Counting letters

Description

default watermark

A substitution cipher is one of the simplest methods for encrypting a message. A unique symbol stands in place of every letter in the hidden message. The symbol set can be just about anything, as long as each symbol maps uniquely to the letters of whatever alphabet you are using for the message you wish to hide.

The usual encryption method is to deploy a different letter from the same alphabet, so that the encrypted message contains the same character set as the hidden message, e.g. the 26 letters of the English alphabet A to Z plus a space.

Letter frequency as decoding method

As known to any Scrabble player, letters in a block of text occur at different frequencies. For example, E is usually more common than Q. Letter frequency provides a clue to decrypting a coded message.

If the frequency of the letter E is 12.1% then that constitutes a probability, i.e. the probability that the letter E will show up in any position in a coded string is 0.121. James Lyons provides a helpful blog post with the frequencies of letters and in various combinations.

As text messages are sequences of characters, it would also be useful to know the probability that any letter will be followed by any other letter. How often is an E followed by another E, or an S or D, etc?

According to Lyonâ??s frequency data, In any block of text you can expect the letters EE to occur in 3.54% of all adjacent pairs of letters, ES occurs 1.32% of the time and ED 1.08% of the time. I calculated these percentages from Lyonsâ?? letter frequency data.

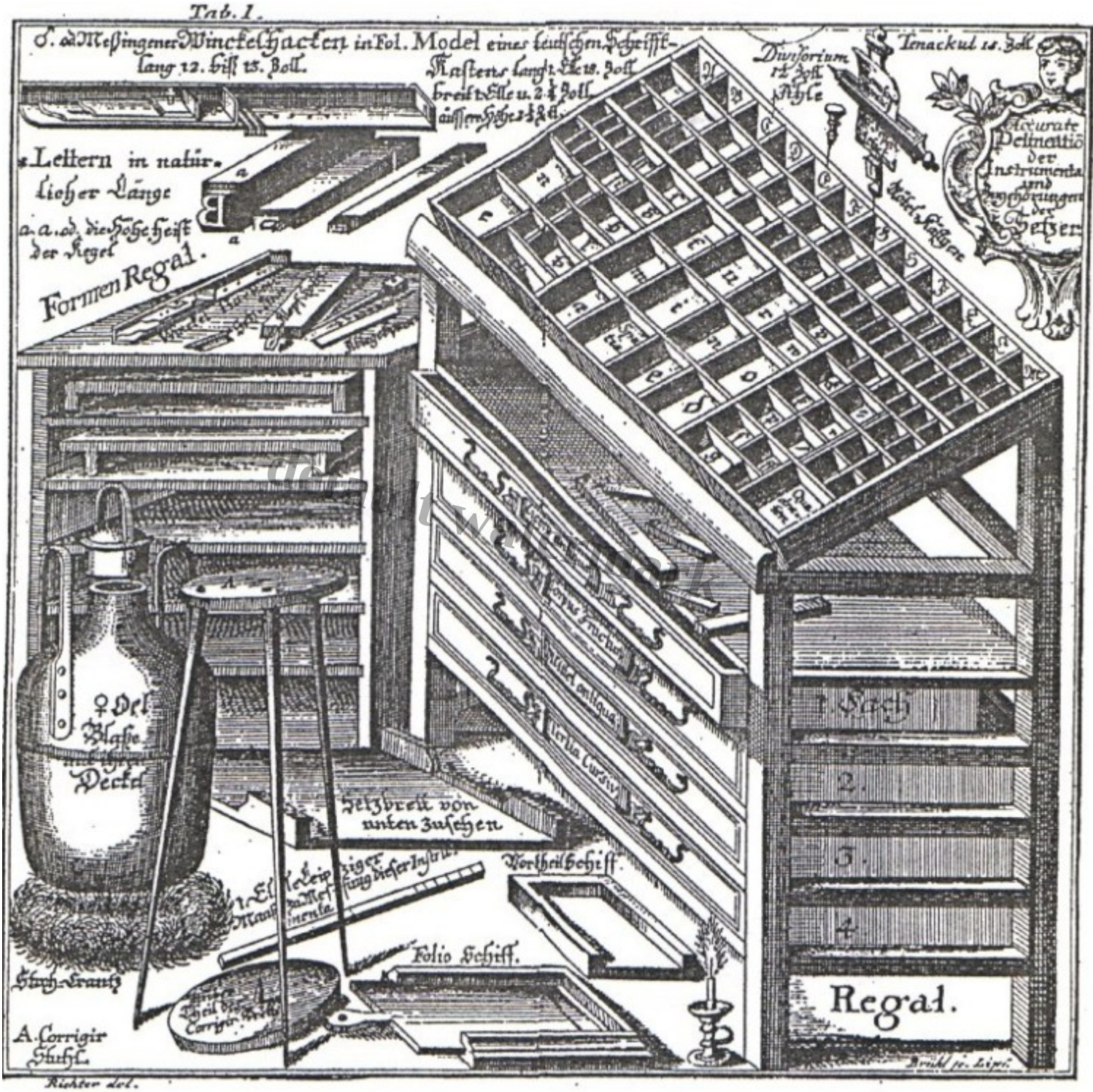
Unfortunately that data doesnâ??t give the frequency with which, given an E, the next letter will also be an E, an S or a D, etc. Itâ??s beyond me at the moment, but that can be calculated (via Bayesâ?? formula) to produce probabilities that populate a transition matrix as in my previous posts, i.e. a table showing all the letters of the alphabet plus the space character, and the probabilities that any letter will be followed by any other, including itself. That information can also be understood as a transition network with 27 nodes and a tangle of connecting arrows connecting each node â?? and with probabilities attached.

When known, this information about the probability that one particular letter will be followed by another particular letter provides the ingredients for a Hidden Markov Model (HMM) formulation of the problem of automatically deciphering a substitution cipher. In HMM terms, the hidden part is the path through a network of 27 letters (plus space) that make up the order of letters in the hidden message. The observation part of the HMM formulation is the encrypted version of the message. Thereâ??s obviously more to be said about how that knowledge helps automate the decoding of a substitution cipher.

default watermark

More frequently used letters

The variation in letter frequency was important in moveable type printing. You needed more Es than Qs. Hereâ??s a public domain image of a â??type caseâ?• that spatialises letter frequency as something that looks almost like architecture.



The image is captioned: â??By Christian Friedrich Gessner â?? Illustration taken from a scan of: Christian Friedrich Gessner, â??Die so nÃ¶thig als nÃ¼tzliche Buchdruckerkunst und Schriftgiesserey : mit ihren Schriften, Formaten und allen dazu gehÃ¶rigen Instrumenten abgebildet auch klÃ¤rlich beschrieben, und nebst einer kurzgefassten ErzÃ¤hlung vom Ursprung und Fortgang der Buchdruckerkunstâ?? 1740, p. 226f, Public Domain, https://commons.wikimedia.org/w/index.php?curid=13538088â??

Calculating the letter frequency in a block of text provides a way of making a good guess at the language itâ??s written in â?? automatically. An interesting website at letterfrequency.org provides the frequency order of letters in a range of languages. For example, the frequency order in English starts â??ETAOIN â?¦â?•. In German it starts â??ENISRAT â?¦â?•. In French itâ??s â??ESAITN â?¦â?•, which coincidentally look like they could be words.

In the 1950s, the pioneering information scientist Herbert Ohlman ran calculations on sets of texts to determine the relative frequencies of letters. He calculated the frequency of letters in different parts of words: the first, second, third, fourth and fifth positions. If I need any further support for the importance of cryptography, Ohlman also asserted:

> â??Coding, or the transforming of information from one guise to another, is one of manâ??s commonest activities. Every picture may be said to be a coding of some real scene and every written word a coding of some utteranceâ??the brain itself is said to work with coded impulsesâ?• (903).

Bibliography

default watermark

- ETA. 2020. Letter frequencies. Available online: http://letterfrequency.org/ (accessed 24 October 2020).
- Lyons, James. 2012. English letter frequencies. *Practical Cryptography*. Available online: http://practicalcryptography.com/cryptanalysis/letter-frequencies-various-languages/english-letter-frequencies/#comment-2753131539 (accessed 23 October 2020).
- Ohlman, Herbert. 1959. Subject-Word Letter Frequencies with Applications to Superimposed Coding. *Proceedings of the International Conference on Scientific Information: Two Volumes* 903-916. Washington DC: National Academies Press.
- Vobbilisetty, Rohit, Fabio Di Troia, Richard M. Low, Corrado Aaron Visaggio, and Mark Stamp. 2017. Classic cryptanalysis using hidden Markov models. *Cryptologia*, (41) 1, 1-28.

Category

1. Voice and text

Tags

1. cryptography
2. HMM
3. Markov process
4. printing
5. text

Date Created
October 24, 2020
Author
rcoyne99