MATHEMATICS OF WORD QUALITY

CHRIS LONG
Bridgewater, New Jersey

Theoretical Results

Over the past hundred years or so, somewhere around one thousand English nine-letter word squares have been discovered, but no legitimate ten-squares (although Jeff Grant has come close). Recently I have been considering forms in general from a probabilistic point of view (see "Mathematics of Square Construction" in the February 1993 **Word Ways**), and I started wondering about the worst words that have appeared as bases in these nine-squares, the best that haven't, and the best base word for a possible ten-square. Toward this goal, given a list of words (which may simply be "all English words") and a form type, we may define the **quality** of a word in a given position to be the product of the frequencies among acceptable words with which letters appear in the appropriate positions, this product taken over all non-redundant intersecting words. For example, letting f(*) equal the frequency with which nine-letter words end in the letter "*", the quality of the word STEELLESS as a base word in a nine-square would equal f(s)f(t)f(e)f(e)f(l)f(l)f(e)f(s). The final term is 1 (not f(s)), since by the definition of a word square, the only acceptable intersecting word in that position is STEELLESS, and thus the final S occurs with frequency 1. (However, the final term would be f(s) if the form were a double square.)

Defined in this manner, quality is approximately directly proportional to the expected number of fills. In other words, the higher the quality of a word, the larger the number of fills we expect to find using that word in that position. For unlikely events finding two or more fills should be much rarer than just finding one, so in this case the quality if also approximately directly proportional to the probability that a fill exists, e.g., if a given word has twice the quality of another word, then there is about twice the probability of a fill existing using the first word than one using the second.

Experimental Results

A program was written which uses a given list of words to first computer letter frequencies for each position in a word, and then to calculate word quality using these results. However, instead of using frequencies directly, it was found that calculations were simplified by using the natural logarithm (ln) of the number of times a given letter occurred in the appropriate position directly, as this allows the use of summations instead of multiplications and the resulting magnitudes of the numbers were also easier to work with. Therefore, when comparing the quality of the words given below, the numbers must first be exponentiated (the constant e raised to the power expressed by the quality number). Furthermore, if any letter did not occur in a particular position (e.g., it was found that no ten-letter word ended in a Q), it was counted as having occurred once to avoid words having a quality of zero (and thus a natural logarithm of negative infinity).

The first run was on a list of 82,794 nine-letter words (solid-form, hyphenated, and dictionary-sanctioned phrases) to find the lowest and highest quality words for each line n in a regular word square. It should be noted that this list lacks many derived forms (e.g., plurals), but I feel that the relative frequencies with which letters occur is still reflective of nine-letter dictionary words in general. None of the best words is a surprise, for all of them consist almost entirely of ADEINRST, the high-frequency letters. Most of the worst words were similarly unsurprising, as they contain many of the tough letters JQZ; the sole exception was the somewhat surprising SKEWWHIFF which was the worst word in position 7. The results are given in the table below.

n	Best	Quality	Worst	Quality
1	ass's steps	71.545	oxybenzyl	53.575
2	acerineae	73.594	jazzstick	47.736
3	recarrier	71.158	zulhijjah	55.077
4	reedenter	71.170	zulhijjah	57.598
5	treegeese	71.747	zulhijjah	55.956
6	triradial	71.391	equivoque	56.713
7	niaiserie	73.159	skewwhiff	55.668
8	ensentede	74.717	hobjobbed	48.401
9	seedsseed	76.862	equivoque	35.708

Another run was now done on the same list to find the highest-quality words for the base position. The resulting output was checked against Murray Pearce's list of words which have appeared as base words in known nine-squares (words below are marked with an asterisk if they haven't been used as bases before).

Word	Quality	Word	Quality	Word	Quality
seedsseed	76.862	ress's test	75.466	deessendo	75.083
nesessest	76.565	stressest	75.466	dessendes	75.083
sessenest	76.565	*setnesses	75.407	esegersee	75.031
seenessel	76.430	rees's test	75.331	seedaseer	74.942
seeresses	76.354	Essernsee	75.316	lesnesses	74.911
assessest	75.954	seernesse	75.316	*senseless	74.911
dyssessed	75.932	serenesse	75.316	reedseeds	74.872
seednesse	75.824	sernesses	75.316	lesserest	74.835
assessees	75.818	teresseen	75.196	sleeresse	74.835
edessenes	75.689	segessera	75.166	assesseth	74.795

Note that out of the thirty best words, only two have not been used as bases before, so it would appear that our definition of quality is a good one. The difference in quality among the best and worst of these 30 words is exp(76.862 - 74.795) = 7.90, a surprisingly large figure. As a quick side-excursion, the worst 10 words which have actually been used as bases are given below.

Word	Quality	Word	Quality	Word	Quality
hingangat	63.431	traderoom	64.935	garoengan	65.721
auger gage	63.498	sea league	65.605	ingestrie	65.844
Sena river	63.659	kissingen	65.622	interlard	65.974
asure card	64.008				

Nine-square seekers will be interested in the best 30 words which have not been used as bases for nine-squares.

Word	Quality	Word	Quality	Word	Quality
setnesses	74.507	Tennessee	74.233	negresses	73.992
senseless	74.911	tenseness	74.233	seertrees	73.947
denseness	74.650	*easelesse	74.164	regressed	73.917
desertest	74.590	eternesse	74.157	seemeless	73.884
detressed	74.590	synereses	74.114	slynesses	73.845
rednesses	74.575	*reedlesse	74.094	seedasere	73.828
redressed	74.499	tense-eyed	74.070	seedstems	73.774
sessement	74.499	sadnesses	74.039	tenseless	73.753
*nesshesst	74.452	seemlesse	74.019	lenenesse	73.737
sestettes	74.264	egrenesse	73.992	medresseh	73.683

I have recently discovered quite a few nine-squares, including the three words marked with asterisks. This suggests that with enough effort any of the above words can be used as the base in the construction of a nine-square.

The next runs were done on a list of 71,671 ten-letter words to find the lowest and highest quality words for each line n in a regular word square; the same caveats apply to this list as applied to the list of nine-letter words.

n	Best	Quality	Worst	Quality
1	assumpsits	77.639	fuzzy-wuzzy	55.734
2	adenoneure	80.939	fuzzy-guzzy	48.689
3	retrorenal	78.232	fuzzy-wuzzy	56.861
4	trebletree	78.746	fuzzy-wuzzy	58.131
5	rerehearse	79.037	fuzzy-wuzzy	56.938
6	araeometer	79.029	fuzzy-guzzy	55.588
7	taratantar	79.578	fuzzy-wuzzy	58.813
8	reinitiate	80.581	fuzzy-wuzzy	59.187
9	rennelesse	82.422	bubblyjock	52.481
10	seedsseeds	84.533	quinquevir	36.243

QUINQUEVIR is, of course, impossible until someone has discovered a ten-letter word ending in Q. As an aid to ten-square seekers, here are the best 30 words for the base position (see next page).The difference in quality between the best and worst words appearing on this list is exp(84.533 - 80.966) = 35.4, which is much higher than the 7.90 figure for the nine-letter words.

Word	Quality	Word	Quality	Word	Quality
seedsseeds	84.533	stressless	82.313	sanenesses	81.445
seednesses	83.918	lesserness	81.982	deadnesses	81.442
serenesses	83.371	senslesnes	81.913	greynesses	81.389
dresser set	83.028	redressers	81.864	slenderest	81.364
eyednesses	82.898	desertress	81.730	desertless	81.298
destressed	82.894	desertness	81.661	degreeless	81.222
reassessed	82.747	dereddened	81.517	meetnesses	81.135
restressed	82.348	searnesses	81.514	steel-edged	81.088
needelesse	82.322	addressees	81.511	sledgeless	80.997
sereneness	82.138	sagenesses	81.509	degendered	80.966

Further Research

There are several areas for futher exploration. It appears that word quality may be used to settle the argument of which order of word placement is best when searching for squares by computer. Preliminary results suggest that neither a pure top-down or bottom-up approach is best, but that for nine-squares it is best to first put in the ninth word, then the eighth, and then the second. Further work in this area should prove enlightening.

Word quality may also aid in the construction of squares directly. For the person attempting construction by hand, a list of the best words which have never appeared as bases before is a good place to start. For the person doing construction by computer, a good approach is to actively seek out high-quality words to add to your word stock. As an example, this is how I found NESSHESST and REEDLESSE, which resulted in new nine-squares. The REEDLESSE square in particular was a nice find, as it turned out that all words were in the Oxford English Dictionary. This is only the second single-dictionary nine-square known, the other being Eric Albert's Webster's Second square described in the November 1991 **Word Ways**. Those interested in this approach can readily determine the quality of a potential nine-letter or ten-letter base word by summing the following letter quality values for each letter of the word, excluding the final one.

	Nine	Ten		Nine	Ten		Nine	Ten
A	7.848	7.588	J	2.197	0.000	S	9.768	9.622
B	4.934	4.394	K	7.139	6.568	T	8.610	8.323
C	7.556	7.779	L	8.114	8.026	U	4.317	3.454
D	9.027	9.003	M	7.703	7.520	V	3.526	1.099
E	9.633	9.415	N	8.594	8.388	W	5.561	5.215
F	5.740	5.537	O	6.390	5.645	X	5.338	5.075
G	8.445	8.453	P	6.477	6.314	Y	8.567	8.602
H	7.520	7.055	Q	0.693	0.000	Z	4.205	2.565
I	6.138	5.236	R	8.519	8.457			