Letter frequency

Letter	Relative frequency in the English language[1]	
	Texts	Dictionaries
A	8.2%	7.8%
B	1.5%	2.0%
C	2.8%	4.0%
D	4.3%	3.8%
E	12.7%	11.0%
F	2.2%	1.4%
G	2.0%	3.0%
H	6.1%	2.3%
I	7.0%	8.6%
J	0.15%	0.21%
K	0.77%	0.97%
L	4.0%	5.3%
M	2.4%	2.7%
N	6.7%	7.2%
O	7.5%	6.1%
P	1.9%	2.8%
Q	0.095%	0.19%
R	6.0%	7.3%
S	6.3%	8.7%
T	9.1%	6.7%

U	2.8%		3.3%	
V	0.98%		1.0%	
W	2.4%		0.91%	
X	0.15%		0.27%	
Y	2.0%		1.6%	
Z	0.074%		0.44%	

Letter frequency is the number of times letters of the alphabet appear on average in written language. Letter frequency analysis dates back to the Arab mathematician Al-Kindi (c. 801–873 AD), who formally developed the method to break ciphers. Letter frequency analysis gained importance in Europe with the development of movable type in 1450 AD, where one must estimate the amount of type required for each letterform. Linguists use letter frequency analysis as a rudimentary technique for language identification, where it is particularly effective as an indication of whether an unknown writing system is alphabetic, syllabic, or ideographic.

The use of letter frequencies and frequency analysis plays a fundamental role in cryptograms and several word puzzle games, including Hangman, *Scrabble*, *Wordle*[2] and the television game show *Wheel of Fortune*. One of the earliest descriptions in classical literature of applying the knowledge of English letter frequency to solving a cryptogram is found in Edgar Allan Poe's famous story *The Gold-Bug*, where the

method is successfully applied to decipher a message giving the location of a treasure hidden by Captain Kidd.[3][*citation needed*]

Herbert S. Zim, in his classic introductory cryptography text "Codes and Secret Writing", gives the English letter frequency sequence as "**ETAON RISHD LFCMU GYPWB VKJXZQ**", the most common letter pairs as "TH HE AN RE ER IN ON AT ND ST ES EN OF TE ED OR TI HI AS TO", and the most common doubled letters as "LL EE SS OO TT FF RR NN PP CC".[4] Different ways of counting can produce somewhat different orders.

Letter frequencies also have a strong effect on the design of some keyboard layouts. The most frequent letters are placed on the home row of the Blickensderfer typewriter, the Dvorak keyboard layout, Colemak and other optimized layouts.

Background

The frequency of letters in text has been studied for use in cryptanalysis, and frequency analysis in particular, dating back to the Arab mathematician al-Kindi (c. 801–873 AD), who formally developed the method (the ciphers breakable by this technique go back at least to the Caesar cipher invented by Julius Caesar[*citation needed*], so this method could have been explored in classical times). Letter

frequency analysis gained additional importance in Europe with the development of movable type in 1450 AD, where one must estimate the amount of type required for each letterform, as evidenced by the variations in letter compartment size in typographer's type cases.

No exact letter frequency distribution underlies a given language, since all writers write slightly differently. However, most languages have a characteristic distribution which is strongly apparent in longer texts. Even language changes as extreme as from Old English to modern English (regarded as mutually unintelligible) show strong trends in related letter frequencies: over a small sample of Biblical passages, from most frequent to least frequent, **enaid sorhm tgþlwu æcfy ðbpxz** of Old English compares to **eotha sinrd luymw fgcbp kvjqxz** of modern English, with the most extreme differences concerning letterforms not shared.[5]

Linotype machines for the English language assumed the letter order, from most to least common, to be **etaoin shrdlu cmfwyp vbgkjq xz** based on the experience and custom of manual compositors. The equivalent for the French language was **elaoin sdrétu cmfhyp vbgwqj xz**.

Arranging the alphabet in Morse into groups of letters that require equal amounts of time to transmit, and then sorting these groups in increasing order, yields **e it san hurdm wgvlfbk opxcz jyq**.[a] Letter frequency was used by other

telegraph systems, such as the Murray Code.

Similar ideas are used in modern data-compression techniques such as Huffman coding.

Letter frequencies, like word frequencies, tend to vary, both by writer and by subject. For instance, ⟨d⟩ occurs with greater frequency in fiction, as most fiction is written in past tense and thus most verbs will end in the inflectional suffix -ed / -d. One cannot write an essay about x-rays without using ⟨x⟩ frequently. Different authors have habits which can be reflected in their use of letters. Hemingway's writing style, for example, is visibly different from Faulkner's. Letter, bigram, trigram, word frequencies, word length, and sentence length can be calculated for specific authors, and used to prove or disprove authorship of texts, even for authors whose styles are not so divergent.

Accurate average letter frequencies can only be gleaned by analyzing a large amount of representative text. With the availability of modern computing and collections of large text corpora, such calculations are easily made. Examples can be drawn from a variety of sources (press reporting, religious texts, scientific texts and general fiction) and there are differences especially for general fiction with the position of ⟨h⟩ and ⟨i⟩, with ⟨h⟩ becoming more common.

Also, to note that different dialects of a language will also

affect a letter's frequency. For example, an author in the United States would produce something in which ⟨z⟩ is more common than an author in the United Kingdom writing on the same topic: words like "analyze", "apologize", and "recognize" contain the letter in American English, whereas the same words are spelled "analyse", "apologise", and "recognise" in British English. This would highly affect the frequency of the letter ⟨z⟩ as it is a rarely used letter by British speakers in the English language.[6]

The "top twelve" letters constitute about 80% of the total usage. The "top eight" letters constitute about 65% of the total usage. Letter frequency as a function of rank can be fitted well by several rank functions, with the two-parameter Cocho/Beta rank function being the best.[7] Another rank function with no adjustable free parameter also fits the letter frequency distribution reasonably well[8] (the same function has been used to fit the amino acid frequency in protein sequences.[9]) A spy using the VIC cipher or some other cipher based on a straddling checkerboard typically uses a mnemonic such as "a sin to err" (dropping the second "r") [10][11] or "at one sir"[12] to remember the top eight characters.

Relative frequencies of letters in the English language

There are three ways to count letter frequency that result in

very different charts for common letters. The first method, used in the chart below, is to count letter frequency in root words of a dictionary. The second is to include all word variants when counting, such as "abstracts", "abstracted" and "abstracting" and not just the root word of "abstract". This system results in letters like ⟨s⟩ appearing much more frequently, such as when counting letters from lists of the most used English words on the Internet. A final variant is to count letters based on their frequency of use in actual texts, resulting in certain letter combinations like ⟨th⟩ becoming more common due to the frequent use of common words like "the", "then", "both", "this", etc. Absolute usage frequency measures like this are used when creating keyboard layouts or letter frequencies in old fashioned printing presses.

An analysis of entries in the Concise Oxford dictionary, ignoring frequency of word use, gives an order of "EARIOTNSLCUDPMHGBFYWKVXZJQ".[13]

The letter-frequency table below is taken from Pavel Mička's website, which cites Robert Lewand's *Cryptological Mathematics*.[14]

According to Lewand, arranged from most to least common in appearance, the letters are: **etaoinshrdlcumwfgypbvkjxqz**. Lewand's ordering differs slightly from others, such as Cornell University Math

Explorer's Project, which produced a table after measuring 40,000 words.[15]

In English, the space character occurs almost twice as frequently as the top letter (⟨e⟩)[16] and the non-alphabetic characters (digits, punctuation, etc.) collectively occupy the fourth position (having already included the space) between ⟨t⟩ and ⟨a⟩.[17]

Relative frequencies of the first letters of a word in English language

Letter	Relative frequency as the first letter of an English word[*citation needed*]			
	Texts		Dictionaries	
A	11.7%		5.7%	
B	4.4%		6%	
C	5.2%		9.4%	
D	3.2%		6.1%	
E	2.8%		3.9%	
F	4%		4.1%	
G	1.6%		3.3%	
H	4.2%		3.7%	
I	7.3%		3.9%	
J	0.51%		1.1%	
K	0.86%		1%	
L	2.4%		3.1%	

M	3.8%		5.6%	
N	2.3%		2.2%	
O	7.6%		2.5%	
P	4.3%		7.7%	
Q	0.22%		0.49%	
R	2.8%		6%	
S	6.7%		11%	
T	16%		5%	
U	1.2%		2.9%	
V	0.82%		1.5%	
W	5.5%		2.7%	
X	0.045%		0.05%	
Y	0.76%		0.36%	
Z	0.045%		0.24%	

The frequency of the first letters of words or names is helpful in pre-assigning space in physical files and indexes.[18] Given 26 filing cabinet drawers, rather than a 1:1 assignment of one drawer to one letter of the alphabet, it is often useful to use a more equal-frequency-letter code by assigning several low-frequency letters to the same drawer (often one drawer is labeled VWXYZ), and to split up the most-frequent initial letters (⟨s, a, c⟩) into several drawers (often 6 drawers Aa-An, Ao-Az, Ca-Cj, Ck-Cz, Sa-Si, Sj-Sz). The same system is used in some multi-volume works such

as some encyclopedias. Cutter numbers, another mapping of names to a more equal-frequency code, are used in some libraries.

Both the overall letter distribution and the word-initial letter distribution approximately match the Zipf distribution and even more closely match the Yule distribution.[19]

Often the frequency distribution of the first digit in each datum is significantly different from the overall frequency of all the digits in a set of numeric data, an observation known as Benford's law.

An analysis by Peter Norvig on words that appear 100,000 times or more in Google Books data transcribed using optical character recognition (OCR) determined the frequency of first letters of English words, among other things.[20]

Relative frequencies of letters in other languages

> This section **may contain citations that do not verify the text**. Please check for citation inaccuracies. *(July 2014)* *(Learn how and when to remove this template message)*

Letter	English [*citation needed*]	French[21]	German[22]	Spanish[23]	Por
a	8.167%	7.636%	6.516%	11.525%	14.6

b	1.492%	0.901%	1.886%	2.215%	1.04
c	2.782%	3.260%	2.732%	4.019%	3.88
d	4.253%	3.669%	5.076%	5.010%	4.99
e	12.702%	14.715%	16.396%	12.181%	12.5
f	2.228%	1.066%	1.656%	0.692%	1.02
g	2.015%	0.866%	3.009%	1.768%	1.30
h	6.094%	0.937%	4.577%	1.973%	1.28
i	6.966%	7.529%	6.550%	6.247%	6.18
j	0.253%	0.813%	0.268%	2.493%	0.87
k	1.772%	0.074%	1.417%	0.026%	0.01
l	4.025%	5.456%	3.437%	4.967%	2.77
m	2.406%	2.968%	2.534%	3.157%	4.73
n	6.749%	7.095%	9.776%	6.712%	4.44
o	7.507%	5.796%	2.594%	8.683%	9.73
p	1.929%	2.521%	0.670%	2.510%	2.52
q	0.095%	1.362%	0.018%	0.877%	1.20
r	5.987%	6.693%	7.003%	6.871%	6.53
s	6.327%	7.948%	7.270%	7.977%	6.80
t	9.056%	7.244%	6.154%	4.632%	4.33
u	2.758%	6.311%	4.166%	3.927%	3.63
v	0.978%	1.838%	0.846%	1.138%	1.57
w	2.360%	0.049%	1.921%	0.027%	0.03
x	0.250%	0.427%	0.034%	0.515%	0.45
y	1.974%	0.708%	0.039%	1.433%	0.00
z	0.074%	0.326%	1.134%	0.467%	0.47

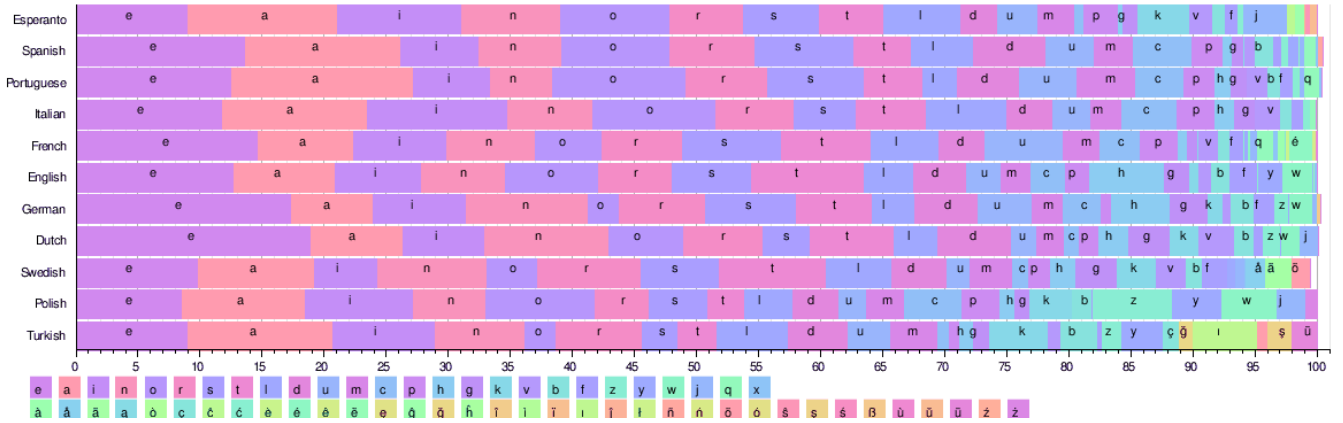
à	~0% [citation needed]	0.486%	0	~0%	0.07
â	~0%	0.051%	0	0	0.56
á	~0%	0	0	0.502%	0.11
å	~0%	0	0	0	0
ä	~0%	0	0.578%	0	0
ã	0	0	0	0	0.73
ą	0	0	0	0	0
æ	~0% [citation needed]	0	0	0	0
œ	~0%	0.018%	0	0	0
ç	~0%	0.085%	0	~0%	0.53
ĉ	0	0	0	0	0
ć	0	0	0	0	0
č	~0%	0	0	0	0
ď	0	0	0	0	0
ð	0	0	0	0	0
è	~0% [citation needed]	0.271%	0	~0%	0
é	~0% [citation needed]	1.504%	0	0.433%	0.33
ê	0	0.218%	0	0	0.45
ë	~0% [citation needed]	0.008%	0	0	0
ę	0	0	0	0	0
ě	0	0	0	0	0

ĝ	0	0	0	0	0
ğ	0	0	0	0	0
ĥ	0	0	0	0	0
î	0	0.045%	0	0	0
ì	0	0	0	0	0
í	0[citation needed]	0	0	0.725%	0.13
ï	~0% [citation needed]	0.005%	0	0	0
ı	0	0	0	0	0
ĵ	0	0	0	0	0
ł	0	0	0	0	0
ľ	0	0	0	0	0
ñ	~0% [citation needed]	0	0	0.311%	0
ń	0	0	0	0	0
ň	0	0	0	0	0
ò	0	0	0	0	0
ö	~0%	0	0.443%	0	0
ô	~0%	0.023%	0	0	0.63
ó	0[citation needed]	0	0	0.827%	0.29
ő	0	0	0	0	0
õ	0[citation needed]	0	0	0	0.04
ø	~0%	0	0	0	0
ř	0	0	0	0	0

ŝ	0	0	0	0	0
ş	0	0	0	0	0
ś	0[citation needed]	0	0	0	0
š	0	0	0	0	0
ß	0	0	0.307%	0	0
ť	0	0	0	0	0
þ	0	0	0	0	0
ù	0	0.058%	0	0	0
ú	0[citation needed]	0	0	0.168%	0.2(
û	~0%	0.060%	0	0	0
ŭ	0	0	0	0	0
ü	~0%	0	0.995%	0.012%	0.02
ű	0	0	0	0	0
ů	0	0	0	0	0
ý	0	0	0	~0%	0
ź	0	0	0	0	0
ż	0	0	0	0	0
ž	0	0	0	0	0

*See İ and dotless I.

The figure below illustrates the frequency distributions of the 26 most common Latin letters across some languages. All of these languages use a similar 25+ character alphabet.



Based on these tables, the '[etaoin shrdlu](#)'-equivalent for each language is as follows:

- French: 'esaitn ruoldc'; (Indo-European: Italic; traditionally, 'esartinulop' is used, in part for its ease of pronunciation[34])
- Spanish: 'eaosrn idltcm'; (Indo-European: Italic)
- Portuguese: 'aeosri dmntcu' (Indo-European: Italic)
- Italian: 'eaionl rtscdu'; (Indo-European: Italic)
- German: 'ensria tdhulg'; (Indo-European: Germanic)
- Swedish: 'eanrts ildomk'; (Indo-European: Germanic)
- Turkish: 'aeinrl ıdkmyt'; (Turkic)
- Dutch: 'enatir odslgv'; (Indo-European: Germanic)[29]
- Polish: 'aioezn rwstcy'; (Indo-European: Balto-Slavic)
- Danish: 'erntai dslogk'; (Indo-European: Germanic)
- Icelandic: 'arnies tulðgm'; (Indo-European: Germanic)
- Finnish: 'aintes loukäm'; (Uralic: Finnic)
- Czech: 'aeonit vsrldk'; (Indo-European: Balto-Slavic)
- Hungarian: 'eatlsn kizroá'; (Uralic: Finno-Ugric)

See also

- Arabic letter frequency
- Corpus linguistics
- Dvorak keyboard layout
- English word frequency
- Etaoin shrdlu
- Letter frequency effect
- Lipogram
- RSTLNE (*Wheel of Fortune*)

Explanatory notes

1. American Morse code was developed in the 1830s by Alfred Vail, based on English-language letter frequencies, to encode the most frequent letters with the shortest symbols. Some efficiency was lost in the reformed version now used: the International Morse Code.

References

1. *Mička, Pavel.* *"Letter frequency (English)"*. *Algoritmy.net.* *Archived* *from the original on 4 March 2021. Retrieved 14 June 2022.* *"Source is Leland, Robert. Cryptological mathematics. [s.l.] : The Mathematical Association of America, 2000. 199 p. ISBN 0-88385-719-7"*none
2. *Guinness, Harry.* *"The Best Starting Words to Win at Wordle"*. *Wired.* *ISSN* *1059-1028*. *Retrieved 2022-02-*

12.none
3. *Poe, Edgar Allan. "The works of Edgar Allan Poe in five volumes". Project Gutenberg*.none
4. *Zim, Herbert Spencer (1961). Codes & Secret Writing: Authorized Abridgement. Scholastic Book Services. OCLC 317853773*.none
5. *Moreno, Marsha Lynn (Spring 2005). "Frequency Analysis in Light of Language Innovation" (PDF). Math. University of California – San Diego. Retrieved 19 February 2015*.none
6. *"British and American spelling - Oxford Dictionaries". Oxford Dictionaries - English. Archived from the original on December 28, 2011. Retrieved 18 April 2018*.none
7. *Li, Wentian; Miramontes, Pedro (2011). "Fitting ranked English and Spanish letter frequency distribution in US and Mexican presidential speeches". Journal of Quantitative Linguistics. **18** (4): 359. arXiv:1103.2950. doi:10.1080/09296174.2011.608606. S2CID 1716455*.none
8. *Gusein-Zade, S.M. (1988). "Frequency distribution of letters in the Russian language". Probl. Peredachi Inf. **24** (4): 102–107*.none
9. *Gamow, George; Ycas, Martynas (1955). "Statistical correlation of protein and ribonucleic acid composition". Proc. Natl. Acad. Sci. **41** (12): 1011–1019. Bibcode:1955PNAS...41.1011G. doi:10.1073/pnas.41.12.1011. PMC 528190*.

PMID 16589789.none
10. Bauer, Friedrich L. (2006). Decrypted Secrets: Methods and maxims of cryptology. p. 57. ISBN 9783540481218 – via Google Books.none
11. Goebel, Greg (2009). The Rise Of Field Ciphers: straddling checkerboard ciphers.none
12. Rijmenants, Dirk. "One-time Pad".none
13. "What is the frequency of the letters of the alphabet in English?". Oxford Dictionary. Oxford University Press. Archived from the original on December 24, 2011. Retrieved 29 December 2012.none
14. Mička, Pavel. "Letter frequency (English)". Algoritmy.net.none
15. "English Letter Frequency (based on a sample of 40,000 words)". cornell.edu. Retrieved 2021-01-24.none
16. "Statistical Distributions of English Text". data-compression.com. Archived from the original on 2017-09-18.none
17. Lee, E. Stewart. "Essays about Computer Security" (PDF). University of Cambridge Computer Laboratory. p. 181.none
18. Ohlman, Herbert Marvin (1959). Subject-Word Letter Frequencies with Applications to Superimposed Coding. Proceedings of the International Conference on Scientific Information. doi:10.17226/10866. ISBN 978-0-309-57421-1.none

19. *Pande, Hemlata; Dhami, H.S. "Mathematical Modelling of Occurrence of Letters and Word's Initials in Texts of Hindi Language" (PDF). JTL. **16**.*none
20. *"English Letter Frequency Counts: Mayzner revisited or ETAOIN SRHLDCU". norvig.com. Retrieved 18 April 2018.*none
21. *"Corpus de Thomas Tempé". Archived from the original on 30 September 2007. Retrieved 15 June 2007.*none
22. *Beutelspacher, Albrecht (2005). Kryptologie (7 ed.). Wiesbaden: Vieweg. p. 10. ISBN 3-8348-0014-7.*none
23. *Pratt, Fletcher (1942). Secret and Urgent: The story of codes and ciphers. Garden City, NY: Blue Ribbon Books. pp. 254–5. OCLC 795065.*none
24. *"Frequência da ocorrência de letras no Português". Archived from the original on 3 August 2009. Retrieved 16 June 2009.*none
25. *Singh, Simon; Galli, Stefano (1999). Codici e Segreti (in Italian). Milano: Rizzoli. ISBN 978-8-817-86213-4. OCLC 535461359.*none
26. *Serengil, Sefik Ilkin; Akin, Murat (20–22 February 2011). Attacking Turkish Texts Encrypted by Homophonic Cipher (PDF). Proceedings of the 10th WSEAS International Conference on Electronics, Hardware, Wireless and Optical Communications. Cambridge, UK. pp. 123–126.*none
27. *"Practical Cryptography". Retrieved 30 October 2013.*none

28. "Frekwencja liter w polskich tekstach - Poradnia językowa PWN".none
29. ^ a b "Letterfrequenties". *Genootschap OnzeTaal*. *Retrieved 17 May 2009*.none
30. "Danish letter frequencies". *Practical Cryptography*. *Retrieved 24 October 2013*.none
31. "Icelandic letter frequencies". *Practical Cryptography*. *Retrieved 24 October 2013*.none
32. "Finnish letter frequencies". *Practical Cryptography*. *Retrieved 24 October 2013*.none
33. "Hungarian character frequencies". *Wolfram Alpha Site*. *Retrieved March 25, 2023*.none
34. Perec, Georges; *Alphabets*; Éditions Galilée, 1976

External links

- *Lewand, Robert Edward*. "Cryptographical Mathematics". *pages.central.edu*. *Archived from the original on 2007-04-02*.none
- "Some examples of letter frequency rankings in some common languages". *www.bckelk.org.uk*.none
- "JavaScript Heatmap Visualization showing letter frequencies of texts on different keyboard layouts". *www.patrick-wied.at*.none
- *Norvig, Peter*. "An updated version of Mayzner's work using Google books Ngrams data set". *norvig.com*.none
- Letter frequency—simia.net

Useful tables

Useful tables for single letter, digram, trigram, tetragram, and pentagram frequencies based on 20,000 words that take into account word-length and letter-position combinations for words 3 to 7 letters in length:

- *Mayzner, M.S.; Tresselt, M.E.; Wolin, B.R. (1965). "Tables of single-letter and digram frequency counts for various word-length and letter-position combinations". Psychonomic Monograph Supplements. **1** (2): 13–32.* OCLC 639975358.none
- *Mayzner, M.S.; Tresselt, M.E.; Wolin, B.R. (1965). "Tables of trigram frequency counts for various word-length and letter-position combinations". Psychonomic Monograph Supplements. **1** (3): 33–78.*none
- *Mayzner, M.S.; Tresselt, M.E.; Wolin, B.R. (1965). "Tables of tetragram frequency counts for various word-length and letter-position combinations". Psychonomic Monograph Supplements. **1** (4): 79–143.*none
- *Mayzner, M.S.; Tresselt, M.E.; Wolin, B.R. (1965). "Tables of pentagram frequency counts for various word-length and letter-position combinations". Psychonomic Monograph Supplements. **1** (5): 144–190.*none