Experimental Analysis of the Dorabella Cipher with Statistical Language Models

Bradley Hauer†, Colin Choi†, Anirudh S Sundar†† [*]
Abram Hindle†, Scott Smallwood‡, Grzegorz Kondrak†
† AMII, Department of Computing Science, University of Alberta, Edmonton, Canada
‡ Department of Music University of Alberta, Edmonton, Canada
†† Dept. of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, USA
†, ‡ {bmhauer,cechoi,hindle1,scott.smallwood,gkondrak}@ualberta.ca
†† asundar34@gatech.edu

Abstract

The Dorabella cipher is a symbolic message written in 1897 by English composer Edward Elgar. We analyze the cipher using modern computational and statistical techniques. We consider several open questions: Is the underlying message natural language text or music? If it is language, what is the most likely language? Is Dorabella a simple substitution cipher? If so, why has nobody managed to produce a plausible decipherment? Are some unusual-looking patterns in the cipher likely to occur by chance? Can state-of-the-art algorithmic solvers decipher at least some words of the message? This work is intended as a contribution towards finding answers to these questions.

1 Introduction

The Dorabella cipher (henceforth, *Dorabella*) is a cipher sent by Edward Elgar to his acquaintance Dora Penny (Figure 1). Elgar was an English composer, best known for works such as *Pomp and Circumstance*, and the *Enigma Variations*. He also had an interest in cryptography, which was an inspiration for some of his compositions.

Prior decipherment attempts have adopted various assumptions. Arguably the most popular assumption is that it is a monoalphabetic substitution cipher (MASC) encoding an English text (Sams, 1970). Given that there was no known key exchange between Elgar and Penny, it is reasonable to assume that the cipher was not intended to be complicated; likewise, given that Elgar was an English composer, it is reasonable to assume that English is the language of the cipher. Another hypothesis is that it is enciphered music (Santa and Santa, 2010). However, no plausible systematic decipherment has been proposed to date, nor a convincing demonstration that it is a hoax.

In this paper, we investigate several hypotheses using modern computational techniques. Our methods are based on *statistical n-gram language models*, which are induced over characters or words from large collections of texts (*corpora*). We apply a state-of-the-art ciphertext language identification algorithm to identify the underlying language of the cipher. We also apply automated decipherment algorithms developed for monoalphabetic substitution ciphers in an attempt to obtain at least partial decipherment. To test the music hypothesis, we develop a transcription encoding scheme that is restricted to 24 distinct musical notes. Finally, we consider whether some statistical properties of the ciphertext support the hoax hypothesis.

Our experiments demonstrate that highly-accurate algorithmic solvers fail to produce any readable decipherment, providing evidence against the hypothesis that Dorabella is a simple MASC encrypting an English text. We do, however, find new evidence that English is one of the most likely languages behind Dorabella. Furthermore, experiments with musical transcriptions suggest that the cipher is unlikely to encode music. Finally, we find evidence of non-random patterns in the ciphertext, which we interpret as evidence against the hoax hypothesis.

This paper is structured as follows: We describe the properties of the Dorabella cipher in Section 2. In Section 3, we summarize prior publications on the topic. The methods and data are described in Sections 4 and 5, respectively. The experimental results are discussed in Section 6.

[*] Sundar's work while at the University of Alberta.

2 Dorabella Symbols

Figure 1 shows the cipher in its entirety. It contains 87 characters, each consisting of one, two, or three semicircles, in one of eight distinct orientations (in increments of 45 degrees), yielding a total of 24 possible symbol types. The orientation of some of the symbols is ambiguous. In our transcription, 20 distinct symbols appear in the cipher, with four hypothetical symbols being unused. The symbols follow a highly non-uniform distribution, with one symbol appearing 11 times, while some appear only once.

This distribution of the number of semicircles is relatively uniform, while the distribution of orientations is not. In our transcription, 29 tokens have one semicircle, 33 have two semicircles, and 25 have three semicircles. On the other hand, one orientation (with the bottom of each semicircle directed at 315 degrees) occurs 23 times in our transcription, while another (with the bottom of each semicircle directed at 0 degrees, e.g. right) occurs only 4 times. In this paper, we make no assumptions or deductions about the meaning of the number and orientation of semicircles; rather, we arbitrarily map each symbol type to an arbitrary lowercase English letter, and treat the resulting transcription as a straightforward substitution cipher.

3 Related Work

In an early decipherment attempt, Sams (1970) analyzes Dorabella using frequency analysis, contact charts, and brute force methods. This work assumes that that the message is partly phonemicized, but not strictly monoalphabetic. The result of this analysis is a decipherment which is not systematic, verifiable, or falsifiable.

Santa and Santa (2010) analyze Dorabella in the broader context of Elgar's work, particularly his *Enigma Variations*. They speculate that Elgar may have used the mathematical constant π, approximated as 3.142, to encipher scale degrees. However, they note that a plausible solution to Dorabella, whether in the form of natural language text or musical notation, is yet to be found.

Schmeh (2018) explores several established techniques for identifying vowels and consonants in monoalphabetic substitution ciphers. The result is a transcription of Dorabella, with some symbols identified as vowels or as consonants. These results are supported by independent analysis on a sample cipher of the same length.

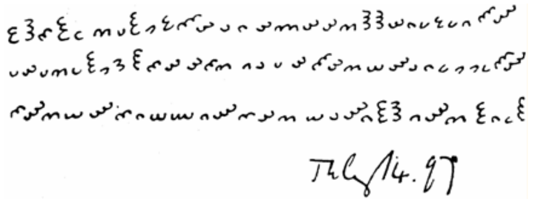

Figure 1: The Dorabella cipher.

The task of computational decipherment of monoalphabetic substitution ciphers is well-studied. Most recent work involves character and/or word language models (Norvig, 2009; Nuhn et al., 2013; Hauer et al., 2014) as well as other techniques, such as electronic dictionaries (Olson, 2007), integer programming (Ravi and Knight, 2008), and Bayesian inference (Ravi and Knight, 2011).

4 Methods

In this section, we describe the cryptographic tools, both previously published and original to this work, which we employ in our analysis of Dorabella.

4.1 Language Models

Our methods are based on statistical n-gram language models, which are induced over characters or words. Language models guide decipherment algorithms by computing the probabilities of various possible decipherments, allowing algorithms to favour decisions which result in more probable solutions. An n-gram language model can be used to compute the probability of a token given the $n-1$ previous tokens. A 3-gram, or *trigram*, character language model, for example, is able to predict that, given the previous characters 'aq', the letter 'u' is more likely to follow than 'e', despite 'e' generally being more common than 'u'. To create language model for our experiments, we use KenLM.[1]

Language models can be applied over a sequence of characters to measure their *perplexity*, which quantifies the extent to which a language model is "surprised" by the text in question. A high perplexity indicates that the sequence of tokens has a correspondingly low probability under the model.

[1] https://github.com/kpu/kenlm

4.2 Computational Decipherment

We experiment with three previously published methods for deciphering monoalphabetic substitution ciphers, which are based on statistical *n*-gram language models.

HILLCLIMB (Norvig, 2009), is a solver that performs a hill-climbing search with multiple random restarts to maximize the probability of the decipherment under a character language model. The best decipherment is selected according to a word language model. We use the implementation provided by the author.[2] Since word language models require word boundaries, we also experiment with HILLCLIMBC, a variant that instead uses a character language model to identify the best decipherment.

TREESEARCH (Hauer et al., 2014) uses a tree search algorithm to find the highest-scoring key. A key scoring function combines word and character n-gram language models of various orders. An initial decipherment based on unigram character frequencies serves as the root of the tree. A key mutation function leverages character repetition patterns to generate a set of children for each key. The solver is reported to have decipherment accuracy on ciphers without spaces (i.e., without word boundaries) of over 92% for length 64, and over 99% for length 128, which represents the state-of-the-art for monoalphabetic substitution decipherment.

UNRAVEL (Nuhn et al., 2015) searches for a mapping of letters that maximize the probability of the decipherment under an *n*-gram character language model. Partial key mappings are structured into a search tree, and a beam search is used to traverse the tree and find the most promising candidates. Unlike TREESEARCH, UNRAVEL does not constrain every node of the search tree to contain a complete decipherment; not all nodes decipher all symbol types. Rather, initially incomplete keys are iteratively expanded, with heuristic search used to guide the expansion until a complete solution is found. We use the version of UNRAVEL that is applicable to deterministic rather than probabilistic ciphers. The experiments presented by the authors focus on word-level decipherment (e.g. identification of lexical translations), without any claims regarding the efficacy of the solver on character-level monoalphabetic substitution ciphers.

We also developed a novel greedy search algorithm with random restarts, which we refer to as GREEDY. Starting with a random key, possible successors are generated by sequentially swapping letters in the current key. Each successor key is assigned a probability using a character trigram language model. The successor which produces the most probable decipherment becomes the new key, provided that its decipherment is more probable than the current key. The key that produces the most probable decipherment over multiple random restarts is returned as the solution.

4.3 Ciphertext Language Identification

Identification of the underlying language of a cipher is crucial for a successful decipherment. For this task, we apply two methods presented by Hauer and Kondrak (2016): UNIGRAM and TRIAL. Both methods are applicable to monoalphabetic substitution ciphers without word boundaries, and require a set of sample texts, each representing one of the candidate languages. Each method iterates over the set of sample texts, computing a score function on each sample. The language of the sample text which maximizes this score function is returned as the identified language of the ciphertext.

The first method, UNIGRAM, leverages the observation that a monoalphabetic substitution does not alter the relative frequencies of characters: the frequency of the *i*-th most frequent character before encipherment is equal to the frequency of the *i*-th most frequent character after encipherment. Given the ciphertext and a sample text, UNIGRAM computes the *sorted symbol distribution* of each. This is a probability distribution over characters $1,\ldots,k$ where k is the length of the longer of the two symbol alphabets, and $P(i)$ is the probability of a randomly selected character being the *i*-th most frequent character in the text. For each language, we compute its score as the distance metric of Bhattacharyya (1943) between the unigram probability distributions of the sample text and the ciphertext.

The second method, TRIAL, is based on the intuition that attempting to decipher a ciphertext into the incorrect language (e.g., deciphering enciphered English into French) will almost certainly not yield a probable text in that language. The method learns a bigram character language model for each language using the corresponding sample

[2] http://norvig.com/ngrams

text. It then applies a hill-climbing decipherment algorithm which seeks to maximize the probability of the decipherment. This algorithm terminates quickly in practice, allowing hundreds of candidate decipherments to be tried. The probability of the best decipherment is returned as the score. It is important to note that Hauer and Kondrak (2016) developed and tested the TRIAL method on ciphers with spaces included, as it was originally designed for the Voynich manuscript.

4.4 Music Decipherment

The algorithms described in the previous section were designed to be applied to natural language texts. Since we wish to test the hypotheses that Dorabella is enciphered music, we seek to apply these algorithms to music as well. This presents multiple challenges, which we discuss here.

Most western music is presented as pitches with duration over time with dynamics, phrasing, and articulations. In terms of pitch, multiple pitches can sound at the same time, resulting in chords, homophony (a primary melody with accompanying chordal notes), or polyphony (simultaneous melodic lines that have independent characteristics, but also outline harmonic motion). A piano is an example of a polyphonic instrument, as with multiple fingers one can play many piano keys at the same time, and each piano string will sound a distinct separate pitch. Thus much music is written and composed in a polyphonic manner. There is no analogue to this phenomenon in natural language text. We therefore need to first serialize the notes and choose an order. To this end, we work with single lines of music rather than polyphonic passages. For example, we would consider only the melodic line of a four-part piece.

Further, music differs from written language in several key ways. Notes do not refer to specific real-world concepts, as words do, and have different intents or meaning. Furthermore, music can be transformed (such as by changing octave or transposing the key of the music) in ways whereby musicians will still understand the music or its origin. Finally, there is no clear equivalent of a sentence or punctuation in music; if such equivalents exist, it is not clear if they can be ignored for the purposes of encipherment and decipherment, as is the case with natural language. [3]

For music to be enciphered it must be first represented as symbols, such as western music notation. Then, we must serialize them, such that one note comes after another, as described above. An example encoding could be the note name, which ignores octave and duration, expressed as space separated notes: `E D C D E E E` (*Mary had a little lamb*). Alternatively we could add duration: `Eq Dq Cq Dq Eq Eq Eq Eq` where `q` would indicate a quarter note. We might also include octave: `E4q D4q C4q D4q E4q E4q E4q E4q` where `C4` is middle C, and `C5` is an octave above that, and so on. Such an encoding would allow us to treat each note (a tuple of pitch and duration) as a symbol. These symbols could then be enciphered or deciphered, just as can be done with the sequence of symbols in a natural language text.

For our experiments we start with music encoded as MIDI files (a digital music communication protocol), which we then pre-process into simpler serial formats. MIDI for our purposes presents notes as pitches that are turned on and off at specified times. In terms of duration, notes are normalized into sixteenth notes, eighth notes, quarter notes, half notes, and whole notes. In terms of pitch, we can decide to look for any 12 notes of the octave, or confine ourselves to a diatonic scale (7 of the 12 notes).

In order to convert these MIDI files into a sequence of symbols, as described above, the files are transposed to the key of C major, and only a single octave of notes is used. Rests, accents, and other symbols that do not signify notes are removed from the sequences; chords are decomposed to their roots. The representation is composed of notes with their respective duration. Three different durations are used for the notes. A duration of 0.5 represents anything shorter than a quarter note, a duration of 1 represents a quarter note, and a duration of 2 represents anything longer than a quarter note.

Our encoding uses 24 unique symbols, the same number of unique symbols that can be made using the Dorabella cipher system. This encoding only uses the eight most frequent notes A, B, C, D, E, F, F♯, and G along with the three durations described above. Notes not among the 8 most frequent notes are moved half a step up or down. For example, a D♯ would be changed to a D and A♭ would be changed to A.

[3] There does exist a musical term of "sentence," which refers to a complete statement that is bigger than a motive or phrase, but shorter than a theme.

5 Data

This section is devoted to the language and music corpora used in our experiments (Hauer et al., 2021). Our natural language corpora include literary prose, newspaper texts, movie subtitles, and multilingual documents. To generate ciphertexts with known solutions for testing purposes, we extract samples from the 19th century fiction works in Project Gutenberg[4], including *The Adventures of Sherlock Holmes*, and *The Letters of Jane Austen*. We chose *Dangerous Connections*, an English translation of an epistolary novel, for deriving character-level language models; and a much larger *New York Times Corpus*[5] for deriving word-level language models. For our language identification experiments, we use a dataset constructed from 380 translations of the *Universal Declaration of Human Rights* (UDHR) (Emerson et al., 2014), and the multilingual *OpenSubtitles* corpus of movie subtitles (Lison and Tiedemann, 2016).

To create test samples used in our experiments, we first normalize the natural language corpora, by removing punctuation, digits, and other non-alphabetic characters, and lower-casing all letters. The test samples we use are 87 letters long, the same length as Dorabella. They are created by first randomly selecting a word in the corpus, and then appending subsequent words until the length of exactly 87 letters is reached. Samples that end with partial words are discarded, and no duplicate samples are admitted. This process ensures that each generated test cipher begins and ends at a word boundary, and contains exactly 87 characters, with no spaces.

Our music corpora consist of monophonic tracts extracted from collections of Elgar and Bach MIDI files.[6] For each composer, we split the collection of MIDI files into testing and training sets.[7] For Bach, the training corpus is composed of 295 MIDI files concatenated together (3.7M notes) with 3 MIDI files (174K notes) held out for testing. For Elgar, the training corpus is composed of 29 concatenated MIDI files (1.2M notes), with 3 MIDI files (24K notes) held out for testing. Samples 87 notes in length are extracted from the test data and enciphered to create sets of music ciphers with known solutions for our experiments.

6 Experiments

In this section we present our applications of the methods described in Section 4, using the data described in Section 5, with the goal of testing several hypothesis regarding the Dorabella cipher. Throughout our experiments, we make the assumption that Dorabella is a monoalphabetic substitution cipher (MASC), which is based on the number and relative frequencies of the characters. For the evaluation of MASC solvers, we compute both *key accuracy*, the proportion of cipher character types which are correctly mapped to the corresponding plain-text character type, and *decipherment accuracy*, the proportion of cipher character tokens which are correctly deciphered.

As a precursor to these experiments, we applied the BION classical cipher type classification programs[8] as used by Nuhn and Knight (2014), to our transcription of Dorabella. Both programs classify the text as a "patristocrat" cipher, which is equivalent to our definition of a monoalphabetic substitution cipher without word divisions. This supports our assumption that Dorabella is a MASC.

6.1 Ciphertext Language Identification

In this section, we apply the ciphertext language identification methods described in Section 4.3 to analyze Dorabella. This includes empirically assessing the reliability of these methods on short ciphers without spaces, as well as examining the output of the state-of-the-art method when applied to Dorabella.

Given an output which induces a ranking of possible classes, the *reciprocal rank* for a given instance is the multiplicative inverse of the position of the correct class, with the highest-ranked class being rank 1. For example, if the correct class is assigned rank 4, the reciprocal rank for that instance is $1/4 = 0.25$. The *mean reciprocal rank* (MRR) is the average of the reciprocal ranks over all instances. A high MRR indicates that the correct class is consistently placed near the top. Closely related to MRR is *average rank* (AvgR), which is simply the mean position of the correct class over all instances (i.e. MRR, without the reciprocal operation). Top-1 accuracy, or simply accuracy (Acc), is the proportion of instances

[4] http://www.gutenberg.org/ebooks/

[5] https://catalog.ldc.upenn.edu/LDC2003T05

[6] https://www.classicalmidi.co.uk/elgar.htm,
http://bestclassicaltunes.com,
http://dardel.info/musique/Bach.html

[7] https://archive.org/download/midi-sources

[8] http://bionsgadgets.appspot.com

Method	Length	Spaces	MRR	AvgR
UNIGRAM	2000	No	0.18	15.7
TRIAL	2000	No	0.94	1.2
TRIAL	2000	Yes	0.96	1.1
UNIGRAM	87	No	0.02	120.0
TRIAL	87	No	0.13	52.6
TRIAL	87	Yes	0.25	33.4

Table 1: Results of the ciphertext language identification methods.

for which the correct class is placed in the first position. For MRR and Acc, higher is better; for AvgR, lower is better. If and only if a method always places the correct class in the first position, its MRR, Acc, and AvgR will all be 1, the maximum/minimum values.

6.1.1 Validation on Synthetic Ciphers

In this experiment, we aim to establish the effectiveness of the current state-of-the-art method of Hauer and Kondrak (2016) on synthetic cipher samples from multiple languages. They report that the TRIAL method achieves over 97% top-1 accuracy; however, their results are on ciphers longer than a thousand characters, which include word boundaries. In contrast, Dorabella is only 87 characters long, and contains no spaces.

We begin by assessing the impact of the cipher length and the presence of word boundaries on ciphertext language identification accuracy. We test 4 cipher variants: long (2000 characters) vs. short (87 characters), with and without spaces. As our data, we use the UDHR dataset (Section 5) for training language models, and the OpenSubtitles corpus for generating test ciphers.

Table 1 shows the results of the experiment. We report the mean reciprocal rank (MRR) and average rank (AvgR) of the correct ciphertext language evaluated over a set of 500 ciphers in 5 distinct languages: English, French, Polish, German, and Italian. The results indicate that even for short ciphers without spaces, the TRIAL method is able to rank the language of the ciphertext much more highly than the UNIGRAM method. Even for Dorabella-like 87-character ciphers without spaces, the TRIAL method consistently assigns a relatively high rank to the correct language. For comparison, a random baseline yields MRR of 0.017, and an average rank of 190.5. From this we conclude that the TRIAL method provides useful information about the language of short ciphers without spaces.

	En	Fr	Pl	De	It	**Avg**
MRR	.68	.68	.72	.69	.87	**.73**
Acc	.49	.48	.55	.50	.78	**.56**

Table 2: MRR and top-1 language identification accuracy on 87-character ciphers The MRR and Acc for each language are the averages over all ciphers for that language.

6.1.2 Impact of Language Sample Size

In our second set of language identification experiments, we investigate whether there is a substantial benefit to increasing the size of the texts used by TRIAL to create language models. Due to the greater difficulty in acquiring larger texts for training language models, we only test on English, French, Polish, German, and Italian, so there are only five possible classifications, rather than 380. For each language we obtain 100M characters from the OpenSubtitles corpus for inducing the language model, and another 20M to create test ciphers. We create 1000 ciphers without spaces for each of the five languages.

The results in Table 2 indicate that TRIAL is able to correctly select, from English, French, Polish, German, and Italian, the language of a ciphertext from one of those languages more than half the time. The MRR values for each language are all well above 0.5, which indicates that the correct language is usually among the top two candidates. We conclude that, given sufficient training data for inducing language models, the TRIAL method can be used to analyze short ciphers without spaces.

6.1.3 Is Dorabella English?

We now explore the hypothesis that the Dorabella cipher represents enciphered English. This hypothesis is based on the observation that the remainder of Elgar's letter, in which the Dorabella cipher is embedded, is written in English. To maximize the number of candidate languages we consider, we again use the UDHR data as a source of language samples. We then apply TRIAL, the more accurate of the two language identification methods, to Dorabella, inducing a ranking of the 380 samples.

Table 3 shows the five highest-scoring languages. The numerical values are the log-probabilities of the best decipherment for each

Rank	Language	LM Score
1	Latin	-217.34
2	Aceh	-221.05
3	English	-221.24
4	Toksave	-222.19
5	Scots	-222.62

Table 3: The highest-scoring Dorabella candidate languages with the TRIAL method.

language, estimated using the corresponding language model. It is notable that this method places English as the third-best choice for the language of the cipher, and the closely related Scots language as the fifth choice. Latin, which is a major source of the English lexicon and its orthography, is ranked first. Given the accuracy of the TRIAL decipherment method, and given the context in which the Dorabella cipher was produced, we conclude that English is the most likely natural language candidate for Dorabella.

6.2 Is Dorabella Music?

Using the music representation described in Section 4.4, and inducing character "language" models over the music corpora described in Section 5, we investigate the hypothesis that Dorabella enciphers music, rather than natural language. To determine the accuracy of our solvers on music, we test two different decipherment programs. The HILLCLIMBC solver and GREEDY solver are chosen for this test because our text decipherment experiments show that these two solvers perform well on short ciphers without spaces, and without a large training corpus.

We created Elgar and Bach language models from the corpora of their music, described in Section 5. The test samples were randomly enciphered with a substitution cipher. Since the accuracy of both solvers on short samples was very low, we instead used very long samples of around 20,000 notes each.

Table 4 shows the results on long ciphers. The best key and decipherment accuracies are only 26.6% and 32.7% respectively, both obtained using our GREEDY method. This indicates that approximately one-third of the notes in each cipher are deciphered correctly, on average. We conclude that deciphering music, in our minimalist representation, is much more difficult than deciphering natural language.

Music	**Solver**	**Key Acc**	**Dec Acc**
Elgar	GREEDY	4.8%	6.4%
Elgar	HILLCLIMBC	7.0%	12.0%
Bach	GREEDY	26.6%	32.7%
Bach	HILLCLIMBC	26.5%	32.0%

Table 4: Key and decipherment accuracy on long music ciphers.

One of the authors of this paper analyzed the notes in the highest-scoring decipherment obtained with the Elgar language model. The notes appear and sound random, containing no clues that would point to an expected tonal center and harmonic progression. Further, no recognizable motives, phrasing, or repetition can be identified. It has nothing to do with Elgar's music, which was more complex and chromatic. We hypothesize that Elgar may have instead enciphered a simple folk-like melody, rather than something comparable to his more mature work. We intend to investigate this direction in future work.

6.2.1 Impact of Perplexity

In this section, we investigate a hypothesis that music has a less predictable structure than natural language, which would make it more difficult to decipher, explaining the results in the previous section. We calculate the relative perplexity of samples of texts vs. samples of music notation encoded using a simple scheme. Both types of encodings have a similar number of distinct symbols: 26 letters vs. the 24 symbols in our encoding of musical notes.

We create language models for Bach music, Elgar music, and English text as in the preceding sections. We then create 100 samples of 87 characters for each of Bach, Elgar, and English. The samples are not included in the training corpora. The English LM is derived from *Dangerous Connections*, while the samples are from *Letters of Jane Austin*. The average perplexity is then calculated for all three sets of samples against the three language models.

We find that English is much more predictable than music, even under our highly simplified encoding scheme. Averaged across the 100 samples, the music of Elgar and Bach have perplexities of 24.40 and 24.52 respectively, while English has a perplexity of only 16.18. We propose this as an explanation of our finding that decipherment algorithms are much less effective on enciphered music compared to enciphered English.

6.2.2 Classifying Text vs. Music

Since rank-based attempts showed some promise in determining the ciphertext language (Sections 6.1.1 and 6.1.3), we decided to create a classifier to determine whether a cipher encodes English text or music. In this section, we describe the classifier, test it on synthetic ciphers, and finally apply it to Dorabella.

We use TRIAL as our classifier, with English and music as candidate languages. We trained the necessary bigram language models on 1M characters of *Dangerous Connections* for English, and *either* 1M symbols of Bach, or 1M symbols of Elgar. This yields two distinct experiments: (1) distinguishing English and Bach music, and (2) distinguishing English and Elgar music.

When tested on the 300 test ciphers from *Letters of Jane Austen*, and 300 samples each of Bach music and Elgar music, we found that TRIAL was able to distinguish between English and Elgar ciphers with 82% accuracy, and between English and Bach with 88% accuracy. These results demonstrate that TRIAL can reliably distinguish between enciphered English and enciphered music.

That established, we applied our classifier to our transcription of Dorabella. We found that TRIAL classifies the cipher as English, compared to both Bach and Elgar music. In the first case, language model log-probabilities of -228.8 and -244.5 are assigned to English and Bach, respectively. The variances on these mean log-probabilities (averaged over ten independently randomized runs) are 11.1 and 13.5, respectively. In the second case, the corresponding average log-probabilities are -226.8 and -248.6, with the variances of 5.0 and 2.0, respectively. We interpret these results as evidence that Dorabella is much more likely to represent English than music.

6.3 Decipherment of English Texts

In this section, we perform validation experiments on several substitution cipher solvers. We compare their accuracy on English MASCs, and attempt to decipher Dorabella with the best-performing solver. Note that *we do not claim to have produced a correct decipherment of or solution to the Dorabella cipher.*

We test five decipherment methods which are described in Section 4.2: TREESEARCH, HILLCLIMB, HILLCLIMBC, GREEDY, and UNRAVEL. To establish the reliability of each of these methods, we measure their accuracy on 87-character ciphertexts without spaces. We use the same set of 300 English ciphers and English training corpus as in Section 6.2.2.

Solver	Key Acc	Dec Acc
TREESEARCH	43.1%	44.9%
UNRAVEL	42.8%	47.8%
GREEDY	69.0%	79.1%
HILLCLIMB	75.8%	84.5%
HILLCLIMBC	78.3%	88.1%

Table 5: Accuracy of substitution cipher solvers on short English ciphers without spaces.

Table 5 shows the average key and decipherment accuracy of the five solvers on the set of 300 test samples. The relatively low accuracy of TREESEARCH is likely due to the small size of the training corpus.[9] Similarly, UNRAVEL did not perform very well on short ciphers without spaces, regardless of the size of the corpus. The remaining three solvers were much more effective. HILLCLIMBC, the variant of HILLCLIMB which is based entirely on a character language model, performed best, reaching nearly 90% average decipherment accuracy.

However, applying HILLCLIMBC to Dorabella does not produce a readable decipherment. The highest-scoring decipherment is as follows:

```
ychswamsopledieveeacceirprult
memarsofsheehaudmeleantdiroorlt
htanthingutheandtuscutasirs
```

Since the other solvers likewise failed to produce any partial decipherment, we conclude that the Dorabella cipher is unlikely to represent English text enciphered with a simple MASC.

6.4 Ciphertext Characteristics

The experiments in this section are aimed at the statistical analysis of two observations made in a video by Keith Massey.[10] The first is that the number of two-symbol sequences in Dorabella which are reflections of one another is greater than chance would allow. The second is that there are

[9] In a separate experiment, we were able to replicate the high decipherment accuracy reported by Hauer et al. (2014), given a larger (but out-of-domain) text corpus.

[10] Keith Massey, *The Dorabella Cipher: Proven to be a Friendly Joke*, 2017-05-29

long series of symbols in Dorabella with no repeat of a symbol with the same number of semicircles. Based on those two observations, it is claimed that the Dorabella cipher is a nonsensical message constructed as a playful joke.

6.4.1 Mirrored Symbols

The Dorabella cipher contains 13 pairs of mirrored symbols. A mirrored pair consists of 2 consecutive symbols that have the same number of semicircles but are facing in opposite directions. (For example, the final pairs of symbols in line 1 and 2 in Figure 1.) How likely is it for a ciphertext of 87 symbols to contain 13 mirrored pairs?

Our procedure is as follows. We randomly extract 100,000 samples of length 87 from *The Adventures of Sherlock Holmes* using the procedure described in Section 5. Given the large number of samples relative to the length of the corpus, there is some overlap between distinct ciphers, however, each starts at a distinct character in the corpus. For each sample, we generate a random key that maps each letter in the sample to a Dorabella symbol. Since there are 26 letters in the alphabet but only 24 Dorabella symbols, up to 2 pairs of letters may share a single symbol. We encode each of the 100,000 samples with Dorabella symbols, and count the number of mirrored symbols that occur in each sample.

The results show that, an English text of length 87 encoded with the Dorabella symbols contains an average 3.64 mirrored pairs. Out of the 100,000 samples, only 123 contained 13 or more mirrored pairs, which implies that a text with 13 mirrored pairs, like Dorabella, has only about a 0.1% chance of occurring by accident.

While these results support Keith Massey's observation, we disagree with the implication that Dorabella is a hoax. Instead, we posit that the mirrored pairs in Dorabella may have some special interpretation, which would support our earlier conclusion that Dorabella is not a simple MASC. For example, the mirrored symbols could have been used by Elgar to represent double letters, such as "ee", in a less conspicuous way.

6.4.2 Longest Non-Repeating Sequence

Each symbol in Dorabella has 1, 2, or 3 semicircles. In each of the three lines of Dorabella, there are sequences of symbols without two consecutive symbols containing the same number of semicircles. The longest such sequence is of length 12. The claim made in the video is that the occurrence of such long sequences with no two adjacent symbols having the same number of semicircles is highly improbable, indicating the Dorabella is a hoax.

We test this claim by applying a similar procedure as in the previous experiment: We encipher 100,000 samples of English with Dorabella symbols using randomly generated keys and count the longest sequence of symbols without repeated semicircles in each sample.

The results show that the average length of the longest sequence of consecutive symbols with different number of semicircles is approximately 10.23. Specifically, 27,472 out of the 100,000 samples contained sequences of 12 or more symbols where there were no repeated semicircles. We conclude that the probability a single occurrence of a sequence of length 12 in Dorabella is about 27.4%. Therefore, while the sequences observed in Dorabella are surprising, they are not sufficiently improbable to dismiss the cipher as a joke. In sum, our investigation of the claim made in this video provide no evidence for the hoax hypothesis.

7 Conclusion

While the short length and lack of word boundaries in the Dorabella cipher present a formidable cryptographic challenge, we have been able to provide evidence for and against various hypotheses via experimental analysis. The failure of several substitution solvers to produce any partially readable decipherment suggests that the cipher is not a simple monoalphabetic substitution cipher that encodes an English text. Our application of a state-of-the-art method for ciphertext language identification provides new evidence for English as the language of the cipher. Furthermore, application of a classifier based on character language models suggests that the underlying message of Dorabella is more likely to be natural language than music. Finally, the occurrence of several pairs of mirrored symbols is unlikely to be due to chance, suggesting that Dorabella is not a hoax.

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Alberta Machine Intelligence Institute (Amii).

References

Anil Bhattacharyya. 1943. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.*, 35:99–109.

Guy Emerson, Liling Tan, Susanne Fertmann, Alexis Palmer, and Michaela Regneri. 2014. Seedling: Building and using a seed corpus for the human language project. In *Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 77–85.

Bradley Hauer and Grzegorz Kondrak. 2016. Decoding Anagrammed Texts Written in an Unknown Language and Script. *Transactions of the Association for Computational Linguistics*, 4:75–86.

Bradley Hauer, Ryan Hayward, and Grzegorz Kondrak. 2014. Solving substitution ciphers with combined language models. In *Proceedings of COLING, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2314–2325, Dublin, Ireland.

Bradley Hauer, Colin Choi, Anirudh S Sundar, Abram Hindle, Scott Smallwood, and Grzegorz Kondrak. 2021. Zenodo: Code and Data for "Experimental Analysis of the Dorabella Cipher with Statistical Language Models", HistoCrypt 2021, May. https://doi.org/10.5281/zenodo.4819086.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929.

Peter Norvig. 2009. Natural language corpus data. In Toby Segaran and Jeff Hammerbacher, editors, *Beautiful Data: The Stories Behind Elegant Data Solution*, pages 219–242. O'Reilly Media.

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

Malte Nuhn, Julian Schamper, and Hermann Ney. 2013. Beam search for solving substitution ciphers. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1568–1576.

Malte Nuhn, Julian Schamper, and Hermann Ney. 2015. UNRAVEL—A Decipherment Toolkit. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP)*, pages 549–553, Beijing, China.

Edwin Olson. 2007. Robust dictionary attack of short simple substitution ciphers. *Cryptologia*, 31(4):332–342.

Sujith Ravi and Kevin Knight. 2008. Attacking decipherment problems optimally with low-order n-gram models. In *Empirical Methods in Natural Language Processing*, pages 812–819. Association for Computational Linguistics.

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

Eric Sams. 1970. Elgar's Cipher Letter to Dorabella. *The Musical Times*, 111(1524):151–154.

Charles Richard Santa and Matthew Santa. 2010. Solving Elgar's Enigma. *Current Musicology*.

Klaus Schmeh. 2018. Examining the Dorabella Cipher with three lesser-known cryptanalysis methods. In *Proceedings of the 1st International Conference on Historical Cryptology (HistoCrypt 2018)*, pages 145–152.