Historical Language Models in Cryptanalysis: Case Studies on English and German

Beáta Megyesi and
Justyna Sikora
Uppsala University
Sweden

Filip Fornmark and
Michelle Waldispühl
University of Gothenburg
Sweden

Nils Kopal and
Vasily Mikhalev
University of Siegen
Germany

Abstract

In this paper, we study the impact of language models (LM) on decipherment of historical homophonic substitution ciphers. In particular, we investigate if decipherment by using hill-climbing and simulated annealing can benefit from LMs generated from historical texts in general and century-specific texts in particular. We carry out experiments on homophonic substitution ciphers with English and German as plaintext languages. We take into account ciphertext length as well as n-gram size of the LMs. We compare the results on decipherment based on historical LMs with large LMs generated from modern texts. The results show that using historical LMs in decipherment of homophonic substitution ciphers leads to significantly better performance on ciphertext produced in the 17th century or earlier, and century-specific language models yield better results on longer and older ciphertexts.

1 Introduction

One of the main components in cryptanalysis of historical ciphertexts is language models of the underlying plaintext. Oftentimes the choice of texts on which the language model is based is opportunistic, i.e. the cryptanalysts choose what texts or models are available. Since collections of (more or less) contemporary texts, such as the collection of the Gutenberg project or the Wikipedia articles are accessible and freely available for many languages, these are often used by cryptanalysts for the generation of LMs, e.g. by Lasry (2018) and Bean (2020). However, using language models derived from contemporary languages might not be optimal for the decipherment of historical ciphertexts since language changes over time.

Before spelling and grammar were normalized for many European written languages in the course of the 18th and 19th century, we find a large variation in spelling in historical texts. The variation can be found not only across regions and writers but the same author could also spell the same word in the same document differently. The use of punctuation marks such as dots and commas was rarer than in modern texts. Further, writers used abbreviations to save space of the expensive paper or parchment. Another important aspect when dealing with historical texts is the fact that language changes over time; new words enter the language while others disappear. Not only words, but also the grammatical structure of the language changes with respect to word order and the internal structure of words (so called morphology). Given the above mentioned reasons we can expect that the usage of LMs generated from contemporary or historical texts have an impact on decipherment accuracy.

In this paper we aim to investigate the role of historical LMs in the decipherment process of historical ciphertexts. In particular, we are interested in finding the answer to the following research questions:

- Do historical LMs have a positive impact on the decipherment of historical ciphers?
- Do historical LMs created from the same century as the cipher originates from lead to an increase in decipherment performance?
- Does increasing the n-gram size of a model result in an increase of accuracy in decipherment of historical manuscripts?

We present a pilot study on English and German ciphertexts and LMs generated from texts originating from the 14th to the 20th centuries of various n-gram sizes. We focus on homophonic substitution ciphers of various lengths. As historical homophonic substitution ciphertexts have been successfully deciphered by using hill-climbing and simulated annealing (see e.g. Lasry et al. (2020)), we chose to apply the same decipherment method in our experiments.

In Section 2, we present previous studies on using historical LMs in decipherment. In Section 3, we give an overview of the data sets used for the experiments on English and German. In Section 4, we describe the method with the experimental setup. In Section 5, the results from the various experiments are presented, and discussed in Section 6. Finally, in Section 7, we conclude our paper.

2 Background

The usage of language models generated from the underlying plaintext language of the cipher is inevitable for successful decipherment. Already in the 9th century, the Arab philosopher and mathematician Al-Kindi described the value of frequency analysis derived from the plaintext and the ciphertext in cryptanalysis. Since then, cryptographers used LMs with information about frequencies of letters and letter co-occurrences to create models of the underlying plaintext. One of the most simplest, efficient and also commonly occurring language models are n-grams that use a statistical approach to predict the probability of a word or character given its context. N-gram models are built by analyzing a sequence of n words or characters in a text corpus and building a probability distribution of the next word or character based on the occurrence of each n-gram in the training data. The distribution of various n-gram orders (e.g. unigrams, bigrams, trigrams) is also reflected on the ciphertext and can thus give more clues into how the text was encrypted (Kahn, 1996) and (Dooley, 2018).

In order to train or generate LMs for decipherment purposes, a wide range of text collections are used. These might include the translation of the Human Rights, texts from Wikipedia, or Google books. When dealing with historical ciphers, the most commonly used corpus is the Gutenberg collection from Project Gutenberg[1]. The Gutenberg project is a digital library of free e-books, collected since 1971. The collection consists of more than 60.000 books in a large number of languages, for which U.S. copyright has been expired; the great majority originating from the 19th century. The collection is publicly available and copyright-free, which explains its popularity to use when building historical language models.

Another collection of historical texts is the HistCorp corpus (Pettersson and Megyesi, 2018) which contains sixteen European languages including Czech, Dutch, English, French, German, Greek, Hungarian, Icelandic, Italian, Latin, Polish, Portuguese, Russian, Slovene, Spanish, and Swedish. The transcriptions of the original manuscripts are diplomatic editions, i.e. the orthography of the original text is kept and mistakes in the original are preserved. The texts are released in a uniform format. Noteworthy is that the number of texts and the data size included for the various languages vary greatly in the collection depending on what kind of historical text corpora are available for the particular language.

To generate language models, character as well as word-based models are common: unigram, bigram, trigram up to sixgram models are used for the purpose of decipherment. Naturally, the higher the order of the n-grams, the more texts are needed to generate suitable models, and the bigger the models become.

Surprisingly, even though historical cryptologists agree on the importance of LMs in cryptanalysis, the role of historical LMs has not been studied before, neither extensively, nor systematically.

A few studies report, however, on the evaluation of various n-gram sizes in decipherment using modern texts. Ravi and Knight (2008) present a method for solving substitution ciphers using low-order character-based n-gram models and show how decipherment accuracy varies as a function of cipher length and n-gram order.

The same authors (Ravi and Knight, 2011) present a Bayesian approach for deciphering complex substitution ciphers and evaluate with different setups of LMs including character-based bigrams and trigrams, as well as word-based trigrams. They conclude that the best decipherment results are achieved with trigram models and a word list.

Nuhn et al. (2014) apply a method for solving substitution ciphers, including the Zodiac-408 cipher, and evaluate n-gram models of orders four, five and six, where the sixgram models performed the best with lowest error rate.

[1] www.gutenberg.org

Hauer et al. (2014) presents an approach for deciphering monoalphabetic substitution ciphers that combines both character-level and word-level based LMs. In the above mentioned studies the language models were generated from contemporary texts, and evaluated in the light of specific approaches.

Indications for using historical sources instead of contemporary texts for the decipherment of historical ciphertexts have been given in the study by Pettersson and Megyesi (2019) for the automatic language identification task in three types of historical ciphertexts. The results showed that historical LMs perform considerably better on the tested languages (German and Italian) and that using models based on historical texts enables to capture old word forms that are not present in modern corpora, despite their larger size.

Historical LMs were also used to identify cleartext and its language in historical ciphertexts (Gambardella et al., 2022). The authors conducted a series of experiments on 214 documents in 8 languages including Dutch, French, Hungarian, Italian, Latin, Portuguese, and Spanish, and tested the ability of the models in various n-gram settings; both character and word based, trained on historical corpora from the HistCorp collection.

More extensive studies on the impact of historical LMs in decipherment are based on two thesis works: one bachelor's thesis in linguistics carried out by Fornmark (2022) on English, and one master's thesis in language technology by Sikora (2022) on German. The thesis works are based on the same method designed by the authors of this paper. In this study, we built upon the two theses and compare the similarities and differences of the decipherments using modern vs historical LMs for the two languages.

3 Text Collections

To carry out experiments on the impact of historical texts in decipherment for English and German, we use the HistCorp collection for the generation of LMs from historical texts, and contemporary texts from the Gutenberg project.

In English, the earliest texts are mostly Biblical sources, for example from The EDGeS Diachronic Bible Corpus (EDGes) (Bouma et al., 2020), whereas the later texts are of different genres and styles, including sources from the Corpus of Late Modern English Texts (DeSmet, 2006) and the Lampeter Corpus of Early Modern English Tracts (Schmied et al., 1999).

The German data consists of texts from the HistCorp collection including material from the Deutsches TextArchiv (DTA) (Textarchiv, 2010), the EDGeS Diachronic Bible Corpus (Bouma et al., 2020), the Nottingham Corpus of Early Modern German Midwifery and Women's Medicine (GeMi) (Whitt, 2016), GerManC (Durrell et al., 2012), Reference Corpus of Middle High German (ReM) (Klein et al., 2016), Reference Corpus of Middle Low German/Low Rhenish (ReN) (Schröder, 2018), and Register in Diachronic German Science (Ridges) (Lüdeling et al., 2016).

The texts from all corpora for both languages were then sorted into centuries to create subdomains per time interval of 100 years to serve as basis for the creation of century-specific LMs.

For English, texts from between the 11th and 13th centuries as well as the 15th century are missing from the source material, and could thus not be included. For German only one time period, namely the 20th century, could not be represented due to data sparseness. More detailed information about the data source and creation is given by Fornmark (2022) for English and by Sikora (2022) for German.

4 Method

To investigate whether LMs generated on historical corpora can lead to a better performance on decipherment of historical ciphers, compared to LMs generated on large modern corpora, we choose to experiment with English and German as the underlying plaintext languages. We use plaintexts from the 11th to the 20th centuries. We first describe the features that might have effect on the decipherment results. Then, we present the experimental design with a walk-through of the various stages.

4.1 Features

For historical ciphers, we select one of the most commonly occurring types, namely homophonic substitution ciphers with a mixture of the number of homophones per plaintext alphabet letter. The reason behind varying the number of homophones for each plaintext letter is to create more authentic ciphers which are similar to original ciphers, as retrieved from European archives and libraries. Previous studies showed that in homophonic substitution ciphers we can find variable number of code elements for different plaintext letters depending on their frequency in the particular language; frequently occurring plaintext letters usually receive two or three code elements, while less frequent letters are assigned one code element (Kahn, 1996) and (Megyesi et al., 2022).

To measure the correlation between decipherment accuracy and the order of LMs on the length of ciphertexts, we apply ciphertexts consisting of 200 and 500 characters. The decision was determined by the assumption that generating shorter ciphers would significantly increase the level of decipherment difficulty. For the investigation of the impact of various n-gram sizes on decipherment, we experiment with trigram, fourgram, and fivegram character-based LMs. The experimental setup of the features and their values that our study is built upon is summarized in Table 1.

Feature	Values
Language:	German, English
Time period (cent.):	11th-20th
N-gram size:	3, 4, 5
Ciphertext length (chars):	200, 500

Table 1: Experimental setup with features and their values.

4.2 Experiment design

Our point of departure is a ciphertext and its corresponding plaintext for evaluation without any access to the cipher key. First, we collect and preprocess the plaintexts in English and German to generate the plaintext alphabet for various centuries for both languages. Since we do not have access to the original ciphertexts (with their corresponding plaintext) we need to generate them to be able to evaluate the decipherment results. We then create the LMs of various order sizes given the plaintexts for the various centuries and languages. Finally, we run cryptanalysis using the various LMs and evaluate the output. The entire process is illustrated in Figure 1.

4.2.1 Alphabet generation

To create century specific LMs, we generate plaintext alphabets for the specific time periods by counting unigram frequencies on the basis of texts in the HistCorp collection for both languages. It turns out that the size of the alphabet varies greatly across centuries, which causes memory problems in the generation of LMs. Therefore, we set several threshold values for letter frequencies: 0.001, 0.01 and 0.05, and linguistically analyze the alphabet output. We decide to use characters more frequent than 0.01% for English and 0.05% for German.

We then normalize the output of the alphabets; all letters are upper-cased, with the exception of the German *SS* letter[2], while punctuation marks, digits and white space are removed. Furthermore, we expand abbreviations and transform characters consisting of a regular letter and a superscript letter into two separate characters.

4.2.2 Key, ciphertext and plaintext generation

To be able to test the impact of historical and contemporary LMs on decipherment, the keys, ciphertexts and their corresponding plaintexts are automatically generated. To generate timely typical historical cipher keys, we studied original cipher keys collected from the DECODE database (Megyesi et al., 2019) with certain characteristics. We decided to derive homophonic substitution keys containing English or German plaintext or cleartext languages with available transcription. Since language identification in cipher keys might be challenging, we studied more carefully the words in the nomenclature list to decide whether the cipher key was created for English or German plaintexts.

To generate plaintext we choose the HistCorp collection. For each language, we split the language specific data set in HistCorp into a training set and a test set in the portion of 80%–20%, respectively. Table 2 presents the number of texts in the training and test sets for English and German. The training set serves for the generation of LMs in the subsequent step while the test set is used for the gold-standard decrypted text to be used for evaluation. To be able to make automatic comparisons between the automatically generated decrypted text and their corresponding gold-standard, the test set is preprocessed; normalized to capitalize all characters, and double spaces, punctuation, and non-letter characters are removed.

The text files for each century are randomly

[2] We keep the letter *SS* in lowercase, since Python returns the upper-cased strings, and transforms *SS* into double *S* - “SS”, which causes problems in frequency analysis.

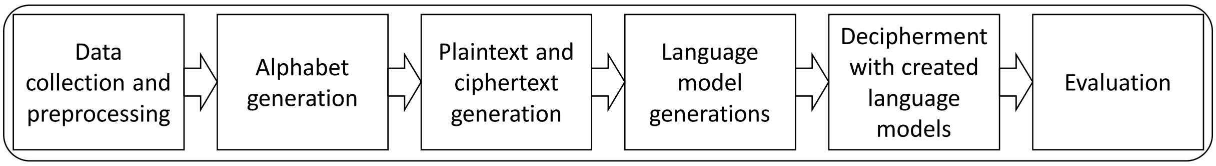

Data collection and preprocessing
Alphabet generation
Plaintext and ciphertext generation
Language model generations
Decipherment with created language models
Evaluation

Figure 1: Method.

Cent.	Training		Test	
	English	German	English	German
11th	–	22	–	6
12th	–	155	–	38
13th	–	98	–	24
14th	60	70	15	18
15th	–	74	–	19
16th	64	39	16	10
17th	122	275	30	69
18th	184	582	46	145
19th	179	529	45	132
20th	46	–	12	–

Table 2: Number of documents in training and test sets in English and German per century.

shuffled to create a random selection of texts. Then, 200 vs 500 characters of each plaintext file are extracted to create plaintexts of various lengths.

Given the keys and the plaintexts, we then generate ciphertexts. For each time period, ciphertexts are created, both of length 200 and 500.

Furthermore, for each ciphertext of a given length and for a given time period, homophonic ciphers with a mixed number of homophones are built. The keys use the percentage value of the measured character-based unigram frequencies described above. A number of homophones between one and five is assigned to each plaintext letter. Between three and five homophones for the most common 14 letters, one homophone for the least common, and between two and three homophones for letters in between. Null characters and nomenclature elements are not considered.

In the experiments, the keys and ciphertexts are all numeric, and use a fixed, uniform length for each ciphertext letter. If the key needs less than 100 homophones, all plaintext letters map to two numbers (00, 01, . . . , 99) in the ciphertext. Otherwise, the plaintext letters map to three numbers (000, 001, . . . , 999) in the ciphertext. An example of a generated cipher key is illustrated in Figure 2.

```
A:[99|18|12|68]       H:[77|08|67]     O:[86|35|41|75]     V:[65]
B:[60]                I:[00|38|34|97]  P:[52]              W:[54]
C:[11]                J:[47]           Q:[30]              X:[20]
D:[29|61]             K:[33]           R:[98|56|24]        Y:[62]
E:[92|13|40|17|19]    L:[88|42]        S:[14|22|79]        Z:[23]
F:[93]                M:[28]           T:[27|84|51|80|96]
G:[72]                N [89|21|74]     U:[45]
```

Figure 2: A generated key; homophones per plaintext letter based on unigram frequencies.

4.2.3 Language model generation

Before generating the models themselves, duplicated texts – the same texts appearing in several centuries – are removed. Then, character-based n-gram models of order 3, 4, and 5 are created from the training set for each century, and a more generic model with all texts available for each language from the Gutenberg collection.

The model format consists of a data and a metadata section. The data section is an array with n-grams and their respective frequencies. The number of occurrences of any particular n-gram is stored as the logarithm of the frequency of that n-gram relative to the full body of text data. The metadata section, which is located in the beginning of the LM file starts with a file identifier “CTLS” (CrypTool Language Statistics), followed by the language code (“EN” for English and ”DE” for German), an integer describing the “n” value of the gram, and the model alphabet.

From the training sets, various character-based models for the three n-gram sizes are created for the different centuries, as well as combined models are generated from all texts.

Lastly, word-based LMs are generated from the texts. The data is cleaned up by removing residual punctuation for both languages. For English, the diacritics were also removed while for German they were kept. The resulting format is a single word per line, with words occurring in the source material. A combined dictionary from the included English and German HistCorp material is also created along with a general German and English dictionary generated from the Gutenberg data.

4.2.4 Decipherment

The cryptanalysis is performed using CrypTool 2[3] (CT2), a freely available open source tool[4] which allows the automatic decipherment of historical and modern ciphers (Kopal, 2018). CT2 contains a component for the cryptanalysis of homophonic substitution ciphers, the so-called Homophonic Substitution Analyzer, see Figure 3. To ease the use of the cryptanalytic algorithm implemented in the component, we extracted the core cryptanalysis algorithm. Thus, it could be used without the need of starting a full-blown CT2 instance. This furthermore speed up the cryptanalysis and allowed us to perform several hundreds of cryptanalysis runs per model needed for our evaluations.

To decipher a given ciphertext, CT2's Homophonic Substitution Analyzer component implemented with hill climbing with simulated annealing (Kopal, 2019) was used. Additionally, a dictionary of common words was given to the algorithm. During the cryptanalysis process, the dictionary is used to already "lock" partially correct decipherments to improve and speed up the further analysis process.

4.2.5 Evaluation

To evaluate the effect the LMs have on decipherment, we calculate decipherment accuracy as defined in the equation below.

$$\mathbf{Correct} = \sum_{\mathbf{i=0}}^{\mathbf{Length-1}} \mathbf{n} \begin{cases} 1, & \text{if D[i] = P[i]} \\ 0, & \text{otherwise} \end{cases}$$

where Length is the length of the ciphertext, D is the deciphered ciphertext, and P is the real plaintext.

We compute the percentage of the correctly deciphered letters of a deciphered ciphertext as defined below.

$$\mathbf{Accuracy} = (Correct/Length) \cdot 100$$

Finally, a LM receives a "point" if it was able to decipher a given ciphertext with $Accuracy \geq 80\%$. Here, only the best of all models received a point. But if more than one model lead to the same Accuracy, all these models obtained a point since they performed equally good. To evaluate the different LMs, we compare the number of points each LM received in our evaluation. The graphs in the Results section show how many points each LM received. We evaluated each LM 500 times for each time period by cryptanalyzing each generated ciphertext using each of the LMs.

5 Results

We provide the results for English in Figures 4 and 5 and for German in Figure 6.

For the English texts we found that texts composed in a certain time period were, in general, best analyzed by the model trained on texts from the same century. This trend was especially clear for the earliest texts, i.e. for texts from the 14th, 16th and 17th centuries. For later texts it became less clear which model performed the best, but there was a marked drop in the performance of the early models. These results can be linked to the development of English orthography, which with the spread of dictionaries in the 18th century became more standardized.

Given the choice of the n-gram order, the results showed that 5-gram models achieved the highest decipherment accuracy on more modern texts, and 4-gram models led to highest performance for older ones, produced in the 17th century or earlier.

Not surprisingly, the longer ciphertexts (500 characters) achieved higher decipherment accuracy in general.

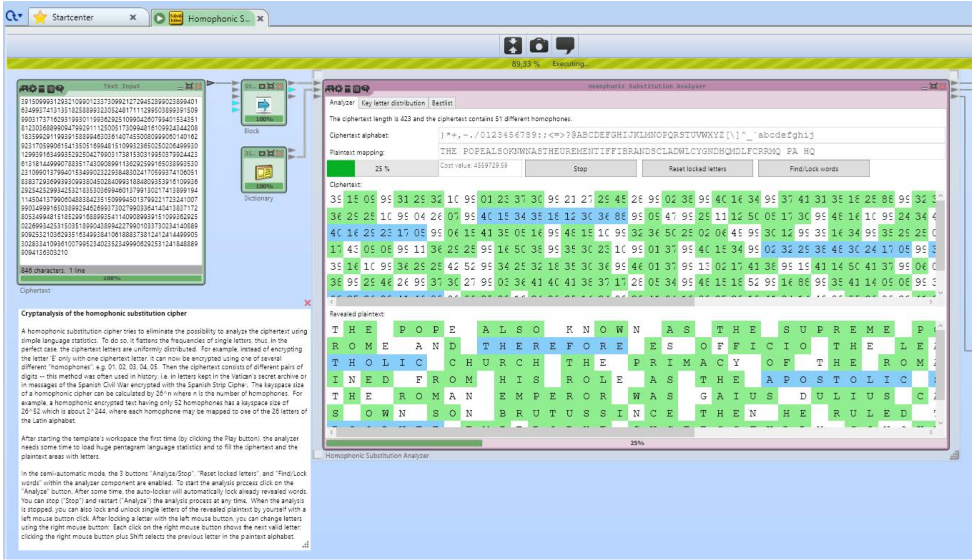
The order size of LMs has also impact on the decipherment performance of the cipherlength. We found that shorter texts require higher order n-grams; the 5-gram models performed better compared to the 3-gram and 4-gram models on the 200 character long text.

Interestingly, using the Gutenberg model compared to the model generated from the combination of all text material (1350-1999), the results are diverse. While Gutenberg 4-and 5-gram models perform best on longer and more modern texts (18th-20th century), the historical 4- and 5-gram merged model yields best result for shorter texts. However, noteworthy is that for all 3-gram models, the historical merged model leads to best decipherment performance.

For German, the results show similar trends, albeit not as pure and straightforward as for English.

[3] https://www.cryptool.org/en/ct2/

[4] https://github.com/CrypToolProject/CrypTool-2



846 characters, 1 line

Cryptanalysis of the homophonic substitution cipher

A homophonic substitution cipher tries to eliminate the possibility to analyze the ciphertext using simple language statistics. To do so, it flattens the frequencies of single letters, thus, in the perfect case, the ciphertext letters are uniformly distributed. For example, instead of encrypting the letter 'E' only with one ciphertext letter, it can now be encrypted using one of several different "homophones", e.g. 01, 02, 03, 04, 05. Then the ciphertext consists of different pairs of digits -- this method was often used in history, i.e. in letters kept in the Vatican's secret archive or in messages of the Spanish Civil War encrypted with the Spanish Strip Cipher. The keyspace size of a homophonic cipher can be calculated by 26^n where n is the number of homophones. For example, a homophonic encrypted text having only 52 homophones has a keyspace size of 26^52 which is about 2^244, where each homophone may be mapped to one of the 26 letters of the Latin alphabet.

After starting the template's workspace the first time (by clicking the Play button), the analyzer needs some time to load huge pentagram language statistics and to fill the ciphertext and the plaintext areas with letters.

In the semi-automatic mode, the 3 buttons "Analyze/Stop", "Reset locked letters", and "Find/Lock words" within the analyzer component are enabled. To start the analysis process click on the "Analyze" button. After some time, the auto-locker will automatically lock already revealed words. You can stop ("Stop") and restart ("Analyze") the analysis process at any time. When the analysis is stopped, you can also lock and unlock single letters of the revealed plaintext by yourself with a left mouse button click. After locking a letter with the left mouse button, you can change letters using the right mouse button: Each click on the right mouse button shows the next valid letter; clicking the right mouse button plus Shift selects the previous letter in the paintext alphabet.

Homophonic Substitution Analyzer

Analyzer | Key letter distribution | Bestlist

The ciphertext length is 423 and the ciphertext contains 51 different homophones.

Ciphertext alphabet:)*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghij

Plaintext mapping: THE POPEALSOKNWNASTHEUREMENTIFFIBRANDSCLADWLCYGNDHQMDLFCRRMQ PA HQ

25 % | Cost value: 4559729,59 | Stop | Reset locked letters | Find/Lock words

Ciphertext:

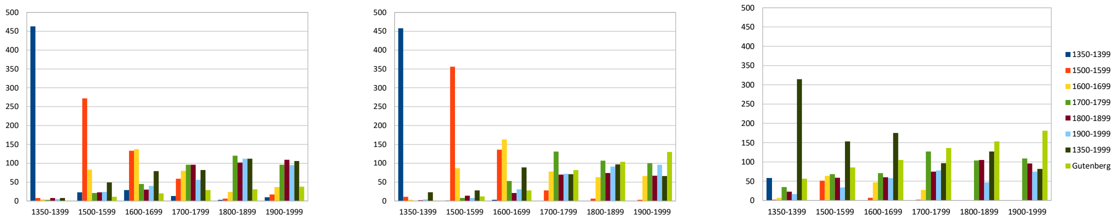
Revealed plaintext:

```
THE POPE ALSO KNOWN AS THE SUPREME P
ROME AND THEREFORE ES OFFICIO THE LE
THOLIC CHURCH THE PRIMACY OF THE ROM
INED FROM HIS ROLE AS THE APOSTOLIC
THE ROMAN EMPEROR WAS GAIUS DULIUS C
S OWN SON BRUTUSSINCE THEN HE RULED
```

25%

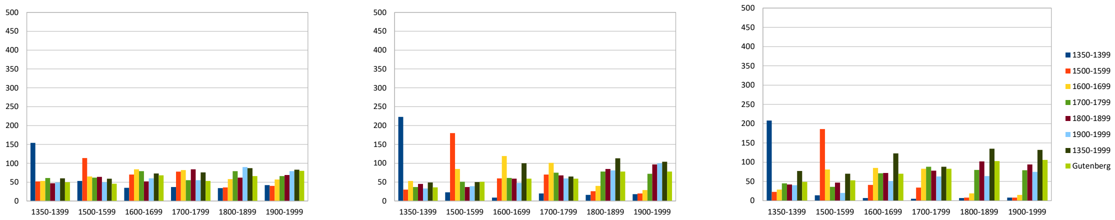
Homophonic Substitution Analyzer

Figure 3: Breaking homophonic substitution ciphers in CrypTool2 .



(a) 3-grams. (b) 4-grams. (c) 5-grams.

Figure 4: Best result for English: 500 character long ciphertexts.



(a) 3-grams. (b) 4-grams. (c) 5-grams.

Figure 5: Best result for English: 200 character long ciphertexts.

Similar to English, historical century-specific LMs achieve best decipherment performance, but for German, that applies to all historical texts produced earlier than the 19th century. The 16th century model yielded most impressive results for ciphertexts produced up to the 18th century.

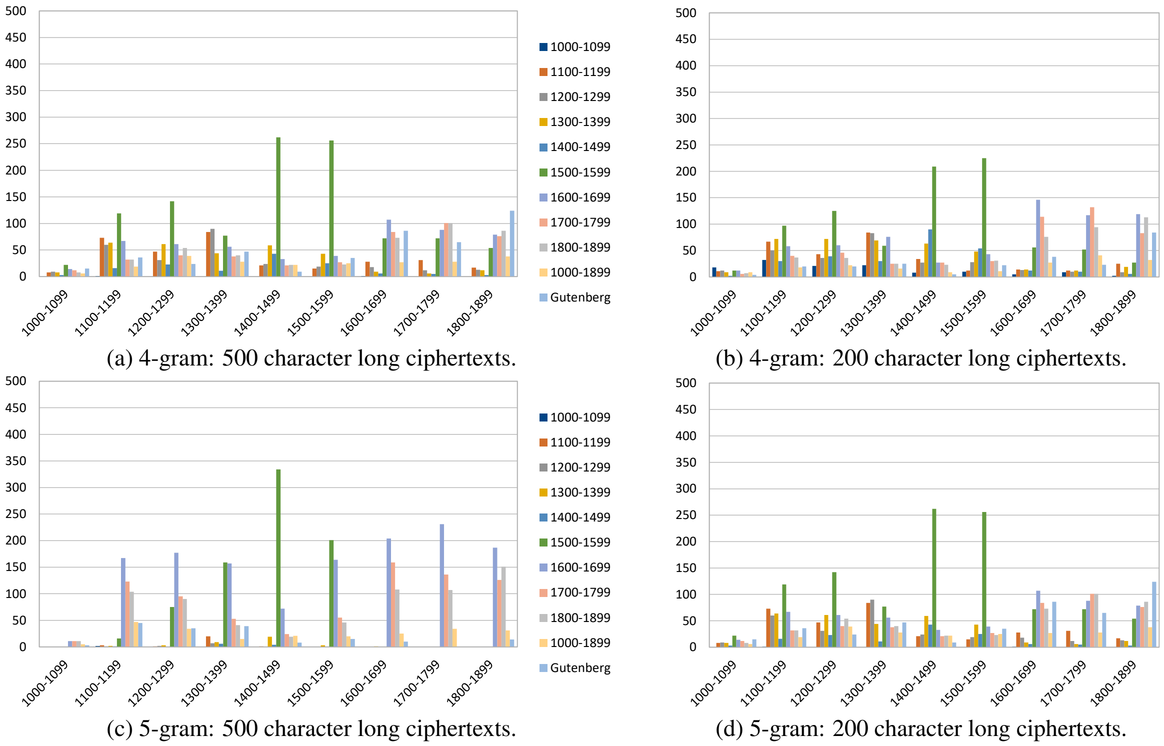
Interestingly, the Gutenberg model outperformed a merged model generated from all historical texts from the HistCorp collection (1000-1899) with the exception of a few cases, see 4-gram and 5-gram models on 500 character long texts from 1200-1299 and 1400-1499, and 200 character long ciphertexts from 1700-1799.

Like English, the longer the ciphertext, the higher decipherment accuracy is. 5-gram models fit best for shorter as well as long ciphertexts if these are century-specific. However, if no century-specific data is available for LM generation, 4-gram models seem to be optimal for longer texts.

6 Discussion

The overall best results for English and German are achieved by applying 4- and 5-gram models, historical 4-grams for texts produced in the 17th century or earlier for English, and 19th century or earlier for German, while 5-gram models are preferable for longer and more modern texts.

The results are not surprising. Shorter cipher-



(a) 4-gram: 500 character long ciphertexts.

(b) 4-gram: 200 character long ciphertexts.

(c) 5-gram: 500 character long ciphertexts.

(d) 5-gram: 200 character long ciphertexts.

Figure 6: Best result for German with 4- and 5-gram models.

texts require more reliable LMs for successful decipherment, which in turn require more specific and larger amounts of input data to achieve cryptanalysis with higher performance. Thus, the results are highly dependent on the amount of available training data, since the larger models tend to perform better in these cases. The conclusion should therefore be considered carefully, and be regarded as potential trends, and further validation with the use of different models, source data, and languages is needed.

Noteworthy is also that accuracies reported in this study might have become higher by applying more restarts and other parameters in the decipherment process. The goal of this study, however, is not to reach the highest decipherment accuracy, but to evaluate the general performance of the models given language data from various time periods.

Considering previous studies discussed in Section 2, such as Nuhn and Knight (2014) and Bean (2020) in which better results are reported on using higher order n-gram models, the data sizes required are significantly larger than for most of the models used in our study. The ciphers themselves are, furthermore, different from the ones analysed here, why a full comparison of the results is not possible.

Our next step is to carry out evaluation of the language models on various cipher types with underlying plaintext of many European languages. While we only investigated two languages, both belonging to the Germanic language family with rather similar structure, we would like to include other, more dissimilar languages of different types, such as the Indo-European Romance and Slavic, as well as Finno-Ugric languages such as Finnish and Hungarian.

Another plan is to investigate the impact of the size and coverage of LMs and their impact on decipherment. The texts which the models were generated from cannot be said to be balanced, and not of the same size with respect to centuries. In addition, for comparison between historical LMs and those generated from other sources, several language models could be used: from unigram up to 6-gram character- as well as word-based models. Apart from the Gutenberg collection, we aim to use the recently released google n-gram models for a wide range of languages from 1 to 5 character-based n-grams generated from printed books of different genres and time periods (Google, 2022)[5].

[5] https://storage.googleapis.com/books/ngrams/books/datasetsv2.html

7 Conclusion

In this paper, we investigated the influence and impact of language models on decipherment of historical ciphertexts. We conducted experiments on English and German to find out if language models generated from historical texts or modern text fit best for decipherment. We ran experiments on texts from the 11th to the 19th centuries. We investigated character-based n-gram models of size 3-, 4-, and 5-grams. We focused on homophonic substitution ciphers of various lengths. For ciphertext length, we experimented with 200 and 500 characters long ciphertext messages. We conducted experiments on homophonic substitution ciphers with a mixture of 1, 2, 3, 4, and 5 code elements to imitate the nature of original ciphers from early modern times. For comparison we generated LMs from contemporary texts derived from the Gutenberg project, and a merged model of all historical century-specific texts.

The experiments clearly indicate that the age and the length of the ciphertext have great influence on the results, and that ciphertext characteristics shall be taken into account when choosing suitable language models for cryptanalysis. Likewise, the amount of available plaintext data serving for the generation of the language models should also be considered when choosing a suitable n-gram order.

The results show that decipherment by hill-climbing and simulated annealing using historical n-gram models perform better on ciphertext produced in the 17th century or earlier for English and in the 19th century or earlier for German, and century-specific language models perform better on longer, older, and less complex ciphertexts. The larger LMs generated from the Gutenberg collection are preferable on ciphers from 18th-19th centuries. Further, experiments on n-gram size show that LMs based on 4-grams and 5-grams achieve highest performance on both English and German; 5-gram models for longer text and more modern texts and 4-gram models for historical ones.

Acknowledgments

This work has been supported by the Swedish Research Council, grant 2018-06074, DECRYPT – Decryption of Historical Manuscripts.

References

Richard Bean. The Use of Project Gutenberg and Hexagram Statistics to Help Solve Famous Unsolved Ciphers. In *Proceedings of the 3rd International Conference on Historical Cryptology HistoCrypt 2020*, number 171, pages 31–35. Linköping University Electronic Press, 2020.

Gerlof Bouma, Evie Coussé, Trude Dijkstra, and Nicoline van der Sijs. The EDGeS Diachronic Bible Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France, May 2020. European Language Resources Association.

Hendrik DeSmet. The Corpus of Late Modern English Texts (extended version), 2006.

John F Dooley. *History of Cryptography and Cryptanalysis*. Springer, 2018.

Martin Durrell, Paul Bennett, Silke Scheible, and Richard J. Whitt. GerManC, 2012. URL `http://hdl.handle.net/20.500.12024/2544`. Oxford Text Archive.

Filip Fornmark. Models, Keys and Cryptanalysis - Evaluating Historical Statistical Language Models in Cryptanalysis of Homophonic Substitution Ciphers, 2022. Bachelor thesis in Linguistics, Gothenburg University, Sweden.

Maria-Elena Gambardella, Beáta Megyesi, and Eva Pettersson. Identifying Cleartext in Historical Ciphers. In *Proceedings of the Workshop on Language Technologies for Historical and Ancient Languages. LT4HALA 2022*, 2022.

Google. The Google Books Ngram Viewer Dataset, link: https://pypi.org/project/google-ngram-downloader/, 2022.

Bradley Hauer, Ryan Hayward, and Grzegorz Kondrak. Solving Substitution Ciphers with Combined Language Models. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, Dublin, Ireland, August 2014.

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

Thomas Klein, Klaus-Peter Wegera, Stefanie Dipper, and Claudia Wich-Reif. Reference Corpus Middle High German (1050–1350) Referenzkorpus Mittelhochdeutsch (1050–1350), version

1.0. https://www.linguistics.rub.de/rem/, 2016. Accessed: 2022-07-30.

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

Nils Kopal. 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. Linköping University Electronic Press, 2019.

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

George Lasry, Beáta Megyesi, and Nils Kopal. Deciphering Papal Ciphers from the 16th to the 18th Century. *Cryptologia*, pages 479–540, 2020. URL https://www.tandfonline.com/doi/full/10.1080/01611194.2020.1755915.

Anke Lüdeling, Carolin Odebrecht, Thomas Krause, Gohar Schnelle, and Catharina Fischer. Ridges-herbology (version 9.0). https://www.deutschestextarchiv.de/, 2016. Accessed: 2022-07-30.

Beáta Megyesi, Nils Blomqvist, and Eva Pettersson. The DECODE Database: Collection of Ciphers and Keys. In *Proceedings of the 2nd International Conference on Historical Cryptology, HistoCrypt19*, Mons, Belgium, June 2019.

Beáta Megyesi, Crina Tudor, Benedek Láng, Anna Lehofer, Nils Kopal, Karl de Leeuw, and Michelle Waldispühl. Keys with Nomenclatures in the Early Modern Europe. *Cryptologia*, 0 (0):1–43, 2022. doi: 10.1080/01611194.2022.2113185.

Malte Nuhn and Kevin Knight. Cipher Type Detection. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1769–1773. Association for Computational Linguistics, 01 2014.

Malte Nuhn, Julian Schamper, and Hermann Ney. Improved Decipherment of Homophonic Ciphers. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, October 2014. Association for Computational Linguistics.

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

Eva Pettersson and Beata Megyesi. Matching Keys and Encrypted Manuscript. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 253–261, Turku, Finland, 30 September – 2 October 2019. Linköping University Electronic Press.

Sujith Ravi and Kevin Knight. Attacking Decipherment Problems Optimally with Low-Order N-gram Models. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 812–819. Association for Computational Linguistics, 01 2008.

Sujith Ravi and Kevin Knight. Bayesian Inference for Zodiac and Other Homophonic Ciphers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, page 239–247. Association for Computational Linguistics, 06 2011.

Josef Schmied, Claudia Claridge, and Rainer Siemund. The Lampeter Corpus of Early Modern English Tracts, 1999. URL http://hdl.handle.net/20.500.12024/2400. ICAME, Oxford Text Archive.

Ingrid Schröder. Reference Corpus of Middle Low German/Low Rhenish (1200–1650). https://corpora.uni-hamburg.de/hzsk/de/islandora/object/text-corpus:ren-1.0, 2018. Accessed: 2022-07-30.

Justyna Sikora. The Influence of Language Models on Decryption of German Historical Ciphers. Master's thesis, Uppsala University, 2022. Master thesis in Language Technology.

Deutsches Textarchiv. Grundlage für Ein Referenzkorpus der Neuhochdeutschen Sprache. Herausgegeben von der Berlin-Brandenburgischen Akademie der Wissenschaften. https://www.deutschestextarchiv.de/, 2010. Accessed: 2022-07-30.

Richard J Whitt. The Nottingham Corpus of Early Modern German Midwifery and Women's Medicine (ca. 1500-1700). 2016.