IS THAT LETTER-STRING REALLY A WORD?

A. ROSS ECKLER
Morristown, New Jersey

Most people would agree that SORNIE looks like a word, but DKYODZ does not. Can subjective judgements like these be codified? Can they be translated into explicit rules that rate the word plausibility of a letter-string? This article suggests a few rules, and speculates how they might be combined to form a super-rule for discriminating between plausible words and impossible combinations. For strings of any length, such rules are likely to be computationally burdensome, requiring a computer to implement them. I propose several rules for letter-strings of length three, and show how well they perform. I then take the best of these rules and evaluate it for letter-strings of length seven.

As a lagnaippe, I show how word plausibility can shed light on an intractable logological problem of long standing: the identification of the most transposable letter-string, currently believed to be AGEINRST (see *Making the Alphabet Dance*). In a pair of articles in this issue, Rex Gooch shows how variable the most fecund letter-string is, depending on the dictionary used.

Consider the 908 three-letter words recognized by the Official Scrabble Players Dictionary--a little under six per cent of the 17,576 possible letter-strings. Any rule worth considering must work better than this--that is, yield a higher fraction of words among strings exceeding a specific numerical value of plausibility.

Many rules can be imagined. I propose three which apparently look at different aspects of word-similitude. The first rule rates a letter-string plausible if its successive bigrams are commonly found in three-letter words, the second looks at vowel-consonant patterns, and the third notes the final letter of the string. Of these, the first seems to be the most powerful discriminant. For longer letter-strings, there is the possibility of generalizing this to successive trigrams, tetragrams, etc.-- to eliminate substrings like BRK for which BR and RK are plausible.

How well does the bigram rule work? On the next page is a table that shows the number of three-letter words containing each of the 676 possibilities (omitting three for AJ, two for IZ, and one each for AZ, DZ, EZ and OZ). Each letter-string is scored by adding the numbers of its two bigrams. Using this table, the highest-scoring ones are TAR 38, ARE ATA TAT TAY 36, ARA RAR TAG TAS TAW 35, and PAR WAR TAD TAN TAP 34. 10 of these (67 per cent) are words. What are the numbers for lower-scoring strings? For letter-strings scoring 20, 30 out of 111 (27 per cent) are words; for those scoring 2 (the lowest possible), 3 out of

BIGRAM TABLE FOR THREE-LETTER WORDS

	A	B	C	D	E	F	G	H	I	K	L	M	N	O	P	R	S	T	U	V	W	X	Y
A	4	12	6	16	8	3	17	8	7	3	11	14	16	1	16	20	17	18	5	5	17	9	18
B	11	1			8				8					12		1			10				3
C	10				3			2						13		1		1	7		1		1
D	9			2	11			1	8					12		1	3		9				1
E	9	5	3	10	16	5	8	1	4	2	12	7	14	1	4	11	7	16	3	4	11	7	8
F	8				9	3			10		2			10		2	2	3	5				
G	13				7		1	2	7				1	10			5						2
H	13				11			1	12			1		12			1	1	8				4
I	2	10	7	12	8	2	10				8	9	19	2	12	6	11	15	1	1		5	
J	6				4				3					6					4				
K	5				10			1	6					4					1				1
L	13	1		2	13	1			7	2	3	1		7	1		1	1	4				4
M	12				6			1	9			2		12	2		1		8				
N	6			2	6		1		7	1			1	10			4	2	5				1
O	9	13	3	14	9	2	10	6	2	3	7	7	13	11	14	13	7	15	6	1	18	8	7
P	14				12			2	10		1			9		2	3	2	9				4
Q																			1				
R	15	1	2	1	16	1	1	1	7	2		1	2	8		1	3	2	6				8
S	12				12			4	10	4	1	1		11	3	1	2	1	5				1
T	18				7			5	7					11		1	3		7		2		2
U	1	10	1	9	6		12	1	3	3	2	9	13	1	9	7	6	13				2	2
V	8				6				5					5					1				1
W	14				8			3	5		2	1	2	9		1			1				1
X					1				1														
Y	9				16				3			1	1	5	2		2		2			1	
Z	3				2				2					3									
	A	B	C	D	E	F	G	H	I	K	L	M	N	O	P	R	S	T	U	V	W	X	Y

205 (1.5 per cent) are words. (The three words in question are CWM, IVY, SYN.)

One can propose another function of the two bigram numbers, such as their product, to be designated the score. The "best" function is the one that maximizes the slope of the line relating score to word-probability.

The second rule evaluates letter-strings by their vowel-consonant patterns. There are eight of these (vvv,cvv,vcv,vvc,ccv,cvc,vcc,ccc), so the discrimination is necessarily coarse. If one counts an initial Y as a consonant but in other positions a vowel, the word counts are given in the table on the next page. The final columns give the possible number of letter-strings and the probability that a string forms a word; the table is arranged so that the most plausible patterns are at the top. For the record, the all-consonant words are CWM, PHT, NTH, SHH and TSK, and the all-vowel ones AYE, EAU and EYE. With a maximum probability of .221, this rule identifies plausible words less well than the bigram rule did.

cvc	556	2520	.221
cvv	131	756	.173
vcv	63	600	.105
vvc	29	600	.048
vcc	84	2000	.042
vvv	3	180	.017
ccv	37	2520	.015
ccc	5	8400	.0006

The third rule evaluates letter-strings by the final letter. The table below gives the relevant statistics, with the third column showing the probability that a string ending with the specified letter is a word. This is the weakest discriminator of all; even a string ending in E is only twice as likely as a random string to be a word.

E	84	.124	M	41	.061	U	16	.024
T	81	.120	W	38	.056	C	15	.022
N	66	.098	A	35	.052	K	14	.021
S	66	.098	R	34	.050	F	9	.013
D	59	.087	X	31	.046	V	5	.007
P	56	.083	L	29	.043	Z	5	.007
G	55	.081	O	28	.041	J	3	.004
Y	53	.078	H	19	.028	Q	0	
B	49	.072	I	17	.025			

To the extent that these various measures of plausibility are uncorrelated, one ought to be able to combine them into a single overall measure having greater predictive value. When plausibility measures are xpressed in numbers (or at least if they can be ranked in order of increasing effectiveness), there exists a body of mathematical knowledge called discriminant analysis that can be brought to bear on the problem. In brief, discriminant analysis creates a linear function of scores like

S = a(score from rule 1) + b(score from rule 2) + c(score from rule 3)

and mathematically determines those values of a,b,c that maximally separate the collection of S-values associated with words from those associated with non-words. Such analysis requires considerable computer power, and seems warranted only if there are important uses to which plausible letter-strings might be put.

What about letter-strings longer than three? A full analysis requires the assistance of a computer, but the following preliminary one is suggestive. Stanley Newman and Daniel Stark's 1996 book *The Crossword Answer Book* groups words of three through seven letters positionally by pairs of letters (for example, eScuDo, iSolDe and pSeuDo). It includes dictionary words, proper names, and a considerable number of non-dictionary combos like EKED OUT, ONLY YOU and A A MILNE. Using this book, one can determine relative frequencies of (say) the bigram ER in seven-letter words by summing the number of words in ER....., .ER....,

..ER..., ...ER.., ...ER. and .....ER. This has been done for 14 common letters in the table below.

BIGRAM TABLE FOR SEVEN-LETTER WORDS

	A	C	D	E	G	I	L	N	O	P	R	S	T	U
A		453	406	58	320	365	864	1227	15	309	1183	655	811	185
C	626			378		162	221	2	571		214	47	174	175
D	283			730	77	465	158	18	336	14	174	191	10	168
E	717	268	1746	449	129	142	654	938	121	203	2535	1946	629	91
G	244		9	335		231	160	79	225	5	219	121	24	157
I	275	460	307	842	250		599	2351	184	189	316	652	536	40
L	754	35	129	1458	24	750		5	571	28	16	284	148	243
N	385	232	492	769	1420	409	43		405	22	31	579	534	122
O	163	206	197	110	155	172	460	914		277	737	369	389	654
P	407		7	584	6	435	251	3	350		251	211	108	164
R	884	92	253	1395	114	815	110	169	681	59		994	276	253
S	335	210	16	760	2	468	145	82	388	238	6		1070	329
T	593	92	17	1245	8	721	178	24	601	16	383	607		242
U	135	154	137	178	133	135	316	514	22	395	465	446	496	
	A	C	D	E	G	I	L	N	O	P	R	S	T	U

Using this table, one can rate the plausibility of a letter-string by the sum of its entries. It requires the assistance of a computer to calculate the probability of a letter-string being a word, graphed as a function of its plausibility. However, a sample of the letter-strings with highest plausibilities include:

ALERING 9443	RESTING 8903	STINGER 8432
STERING 9436	TEERING 8815	TINGERS 8356
ATERING 9177	LINGERS 8569	ALINGER 8260
LEERING 9038	LERSING 8522	

* * *

The only application of letter-string plausibility I am aware of is one described elsewhere in this issue--the determination of the most-transposable collection of letters, as described in two articles by Rex Gooch. I suggest that one can select the most transposable set of letters by summing the scores of the letter-strings formed by the set (n! such strings if the letters are all different), and selecting that set of letters for which this is maximized.

The three-letter results are instructive. A full transposal set contains six different bigrams, each used twice. Summing the scores of the six bigrams, one finds that the largest totals occur for the following sets: AER 79, AET 76, ART 74, AYR 70, AST 69, AYE 68, AWE AIT 67. In fact, all six transposals of AER are words in Webster's Second (RAE is a Scots variant of ra or roe, besides being a feminine given name). AET is almost as good, with five words in Webster's Second and AET itself in

the Oxford English Dictionary as a variant of ete (obsolete variant of eat) or an obsolete form of at.

However, the seven-letter results are more useful, since it is expected that the largest number of transposals will occur for words of seven or eight letters. In a computer-generated listing of transposals of Webster's Second words together with derived forms (plurals, past tenses, etc.), the set AEINRST generated ten (asterin, eranist, ratines, restain, retains, retinas, sainter, stainer, starnie, stearin), followed by EOPRSTU, AELPRST, AEILRST and AEELRST with seven transposals each. AEINRST was in the highest or second-highest group for all five of Gooch's dictionary sources as well.

Summing the bigram scores--there are now 42 to consider instead of six--one discovers that AEINRST totals 29412, followed by EOPRSTU 21274, AELPRST 24363, AEILRST 28149 and AEELRST 22147. Checking other plausible combinations confirms the superiority of AEINRST--besides AEILRST, the closest rivals are ADEINRS 27064, AEILNRS 27640, AEILNRT 26712.

One can use the bigram statistics for seven-letter words to evaluate the relative transposabilities of eight-letter sets. Here there is a surprise. Ever since Dmitri Borgmann, it has been believed that AGEINRST is the best eight-letter set; adding the G enables one to consider words ending -ING. But when one checks the bigrams, AEGINRST 32908 is bested by ADEINRST 34520 and AEINORST 34662. The -ED ending appears more useful than the -ING one. In general, vowel-consonant and consonant-vowel bigrams are commoner than vowel-vowel or consonant-consonant ones, dictating the addition of the fourth vowel, O.

AEINORST, the apparently-best set according to the bigram analysis, has six transposals in the aforementioned computer-generated Webster's Second listing, but so do seven other sets; three sets, in fact, have seven transposals. AEINORST appears only in Rex Gooch's Webster and Pulliam sources. ADEINRST has only four transposals in the computer listing, and appears nowhere in the Gooch sources. On the other hand, AGEINRST has five transposals in the computer listing, and is the undisputed champion in three of Gooch's sources (Official Scrabble Players Dictionary. Oxford English Dictionary, Stedman). The Websterian transposals are listed below.

AEINORST	AGEINRST	ADEINRST	ACEILNOR	AEEGNRST	EEGINRST
arsonite	astringe	detrains	acrolein	estrange	energist
asterion	ganister	stradine	arecolin	grantees	gentries
oestrian	gantries	strained	Caroline	greatens	ingester
rosinate	granites	tarnside	colinear	reagents	integers
senorita	ingrates		Cornelia	segreant	reesting
serotina			creolian	sergeant	sigenter
			lonicera	sternage	steening