection includes datasets for 10 languages (Czech, Dutch, English, French, Hungarian, Icelandic, Italian, Polish, Spanish, Swedish) spanning Indo-European and Uralic families, covering various periods from the 12th to the 19th century. All texts are written in the Latin script, though many of the languages historically em-

	Train		Validation		Test	
	pt	ct	pt	ct	pt	ct
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

Table 2: Document counts for plaintext (pt) texts and their corresponding ciphertext (ct) variants across train, validation, and test splits for each language in HistCiph.

ployed additional characters, diacritics, and ligatures. The dataset captures both pre-standard and post-standardization stages across languages, enabling diachronic analysis of orthographic variation. A summary of all texts included in each training/validation/test split is included in Table 2. The collection and all datasets are publicly available on HuggingFace[3].

3.1. Plaintext Collection & Normalization

All plaintext data was sourced from various subcorpora collected by the HistCorp corpus, a large-scale collection of historical texts covering multiple languages, time periods, and genres (Pettersson and Megyesi, 2018).

Prior to encryption, all plaintext data underwent normalization and cleaning. This process included the removal of whitespace, formatting characters, and punctuation. Texts were further filtered to minimize the presence of extended foreign-language passages in order to maintain as close to monolingual content as possible. Approximately equal quantities of text were sampled for each available 100-year interval to support diachronic balance within languages. A description of the breakdown of text by year range is available in Appendix A.

3.2. Encryption Procedure

Ciphertexts are generated as homophonic substitution ciphers encrypted with digits utilizing the ChronoFidelius[4] toolkit (Bruton and Megyesi, 2025). For the decryption description, let Σ denote the plaintext character inventory and $\Gamma = \{0, \ldots, 9\}$ the digit alphabet. For each plaintext character $c \in \Sigma$, a homophone set $H_c \subset \Gamma^3 \cup \Gamma^4$ is sampled, where $1 \leq |H_c| \leq 5$. Each element of H_c is a unique 3- or 4-digit sequence.

[3] huggingface.co/collections/mbruton/HistCiph
[4] github.com/mbruton0426/ChronoFidelius

Encryption is defined as:

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

that is, each occurrence of c is replaced by a randomly selected code from H_c with equal probability. Keys are generated independently for each text.

Ciphertext generation proceeds left-to-right over plaintext characters. In variants including transcription noise, after sampling a homophonic code for a plaintext character, an error operation is applied with probability $p = 0.05$. When an error occurs, either (i) a deletion operation removes the sampled ciphertext token, or (ii) an insertion operation adds an additional ciphertext token sampled uniformly from the set of tokens already present in the current ciphertext.

Insertions and deletions are applied independently at each plaintext position. At each plaintext position, at most one error operation is applied. To preserve alignment information, the corresponding plaintext variant is produced in parallel and the affected position is marked with the symbol '#' to indicate the presence of a transcription error.

3.3. Dataset Fields

Each 'document' in HistCiph is represented as a structured record containing multiple plaintext, ciphertext, and key variants, as well as associated metadata. An overview of the field structure is shown in Table 3.

For each text, we provide a clean, normalized plaintext version and additional variants with injected transcription errors. Error-marked variants use the symbol `#` within to denote character-level corruption. Plaintext variants are aligned with the corresponding ciphertext regimes.

Ciphertexts are generated using homophonic substitution under two independent noise types that result in four unique ciphertext variants per document: `with_` or `without_errors`, denoting the inclusion of transcription errors, and `with_`

Variant	Examples	Description
`plaintext`	**EN**: THEGOSPE...U **HU**: ESMONDAN...K **SV**: PERBRAHE...S	Clean, normalized plaintext; matches both ciphertext variants without injected transcription errors
`plaintext_with_errors_with_mix`	**EN**: THEGOSPE...U **HU**: E#SMONDA...K **SV**: PERB#RAH...S	Plaintext with injected transcription errors (#) and matching ciphertext variant including variable length (3- and 4-digit) codes
`plaintext_with_errors_without_mix`	**EN**: THEGOSPE...U **HU**: ESMONDAN...K **SV**: PERBR#AH...S	Plaintext variant with injected transcription errors (#) and matching ciphertext variant including fixed length (4-digit) codes
`ciphertext_with_errors_with_mix_code`	**EN**: 8148 0511 961 2209 ... 079 **HU**: 726 2488 3556 8306 ... 2219 **SV**: 9052 553 7159 4047 ... 0145	Ciphertext variant both including transcription errors and variable length (3- and 4-digit) codes
`ciphertext_with_errors_without_mix_code`	**EN**: 7533 0280 6091 3906 ... 4567 **HU**: 8376 7164 2700 8866 ... 7682 **SV**: 6152 5841 5701 4518 ... 6670	Ciphertext variant including transcription errors and fixed length (4-digit) codes
`ciphertext_without_errors_with_mix_code`	**EN**: 5903 1187 840 2910 ... 322 **HU**: 662 9650 7194 390 ... 2020 **SV**: 9330 111 3860 3056 ... 7800	Ciphertext variant without transcription errors and variable length (3- and 4-digit) codes
`ciphertext_without_errors_without_mix_code`	**EN**: 7592 3475 6912 6940 ... 0382 **HU**: 2739 8312 0238 5634 ... 4605 **SV**: 1329 8204 3911 3386 ... 6584	Ciphertext variant without transcription errors and fixed length (4-digit) codes
`key_with_errors_with_mix_code`	**EN**: A:[284, 123], B:[7876], ... **HU**: A:[554, 406, ...], D:[6176, 3051], ... **SV**: A:[958, 220, ...], B:[4047], ...	Homophonic substitution keys corresponding to the respective ciphertext variant
`key_with_errors_without_mix_code`	**EN**: A:[8824, 3381], B:[1579], ... **HU**: A:[5102, 7559, ...], D:[3199, 4267], ... **SV**: A:[6518, 2717, ...], B:[4518, 9300, ...], ...	
`key_without_errors_with_mix_code`	**EN**: A:[081, 271], B:[4428], ... **HU**: A:[286, 769, ...], D:[7264, 6913], ... **SV**: A:[148, 468, ...], B:[7360, 3056], ...	
`key_without_errors_without_mix_code`	**EN**: A:[1499], B:[6659], ... **HU**: A:[5044, 5026, ...], D:[3396, 3294], ... **SV**: A:[6874, 4566, ...], B:[4152, 3386], ...	
`year`	**EN**: 1568 **HU**: 1400 **SV**: 1585	Year of composition
`year_range`	**EN**: 1500-1599 **HU**: 1400-1499 **SV**: 1500-1599	Century-level bin for temporal filtering
`text_length`	**EN**: 50 **HU**: 50 **SV**: 50	Plaintext length bin (characters)
`text_id`	**EN**: text_50_...4 **HU**: text_50_...6 **SV**: text_50_...4	Unique document identifier

Table 3: Multilingual example illustrating the document structure of HistCiph. For each language (EN, HU, SV), we show truncated excerpts of plaintexts, ciphertexts, and corresponding homophonic substitution keys across different variants. Not all errors in each text are visible due to the truncation. Metadata fields are included to support diachronic and length-based filtering.

or `without_mix_code`, denoting the inclusion of variable length code.

Each ciphertext variant is accompanied by its corresponding homophonic substitution key, represented as a mapping from plaintext characters to 1-5 numeric codes. Key names reflect the same noise and code-length conditions as their paired ciphertext.

Additionally, each record contains:

- `year`: year of composition;
- `year_range`: century-level bin for temporal filtering;
- `text_length`: plaintext length bin (in characters), and;
- `text_id`: unique document identifier.

Text lengths within each range are allowed to be up to 10 characters shorter than the defined limit to allow for variation. Ciphertext variants including transcription noise may exceed the matching plaintext length due to insertions. Depending on the `text_length` category, up to 50 additional ciphertext codes may be introduced.

4. Dataset Statistics & Analyses

4.1. Plaintext Character Inventory and Entropy

Across the ten languages, total plaintext volume ranges from 3.8M (French) to over 254M (English) characters, providing substantial statistical support for character-level analysis. Character inventory sizes vary from 37 (Polish) to 138 (Spanish) unique characters, with most languages falling between 39 and 83 distinct symbols.

Shannon entropy over full plaintext corpora ranges from 4.03 (Dutch) to 4.74 (Czech) bits per character, with average across languages of 4.32 bits. Despite orthographic and diachronic variation, entropy therefore remains within a relatively narrow band, suggesting that the information density of running historical text is broadly stable across languages.

Theoretical maximum entropy $H_{\max} = \log_2 |\Sigma|$ varies as a function of character inventory size, producing redundancy values $R = H_{\max} - H$ between 0.65 (Polish) and 2.93 (Spanish) bits across languages. This represents more than a fourfold difference in redundancy, driven primarily by inventory size rather than large differences in empirical entropy; average redundancy across all languages is 1.55.

Languages with larger character inventories exhibit increased theoretical entropy ceilings, but much of this capacity lies in extremely low-frequency characters. These rare characters contribute negligibly to cumulative probability mass ($< 0.005\%$) while expanding the theoretical alphabet and therefore inflating redundancy estimates. The long-tail behaviour observed in languages with large inventories reflects residual multilingual material not entirely eliminated during normalization and historical orthographic variation. However, their minimal cumulative mass indicates negligible impact on empirical entropy values.

4.2. Ciphertext Entropy and Cryptographic Properties

Ciphertexts are generated as homophonic substitution ciphers based on 3- or 4-digit numeric codes drawn uniformly from a global pool of 11,000 possible sequences. The encryption scheme allocates up to five homophones per plaintext character, with ciphertext tokens sampled uniformly from character-specific homophone sets.

The theoretical upper bound on ciphertext entropy is therefore:

$$H_{\max}^{CT} = \log_2(11000) \approx 13.43 \text{ bits.}$$

Observed ciphertext entropy achieves between 99.46% and 99.80% of this theoretical maximum across languages, with an average of 13.37 bits. This confirms effective frequency masking as plaintext frequency imbalances are substantially suppressed, producing near-uniform ciphertext token distributions independent of language-specific orthographic structure.

Because keys are generated independently, ciphertext entropy is not artificially constrained by global code reuse. Entropy estimates therefore reflect intrinsic properties of the encryption design rather than corpus-level artifacts.

Across languages, homophone set size is bounded by a maximum of 5 codes per characters, but allocation is frequency-dependent within each text. The mean allocation increases systematically with text length, from 2.12 codes per character for 50-character texts to 4.50 codes for 1000-character texts. Within each text length, frequently occurring characters receive the maximum of 5 codes, while least frequent characters receive a single code. Intermediate lengths exhibit monotonic growth (2.79 at 100 characters, 3.50 at 200, 4.07 at 400, 4.31 at 600, and 4.43 at 800).

This length-dependent allocation arises because characters must exceed frequency thresholds within a text to receive additional homophones. As text length increases, more characters cross these thresholds, expanding homophone sets and increasing key entropy.

4.3. Unicity Distance

Unicity distance (UD), defined as the expected ciphertext length required to uniquely determine the encryption key, is computed as:

$$U = \frac{H(K)}{R}$$

where $H(K)$ is the key entropy and $R = H_{\max} - H$ is plaintext redundancy. Mean UD values vary across languages as a function of redundancy.

Because keys are generated independently per text, key entropy increases with realized alphabet coverage and therefore scales with text length because more distinct plaintext characters are instantiated and therefore require allocated homophone sets.. Averaged across texts, mean unicity distance ranges from 8.73 (Spanish) to 43.34 (Polish) characters. Languages with higher redundancy (e.g., Spanish) exhibit smaller UD, whereas lower-redundancy languages (e.g., Polish) require longer ciphertexts to uniquely determine keys.

When examined by text length, realized UD increases systematically as more plaintext characters are instantiated. Spanish consistently exhibits the smallest UD values across all lengths, while Polish exhibits the largest. For 50-character texts, UD ranges from 2.25 characters (Spanish) to 37.03 characters (Polish). For 1000-character texts, UD ranges from 14.58 characters (Spanish) to 119.36 characters (Polish). Intermediate lengths show monotonic growth in the same pattern.

Across all languages and length categories (50–1000 characters), ciphertext length generally exceeds the corresponding unicity distance. The dataset therefore operates predominantly in the theoretically uniquely solvable regime, with shorter texts approaching transitional thresholds but rarely falling below their expected UD.

These results demonstrate that recoverability is governed jointly by language-specific redundancy and realized key entropy. While orthographic inventories differ substantially across languages, the encryption procedure yields predictable and internally consistent cryptographic behaviour across the dataset.

5. Discussion

Historical ciphertext decryption constitutes a genuinely low-resource task. Unlike modern NLP benchmarks, no large-scale publicly available corpora of aligned historical plaintext–ciphertext pairs exist. Real-world archival ciphertexts are scarce, unevenly distributed across languages and periods, and often inaccessible due to institutional restrictions. As a result, prior work has relied almost exclusively on self-generated synthetic datasets that are not released or not historically grounded.

The present dataset addresses this gap by combining (i) diachronically balanced historical plaintext corpora, (ii) independently generated simple and homophonic substitution keys for encryption per text, and (iii) controlled transcription noise. This enables systematic experimentation under conditions that approximate real-world archival material.

5.1. Modelling Implications

From a modelling perspective, several properties are particularly relevant. Unlike standard NLP sequence modelling, ciphertext tokens are not linguistically meaningful units and do not correspond to morphemes, words, or whitespace-delimited segments. In variable-length regimes, segmentation ambiguity arises from mixed 3- and 4-digit codes, and in archival practice whitespace was often inconsistently used or omitted . This contrasts with many modern NLP benchmarks, which assume stable token boundaries and error-free segmentation.

Character-Level vs. Code-Level Modelling Plaintext entropy varies modestly across languages, while ciphertext entropy approaches theoretical maxima due to homophonic substitution. As a result, ciphertext code distributions are nearly uniform and largely language-independent. This reduces the effectiveness of naive frequency-based methods and encourages models to exploit structural regularities and longer-range dependencies.

Effect of Homophones The use of up to five homophones per plaintext character increases key entropy and suppresses plaintext frequency leakage. Because homophone allocation is frequency-dependent within each text, longer texts instantiate larger homophone sets and therefore higher key entropy. Keys are generated independently for each text, models cannot rely on cross-text code reuse, preventing trivial memorization strategies and promoting generalizable cryptanalytic learning.

Variable-Length Codes and Noise The inclusion of mixed 3- and 4-digit codes and stochastic insertion/deletion noise introduces segmentation ambiguity and alignment uncertainty. These factors approximate challenges observed in manually transcribed archival ciphers and provide a controlled setting for evaluating robustness to transcription artifacts.

Cross-Linguistic Redundancy Differences Unicity distance varies across languages as a function of measured redundancy. Languages with higher redundancy theoretically require shorter ciphertexts to uniquely determine keys, whereas lower-redundancy languages require

longer sequences. This introduces measurable cross-linguistic differences in intrinsic cryptanalytic difficulty, enabling comparative evaluation of model performance across typologically distinct settings.

Together, these properties position the dataset as a controlled test-bed for studying: ii) cross-lingual transfer; iii) robustness to orthographic variation, and; iv) alignment and/or decryption under partial-information regimes.

- data efficiency in low-resource cryptanalysis;
- cross-lingual transfer;
- robustness to orthographic variation, and;
- alignment and/or decryption under partial-information regimes.

5.2. Generalizability

Beyond historical cryptanalysis, the methodology presented in this paper provides a general framework for principled synthetic data generation in NLP. Rather than producing artificial data solely to increase quantity, we derive synthetic instances from empirically grounded properties of real-world sources and explicitly validate their distributional and structural characteristics. This two-step approach—(i) constraining generation through domain-informed rules and (ii) quantitatively validating whether the resulting data reproduces key statistical signatures of the target setting—can be transferred to a wide range of other low-resource NLP scenarios, including machine translation, morphological analysis, and sequence labelling tasks.

Tasks involving noisy OCR text, dialectal variation, code-switching, or historical language stages could benefit from synthetic augmentation calibrated to observed entropy, token distributions, error profiles, or structural constraints in authentic corpora. In low-resource machine translation, synthetic parallel data could be generated by modelling empirically observed alignment patterns, sentence length distributions, and morphological productivity in authentic corpora, and then validating whether the synthetic pairs preserve cross-lingual entropy relationships and token-frequency profiles. Similarly, in morphological analysis, particularly for highly agglutinative or polysynthetic languages, synthetic word forms could be constructed by modelling attested morpheme inventories, combinatorial constraints, and positional regularities. In such cases, morphemes function analogously to structured “codes”, and controlled recombination can produce scalable training data while preserving typologically grounded constraints.

By coupling controlled generation with quantitative validation, researchers can construct synthetic benchmarks that are not only scalable and reproducible, but also demonstrably representative of the phenomena they aim to model. Such an approach helps mitigate the risk of oversimplified artificial tasks and supports the development of models that generalize more reliably to real-world data.

Crucially, when transferring this methodology beyond historical cryptanalysis, validation remains essential. Synthetic augmentation must be evaluated against real-world distributions and informed by expert or native-speaker knowledge to avoid unintended biases or implausible patterns. The central contribution of this work is therefore not only a dataset for historical cipher research, but also a reproducible framework for quality-controlled synthetic data generation in low-resource settings. Information-theoretic diagnostics such as entropy, redundancy, and related complexity measures may serve as general evaluation tools in low-resource domains where underlying distributions are partially known but large-scale supervision is unavailable.

6. Conclusion

We introduce HistCiph, the first publicly available multilingual dataset of historically grounded plaintext–ciphertext pairs for classical homophonic ciphers. The collection spans ten languages, multiple centuries, and diverse orthographic traditions, combining diachronic plaintext corpora with independently generated homophonic encryption keys and controlled transcription noise.

Through quantitative analyses of plaintext entropy, redundancy, ciphertext entropy, homophone allocation, and unicity distance, we show that the dataset balances linguistic diversity with cryptographic rigour. Ciphertexts exhibit near-maximal entropy and effective frequency masking, while cross-linguistic differences in redundancy yield measurable variation in intrinsic decryption difficulty.

By releasing aligned plaintext, ciphertext, key information, and metadata, this resource enables reproducible experimentation in a previously underserved low-resource setting. We hope HistCiph facilitates research in neural cryptanalysis, cross-lingual transfer, robustness to orthographic variation, and historically informed decryption models, and serves as a foundation for future work on large-scale archival cipher recovery.

More broadly, our work underscores the importance of grounding synthetic data generation in empirically attested properties of real-world sources. By explicitly modelling structural characteristics of historical cryptographic material, such as homophonic allocation strategies, orthographic variation, and transcriptional irregularities, we ensure that synthetic ciphertexts capture not only surface statistical patterns but also the procedural constraints

shaping authentic encryption practices. This alignment between structure and computational generation is crucial for constructing representative benchmarks, reducing the risk of oversimplified artificial tasks, and supporting methods that generalize more robustly to real-world material.

7. Limitations

The dataset uses synthetic encryption of real-world historical plaintext, rather than real-world historical plaintext–ciphertext pairs. While this enables controlled experimentation, large-scale generation, systematic manipulation of cipher parameters, and appears to match the properties of these real world ciphertexts; it does not capture all properties of authentic archival material. In particular, the encryption procedure does not model semantic obfuscation, key reuse across documents, or historically idiosyncratic codebook design.

Historical cipher systems often incorporated additional layers of concealment beyond character-level substitution. For example, sensitive names, locations, or political terms were frequently replaced with dedicated codebook entries, null symbols were inserted strategically to mislead frequency analysis, and abbreviatory conventions could compress semantically salient expressions into single high-value codes. Such practices introduce structured irregularities and semantic asymmetries that are not fully reproduced by uniform homophonic substitution.

A further difference concerns key reuse. Archival collections commonly show the same key being reused across multiple documents, sometimes over extended periods. This creates cross-document statistical dependencies that may facilitate cryptanalysis through comparative frequency analysis, crib-dragging, or partial key reconstruction. In contrast, HistCiph generates independent keys per text to ensure experimental control and avoid unintended information leakage between dataset splits. While this design supports reproducibility and clean evaluation, it removes an important dimension of historical realism by eliminating opportunities for modelling cross-document attacks.

Codebook design presents an additional limitation. Authentic historical keys were rarely constructed according to uniform allocation principles. Surviving examples show uneven homophone distributions, special-purpose symbols, hierarchical code groups, and ad hoc expansions introduced over time. These idiosyncrasies reflect operational constraints, scribal practice, evolving security needs, and pragmatic adjustments. Our controlled homophonic allocation strategy abstracts from such variability in order to isolate the effects of redundancy, entropy, and code multiplicity. Future extensions could incorporate empirically derived key distributions to better approximate archival complexity.

Noise modelling also remains simplified. Transcription noise is applied at a fixed rate ($p = 0.05$), whereas real-world error rates vary substantially depending on manuscript condition, editorial conventions, and transcription methodology. Introducing variable or corpus-calibrated noise models would increase ecological validity and enable finer-grained robustness testing.

Finally, large character inventories observed in some languages partly reflect residual multilingual material, loanwords, and rare graphemic variants present in historical corpora. Although the probability mass of these low-frequency characters is negligible, they increase theoretical redundancy estimates and may slightly affect derived unicity values. Further corpus-level cleaning or frequency-thresholding strategies could mitigate this effect in future releases.

8. Acknowledgments

This work has been supported by Riksbankens Jubileumsfond, grant M24-0028: Echoes of History: Analysis and Decipherment of Historical Writings (DESCRYPT).The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725. Additionally, we would like to thank Nils Kopal for his input on cryptanalytic analyses.

9. Bibliographical References

Gulinigeer Abudouwaili, Wayit Ablez, Kahaerjiang Abiderexiti, Aishan Wumaier, and Nian Yi. 2023. Strategies to Improve Low-Resource Agglutinative Languages Morphological Inflection. In *Conference on Computational Natural Language Learning*, pages 508–520, Singapore. Association for Computational Linguistics.

Ezat Ahmadzadeh, Hyunil Kim, Ongee Jeong, Namki Kim, and Inkyu Moon. 2022. A Deep Bidirectional LSTM-GRU Network Model for Automated Ciphertext Classification. *IEEE Access*, 10:3228–3237.

Ezat Ahmadzadeh, Hyunil Kim, Ongee Jeong, and Inkyu Moon. 2021. A Novel Dynamic Attack on Classical Ciphers Using an Attention-Based LSTM Encoder-Decoder Model. *IEEE Access*, 9:60960–60970.

Eugen Antal, Pavol Marák, Pavol Zajac, Tünde Lengyelová, and Diana Duchoňová. 2023. Encrypted Documents and Cipher Keys From the 18th and 19th Century in the Archives of Aristocratic Families in Slovakia. In *International Conference on Historical Cryptology*, Germany. Linköping University Electronic Press.

Micaella Bruton and Beata Megyesi. 2025. From Statistics to Neural Networks: Enhancing Ciphertext-Plaintext Alignment in Historical Substitution Ciphers for Automatic Key Extraction. In *International Conference on Historical Cryptology*, Poland. Tartu University Library.

Santwana Chimalamarri and Dinkar Sitaram. 2021. Linguistically enhanced word segmentation for better neural machine translation of low resource agglutinative languages. *International Journal of Speech Technology*, 24(4):1047–1053.

Ona de Gibert, Joseph Attieh, Teemu Vahtola, Mikko Aulamo, Zihao Li, Raúl Vázquez, Tiancheng Hu, and Jörg Tiedemann. 2025. Scaling Low-Resource MT via Synthetic Data Generation with LLMs. In *Empirical Methods in Natural Language Processing*, pages 27674–27692, Suzhou, China. Association for Computational Linguistics.

Camille Desenclos and George Lasry. 2024. An Early French Digit Dipher: Deciphering a Letter from the King of France to the Duke of Nevers (1592). In *International Conference on Historical Cryptology*, United Kingdom. Tartu University Library.

Chandra Kiran Evuru, Sreyan Ghosh, Sonal Kumar, Ramaneswaran S, Utkarsh Tyagi, and Dinesh Manocha. 2024. CoDa: Constrained Generation based Data Augmentation for Low-Resource NLP. In *North American Chapter of the Association for Computational Linguistics*, pages 3754–3769, Mexico City, Mexico. Association for Computational Linguistics.

Nino Fürthauer, Vasily Mikhalev, Nils Kopal, Bernhard Esslinger, Harald Lampesberger, and Eckehard Hermann. 2022. Evaluating Deep Learning Techniques for Known-Plaintext Attacks on the Complete Columnar Transposition Cipher. In *International Conference on Historical Cryptology*, The Netherlands. Linköping University Electronic Press.

Unnikrishnan Gopinathan, David S. Monaghan, Thomas J. Naughton, and John T. Sheridan. 2006. A Known-Plaintext Heuristic Attack on the Fourier Plane Encryption Algorithm. *Optics Express*, 14(8):3181–3186.

David Kahn. 1996. *The Codebreakers: The Comprehensive History of Secret Communication from Ancient Times to the Internet*. Scribner, New York, NY.

Nishant Kambhatla. 2024. *Augmented Input Representations in Sequence Generation Models for Decipherment and Translation*. Doctoral Thesis, Simon Fraser University, Canada.

Hayato Kimura, Keita Emura, Takanori Isobe, Ryoma Ito, Kazuto Ogawa, and Toshihiro Ohigashi. 2022. Output Prediction Attacks on Block Ciphers Using Deep Learning. In *Applied Cryptography and Network Security Workshops*, volume 13285 of *Lecture Notes in Computer Science*, pages 248–276, Cham. Springer.

Nils Kopal and Michelle Waldispühl. 2022. Deciphering Three Diplomatic Letters Sent by Maximilian II in 1575. *Cryptologia*, 46(2):103–127.

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

Ernst Leierzopf, Nils Kopal, Bernhard Esslinger, Harald Lampesberger, and Eckehard Hermann. 2021a. A Massive Machine-Learning Approach For Classical Cipher Type Detection Using Feature Engineering. In *International Conference on Historical Cryptology*, pages 111–120.

Ernst Leierzopf, Vasily Mikhalev, Nils Kopal, Bernhard Esslinger, Harald Lampesberger, and Eckehard Hermann. 2021b. Detection of Classical Cipher Types with Feature-Learning Approaches. In *Data Mining*, volume 1504 of *Communications in Computer and Information Science*, pages 152–164, Singapore. Springer.

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 *Proceedings of the 6th International Conference on Historical Cryptology (HistoCrypt 2023)*, pages 120–129.

Beáta Megyesi, Nils Blomqvist, and Eva Pettersson. 2019. The decode database: Collection of historical ciphers and keys. In *Proceedings of the 2nd International Conference on Historical Cryptology, HistoCrypt 2019*. NEALT Proceedings Series 37, Linköping Electronic Press.

Beáta Megyesi, Alicia Fornés, Nils Kopal, Benedek Láng, Michelle Waldispühl, Vasily Mikhalev, and Bernhard Esslinger. 2024a. *Historical Cryptology*. Artech House.

Beáta Megyesi, Crina Tudor, Benedek Láng, Anna Lehofer, Nils Kopal, Karl de Leeuw, and Michelle Waldispühl. 2024b. Keys with nomenclatures in the early modern europe. *Cryptologia*, 48(2):97–139.

Aloys Meister. 1902. *Die Anfänge der modernen diplomatischen Geheimschrift.* Paderborn: Ferdinand Schöningh.

Aloys Meister. 1906. *Die Geheimschrift im Dienste der Päpstlichen Kurie von Ihren Anfängen bis zum Ende des XVI. Jahrhunderts*, volume 11. F. Schöningh.

Francois Meyer and Jan Buys. 2024. Triples-to-isiXhosa (T2X): Addressing the Challenges of Low-Resource Agglutinative Data-to-Text Generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16841–16854, Torino, Italia. ELRA and ICCL.

Mihai Nadăş, Laura Dioşan, and Andreea Tomescu. 2025. Synthetic Data Generation Using Large Language Models: Advances in Text and Code. *IEEE Access*, 13:134615–134633.

Eva Pettersson and Beáta Megyesi. 2018. The histcorp collection of historical corpora and resources. In *Proceedings of the Third Conference on Digital Humanities in the Nordic Countries*.

Sevitha Simhadri, Raghavendra, and B R Purushothama. 2025. AI-Powered Cryptanalysis: Identifying Encryption Algorithms and Recovering Plaintext. In *International Conference on Networks and Cryptology*, pages 1522–1527, India.

Xusen Yin, Nada Aldarrab, Beáta Megyesi, and Kevin Knight. 2019. Decipherment of Historical Manuscript Images. In *International Conference on Document Analysis and Recognition*, pages 78–85.

A. Dataset Breakdown By Year Range

	Train			Validation			Test		
	Min	**Max**	**Spread**	**Min**	**Max**	**Spread**	**Min**	**Max**	**Spread**
Czech	37,922	38,079	0.41	4,740	4,756	0.34	4,750	4,768	0.38
Dutch	98,384	98,631	0.25	12,293	12,325	0.26	12,306	12,338	0.26
English	240,507	241,217	0.30	30,050	30,151	0.34	30,065	30,159	0.31
French	4,184	4,813	15.03	523	601	14.91	525	608	15.81
Hungarian	29,132	29,153	0.07	3,642	3,644	0.05	3,648	3,652	0.11
Icelandic	11,200	11,219	0.17	1,397	1,402	0.36	1,400	1,409	0.64
Italian	40,948	40,975	0.07	5,117	5,119	0.04	5,128	5,129	0.02
Polish	130,046	131,688	1.26	16,320	16,472	0.93	16,217	16,477	1.60
Spanish	117,126	117,224	0.08	14,626	14,650	0.16	14,651	14,662	0.08
Swedish	67,988	68,210	0.33	8,492	8,524	0.38	8,494	8,531	0.44

Table 4: Minimum, maximum, and relative spread of document counts across 100-year intervals for each dataset split.