Excursus 3

Cryptanalysis and the validation of deciphered texts

Cryptanalysis refers to the deciphering of a cryptogram by someone who does not have access to the key: a codebreaker. For example, if an enemy courier with an encrypted message is intercepted, an attempt may be made to decipher the message without access to the cryptographic key. In the first chapter, a simple example of cryptanalysis was provided in Fig. 1.2, which for convenience is repeated below in Fig. E3.1. In the discussion of that example, we found that the absolute rate of language was 24^{12}, equal to approximately 36,520 trillion. We then estimated the number of 12-letter English words to be 20,000. The probability, then, of a serendipitous deciphering was determined by dividing the number of 12-letter English words by the absolute rate of language, and then multiplying by the number of potential keys, which is 23. We obtained a probability of one in 79 billion.

Ciphertext:	Z	O	U	M	Q	L	D	O	X	M	E	U
Key = 1:	A	P	W	N	R	M	E	P	Y	N	F	W
Key = 2:	B	Q	X	O	S	N	F	Q	Z	O	G	X
Key = 3:	**C**	**R**	**Y**	**P**	**T**	**O**	**G**	**R**	**A**	**P**	**H**	**Y**
Key = 4:	D	S	Z	Q	U	P	H	S	B	Q	I	Z
Key = 5:	E	T	A	R	W	Q	I	T	C	R	K	A

(Keys 6 through 23 are omitted)

Fig. E3.1 Cryptanalysis of Caesar shift cipher

The validation of cryptograms comes down to the following simple principle. There are two circumstances that could have produced the plaintext message: either someone actually encrypted the message using the key or, by some freakish chance, the plaintext serendipitously emerged. If one can show that the probability of the second circumstance—that the cryptanalytic process accidentally generated a valid plaintext—is sufficiently remote, then the plaintext must be the encipherer's authentic and intended message. A cryptogram is validated by showing that the chance of its accidental generation is essentially nil.

In the above example, the message was enciphered using a simple Caesar shift. This, of course, provides very little security because the key has a range of only 1 to 23. Your bank would never allow you to choose such a short key, but it might allow you to choose a 4-digit PIN, with a range of 0 to 9,999. We next consider a polyalphabetic cipher (discussed in “The Cryptography of the *Polygraphia*” section of Chapter 2): we encipher the 12-letter word CRYPTOGRAPHY using 4 Caesar shifts (limited to shifts from between 0 and 9), each applied to a group of 3 characters. If we arbitrarily select our PIN to be 1324, we then subtract 1 from the first three letters (CRY is enciphered as BQX), we then subtract 3 from the next three letters (PTO is enciphered as MQL), we then subtract 2 from the next three letters (GRA is enciphered as EPY), and finally 4 from the last three letters (PHY is enciphered as LDT). In all, CRYPTOGRAPHY is enciphered as BQXMQLEPYLDT.

To decipher our enciphered message, we add, rather than subtract, the PIN. Fig. E3.2 shows the deciphering of BQXMQLEPYLDT using the PIN or key, 1324, which produces the plaintext CRYPTOGRAPHY (which appears in bolded letters in four different rows). An unauthorized decipherer (cryptanalyst or codebreaker) could crack the key by iterating through all 10,000 possible keys (0000 to 9999), or by iterating through a smaller set of possibilities based on the likelihood of the appearance of various trigrams (a sequence of three letters).

Ciphertext:	B	Q	X	M	Q	L	E	P	Y	L	D	T
Key = 1	**C**	**R**	**Y**	N	R	M	F	Q	Z	M	E	U
Key = 2	D	S	Z	O	S	N	**G**	**R**	**A**	N	F	W
Key = 3	E	T	A	**P**	**T**	**O**	H	S	B	O	G	X
Key = 4	F	U	B	Q	U	P	I	T	C	**P**	**H**	**Y**

Fig. E3.2 Cryptanalysis of polyalphabetic cipher

We may now ask: what is the probability, given a 12-letter plaintext enciphered by a key whose range is 0 to 9,999, that an unintended plaintext word is generated? The key range of 10,000 is much larger than the previous example, in which it was 23. When we perform the probability calculation once again, using 10,000 instead of 23, the result is a probability of one in 183 million. This is more likely than the previously calculated odds of 1 in 79 billion, but still very improbable.

Now, let us consider the most extreme example of key range, in which the key consists of 12 independent numbers, with a range of 1 to 24. This allows any ciphertext to generate any plaintext because the range of the key is as large as the absolute rate of language. In Fig. E3.1, the ciphertext ZOUMQLDOXMEU, when deciphered using a key = 333333333333, produced the plaintext CRYPTOGRAPHY. Below, I have begun with the same ciphertext, ZOUMQLDOXMEU, but chosen a key to produce the plaintext WICKETKEEPER; indeed, I could have chosen another key and produced any 12-letter word that I desired.

Ciphertext:	Z	O	U	M	Q	L	D	O	X	M	E	U
Key:	21	19	7	22	13	8	6	15	7	3	24	21
Plaintext:	**W**	**I**	**C**	**K**	**E**	**T**	**K**	**E**	**E**	**P**	**E**	**R**

Cryptographers call this "perfect secrecy," because no cryptanalysis is possible. This modern-day method is commonly used for diplomatic messages and is called "a one-time pad" (because keys are used just once). As a result of the key having the same amount of information (or range) as the message, the cipher is indecipherable, except by someone who knows the key.

The fundamental principle demonstrated here is that the complexity or length of the key must be factored into any calculation of the security of a cipher, or as concerns us here, of the validity of a cryptanalytic (unauthorized) deciphering. If the key is relatively short, as is the case with the simple Caesar shift shown in Fig. E3.1, then the cipher is not secure, and one can easily guess the key. In such situations, validation of the unauthorized deciphering is immediately apparent. The coherent plaintext message must certainly be the intended message of the decipherer because the probability of it occurring by chance is extremely low (1 in 79 billion, as previously calculated). In contrast, if the key is long and approaches or is equal to the absolute range of language of the ciphertext, as in the example of perfect secrecy above, then any unauthorized deciphering of the message is either impossible or unreliable. Cryptographers validate unauthorized decryptions by comparing the range of the key (known as "key equivocation") to the ratio of the absolute rate of language of the ciphertext, divided by the range of the plaintext. If the range of the key is significantly smaller than the range of the absolute rate of language divided by the range of the plaintext, then the decryption is valid; if not, it is suspect.

In calculating the probability that a decryption is valid, we made a simplifying assumption for convenience: that the 12-letter message is one of 20,000 possible 12-letter English words. Of course, the message could consist of multiple words rather than a single word. Also, many of the 20,000 12-letter words on our list are incredibly unlikely choices. Messages usually have some implicit context. For example, in the context of this study, the plaintext CRYPTOGRAPHY is relevant, but the plaintext WICKETKEEPER is not. Indeed, most words on the 12-letter word list that I consulted, including the first, ABANDONMENTS, and the last, ZYMOTECHNICS, are not relevant: this study has nothing to do with playing cricket, abandonments, or brewing beer.

In order to correctly validate cryptograms, one must account for the use of multiple words and the relevance of the message. This can be accomplished using Information Theory, a foundational theory in computer science formulated by Claude Shannon (1916–2001). He recognized that information could be quantified and measured in logarithmic units (either base 2 logarithms—bits of information—or base 10 logarithms). For example, in a 16-letter (equiprobable[61]) alphabet, each letter contains 4 bits of information because exactly 4 bits of information are required to represent any letter ($2^4 = 16$). The value of 4 is then taken to be the quantitative measure of a character's information content. This measure of information is a fundamental mathematical quantity that is used in the analysis of cryptographic systems, and it can be used to determine whether an unauthorized deciphering of a cryptogram is valid.[62]

Shannon developed the means to measure how often a random sequence of letters would be a valid expression in a natural (human) language such as English. He envisioned a series of steps by which random letters approach natural language, as shown in Fig. E3.3. The labels inside the trapezoid figure describe the level to which an English text is approximated; the messages to the right of the trapezoid are sample texts for each level of approximation. At the base of the trapezoid, the letters are random. At the second level of the trapezoid, Shannon's first level of approximation to English, he duplicated the individual letter frequencies of English. At the third level, Shannon increased his approximation to English by duplicating trigram (3-letter group) frequencies. At the fourth level of the trapezoid, only sequences of letters that are valid English words are included. The fifth level improves resemblance to English by mimicking English word order, but the text is still nonsensical. Not until the sixth level of the trapezoid, do we have a meaningful, grammatical English sentence. In climbing up each level of the trapezoid, the number of qualifying texts is winnowed down exponentially. Only an incredibly small percentage of the sequences of random letters from the first level of the trapezoid qualify as valid English at the sixth level of the trapezoid.

Finally, the seventh and top level of trapezoid in Fig. E3.3 adds a further important qualification when validating a message: contextual relevance. Notice that the sample text at the sixth level concerns the origin of the Homeric poems, a matter irrelevant to this study. In contrast, the sample text at the seventh level is descriptive of the chart itself. When a cryptogram is deciphered, one expects the deciphered text to have some contextual relevance. For example, if one intercepts an enemy's military communications, one would not expect the subject of the message to be Homer. In fact, the range of texts one would expect is rather narrow, perhaps something like "Send the armored division to Bastogne." This contextual relevance is also applicable to our deciphering of the Puzzle Sonnet's cryptogram. Any deciphered message ought to be very germane to the context in which the Puzzle Sonnet is presented. This greatly narrows the scope of valid messages.

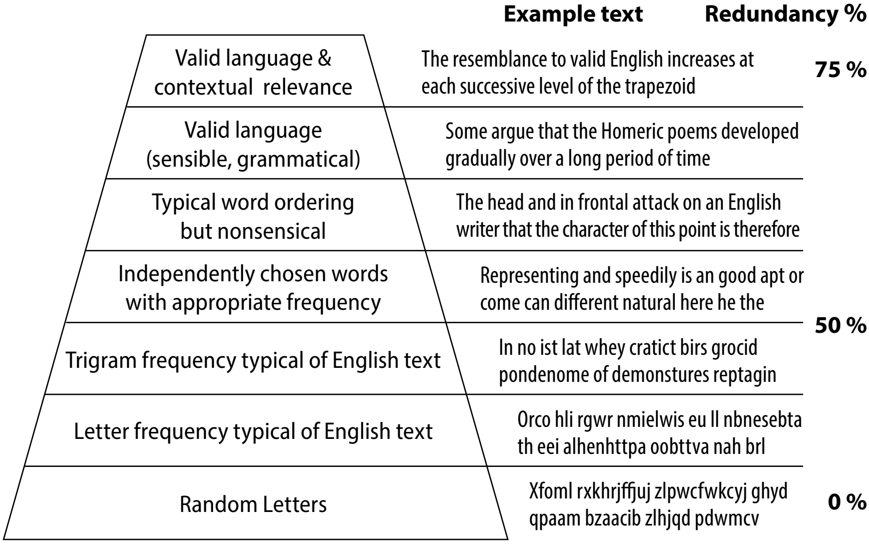
Example text
Redundancy %
Valid language & contextual relevance
The resemblance to valid English increases at each successive level of the trapezoid
75 %
Valid language (sensible, grammatical)
Some argue that the Homeric poems developed gradually over a long period of time
Typical word ordering but nonsensical
The head and in frontal attack on an English writer that the character of this point is therefore
Independently chosen words with appropriate frequency
Representing and speedily is an good apt or come can different natural here he the
50 %
Trigram frequency typical of English text
In no ist lat whey cratict birs grocid pondenome of demonstures reptagin
Letter frequency typical of English text
Orco hli rgwr nmielwis eu ll nbnesebta th eei alhenhttpa oobttva nah brl
Random Letters
Xfoml rxkhrjffjuj zlpwcfwkcyj ghyd qpaam bzaacib zlhjqd pdwmcv
0 %

Fig. E3.3 Shannon Information: Successive approximations to English

Shannon calculated the information content at various levels of the trapezoid. He found that for the top level the information component of a natural language is approximately 25%. The remaining 75% is called "redundancy." This means that a theoretical language that has perfect concision (i.e., fully compressed) could represent every possible message with only 25% of the information present in a natural language.[63] This ratio of 25% information content to 75% redundancy is a critical number because it allows one to answer the following essential question: what is the probability that a randomly derived string of characters is a valid text? In the previous examples, we used the value of 20,000, the number of 12-letter words, as a count of the number of possible messages. But that assumption is unwarranted for two reasons: it overestimates that count by including irrelevant and low probability words such as "wicketkeeper." On the other hand, it underestimates the count by failing to allow for 12-letter phrases composed of multiple words. Instead, we should use a count based on Shannon's estimate of the information content of English, which is 25%. Because "Shannon information" is measured logarithmically, 25% represents a small fraction of randomly generated texts. If we were to start with a trillion or 10^{12} possible texts at the lowest level of the trapezoid (random letters), and then apply 25% to the exponent, 12, the result is 3. Then, at the seventh level of the Trapezoid, we would have 10^3 or 1,000 relevant English statements. In this example, only one out of a billion (one thousand / one trillion) sequences of random letters qualify at the seventh level as relevant English messages.

In the previous discussion of Fig. E3.2, the ciphertext BQXMQLE-PYLDT was deciphered to CRYPTOGRAPHY using the key 1324. In the validation of that deciphering, the number of possible plaintext messages was taken to be 20,000, the number of 12-letter English words. That standard was flawed for two reasons: first, most of the 20,000 words are improbable (e.g., zymotechnics) and information should be measured in "equiprobable" events; second, the number of possible plaintext messages should include not only 12-letter words but multiword messages. The appropriate number to use when calculating the range of possible plaintext messages is a value based on Shannon information. Let us apply Shannon's 25% number for information content to our 12-letter plaintext message. The result is that the number of relevant English texts is equal to $24^{(25\% \text{ of } 12)} = 24^3 = 13{,}824$. As it happens, this result is not that dissimilar from the earlier value of 20,000. The reason for this modest difference is that the two countervailing factors nearly cancel each other out: the value of 20,000 is too high because it is not an equiprobable wordcount; the value of 20,000 is too low because messages with multiple words were not counted. Shannon information is an essential tool because it allows us to estimate the number of coherent and relevant messages that may be contained in a message of a given length.

In the cryptanalytic example of Fig. E3.2, a codebreaker (an unauthorized decipherer) had to guess at the method (a Caesar shift every three characters) and the key (1324)—arbitrary assumptions. Yet, the range of the key (10,000) is small compared to the coherence achieved in arriving at the valid 12-letter plaintext message, CRYPTOGRAPHY. The probability of a false result, after factoring in the range of the key, was calculated to be extremely small: 1 in 183 million (performed with the earlier estimate of 20,000 valid plaintexts). Thus, in the course of solving a cryptographic puzzle, one should tally the arbitrary assumptions that are made. Then, at the conclusion of the puzzle, when the solution is in hand, its validity can be measured by comparing the order or coherence of the solution to the range of the arbitrary assumptions made in arriving at the solution. In the case of the plaintext message CRYPTOGRPAHY, there is a high level of order because there are only 20,000 valid plaintexts as compared to 36,520 trillion random sequences of 12 letters, a ratio of one in 1.8 trillion. Now this must be compared to the range of arbitrary assumptions made in arriving at the solution. That range is the number of key values that may be considered, which is the number of four-digit PINs, equal to 10,000. This number is small compared to 1.8 trillion, and this assures the cryptanalyst that the deciphered message, CRYPTOGRAPHY, is valid.

In the process of cryptanalysis, the codebreaker makes many speculations, some based on a hunch and others quite arbitrary. As in navigating a labyrinth, the validity of choices is not known when they are made, but only after validation is confirmed at the end of the process. In the course of using cryptanalysis to decipher a presumptive plaintext message, the range (or information content) of the arbitrary decisions undertaken may be tallied and later compared against the likelihood of a valid message arising out of chance. With too much range in the key (as in perfect secrecy) or other arbitrary assumptions, no validation is possible. If the information content or range of the key is reasonably low, then the deciphering is valid.

60 See Borris, *Visionary Spenser and the Poetics of Early Modern Platonism* (Oxford: Oxford University Press, 2017), 148. Borris references Fraunce's *Countesse of Pembrokes Yuychurch. Part* 3 (1592), STC 11341, 3v-4r, and Harrington's introduction to *Orlando furioso in English Heroical Verse* (1591), STC 746, 4r-v.

61 In an equiprobable alphabet, all letters appear with the same frequency. For the purposes of this basic introduction to Information Theory, we will sometimes overlook the defined equiprobability of information in our examples to avoid complexity.

62 Shannon developed the theoretical framework under which a cryptogram may be validated in his "Communication Theory of Secrecy Systems," *Bell System Technical Journal*, 28 (1949): 656–715. Shannon's seminal work in the field of Information Theory can be found in "A Mathematical Theory of Communication," *Bell System Technical Journal*, 27 (July and October 1948): 379–423 and 623–56. This is reprinted in *The Mathematical Theory of Communication* (Urbana: University of Illinois Press, 1964), along with a helpful introduction by Warren Weaver.

63 See Claude Shannon, "Prediction and Entropy of Printed English," *Bell System Technical Journal* 30.1 (1951): 50-64.

64 *Rhetoric, Hermeneutics and Translation in the Middle Ages* (Cambridge: Cambridge University Press, 1991), 3.

65 Ibid., 112.

66 Ibid., 70.

67 *Andreas Capellanus, Scholasticism, and the Courtly Tradition* (Washington, DC: Catholic University of America Press, 2005), 88.

68 M. B. Parkes, "The Influence of the Concepts of *Ordinatio* and *Compilatio* on the Development of the Book" in *Medieval Learning and Literature: Essays Presented to Richard William Hunt*. Ed. J. J. G. Alexander and Margaret Gibson (Oxford: Clarendon Press, 1976), 116N1.

69 *The Book of Memory: A Study of Memory in Medieval Culture*, 2nd ed. (Cambridge: Cambridge University Press, 2008), 222.

70 *Rhetoric, Hermeneutics and Translation in the Middle Ages*, 203.

71 Ibid., 202.

72 Ibid., 203.

73 *Philip Sidney and the Poetics of Renaissance Cosmopolitanism* (London: Routledge, 2008), ix.

74 Ibid., 359–60. See also James Coulter, *The Literary Microcosm: Theories of Interpretation of the Later Neoplatonists* (Leiden: Brill, 1976), 95–126.

75 Kathy Eden, *Hermeneutics and the Rhetorical Tradition: Chapters in the Ancient Legacy and its Humanist Reception* (New Haven: Yale University Press, 2009), 71.

76 Ibid., 57.

77 *De doctrina christiana*, 3.9.13; 3.10.14. Tr. Rev. J. F. Shaw.

78 *Hermeneutics and the Rhetorical Tradition*, 62–63.

79 *De doctrina christiana*, 3.10.15–16.

80 *Rhetoric, Hermeneutics and Translation in the Middle Ages* (Cambridge: Cambridge University Press, 1991), 44.

81 *Hermeneutics and the Rhetorical Tradition*, 39.

82 *Rhetoric, Hermeneutics and Translation in the Middle Ages*, 61.