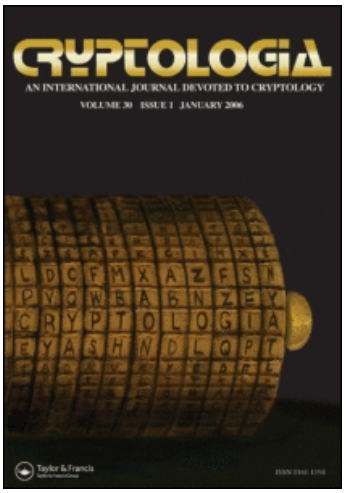


Cryptologia
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t725304178

Breaking Short Vigenère Ciphers

Tobias Schrödel

Online Publication Date: 01 October 2008

To cite this Article Schrödel, Tobias(2008)'Breaking Short Vigenère Ciphers',Cryptologia,32:4,334 — 347
To link to this Article: DOI: 10.1080/01611190802336097
URL: http://dx.doi.org/10.1080/01611190802336097

Cryptologia, 32:334–347, 2008

ISSN: 0161-1194 print
DOI: 10.1080/01611190802336097

Taylor & Francis
Taylor & Francis Group

Breaking Short Vigenère Ciphers

TOBIAS SCHRÖDEL

Abstract Vigenère ciphers can be broken, if the key length is known. In trying to break the Vigenère cipher, Charles Babbage and Friedrich Wilhelm Kasiski found the length of the key by searching for periodical repetitions in the ciphertext to split the cipher into multiple Caesar ciphers. William Friedman's, "index of coincidence," also requires an adequate length of the ciphertext to retrieve the key length. Both methods lack, if the ciphertext is short or does not include repetitions and no other effective linguistic solution to break short Vigenère ciphers is known. Massively decreasing the solution space by logic, reverse digram frequency, and language properties allows breaking short and long Vigenère ciphers with and without repetitions.

Keywords Babbage, digram, Kasiski, repetition, running key cipher, trigram, Vigenère

Historical Context

After Blaise de Vigenère invented the so called "chiffre indechiffrable" around 1586, from Leon Battista Alberti, Abbot Trithemius, Giovanni Battista Belaso, and Giovanni Battista Porta's preliminary work, it took almost 300 years until a solution to break the cipher was published. Giacomo Casanova claimed that he solved a Vigenère cipher earlier, but never described his solution [5, p. 153].

Charles Babbage invented the first established method approximately ten years earlier than Friedrich Wilhelm Kasiski but never published his results. When Kasiski, a former Prussian infantry officer [5, pp. 207–208], published his findings in 1864, in "Geheimschriften und die Dechiffrirkunst" [6, p. 41 ff], he made history, even if it was not recognized at that time (Figure 1).

The Vigenère Cipher

The Vigenère cipher is a so-called polyalphabetical substitution. Each character of the plaintext is encrypted using a different alphabet. Vigenère allows for the use of 26 different alphabets, each created by shifting the alphabet by one each time. To determine which alphabet is used for the encipherment of each character of the plaintext, one needs to know the key. When n is the position of the current character of the key in the regular alphabet, the cryptograph uses the alphabet that is shifted by n-characters. In a matrix of 26×26 characters, taking the rows to be the currently used characters for key and the columns to be characters of the text, the character of the ciphertext is retrieved by finding the intersection of the appropriate row

Address correspondence to Tobias Schrödel, Connollystr. 20/EG, München (Munich), Bayern (Bavaria), D-80809, Germany. E-mail: Tobias.Schroedel@t-online.de

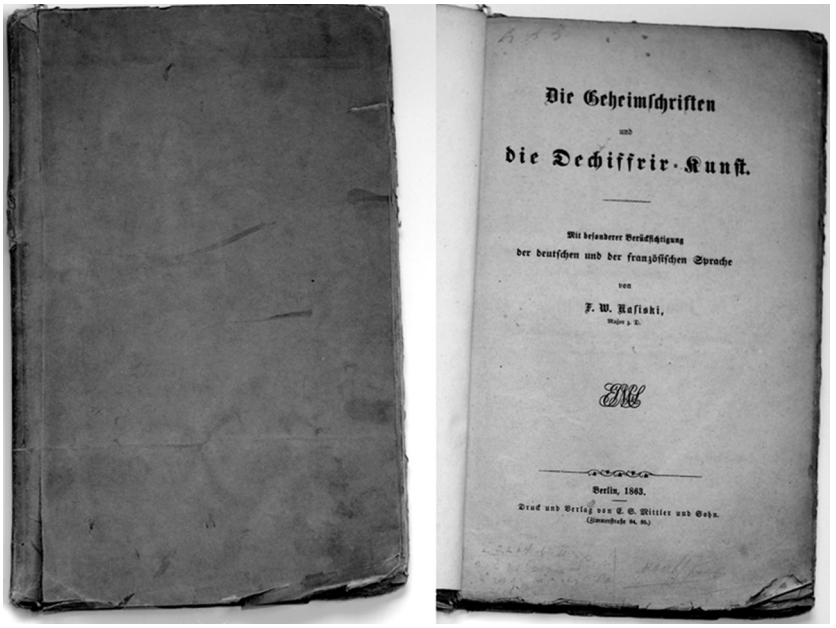


Figure 1. Cover and 1st page of Kasiski's book.

and column. If the length of the key is shorter than the text to encrypt, the key is appended to itself as often as required [5, p. 148 ff] [12, p. 45 ff].

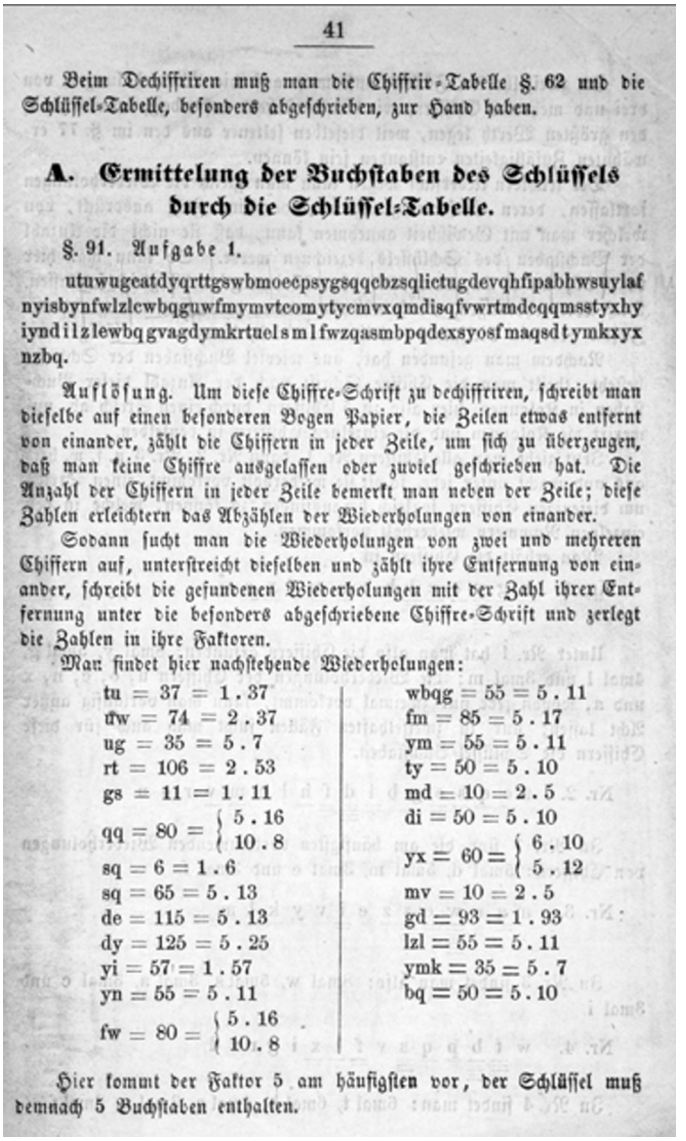
The Strength of the Vigenère Cipher

The Vigenère cipher was called the ''unbreakable chiffre'' because the number of possible solutions grows with the length of the text by a power of 26. For example, a text with 5 characters has $26*26*26*26*26 = 11{,}881{,}376$ possible solutions, a text with 10 characters has 141,167,095,653,376. Even for an up-to-date computer, calculating all possibilities for a longer text would take years.

Babbage and Kasiski's Method

Solving the Vigenère cipher can be done by finding repetitions of at least two consecutive characters (digram) in the ciphertext. Repetitions occur when the key is repeated to match the length of the plaintext and equal characters of the plaintext are encrypted with the same alphabet. As a regular text in whatever language does have digrams or trigrams that occur more often than others (e.g., ''the'' or ''ing'' for English texts), it is very likely that they are more than once encrypted using the same characters of the key, and therefore, with the same shifted alphabet. Finding a sufficient amount of these repetitions allows retrieving the length of the used key by calculating the biggest divisor of their distances (Figure 2).

A second method to retrieve the length of the key was invented in the early years of the 20th century by William Friedman. He found a formula that calculates the length of the key—the so called ''index of coincidence.'' When the cryptanalyst knows the length of the key, he is able to split the ciphertext into multiple Caesar



41

Beim Dechiffriren muß man die Chiffrir-Tabelle §. 62 und die Schlüssel-Tabelle, besonders abgeschrieben, zur Hand haben.

A. Ermittelung der Buchstaben des Schlüssels durch die Schlüssel-Tabelle.

§. 91. Aufgabe 1.

utuwugeatdyqrttgswbmoecpsygsqqcbzsqlictugdevqhfipabhwsuylaf nyisbynfwlzlewbqguwfmymvtcomytycmvxqmdisqfvwrtmdeqqmsstyxhy iyndilzlewbqgvagdymkrtuelsmlfwzqasmbpqdexsyosfmaqsdtymkxyx nzbq.

Auflösung. Um diese Chiffre-Schrift zu dechiffriren, schreibt man dieselbe auf einen besonderen Bogen Papier, die Zeilen etwas entfernt von einander, zählt die Chiffern in jeder Zeile, um sich zu überzeugen, daß man keine Chiffre ausgelassen oder zuviel geschrieben hat. Die Anzahl der Chiffern in jeder Zeile bemerkt man neben der Zeile; diese Zahlen erleichtern das Abzählen der Wiederholungen von einander.

Sodann sucht man die Wiederholungen von zwei und mehreren Chiffern auf, unterstreicht dieselben und zählt ihre Entfernung von einander, schreibt die gefundenen Wiederholungen mit der Zahl ihrer Entfernung unter die besonders abgeschriebene Chiffre-Schrift und zerlegt die Zahlen in ihre Faktoren.

Man findet hier nachstehende Wiederholungen:

tu = 37 = 1 . 37
ftw = 74 = 2 . 37
ug = 35 = 5 . 7
rt = 106 = 2 . 53
gs = 11 = 1 . 11
qq = 80 = { 5 . 16 / 10 . 8
sq = 6 = 1 . 6
sq = 65 = 5 . 13
de = 115 = 5 . 13
dy = 125 = 5 . 25
yi = 57 = 1 . 57
yn = 55 = 5 . 11
fw = 80 = { 5 . 16 / 10 . 8

wbqg = 55 = 5 . 11
fm = 85 = 5 . 17
ym = 55 = 5 . 11
ty = 50 = 5 . 10
md = 10 = 2 . 5
di = 50 = 5 . 10
yx = 60 = { 6 . 10 / 5 . 12
mv = 10 = 2 . 5
gd = 93 = 1 . 93
lzl = 55 = 5 . 11
ymk = 35 = 5 . 7
bq = 50 = 5 . 10

Hier kommt der Faktor 5 am häufigsten vor, der Schlüssel muß demnach 5 Buchstaben enthalten.

Figure 2. Page 40/41 of Kasiski's book.

ciphers and can then find each alphabet-shift using the well known technique of frequency tables [12, p. 63 ff] (Figure 2).

Advantage of Babbage and Kasiski's Method

The method described by Babbage and Kasiski does have advantages for the cryptanalyst. Although it was common to use an easy to remember word as key, it was not required for deciphering the message. Babbage and Kasiski's method works as well if the key consists only of random characters or a complete sentence. In addition, it was not necessary to know the language of the text to find the length

of the used key, while it was a requirement for Friedman's "index of coincidence." Even the multiple Caesar ciphers can be arranged with a maximum of 25 additional shifts in case of an unknown language.

Shortfall of Babbage and Kasiski's Method

Babbage and Kasiski's method needs repetitions of digrams and/or trigrams. If the ciphertext does not have any or enough repetitions the method fails. This would happen, if the key is as long as the plaintext (or a little shorter), as repetitions would then only be by accident and not due to same shifts in the used alphabet.

The great Austrian cryptanalyst Andreas Figl asserted in his 1927 manuscript, "Systeme des Dechiffrierens," that "... short texts enciphered using Tritheim's method normally remain unbreakable. [...] If these ciphers are encrypted with long key words or phrases [...] they remain absolutely unbreakable." ("... bleiben kurze Schriften nach Tritheims in der Regel unlösbar. Sind solche Schriften zudem mit langen Schlüsselwörtern oder -Phrasen [...] geschlüsselt [...] so wird diese regelmäßige Unlösbarkeit zur unbedingten.") [2, p. 185]. It will also fail in the unlikely [9] event of a large amount of accidental repetitions found in the ciphertext (Figure 3).

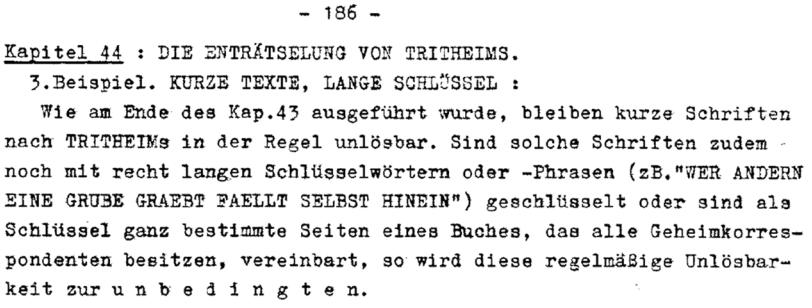
Breaking the Vigenère Cipher without Knowing the Key Length

Although longer Vigenère ciphers can be solved using linguistic tools such as probable words, repeated patterns, and frequency analysis, Brawley and Levine published a mathematical solution [1]. They solved a Vigenère cipher by finding "Equivalences of Vigenère Systems" over finite groups.

In the June/July 1943, edition of "The Cryptogram," an article was published that describes "Solving Vigenère by the Trigram Method" [4]. The author, Richard Hayes, shows how a Vigenère can be solved by testing the most common trigrams to find portions of the key. This idea is based on a linguistic method described by Helen Fouché Gaines in 1939 [3, p. 138 ff] and requires luck and many trial and errors.

But why not go in a direction contrary to this method and **leave out** all of the 26 possible text and key pairs derived from a cipher character that would never appear in English texts? Could one then find the one and only key and plaintext by eliminating the impossible or unlikely solutions step by step? Would that be sufficient enough, or can we otherwise decrease the solution space to a minimum?

With these thoughts in mind, an attempt will be made to break the following short Vigenère cipher:



- 186 -

Kapitel 44 : DIE ENTRÄTSELUNG VON TRITHEIMS.
3.Beispiel. KURZE TEXTE, LANGE SCHLÜSSEL :
Wie am Ende des Kap.43 ausgeführt wurde, bleiben kurze Schriften nach TRITHEIMs in der Regel unlösbar. Sind solche Schriften zudem noch mit recht langen Schlüsselwörtern oder -Phrasen (zB."WER ANDERN EINE GRUBE GRAEBT FAELLT SELBST HINEIN") geschlüsselt oder sind als Schlüssel ganz bestimmte Seiten eines Buches, das alle Geheimkorrespondenten besitzen, vereinbart, so wird diese regelmäßige Unlösbarkeit zur u n b e d i n g t e n.

Figure 3. Figl's manuscript.

IZRUOJVREFLFZKSWSE.

The method should be solid, reproducible, and be able to be automated with a (software) tool to exclude the need of a cryptanalyst's mind and eyes.

The example cipher is short and does not have any repetitions of digrams or trigrams in the ciphertext. The repetitions of single letters do not help and neither key nor text start with one of the ninety-eight most frequent trigrams [3, Appendix].

Reducing the Solution Space of Text/Key Pairs

Reducing the solution space should rely on more than a statistic. Based on experience and assumptions, a statistic could separate out the correct text/key pair and all following steps would be made without success. We first need a mathematical or logical solution.

The first letter of our example ciphertext "**I**" can be created by 26 different pairs of characters from key and plaintext. They are **A-I B-H C-G D-F E-E F-D G-C H-B I-A J-Z K-Y L-X** and so on.

Taking a look at pair 2 and 8 shows, that the letter **I** is created by the letter **B** in the plaintext and the letter **H** in key, as well as the letter **H** in the plaintext and the letter **B** in the key. These pairs (**B-H** $\rightarrow$ **I** and **H-B** $\rightarrow$ **I**) will be referred to as *dupes* (from duplicates) in the further text.

If we do not need to know at this point, which of the two characters belong to the plaintext or the key, we can disregard pair 2 or pair 8 in the following steps. We can also disregard pair 1 or 9, pair 3 or 7, as well as pair 4 or 6 (and so on for all other of the 26 possibilities).

Pair 5 in our example is a special one, as the character of the key is equal to the character of the plaintext. This is possible only for cipher characters being on odd positions in a regular alphabet (**A C E G I K M O Q S U W Y**). We need to take this pair into account.

This reduces the solution space for the first character from 26 to $26/2+1 = 14$ in case of an "odd first character" of the cipher, which is a reduction of 46%. All other first characters of the cipher do have a reduction of 50%. This difference will be equated by the second cipher character, as we will see later.

Conclusion: If, at this point of the cryptanalysis, we do not want to know what is the text and what is the key, the number of possibilities for a Vigenère cipher decreases to 50%. For example, a plaintext of 5 characters then has a number of $13^*26^*26^*26^*26 = 5{,}940{,}688$ instead of 11,881,376 possibilities.

The Second Character of the Ciphertext

After leaving out the dupes, we need to append all 26 characters and their equivalent of the key for the second cipher character **Z** to any of the remaining 14 pairs. This results in a total of $14^*26 = 364$ pairs of digrams (instead of $26^*26 = 676$) (Table 1).

More Duplicates

In case the first character of the ciphertext is one of the thirteen characters at odd positions in a regular alphabet the second character will create new dupes. (see Table 1: bold, strike-through).

Table 1. All possible digrams for the first two characters of the cipher"IZ"after removing dupes

AA-IZ	BF-HU	CK-GP	DP-FK	~~EU-EF~~	JZ-ZA	LE-XV	MJ-WQ	NO-VL	OT-UG	PY-TB	**RD-RW**
AB-IY	BG-HT	CL-GO	DQ-FJ	~~EV-EE~~	KA-YZ	LF-XU	MK-WP	NP-VK	OU-UF	PZ-TA	**RE-RV**
AC-IX	BH-HS	CM-GN	DR-FI	~~EW-ED~~	KB-YY	LG-XT	ML-WO	NQ-VJ	OV-UE	QA-SZ	**RF-RU**
AD-IW	BI-HR	CN-GM	DS-FH	~~EX-EC~~	KC-YX	LH-XS	MM-WN	NR-VI	OW-UD	QB-SY	**RG-RT**
AE-IV	BJ-HQ	CO-GL	DT-FG	~~EY-EB~~	KD-YW	LI-XR	MN-WM	NS-VH	OX-UC	QC-SX	**RH-RS**
AF-IU	BK-HP	CP-GK	DU-FF	~~EZ-EA~~	KE-YV	LJ-XQ	MO-WL	NT-VG	OY-UB	QD-SW	**RI-RR**
AG-IT	BL-HO	CQ-GJ	DV-FE	JA-ZZ	KF-YU	LK-XP	MP-WK	NU-VF	OZ-UA	QE-SV	**RJ-RQ**
AH-IS	BM-HN	CR-GI	DW-FD	JB-ZY	KG-YT	LL-XO	MQ-WJ	NV-VE	PA-TZ	QF-SU	**RK-RP**
AI-IR	BN-HM	CS-GH	DX-FC	JC-ZX	KH-YS	LM-XN	MR-WI	NW-VD	PB-TY	QG-ST	**RL-RO**
AJ-IQ	BO-HL	CT-GG	DY-FB	JD-ZW	KI-YR	LN-XM	MS-WH	NX-VC	PC-TX	QH-SS	**RM-RN**
AK-IP	BP-HK	CU-GF	DZ-FA	JE-ZV	KJ-YQ	LO-XL	MT-WG	NY-VB	PD-TW	QI-SR	~~RN-RM~~
AL-IO	BQ-HJ	CV-GE	**EA-EZ**	JF-ZU	KK-YP	LP-XK	MU-WF	NZ-VA	PE-TV	QJ-SQ	~~RO-RL~~
AM-IN	BR-HI	CW-GD	**EB-EY**	JG-ZT	KL-YO	LQ-XJ	MV-WE	OA-UZ	PF-TU	QK-SP	~~RP-RK~~
AN-IM	BS-HH	CX-GC	**EC-EX**	JH-ZS	KM-YN	LR-XI	MW-WD	OB-UY	PG-TT	QL-SO	~~RQ-RJ~~
AO-IL	BT-HG	CY-GB	**ED-EW**	JI-ZR	KN-YM	LS-XH	MX-WC	OC-UX	PH-TS	QM-SN	~~RR-RI~~
AP-IK	BU-HF	CZ-GA	**EE-EV**	JJ-ZQ	KO-YL	LT-XG	MY-WB	OD-UW	PI-TR	QN-SM	~~RS-RH~~
AQ-IJ	BV-HE	DA-FZ	**EF-EU**	JK-ZP	KP-YK	LU-XF	MZ-WA	OE-UV	PJ-TQ	QO-SL	~~RT-RG~~
AR-II	BW-HD	DB-FY	**EG-ET**	JL-ZO	KQ-YJ	LV-XE	NA-VZ	OF-UU	PK-TP	QP-SK	~~RU-RF~~
AS-IH	BX-HC	DC-FX	**EH-ES**	JM-ZN	KR-YI	LW-XD	NB-VY	OG-UT	PL-TO	QQ-SJ	~~RV-RE~~
AT-IG	BY-HB	DD-FW	**EI-ER**	JN-ZM	KS-YH	LX-XC	NC-VX	OH-US	PM-TN	QR-SI	~~RW-RD~~
AU-IF	BZ-HA	DE-FV	**EJ-EQ**	JO-ZL	KT-YG	LY-XB	ND-VW	OI-UR	PN-TM	QS-SH	~~RX-RC~~
AV-IE	CA-GZ	DF-FU	**EK-EP**	JP-ZK	KU-YF	LZ-XA	NE-VV	OJ-UQ	PO-TL	QT-SG	~~RY-RB~~
AW-ID	CB-GY	DG-FT	**EL-EO**	JQ-ZJ	KV-YE	MA-WZ	NF-VU	OK-UP	PP-TK	QU-SF	~~RZ-RA~~
AX-IC	CC-GX	DH-FS	**EM-EN**	JR-ZI	KW-YD	MB-WY	NG-VT	OL-UO	PQ-TJ	QV-SE	
AY-IB	CD-GW	DI-FR	~~EN-EM~~	JS-ZH	KX-YC	MC-WX	NH-VS	OM-UN	PR-TI	QW-SD	
AZ-IA	CE-GV	DJ-FQ	~~EO-EL~~	JT-ZG	KY-YB	MD-WW	NI-VR	ON-UM	PS-TH	QX-SC	
BA-HZ	CF-GU	DK-FP	~~EP-EK~~	JU-ZF	KZ-YA	ME-WV	NJ-VQ	OO-UL	PT-TG	QY-SB	
BB-HY	CG-GT	DL-FO	~~FQ-EJ~~	JV-ZE	LA-XZ	MF-WU	NK-VP	OP-UK	PU-TF	QZ-SA	
BC-HX	CH-GS	DM-FN	~~ER-EI~~	JW-ZD	LB-XY	MG-WT	NL-VO	OQ-UJ	PV-TE	**RA-RZ**	
BD-HW	CI-GR	DN-FM	~~ES-EH~~	JX-ZC	LC-XX	MH-WS	NM-VN	OR-UI	PW-TD	**RB-RY**	
BE-HV	CJ-GQ	DO-FL	~~ET-EG~~	JY-ZB	LD-XW	MI-WR	NN-VM	OS-UH	PX-TC	**RC-RX**	

As these dupes can be discarded, we now have $(26^*26)/2 = 338$ remaining pairs. The same value as if the first cipher character would have been on an even position in the alphabet: 50%. If the second cipher character is also one of the odd positioned characters, two additional pairs remain in the list. They represent cipher characters created by the same characters in text and key (e.g., **JL-JL** and **WL-WL**).

The list with either 338 or 340 elements represents all possible digram pairs for the beginning of the cipher **IZ**...

Discard Even More Pairs

We can then search for digrams that do appear more often than others, but that is trial and error. Instead we can also leave out digrams that have an acceptable probability of not appearing in either text or key and continue our research on the remaining pairs.

The question was: If we search for digrams that are extremely infrequent in English texts, can we discard any of our 338 digram pairs, containing at least one of these infrequent digrams? Or can we only discard those where both of the digrams in a pair are infrequent? Or can't we discard any at all?

Naturally, a text starts with a word, and a Vigenère is mainly used with the key being a word as well. We can therefore discard any pair where at least one of the digram pairs is known to be unlikely.[1] The text may also start with a one-letter-word such as "**I** quit..." or "**A** journey..." and that may cause infrequent digrams, so we should not discard any pairs starting with an "**I**" (**IQ**) or an "**A**" (**AJ**).

Conclusion: We are currently examining two digits of the ciphertext, and they reveal for sure two digits of the text and also for the key; in other words, we are currently examining parts of two words. This information allows disregarding pairs of digrams where at least one of the digrams never occurs in the language of the text message. We now learn as well, that we need to know the language of the original text.

Digram Statistics

We are searching for the **least** frequent digrams to eliminate an adequate number of pairs. A research on different English texts with a total of 3,310,696 characters in 688,321 words (~30,000 unique) showed that 146 digrams (~22%) never appeared (most of them included JQVX or Z). But are we looking at a correct statistic?

The digram-pairs stand for the beginning of two words (text and key) and we should therefore take a look at a statistic for digrams positioned at the **beginning of words**, so-called leading digrams. And this reveals that there are 363 (~54%) digrams that never occur at the beginning of the examined words.

Taking into account that digrams more seldom than 1% refer from abbreviations (such as **HB** is coming from **HBO**), the number increases by 45 to 408 (~60%) (Figure 4). In our example, this eliminates another 281 of the remaining 338 pairs and leaves only 57 for further examination. This is a reduction of 619 pairs (~90%).

[1]This is an assumption that could fail to decrypt a cipher, should it be a case when the key is not a word.

Identify the Correct Digram

Appending the 26 character pairs for the next cipher character **R** to the remaining 57 pairs gives 57*26 = 1,428 pairs of trigrams. They represent all possibilities for the beginning of the cipher **IZR.** The remaining pairs are currently unsorted, despite an unhelpful alphabetical order. So scoring and sorting the pairs seems to be the next step to take into account. What would be the most promising pair to start the next step with? Can we sort the pairs from the highest to lowest chance of both digrams being the beginning of two commonly used words?

In order to break a certain message by hand in the past, it sometimes took days for the "Black Chambers" [2, 13], even though a team of clerks did the routine work. Digram and trigram statistics were available [3, 6, 7], even if they were based on a lower number of words. In our example it would have been necessary to score 13*26*26 = 8,788 digrams and trigrams. Based on 5 seconds to score a pair by hand, this step would have taken over 12 hours. It would be more effective to start with digrams that have a high chance of being the correct pair.

It seems adequate to rate pairs with two medium chances higher than pairs with one extremely high chance while the chance of the second digram of the pair is low. If we apply, for example, the value of each leading digram occurrence from the statistic and then multiply both values we score each pair (like SPAM-detectors do with email). Also percentages, averages, or any mixture of scoring values can be chosen, as the sorting order is not vivid for the success of the solution. Nevertheless, a good sorting order minimizes the remaining calculations. Tests on what would be the best values to score digram pairs have not been made for this article and may be a goal for further investigations.

Tests with Different Ciphers

To isolate a smaller solution space for the correct digram pair, it was necessary to test the position of digrams and trigrams after scoring and sorting in different ciphers first. In order to create these ciphers, I used four different texts from the Project Gutenberg website [10] – A Christmas Carol (Charles Dickens); The Comedy of Errors (William Shakespeare); Tom Sawyer (Mark Twain); and Metamorphosis (Franz Kafka). The key was always a different random word with at least 5 characters from one of the four texts. Stepping through all four texts word by word, encrypting 20 characters, created 111,303 ciphers. The complete list of ciphers and the position of each pair can be downloaded [11].

In 83.80%, the correct **digram** was listed in the top ten of the sorted list and, in 90% of all ciphers the number of remaining pairs was between 30 and 70. The correct **trigram** pair was found in the top ten of the sorted list in 73.39%, and the number of remaining trigram pairs lay between 50 and 155 in nine out of ten ciphers (Figure 5). Common words such as "there," "could," or "while" were randomly chosen more often as key than rare words, so the statistic is more accurate for "running key ciphers" than for Vigenère ciphers. In "running key ciphers," the key is defined as a sentence from a text, containing such words as we used for the test.

A key for a Vigenère cipher would rather be a rare word such as "amphitheatres" and here, 12 out of 18 digrams and 7 out of 18 trigrams were in the top ten of the sorted list. The number of remaining digrams was between 26 and 76, while it was between 22 and 189 for trigrams (Figure 6).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	3	2219	6621	1681	1724	3252	1300	7518	0	421	56	4156	9662	2053	61	6000	0	2332	3064	2080	0	1550	8428	2	32	10
B	2063	0	0	0	39	2	2	3	1	0	0	1	13	0	883	0	0	0	4	0	0	0	0	0	0	0
C	2698	0	0	0	260	0	0	0	54	0	0	0	76	0	917	0	0	0	1241	0	0	0	0	0	0	0
D	1901	0	3	0	365	0	0	0	256	0	0	0	0	3	57	0	0	4	1	0	7	0	0	0	0	0
E	70	11533	2462	6458	21	2169	1676	4424	0	519	443	3405	4257	4099	111	3673	0	10997	6325	2702	1	2026	6672	1	1078	46
F	1735	0	10	0	534	2	0	0	1189	0	0	0	0	0	39381	1	0	0	0	0	0	0	0	0	0	0
G	1638	0	0	0	174	0	0	0	49	0	0	0	0	0	23	21	0	0	0	0	3	0	0	0	0	0
H	15	0	4015	0	5	0	30	0	0	0	4	0	0	0	93	949	0	168	3828	89259	0	1	10604	0	0	0
I	372	1025	1166	6284	418	5139	1108	4005	172	3	1269	4602	2769	302	96	811	0	1734	3035	2719	0	1620	7477	104	68	29
J	0	0	0	2	3	1	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
K	8	0	0	0	0	0	0	0	0	0	0	2	0	0	21	0	0	0	924	0	1	0	0	0	0	0
L	4915	1237	1911	0	1411	1156	693	1	1257	0	28	15	0	0	503	2078	0	0	1009	0	454	0	0	0	0	0
M	2340	0	4	0	1033	0	0	0	1800	0	0	0	2	0	36	0	0	0	661	114	7	1	6	0	0	0
N	28036	0	1	0	3367	0	5	0	27571	0	1026	0	0	0	7161	32	0	0	102	0	3922	0	0	0	0	0
O	32	3324	15800	2382	13	9924	2132	2678	109	937	16	3004	5299	5785	19	4166	0	1683	5877	18217	0	623	3190	0	1034	58
P	1945	0	0	0	343	0	0	0	2	0	0	0	3	1	1440	305	0	0	2702	0	1770	0	0	0	1	0
Q	21	0	0	0	288	0	0	0	0	0	0	0	0	0	0	0	0	0	113	0	0	0	0	0	0	0
R	6048	2416	1565	1311	301	7111	3655	0	523	1	4	0	100	0	7136	8691	0	0	0	3602	192	1	455	0	1	0
S	6630	0	4	0	1178	0	0	0	8631	0	0	1	8	0	312	39	0	0	2	10	1693	7	0	0	1	0
T	4270	0	1	0	166	2	0	20	8269	0	0	12	1	0	1550	15	0	0	6471	0	164	0	0	0	0	0
U	710	3445	1111	1327	293	958	640	1012	0	1003	11	328	2219	768	1829	1838	999	724	6290	1329	0	16	1	0	14	6
V	247	0	1	1	2085	1	0	0	83	0	0	0	0	0	1034	0	0	0	1	0	1	0	0	27	0	0
W	289	0	0	50	3	0	0	0	0	0	0	0	0	0	571	0	0	0	618	1180	0	0	14	0	0	0
X	135	0	0	0	4262	0	0	0	47	0	0	0	0	0	68	0	0	0	0	2	0	0	0	22	0	0
Y	3	5902	229	69	198	0	9	309	0	0	2	427	252	5	9	174	0	9	1220	304	0	0	14	0	0	3
Z	5	0	9	0	0	0	0	0	0	0	0	0	0	0	5	0	0	1	0	0	0	0	0	0	0	0
669741	66129	31101	34913	19565	18484	29717	11250	19971	50013	2884	2859	15953	24661	13016	63317	28793	999	17652	43488	121518	8215	5845	36861	156	2229	152

To learn the frequency of any leading digram, find ist first letter at the top, find ist second letter at the side. (Frequency for AE = 70; for EA = 1724)

0 = frequency is zero 29 = frequency is below 30 2204 = frequency is above 29

Figure 4. Statistic of leading digrams.

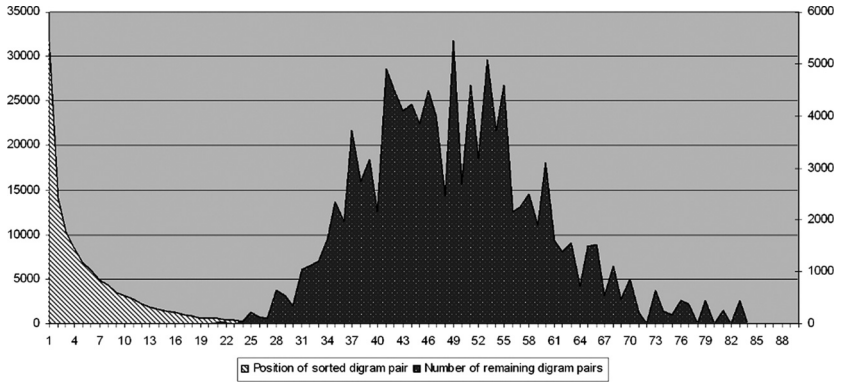


Figure 5. Diagraph for digrams.

Score and Sort the Digram Pairs

Sorting the pairs by the result of their multiplied scores does give a good starting point for the next step. Pairs beginning with a two letter word in both digrams can be sorted to the end of the list, as they don't seem to be very promising (Table 2).

The Third Character of the Ciphertext

From the now sorted list, we take the highest scored pair **AN-IM** and append all 26 characters and their equivalent of the key as we did before with the second character. We now have the first 26 trigram pairs that were derived from the first three characters of the cipher **IZR** (Table 3)...

To discard trigrams, we need a statistic of leading trigrams. 20 of the 26 can be discarded as either the first or the second trigram does not lead any other of the 688,312 words we investigated for the statistic (Table 4). And again, we need to think of word breaks, as the text may have started with a single-letter-word or a two-letter-word and these will generate seldom trigrams.

Fortunately, there are only 24 common two-letter-words in English language:

AD AM AN AS AT BE BY DO GO HE IF IN IS IT ME MY NO OF OK ON SO TO UP US

			Digram		Trigram	
CIPHER	TEXT	KEY	Pos	Max	Pos	Max
IZRUOJVREFLFZKSWSE	- the cipher of the article -	- unknown -	14	57	24	124
EZUVIDBUAAPMGZM	TROOPSTOTHEEAST	LIGHT	5	48	11	69
SWIEGQPQABGOXWHPP	GIVEPOISONTOGUARD	MONARCHY	12	61	4	148
YVVTJXXWUIFMNGZRVTSUVV	THEPRESIDENTISINDANGER	FOREST	1	37	1	69
ISEITWOSXWPH	HELPISNEEDED	BOTTLE	11	66	23	189
TBGVGSL	BUYGOLD	SHIP	8	48	65	135
IYASIOEBWJHOP	MYHORSEISSLOW	WATER	23	65	37	146
MIXPYET	KILLHER	CAMERA	14	76	57	183
SAKMJKVXV	WARISOVER	WATER	2	53	5	123
OUZONJVLBIZMYEQMKJIOXSWB	WAITFORTHREEDAYSTOATTACK	SURVIVE	2	53	12	98
RHUIPBTSWZVOTB	ATTACKFROMEAST	ROBIN	8	28	8	81
PSFGTKEFJAKH	DONOTEATFISH	MESSAGE	10	50	67	150
YXXPALRGQGFVAFRWPDBVASVZEHRQVM	OPENTHEWINDOWSHOWBURNINGCANDLE	KITCHEN	4	26	5	22
LHZPSTSNLWSJANBSSNANLZK	THREECANDLESINTHEWINDOW	SAILOR	1	37	4	112
CEHUFMQFIPJTZFV	ATTACKFROMHILLS	CLOUD	9	46	4	110
VEJEKWKTFCEHQBDQUXIUTSH	BRINGWATERANDAMMUNITION	UNBREAKABLE	6	42	33	82
WUQVIRGQNWIWDGRZCSDHJAA	ENIGMACODEBOOKAVAILABLE	SHIPWRECK	5	60	66	91
BGGULHDRHHLWW	BURNDOWNHOUSE	AMPHITHEATRES	5	36	11	92

Figure 6. Table of different positions of digrams and trigrams.

Table 2. Sorted list of remaining diagram pairs

AN-IM	PS-TH	MR-WI	EF-EU	KL-YO	CF-GU
DI-FR	EM-EN	AS-IH	OM-UN	EG-ET	AE-IV
PL-TO	BL-HO	AL-IO	AH-IS	AV-IE	OG-UT
PR-TI	PI-TR	AT-IG	AC-IX	OI-UR	PN-TM
AG-IT	DO-FL	AB-IY	EI-ER	EL-EO	AQ-IJ
CO-GL	CR-GI	AI-IR	LE-XV	AJ-IQ	MB-WY
BR-HI	MI-WR	AP-IK	AW-ID	PH-TS	AM-IN
DR-FI	EC-EX	AD-IW	EE-EV	OO-UL	
CI-GR	AR-II	AF-IU	AO-IL	EB-EY	
CL-GO	AU-IF	OH-US	OK-UP	AX-IC	

Trigrams starting with one of these digrams need to be verified if a message can logically start with it, and should then not be discarded.

What do we do with trigrams starting with **I** or **A**? In addition to scoring the trigram, we take a look at the digram presented by the **second and third** character and score them with the leading **di**gram statistic. **ANQ1-IMB,** for example, scores zero because the combination of **NQ** and **MB** is below 1% in the leading digram statistic. We can discard the complete trigram, even if it starts with a single-letter-word. Actually the pair with the high weighted **THE** falls out due to its related trigram **PSN**.

Dictionary Assistance

Expanding all remaining digrams and discarding trigrams would result in a total of ~125 possible solutions. All of them represent the first three characters of two words, although we have to take short words into account and treat them differently. We may now look out for all words starting with the trigrams of the currently investigated trigram pair and check them for being text or key.

That list may be very long and time consuming and one would normally survey the complete solution space. Once again, linguistic tools can help to decrease it. As the pronunciation of a language is based on its roots and these differ from hemisphere to hemisphere, native English does not allow to articulate a word with **THRB**. Statistics for English texts show that **B** can follow **R** like in bar**b**ecue [3, Appendix], but **RB** will never follow **TH**.

We can take an English dictionary (for our example we use [8]) and see which words start with our trigram pairs, then make a list of **all different characters** at the **fourth** position of these words. The fourth character of the ciphertext

Table 3. First 26 trigram pairs derived from the digram AN-IM

ANA-IMR	ANF-IMM	ANK-IMH	ANP-IMC	ANU-IMX	ANZ-IMS
ANB-IMQ	ANG-IML	ANL-IMG	ANQ-IMB	ANV-IMW	
ANC-IMP	ANH-IMK	ANM-IMF	ANR-IMA	ANW-IMV	
AND-IMO	ANI-IMJ	ANN-IME	ANS-IMZ	ANX-IMU	
ANE-IMN	ANJ-IMI	ANO-IMD	ANT-IMY	ANY-IMT	

Table 4. Trigram pairs derived from AM-IN

~~ANA-IMR~~	**ANF-IMM**	~~ANK-IMH~~	~~ANP-IMC~~	~~ANU-IMX~~	~~ANZ-IMS~~
~~ANB-IMQ~~	~~ANG-IML~~	~~ANL-IMG~~	~~ANQ-INB~~	~~ANV-IMW~~	
ANC-IMP	~~ANH-IMK~~	~~ANM-IMF~~	**ANR-IMA**	~~ANW-IMV~~	
~~AND-IMQ~~	~~ANI-IMJ~~	**ANN-IME**	~~ANS-IMZ~~	~~ANX-IMU~~	
~~ANE-IMN~~	**ANJ-IMI**	~~ANO-IMD~~	**ANT-IMY**	~~ANY-IMT~~	

is **U**. Can it be created by any combination of the fourth characters found in the words?

Some Example Trigram Pairs

Example: **PSA-THR**
We find three words starting with **PSA** (Psa**l**m, Psa**l**mist and Psa**m**motherapy). This gives us two different characters at the fourth position [**LM**]. Checking **THR** gives us many words, but only five different characters at the fourth position of these words [**AEIOU**].

Can one combination of **[LM] + [AEIOU]** create the cipher character **U**? The answer is yes, as **[M] + [I] = U** and this leaves only one word: **PSAM**MOTHERAPY for **PSAM*** and a couple of words starting with **THRI*** on the other side.

We find two dozen words starting with **THRI*** but only five different characters at the fifth position, so we check if **[CFLPV] + [M] = O**. Still one match **[C] + [M] = O**. That leaves **THRIC**E and **PSAMM**OTHERAPY as the only possible solutions for the first five characters of the cipher derived from the trigram pair **PSA-THR**.

Both words lead into a dead end, as **[E] + [O] ≠ J** at the 6th position. This dead end situation will happen very often, mostly at the 4th and the 5th position. The trigram pair can be discarded.
Example: ***ANC-IMP***

We find many words starting with **ANC** giving four different characters at the 4th position **[EHIO]**. For **IMP** we find many words as well, but only seven different characters at the 4th position **[AEILORU]**. No combination of **[EHIO]** with **[AEILORU]** creates **U** in a Vigenère square. **[EHIO] + [AEILORU] • U**, so the trigram pair **ANC-IMP** can be discarded.

Solve the Message

After checking 23 trigrams, all of them leading in dead ends, we find **BLA-HOR** as the next trigram pair. This pair gives 11 characters **[BCDIMNSTZ]** from different words for **BLA** and seven **[ADIMNRS]** for **HOR**.

[BCDIMNSTZ] + [ADIMNRS] generates **U** from **C←→S D←→R I ←→M** and **M←→I**

So we need to check all word pairs starting with

C←→S--- ---D←→R	**I←→M--- ---M←→I**

BLAC HORS	**BLAD HORR**	**BLAI HORM**	**BLAM HORI**

This gives:

BLACK* **HORSE***	**BLADD*** **HORRE***	**BLAIN** **HORME***	**BLAMA*** **HORIZ***
	BLADE* **HORRI***	**BLAIR*** **HORMI***	**BLAME***
	BLADI* **HORRO***	**HORMO***	

Starting alphabetically, we find the words **BLACK** and **HORSE**. Both words do not necessarily stand alone, but they can – and should, therefore, be tested to be text or key. As it is still not clear, what would be text and key, it is essential to investigate both words as key and verify the text to be a valid message text.

Cipher: IZRUOJVREFLFZKSWSE
Key: BLACKBLACKBLACKBLA
Text: HORSEIGRGPMQ...

It seems obvious, that **BLACK** is not the key, so try **HORSE** instead.

Cipher: IZRUOJVREFLFZKSWSE
Key: HORSEHORSEHORSEHOR
Text: BLACKCHAMBERISOPEN

And this attempt solves the cipher. The ciphertext **IZRUOJVREFLFZKSWSE** was encrypted with the key **HORSE** and means "**Black chamber is open**."

A Word on "Running Key Ciphers"

The approach described in this article is adequate to find a "running key". If the first word for key and text was identified and the cipher turns out to be a "running key cipher" two situations can occur that require different proceedings.

(a) The first words of key and text have equal lengths
A rerun of the complete procedure needs to be done with the undecrypted part of the cipher.
(b) The first words differ in word length
In this case, the longer word and the cipher generate one or more correct starting letters for the unknown word. They can be expanded to a trigram and then the cryptanalyst can continue with the dictionary assistance.

Shortfall of this Method

The described method will fail, if the key is not a word and consists only of random characters, although this cannot happen for running key ciphers. Texts starting with short words will create a bad sorting order and the cryptanalyst needs to perform more calculations for success. Automated routines need excellent algorithms to succeed with short words.

A meaningful statistic for leading digrams and leading trigrams is essential for this method, as the statistic dictates if a digram/trigram will be discarded or not. It is also dependent upon a good dictionary, as automated routines will fail if the used key word is not listed. To be able to work with the correct statistics and dictionaries, the cryptanalyst needs to know the language of the cipher.

Automation

A proof-of-concept tool is available at my homepage [11]. It lists all possible trigram pairs for Vigenère ciphers and tries to solve the message with a short dictionary.[2] The reader can create a message by entering key and text as well as only a cipher. The tool currently supports English language only.

Acknowledgments

My warm thanks go to Dr. Klaus Pommerening. His article [8] inspired me, when I wanted to prove him wrong and vainly tried to find a cipher with solely accidental repetitions. Thanks to my family for giving me the time to spend with my weird hobby. This article is dedicated to my friend, encryption specialist Sascha Hanke, who died far too young in January 2007.

About the Author

Tobias Schrödel lives and works in Munich, Germany as consultant for software implementation, and programming at T-Systems. The IT specialist (CoC) is examiner at the Chamber of Commerce and performs live-hacking shows at customer events. He collects historic books on cryptanalysis and trains elementary school children in simple encryption methods. His Homepage is http://www.sichere.it

References

1. Brawley, J. V. and J. Levine. 1977. "Equivalences of Vigenere Systems," *Cryptologia*, 1:338–361.
2. Figl, A. 1927. *Systeme des Dechiffrierens*, Wien: Never published.
3. Gaines, H. F. 1939. *Elementary Cryptanalysis*, Boston: American Photographic Publishing.
4. Hayes, R. (aka TRYIT) 1943. "Solving Vigeneres by the Trigram Method," *The Cryptogram*, JJ, 1943:46–49.
5. Kahn, D. 1967. *The Codebreakers*, New York: Scribner.
6. Kasiski, F. W. 1863. *Geheimschriften und die Dechiffrirkunst*, Berlin: Mittler und Sohn.
7. Langie, A. 1918. *De la Cryptographie*, Paris: Payot & C[ie].
8. LEO GmbH. *Online Dictionary*, http://dict.leo.org (Accessed Feb 2008).
9. Pommerening, K. 2006. "Kasiski's Test: Couldn't the repetitions be by accident?" *Cryptologia*, 30:346–352.
10. Project Gutenberg Literary Archive Foundation. *Project Gutenberg*, http://www.gutenberg.org (Accessed Feb 2008).
11. Schrödel, T. *Sichere IT*, http://www.sichere.it/cryptologia.php?language=EN. (Accessed Aug 2008).
12. Singh, S. 1999. *The Code Book*, London: Fourth Estate.
13. Yardley, H. O. 1931. *The American Black Chamber*, Indianapolis: The Bobbs-Merrill Company.

[2]The software has no support for ciphers beginning with 1–3 letter words and will only find the first key word in case of a "running key cipher."