Examining The Dorabella Cipher with Three Lesser-Known Cryptanalysis Methods

Klaus Schmeh
Freelanced Journalist
klaus@schmeh.org

Abstract

Most mono-alphabetic substitution ciphers (MASCs) can be solved with well-known techniques, like frequency analysis, or hill climbing. However, there are exceptions. It therefore makes sense to look around for additional MASC solving techniques—like the ones described in the book *Cryptanalysis* by Helen Fouché Gaines. These techniques—vowel detection, digram analysis, and consonant lining—have been almost forgotten since the advent of computer technology. In this paper, two of these methods (the third one is not suitable) will be applied on a famous unsolved cryptogram, the Dorabella Cryptogram, and on an unencrypted comparison text. Although this paper will not present a solution of the Dorabella Cryptogram, a number of interesting insights will be introduced. Especially, it will be shown that one of the methods applied works surpsingly well on the comparison text and that there are still ways to improve this technique. In addition, some interesting properties of the Dorabella Cryptogram will be presented, which might be helpful for further cryptanalysis.

1 Introduction

With the advent of computer technology and suitable software (especially, the open source tool CrypTool), breaking a mono-alphabetic substitution cipher (MASC) has become quite easy. Frequency analysis, which once was the most important way to break a MASC, is often not even necessary any more, as a word pattern search conducted by a software is usually more effective. In addition, hill climbing—another technique that requires computer support—has proven a powerful technique to break MASCs.

In spite of all this progress, breaking a MASC encryption still may be difficult, especially if some of the following pre-conditions are given:

- The ciphertext is especially short.
- The cleartext language is not known.
- The word boundaries are not indicated.
- The cleartext contains unusual expressions or abbreviations.
- The cryptogram is hard to read because of bad penmanship or bad reproduction.

I am aware of over 20 cryptograms that have the appearance of a MASC (of course, one does not know before the cipher is broken) and that are still unsolved, potentially for one or several of the named reasons. The following are the most important of these cryptograms:

- The Dorabella Cryptogram (Wikipedia, 2018)
- The MLH cryptogram (Schmeh1, 2017)
- The Voynich Manuscript (Wikipedia, 2018)
- The Rayburn cryptogram (Schneier, 2006)
- Cigaret Case Cryptogram (Schmeh2, 2017)

For these cryptograms, frequency analysis, word pattern analysis, word guessing, and hill climbing have failed so far. An interesting question is whether there are other MASC breaking methods that can be applied in such a case. In fact, there are. The book *Cryptanalysis* by Helen Fouché Gaines mentions three MASC breaking methods that are worth considering (Fouché Gaines, 1939): vowel detection, digram analysis, and consonant lining. All three methods are as good as not mentioned in the literature that has

Proceedings of the 1st Conference on Historical Cryptology, pages 145– 152,
Uppsala, Sweden, 18-20 June, 2018

been published since computer technology came up.

The goal of this paper is to apply the three methods mentioned by Fouché Gaines on the aforementioned Dorabella Cryptogram. The Dorabella Cryptogram was created by British composer Edward Elgar (1857-1934). In 1897, Elgar, who had a strong interest in cryptology, sent an encrypted message to a female friend named Dora Penny. This cryptogram is written in symbols consisting of one, two or three bows (see figure 1)—probably an alphabet of Elgar's own creation.

The Dorabella Cryptogram has never been solved, although many experts have tried. It is covered in virtually every famous unsolved cipher list, e.g., on Elonka Dunin's website (http://elonka.com/UnsolvedCodes.html), in Craig Bauer's book *Unsolved!* (Bauer, 2017), in Klaus Schmeh's *Nicht zu knacken* (Schmeh, 2012), and in Richard Belfield's *Can You Crack the Enigma Code?* (Belfield, 2006).

To check whether the three cryptanalysis methods work, I will apply them not only on the Dorabella Cryptogram but also on a non-encrypted comparison text. When looking for a suitable text, I came across the novel *The Gadfly* by Ethel Boole (Boole 1897), the wife of Wilfrid Voynich, who is known to crypto history scholars as the person the Voynich Manuscript is named for. *The Gadfly* was published in 1897, the same year as the Dorabella cryptogram was created. I chose the following 87 letter excerpt (same length as the Dorabella Cryptogram) from the *The Gadfly* as the comparison text (I will ignore the spaces because the Dorabella Cryptogram doesn't contain any):

```
THEYW ENTOU TINTO THEST ILLSH ADOWY
CLOIS TERGA RDENT HESEM INARY OCCUP
IEDTH EBUIL DINGS OFANO LDDOM IN
```

Here's a transcription of the Dorabella Cryptogram:

```
ABCDE FGDHA IJKLJ MJJFB BJNGO GNIP
GJGFQ DHRSC JJCFN KGJIJ FTPKL QHHQI P
CPFUP CLUUN PCJFU KPNDB NPFDL ED
```

2 Basic Examinations

As both Elgar and Penny spoke English, I will assume that the Dorabella Cryptogram is written in English. As proposed by Fouché Gaines in her book, I started my examinations with a frequency count. I examined the comparison text first:

A	B	C	D	E	F	G	H	I
4	1	3	6	9	1	2	5	8

J	K	L	M	N	O	P	Q	R
-	-	5	2	7	8	1	-	3

S	T	U	V	W	X	Y	Z
5	9	3	-	2	-	3	-

According to Fouché Gaines, the nine most frequent letters of the English languange (E, T, A, O, N, I, R, S, H) make up about 70 percent of an English text. In the comparison text the nine most frequent letters are E, T, D, O, N, I, L, S, and H. These are not exactly the letters we have expected. However, they appear 62 times (71.3%), which is a pretty good fit.

Here's the frequency analysis of the Dorabella Cryptogram:

A	B	C	D	E	F	G	H	I
2	4	6	6	2	8	6	4	4

J	K	L	M	N	O	P	Q	R
11	4	4	1	6	1	8	3	1

S	T	U
1	1	4

The seven most frequent letters (J, F, P, C, D, and N) together appear 45 times (51.7%). In addition, there are four letters with a frequency of 6 (6.9%) each. If we take two of the latter we get 57 appearances (65.5%) for the nine most frequent letters. This is reasonably close to the 70% postulated by Fouché Gaines. We can take this as an indication that the Dorabella Cryptogram is a MASC encryption of an English, not a hoax.

To follow Fouché Gaines' approach, we now need to identifiy three groups of letters: the frequent, the less frequent, and the rare ones. Fouché Gaines' book does not define exactly the borders between these groups. I will work with the following definitions:

- *High frequency group (6 or more appearances)*: In the comparison text the following letters belong to this group: D, E, I, N, O, T. In the Dorabella Cryptogram the high-frequency letters are F, J, P, C, D, G, and N.
- *Medium frequency group (3-5 appearances)*: The medium-frequent letters of the com-

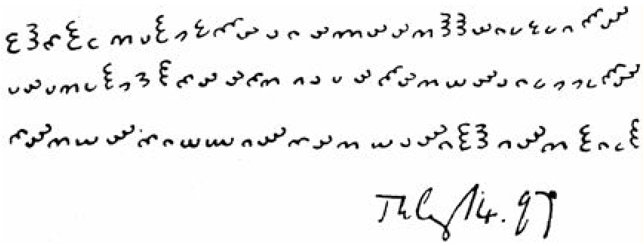

Figure 1: The Dorabella Cryptogram is an unsolved ciphertext that has the appearance of a monoalphabetic substitution cipher (MASC).

parison text are H, L, S, C, R, and Y. The medium-frequent letters of the Dorabella Cryptogram are B, H, I, K, L, U, and Q.

- *Low frequency group (1-2 appearances)*: Letters with a low frequency are B, F, G, M, P, and W (comparison text) and A, E, M, O, R, S, and T (Dorabella Cryptogram).

As the next step, we determine the contacts of each letter. A contact is defined as a letter that stands directly before or behind a certain character. As the Dorabella Cryptogram doesn't indicate spaces (and as we ignore the spaces in the comparison text), each of the 87 letters, except the first and the last one, has two contacts. The examples described by Fouché Gaines contain spaces, which means that her contact analysis is a little different from the one performed here.

Here's the contact analysis for the comparison text (below each letter the left and the right contacts are isted, one contact pair per line):

A	B	C	D	E	F	G	H
HD	EU	YL	AO	HY	OA	RA	TE
GR		OC	RE	WN		NS	TE
NR		CU	ET	HS			SA
FN			LI	TR			TE
			LD	DN			TE
			DO	HS			
				SM			
				ID			
				HB			

I	J	K	L	M	N	O	P
TN	-	-	IL	EI	ET	TU	UI
TL	-	-	LS	OI	IT	TT	
OS	-	-	CO		ET	DW	
HN	-	-	ID		IA	LO	
PE	-	-	OD		IG	YC	
UL	-	-			AO	SF	
DN	-	-			I-	NL	
MN	-	-				DM	

Q	R	S	T	U	V	W	X
-	EG	ET	-H	OT	-	YE	-
-	AD	LH	NO	CP	-	OY	-
-	AY	IT	UI	BI	-		-
-		EE	NO		-		-
-		NS	OH		-		-
-			SI		-		-
-			SE		-		-
-			NH		-		-
-			DH		-		-

Y
EW
WC
RO

Here's the contact analysis for the Dorabella Cryptogram:

A	B	C	D	E	F	G
-B	AC	BD	CE	DF	EG	FD
HI	FB	SJ	GH	LD	JB	NO
	BJ	JF	QH		GQ	ON
	DN	PP	NB		CN	PJ
		PL	FL		JT	JF
		PJ	E-		PU	KJ
					JU	
					PD	

H	I	J	K	L	M	N
DA	AJ	IK	JL	KJ	JJ	JG
DR	NP	LM	NG	NG		GI
QH	JJ	MJ	PL	PL		FK
HQ	QP	JF	UP	UP		UP
		BN				PD
		NC				
		KN				
		NF				

O	P	Q	R	S	T	U
GG	IG	FD	HS	RC	FB	FP
	TK	LH				LU
	IC	HI				UN
	CF					FK
	UC					
	NC					
	KN					
	NF					

3 Vowel Detection Method

The vowel detection method is the first one described by Fouché Gaines. It is based on eight criteria (Fouché Gaines calls them pointers) that can be used to identifiy vowels in a ciphertext. The idea of this method is to use the vowels identified for further investigations with other cryptanalysis methods (i.e., the vowel detection method alone will usually not break a cipher, but it can help to do so). Fouché Gaines' vowel detection method should not be confused with the Shukotin algorithm (Guy, 1991), which has the same purpose.

3.1 Pointer 1: High frequency of vowels A, E, I, and O

What Fouché Gaines writes: The vowels A, E, I, and O are normally found in the high-frequency section of a cryptogram.

Does this hold for the comparison text? It is true for the vowels E, I, and O. The letter A, however, has a lower frequency than expected.

What does this mean for the Dorabella Cryptogram? If Fouché Gaines is correct the ciphertext letters F, J, P, C, D, G, and N contain the cleartext vowels A, E, I, and O.

Vowel candidates in the comparison text: D, E, I, N, O, T

Vowel candidates in the Dorabella Cryptogram: F, J, P, C, D, G, N

3.2 Pointer 2: Letters contacting low-frequency letters

What Fouché Gaines writes: Letters contacting low-frequency letters are usually vowels.

Does this hold for the comparison text? Yes. The letters with frequency 1 are contacted by A, E, I, O, and U. The letters with frequency 2 are contacted by A, N, R, S, E, O, Y, and Y. This means that 10 of 13 letters contacting low-frequency letters are vowels. If we count only the contacts that appear more than once or that contact a letter that appears only once we get exactly A, E, I, O, U, and Y.

What does this mean for the Dorabella Cryptogram? The letters contacting low-frequency letters are B, B, D, D, F, H, I, J, J, L, G, G, H, S, R, C, and F. If we count only the contacts that appear more than once or that contact a letter that appears only once we get B, D, J, G, H, S, R, C, F, and B. As can be seen, the vowel detection doesn't work here as good as for the comparison text.

Vowel candidates in the comparison text: A, E, I , N, O R, S, and Y

Vowel candidates in the Dorabella Cryptogram: B, D, F, H, I, J, L, G, H, S, R, C, F

3.3 Pointer 3: Wide variety in contact letters

What Fouché Gaines writes: Letters showing wide variety in their contact letters are vowels.

Does this hold for the comparison text? Yes. Fouché Gaines does not define exactly what wide variety means. However, if we look at the three letters with the widest variety we see that these are the vowels E, I and O.

What does this mean for the Dorabella Cryptogram? The three letters with the widest variety are J, F, and P.

Vowel candidates in the comparison text: E, I, O.

Vowel candidates in the Dorabella Cryptogram: J, F, P.

3.4 Pointer 4: Repeated digrams

What Fouché Gaines writes: In repeated digrams (in immediate succession), one letter is usually a vowel.

Does this hold for the comparison text? There is no repeated digram in the comparison text.

What does this mean for the Dorabella Cryptogram? There is no repeated digram in the Dorabella Cryptogram.

Vowel candidates in the comparison text: -

Vowel candidates in the Dorabella Cryptogram: -

3.5 Pointer 5: Reveresed digrams

What Fouché Gaines writes: In reversed digrams, one letter is usually a vowel.

Does this hold for the comparison text? Yes. The reversed digrams of the comparison text are TO/OT, DE/ED, LO/OL, NA/AN, YW/WY, and SE/ES. Each of these six pairs consists of a vowel and a consonant.

What does this mean for the Dorabella Cryptogram? The reversed digrams are GJ/JG, JI/IJ, DE/ED, GF/FG, MJ/JM, GN/NG, FG/GF, JC/CJ, KP/PK, QH/HQ, PC/CP, and PN/NP. 9 of these 12 digrams contain one of the vowel candidates identified above (J, F, P). It is important to note that the Dorabella Cryptogram has double as many reversed digrams as the comparison text. It is beyond the scope of this paper to examine whether this high number of reversed diagrams is still consistent with an English text or whether it is evidence for the Dorabella Cryptogram being something else as English text.

Vowel candidates in the comparison text: -

Vowel candidates in the Dorabella Cryptogram: -

3.6 Pointer 6: Doubled consonants

What Fouché Gaines writes: Doubled consonants are usually flanked by vowels, and vice-versa.

Does this hold for the comparison text? The comparison text contains three doubled letters: LL, CC, and DD. Only CC appears inside a word (OCCUPY), while the other two stand at the end of a word (STILL) or are spread to two words (OLD DOMIN). While CC is flanked by two vowels, LL and DD aren't. It seems that this pointer is not applicable, if the word boundaries are not known.

What does this mean for the Dorabella Cryptogram: The digram JJ appears twice. HH and UU are two more doubled letters. No conclusions can be drawn from these facts at this stage.

Vowel candidates in the comparison text: -

Vowel candidates in the Dorabella Cryptogram: -

3.7 Pointer 7: Five consonants in succession

What Fouché Gaines writes: It is unusual to find more than five consonants in succession.

Does this hold for the comparison text? Yes.

What does this mean for the Dorabella Cryptogram? No conclusions can be drawn at this stage.

Vowel candidates in the comparison text: -

Vowel candidates in the Dorabella Cryptogram: -

3.8 Pointer 8: Vowels contacting each other

What Fouché Gaines writes: Vowels do not often contact one another.

Does this hold for the comparison text? Yes.

What does this mean for the Dorabella Cryptogram: There are five contacts between two of the supposed vowels J, F, and P. Further work might examine whether this information is of any use for breaking the Dorabella Cryptogram.

Vowel candidates in the comparison text: -

Vowel candidates in the Dorabella Cryptogram: -

3.9 Result of vowel detection

Only three of the eight criteria can be used to search for vowel candidates in the two texts examined in this paper. In the comparison text there are three letters that fulfill all three criteria: E, I, and O. This means that this method works to a certain degree for the comparison text, although the vowels A, U, and Y remained undetected.

Only two letters of the Dorabella Cryptogram, J and F, fulfill all three criteria. The letter P fulfills two of them. Although this is an interesting result, it is not enough to solve the Dorabella Cryptogram.

4 Digram Method

The digram solution method is the second one proposed by Fouché Gaines. However, Fouché Gaines writes that 80 letters are not enough for this technique. For this reason she demonstrates it on a 235 letter message. The 87 letters in the Dorabella Cryptogram and the comparison text are clearly not enough for this method to work. I therefore skip it in this paper.

5 Consonant Line Method

The third method Fouché Gaines introduces is the consonant line method. To apply this technique we need to assemble a table that lists the letters of the text along with their frequencies and number of different letters contacting them. Here's the table for the comparison text:

O	E	I	T	D	N	S	A	U	C
8	9	8	9	6	7	5	4	3	3
12	11	11	8	8	6	6	6	6	5

R	Y	G	H	M	W	B	F	P
3	3	2	5	2	2	1	1	1
5	5	4	4	3	3	2	2	2

According to Fouché Gaines, the lowest 20 percent of the total number of contacts are consonants. The letters G, M, W, B, F, P, and H together have 20 appearances. As we have a total of 106 contacts, we can determine these (correctly) as consonants. Here's the same table for the Dorabella Cryptogram:

J	P	B	C	D	F	G	H	I	K	U
11	8	4	6	6	8	6	4	4	4	4
10	8	7	7	9	11	7	5	5	6	6

L	N	Q	E	A	R	S	T	M	O
4	6	3	2	2	1	1	1	1	1
7	9	5	4	3	2	2	2	1	1

Here we have 117 accumulated contacts. 20 percent of 117 are 23. The letters M, O, R, S, T, A, and E together have 15 appearances, so they should be consonants. The next candidates are H, I and Q (five appearances each). If we include all of them we are above 20 percent. It seems best to omit all three.

It is important to note that Fouché Gaines' statements about contacts at this place refer to a text of roughly 100 letters. They make no sense for a much longer or shorter text. It is therefore an interesting question how the frequency of consonants can be measured in a way that is independent from the message length. This question, which is out of scope in this paper, is addressed in (Schmeh3, 2017).

In the next step, I write all the consonants I have identified in a line (see figure 2, ignore the underlined letters for now). Below, I write the left contacts of each letter left of the line, and the right contacts right of the line (see figure 2, again ignore the underlined letters).

Now, according to Fouché Gaines, all letters that don't show up as contacts left or right of the line can be identified as consonants, as well. These are C, D, L, and T (comparison text) and C, K, N, P, Q, and U (Dorabella Cryptogram). Note that this method has now correctly identified 11 consonants in the comparison text. The newly detected values are included in the upper line of consonant line diagram (underlined), and the contacts of these letters are added left and right of the vertical line (underlined).

According to Fouché Gaines, the N can now be identified, as it is a frequent letter that stands almost always left of the consonant line. The H can be identified, as it almost always stands right of it. Both identifications works pretty well for the comparison text, though the N could be confused with the T.

Following Fouché Gaines' instructions, we can now also indentify a few vowels. The vowels A, E, I, and O (i.e., the frequent ones) are expected

GMWBFPHCDLT	
RR	
A	AAA
NNNNN	
SSS	SS
EEE	EEEEEEE
II	IIIIIII
OOOOOO	OOOOOO
YY	Y
UU	UUU
TTTT	
LL	L
	T
DD	DDD
C	
	HHHH

EARSTMOCKNPQU	
DD	DDD
	FF
LLLL	LL
	BBB
HHHH	H
II	IIII
	SS
FFFFF	FF
JJJJ	JJJ
GG	GGGG
BB	
PPPPP	PPPPP
NN	
UUUU	U
K	KKK
T	
C	CCC
N	NN

Figure 2: Consonant lines for the comparison text (left) and the Dorabella Cipher (right).

to appear in a high frequency and on both sides of the line. Looking at the comparison text, this is absolutely correct for the O and partially holds for E, I, and U, while the A behaves a little different than expected. This attempted vowel identification is far from perfect, but it would certainly provide helpful evidence when combined with other cryptanalysis methods.

The information we have gathered by now would be sufficient to break the comparison text, if it were a cryptogram. It is very likely that the consonant that precedes the H four times is T (TH is the most frequent consonant pair in the English language). In addition, it is clear that the vowel that follows TH four times in the text is E (THE is the most frequent trigram in English texts). Knowing the four letters T, H, E, and N, the rest would be routine cryptanalysis work.

In the case of the Dorabella Cryptogram, things are less clear. There are several candidates for the N (L, H, and U) and several for the H (D, I, and K). The most promising vowel candidates are J, F, and P (the same candidates as determined using the vowel detection method). Note that the letter P was identified as a consonant in the second step of the consonant line method (because it doesn't touch any of the vowels determined in the first step), though its appearance on the consonant line (five left and five right contacts) makes it look like a vowel. Figure 3 shows the Dorabella Cryptogram with marked vowels and consonants (this is the next step recommended by Fouché Gaines).

In my view, there is no obvious way to proceed from here. So, I will leave further steps (for instance, checking if one of the different candidates for N and H makes sense) to future research.

All in all, it can be said that the consonant line method, which works surpisingly well on an ordinary English text of the same length and written in the same year, doesn't render a clear result for the Dorabella Cryptogram.

6 Conclusion

To my regret, none of the three methods described in this paper has led to a solution of the Dorabella Cryptogram. Nevertheless, there are a number of interesting conclusions that can be drawn from the examinations described in the previous paragraphs. Here are some general ones:

- The consonant line method has worked surprisingly well on the comparison text. It not only correctly identified G, M, W, B, F, P, H, C, D, L, and T as vowels but also found the letters N and H. This information would be enough to break a MASC-encrypted text. The consonant line method therefore appears to be an interesting alternative, whenever other methods (including frequency analysis and word pattern guessing) don't work.

- Digram analysis doesn't work on a cryptogram consisting of 87 letters. As virtually all unsolved MASC encryptions known to me are of about this size or smaller, this method will not be of much value.

- Vowel detection has worked on the comparison text. It correctly identified E, I, and O as vowels. While this is, of course, not enough to break a cipher, it might be an interesting aid.

- Generally, the concept of letter contacts seems to be an interesting tool in cryptanalysis, which has been underestimated so far. As far as I know, this concept is not mentioned at all in the crypto history literature of the last decades.

The following conclusions can be drawn about the Dorabella Cryptogram:

ABCDE **F**GDHA I**J**KL**J** M**JJF**B B**J**NGO GNI**P**

G**J**G**F**Q DHRSC **JJ**C**F**N KG**J**I**J** **F**T**P**KL QHHQI **P**

C**PF**U**P** CLUUN **P**C**JF**U K**P**NDB NP**F**DL ED

Figure 3: The Dorabella Cipher transcription with marked vowels (bold) and consonants (underlined). Further research may show whether there are conclusions that can be drawn from this.

- The frequency count and the contact counts of the Dorabella Cryptogram are consistent with the English language.
- The following letters in the transcribed Dorabella Cryptogram could be vowels: F, J, and P.
- The number of reversed digrams in the comparison text is double as high as in the Dorabella Cryptogram.
- Candidates for the letters H and N have been found in the Dorabella Cryptogram.
- The vowel detection method and the consonant line method work less good on the Dorabella Cryptogram than on the comparison text.

Here are some some ideas for future work:

- Some of the instructions given by Fouché Gaines are a little fuzzy. Especially, the definition of letter frequency classes and the quantification of contact variety is not very precise. This leaves room for further research.
- Fouché Gaines' methods assume that the ciphertext examined contains spaces. As this is not the case in many cases, the methods should be adapted to cryptograms without known word boundaries.
- The concept of letter contacts should be made more popular in cryptananalysis.
- The consonant line method should be applied on other cryptograms, as well.
- Further tests whether the Dorabella Cryptogram is a real text should be made.
- The consonant line method should be adapted to other languages.
- The concept of reversed digrams might be helpful for cryptanalysis. It should be explored.

I am optimistic that some of the unsolved cryptograms mentioned in the introduction can be solved with the methods covered here. I hope that additional research in this direction will be conducted.

References

Bauer, Craig. 2017. *Unsolved!*. Princeton University Press, Princeton, NJ.

Belfield, Richard. 2006. *Can You Crack The Enigma Code*. Orion Publishing, London, UK.

Fouché Gaines, Helen. 1939. *Cryptanalysis*. Dover Publications, New York, USA.

Guy, Jacques B. M. 1991. *Vowel Identification: An Old (but Good) Algorithm*. Cryptologia (4), 258-262 (1991)

Schmeh, Klaus. 2012. *Nicht zu Knacken*. Hanser, Munich, Germany.

Schmeh, Klaus 2017. *The Top 50 unsolved encrypted messages: 31. The MLH cryptogram* Klausis Krypto Kolumne, 2017-05-23

Schmeh, Klaus 2017. *Who can decipher this encrypted inscription on a cigaret case?* Klausis Krypto Kolumne, 2017-11-26

Schmeh, Klaus 2017. *Mathematical formula needed* Klausis Krypto Kolumne, 2017-12-22

Schneier, Bruce 2006. *Handwritten Real-World Cryptogram* Schneier on Security, 2006-01-30

Wikipedia. *Dorabella Cipher*. Retreived 2018-01-08.

Wikipedia. *Voynich manuscript*. Retreived 2018-01-08.