Hangman

What is the best word to play in Hangman?

In Hangman, there are two apparently contradictory strategies for picking a word: choose a word with only rare letters, or choose a word with only common letters. Both strategies have something to argue for them: if you choose only rare letters, your opponent will have to spend time asking for common letters. On the other hand, if you choose common letters, your opponent will have to test for all the words that can contain these letters. Which strategy is better?

When presented with a difficult problem, one approach is to solve a simple version of the problem. In Hangman, you can reduce the number of letters in the word and simplify the game. Suppose we start with the simplest possible game of Hangman: one-letter words.

Contrary to popular belief, there are 26 one-letters words in English. The word consisting of a single letter is the name for that letter. Thus, for example, "z" is the word for the letter z. Some of these words also have other meanings, but each one-letter word has at least this meaning.

The strategy for one-letter Hangman is not complicated. The frequency of letters makes no difference, since each word contains exactly one unique letter. The distribution of letters within a word makes no difference, since each letter occurs in the only available position in each word. The best that your opponent can do is get your word in one guess; the worst is 26 guesses; the average is 13.5 guesses. It makes no difference which one-letter words you pick.

Moving up to two-letter words, the situation is a bit more complex. There are about 330 two-letter words; that is too many to analyze simply, so reduce that number to a more manageable level, say, seven. Consider the best word to pick when playing Hangman with the lexicon: am, an, be, no, on, of, we.

This is the count of words containing a specific letter in a specific position in this list:

letter	missing	first position	second position
a	5	2	0
b	6	1	0
e	5	0	2
f	6	0	1
m	6	0	1
n	4	1	2
o	4	2	1
w	6	1	0

You may know that your opponent always guesses vowels, or has a limited vocabulary, or some other piece of specific knowledge, but for the sake of this analysis assume that your opponent knows this lexicon and will play optimally. This means that your opponent will attempt to get as much information as possible out of each guess. This is done by using the guess to split the lexicon into sets that are as nearly equal in size as possible. Further guesses will be used to subdivide each set, and subdivide again, and so on until there is only one word left in each set.

For example, suppose the first guess is the letter n. This subdivides the lexicon into these sets:

no n	n in the first position	n in the second position
am	no	an
be		on
of		
we		

The first of these sets (no n) requires two more guesses to subdivide down to one word. The second set (n in the first position) requires no more guesses. The third set (n in the second position) requires only one more guess. The maximum total number of guesses required is three.

Consider the case where the first guess is o instead of n; the corresponding sets are:

no o	o in the first position	o in the second position
am	of	no
an	on	
be		
we		

Again, the maximum total number of guesses required is three.

What happens if the first guess is the letter b? The sets are:

no b	b in the first position	b in the second position
am	be	
an		
no		
on		
we		

B is not a good choice for the first guess, because it requires three more guesses to break the no-b set down into single words, and thus it requires four guesses overall.

Now we have a way of deciding which letters are good first guesses, and which letters are not. Good guess letters are letters that divide the lexicon up into more sets of more equal sizes. Such divisions give our opponent more information. Can we reduce this idea to a formula? We can.

If a letter divides a set into two exactly equal subsets, we'll say this letter gives us one bit of information. If a letter divides a set into four equal subsets, that's the same as dividing it into two twice, so that should equal two bits of information. If a letter divides a set into eight equal subsets, that should be three bits, and so on. Is there a formula that, given the sizes of these subsets, computes the number of bits of information? Fortunately, there is a well-known formula from information theory that does just this. We won't go into the details of the formula here, but using it we get the following information associated with each letter:

letter	subsets sizes	information
a	5, 2, 0	.86
b	6, 1, 0	.59
e	5, 0, 2	.86
f	6, 0, 1	.59
m	6, 0, 1	.59
n	4, 1, 2	1.38
o	4, 2, 1	1.38
w	6, 1, 0	.59

From this table, you can deduce which letter your opponent will pick for the first guess. Once this letter is picked, your opponent will apply the same formula again to deduce the next best letter to pick, given the results of the first letter. For example, if the first guess is n and if your word contains no n, your opponent will have to decide between "am," "be," "of," and "we." For these words the information content of each letter is:

letter	subsets sizes	information
a	3, 1, 0	.81
b	3, 1, 0	.81
e	2, 0, 2	1.00
f	3, 0, 1	.81
m	3, 0, 1	.81
o	3, 1, 0	.81
w	3, 1, 0	.81

So e has the highest information content, and your opponent will pick it. Your opponent

will continue in this way, picking the most informative letters, until your word is uncovered.

Your best strategy is to pick words that resist this procedure until the bitter end. Which word out of the seven is the best to pick? The answer is: any word in the no-n, no-o set, namely, "am," "be," or "we." We see now why the two strategies mentioned at the start of the article are neither quite right. If you pick only rare letters, your opponent will quickly discover this. If you pick only common letters, your opponent will get them all in the first few guesses. The trick is to mix common letters with rare letters. Also, although we cannot see this from the simple examples we have looked at, you want to pick words that have common letters in common patterns, so that the distribution of common letters does not give away too much.

Moving now from our toy lexicon to the real English lexicon, if we use this algorithm to deduce the best words to pick in Hangman for each word length, we get these results:

2: ys, zs
3: wat, wes, xat, yes
4: yill, zill
5: yills, zills
6: coxing, cozing, wagged, zagged
7: tilling, willing
8: bummings, buzzings
9: quakering, quavering

One caveat: If your opponent knows that you are using this procedure to pick words, it is possible that in some cases it is no longer the best procedure and requires some modification. This becomes a problem in game theory. Specifically, if your opponent has read this article, don't use these words!

ex-servicewoman—hyphenated word containing most consecutive letters in the alphabet after letter v (3-t)—*former servicewoman*

extemporaneously—common word with the lowest ratio length to number of different letters for length 16 (1.23-t)

exterminations—see "determinations"

extractors—longest common double progressive beheadment [extractors → tractors → actors → tors → rs] (10-t)

extracurricular—common word containing the most repeats of letter r (4-t)

extraordinaire—longest common word that starts and ends with letter e (14)

extraterrestrial—common word containing the most repeats of letter r (4-t)

extraterrestrials—longest well-mixed common words with same letter sum [electroencephalography] (22)

extraterritorial—common word containing the most repeats of letter r (4-t)

extraterritoriality—common word containing the most repeats of letter r (4-t)

eyas—example of an uncommon word formed by borrowing or lending letters to frequent companions, beginning with letter e [a neyas → an eyas]—*baby hawks*

eye—longest common head 'n' tail centered on letter y (3)—longest common palindrome beginning with letter e (3-t)—shortest common word that starts and ends with letter e (3-t)

eyre—see "aer"—*court with a traveling judge*