SHIFTGRAMS CIPHERED, ANALYZED

RICHARD SABEY
Chelmsford, England
richard_8978_sabey@hotmail.com

A list of the shiftwords or shiftgrams in some word stock, classified by the amount of shift, such as Leonard Gordon's "Letter-shift Words in the OSPD" (WW 2.1990-61), shows that some shifts produce more word pairs than others. To get some insight as to why, this article surveys the whole stock of shiftwords and shiftgrams of words of lengths 3-12 from a specified word stock, before going on to analyze words of some specified lengths in further depth.

My word stock is the union of the Air Force list of words from Webster's 2nd, and UKACD16 (the 16th edition of the United Kingdom Advanced Cryptics Dictionary). Only one representative was kept from each set of homographs, for example, "japanner", which may be given an upper- or lower-case J. Thus the fact that both alike 4-shiftgram to INVERTER counts as only one shiftgram, not two. However, if words are anagrams of each other, their respective shiftgrams are all counted, so JAPANNER(4)REINVERT,TRINERVE counts as two more.

n-shiftwords and *n*-shiftgrams compared

The numbers of *n*-shiftwords and *n*-shiftgrams of each *n* for each word length *l* are given here.

Number of shiftwords							Number of shiftgrams										
3	**4**	**5**	**6**	**7**	**all *l***	***n***	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**all *l***
66	42	10	1		**119**	**1**	522	1192	740	290	146	31	10	1			**2932**
64	35	1	1		**101**	**2**	586	1176	944	536	195	49	13	1			**3500**
118	92	6		1	**217**	**3**	700	1494	1016	495	215	104	20	4			**4048**
197	154	34	2		**387**	**4**	987	3120	2860	1588	667	253	51	17	3		**9546**
38	14	4			**56**	**5**	434	675	362	162	69	7	1	1			**1711**
253	197	88	6		**544**	**6**	1004	2122	2013	1218	467	162	47	9	5		**7047**
97	69	18	3	1	**188**	**7**	724	1848	1534	946	393	116	35	8	1		**5605**
139	104	16			**259**	**8**	824	2037	1504	607	217	59	11				**5259**
84	67	16	4		**171**	**9**	556	1236	1172	694	292	78	10	4	1		**4043**
183	102	20	3		**308**	**10**	776	1874	1772	914	367	137	25	4			**5869**
79	68	16	1		**164**	**11**	727	2028	1871	1114	686	240	80	23	6	2	**6777**
211	152	17	1		**381**	**12**	993	2159	1652	769	314	92	23	3	1		**6006**
66	74	19	6	2	**167**	**13**	405	1111	1406	1122	552	197	52	22	1		**4868**
1595	**1170**	**265**	**28**	**4**	**3062**	**all**	**9238**	**22072**	**18846**	**10455**	**4580**	**1525**	**378**	**97**	**18**	**2**	**67211**

In all cases, the 4-shift is good, and the 1-, 2- and 5-shifts are bad, especially the 5-shift.

The 6-shift does well (but is not the best shift) at shiftgrams. At shiftwords, it is best. This better performance is because the 6-shift ciphers IOUY as OUAE. The fact that it ciphers four vowels as vowels makes it especially good, because it is likely that a word with a common vowel-consonant sequence will cipher into a string with a common vowel-consonant sequence. It also helps that the 6-shift ciphers S as Y, and that E, S and Y are all very common at the end of a word. This is why the 6-shift does so much better than the 4-shift, which one might have thought would do almost as well, because it ciphers AEU as EIY. This is not of so much help at the end of the word, as A and I are rare enders; the end-letters used most frequently by 4-shiftwords are N-R and O-S (even though O is rare as an ender).

The 11-shift is a close second for shiftgrams of 8-letter words, and is even more frequent in longer shiftgrams. It is good mainly because TAI 11-shifts to ELT. It is better with the longer words partly because C and P are more frequent in words of length 9-12 than in 8-letter words; ER 11-shifts to PC, which 11-shifts to AN.

The 13-shift is only middling with short words. However, it does better with longer shiftgrams. This is partly because there are more long words than short ones with one of the affixes RE-, -ER and -LY (R and E 13-shift to each other, as do L and Y).

Shiftgrams

What makes a shift cipher fruitful? It is useful if common letters map to common letters. It is less important how the rare letters map. The commonness or rareness of letters is assessed by calculating their relative frequency in the word stock. I quantify the fruitfulness of the cipher by the probability that, given two letters picked at random from the word stock, shift-ciphering the first produces the second, as follows:

$$shiftgrammerit(n) = \sum_{\alpha} \left[freq(\alpha) freq(\alpha + n) \right]$$

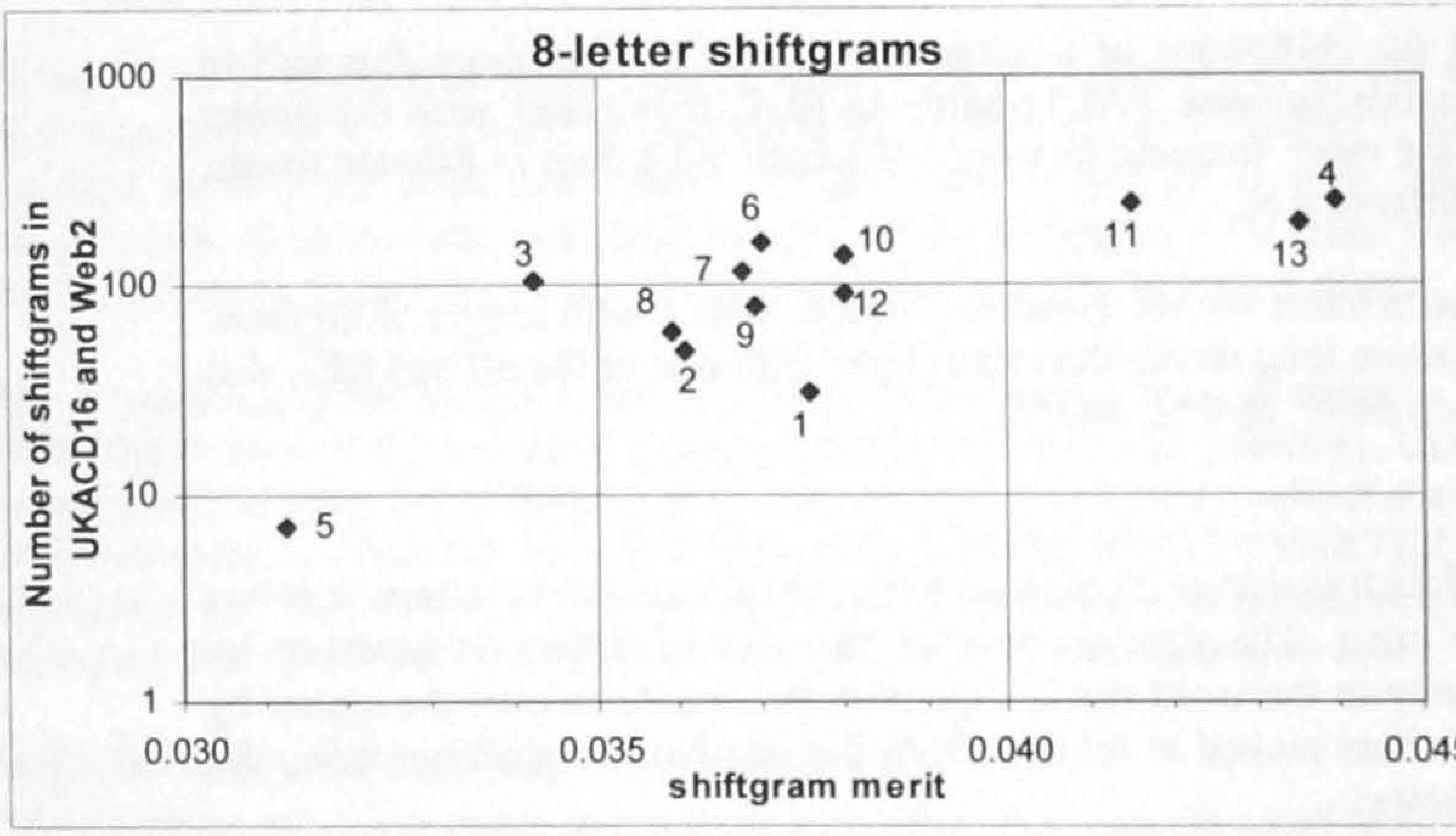
α means a letter of the alphabet, and the sum over all α thus means the sum for all letters of the alphabet. $freq(\alpha)$ means the frequency of α in the word stock. n is the amount of shift. Thus for example if α is the letter Z, then $\alpha+1$ is the letter A. This formula takes no account of the word length or the freedom to permute the letters, let alone the fact that the number of permutations depends on the letter-pattern of the words. It is designed only as a quantity with which to correlate experimental results, not as a predictor of those results.

This word stock produces the following statistics for the shiftgrams of 8-letter words.

Number of occurrences of each letter								Number of shiftgrams for each shift			
a	33522	**h**	9942	**o**	25505	**v**	3595	**1**	31	**8**	59
b	8357	**i**	31633	**p**	11312	**w**	4511	**2**	49	**9**	78
c	14839	**j**	772	**q**	661	**x**	1105	**3**	104	**10**	137
d	15068	**k**	4987	**r**	27362	**y**	6701	**4**	253	**11**	240
e	43776	**l**	22340	**s**	30451	**z**	1449	**5**	7	**12**	92
f	5649	**m**	11543	**t**	23394	388840 letters, 48605 words		**6**	162	**13**	197
g	10753	**n**	25178	**u**	14435			**7**	116	Total 1525	

The relation between theory and experiment is shown in the following scatter diagram. Each point's label is the amount of shift. The y-axis has a logarithmic scale. The x-axis has a linear scale.

Experimental statistics are roughly in line with theory, with a couple of exceptions. Why are there so few 1-shiftgrams? Theory predicts many 1-shiftgrams, mainly because the common letters RSDNTH 1-shift to STEOUI. However, words need vowels, and AEIOUY 1-shift to BFJPVZ, all awkward consonants. Why are there so many 3-shiftgrams? It seems to be because a few common morphemes produce good letters for the other word. There are 12 examples of ABLE to DEOH, 11 of LSBO to OVER, 7 of TLLA to WOOD.



Shiftwords

The above formula used letter frequencies without reference to the position of the letter in the word. This was appropriate when discussing shiftgrams. Different considerations apply when considering shiftwords. When a word is shift-ciphered, each letter in that word ciphers to a letter in the same position in the ciphered word. This consideration means that frequencies and merits are calculated for each letter position separately. The overall merit is found by taking the product of the separate merits of each letter position, as follows:

$$shiftmerit(n,l) = \prod_{i=1}^{l} \sum_{\alpha} \left[freq(\alpha,i)\, freq(\alpha+n,i) \right]$$

i means a position of a letter in the word, and $freq(\alpha, i)$ means the frequency of α among the letters in position i of the words in the word stock.

The 4-letter shiftwords in UKACD16 and Webster's 2nd produce the following statistics.

Number of occurrences of each letter in each position										Number of shiftwords for each shift	
a	404	1245	603	546	n	203	145	529	355	1	42
b	422	43	161	112	o	191	1147	399	320	2	35
c	355	70	208	42	p	395	70	187	205	3	92
d	332	61	219	339	q	27	2	2	2	4	154
e	196	838	553	807	r	301	318	571	236	5	14
f	260	15	112	117	s	588	56	327	950	6	197
g	338	52	187	163	t	423	90	323	500	7	69
h	290	209	90	178	u	85	671	305	132	8	104
i	119	807	441	239	v	113	35	90	11	9	67
j	145	8	26	5	w	264	65	136	76	10	102
k	227	58	170	299	x	9	22	39	53	11	68
l	321	278	458	306	y	145	164	97	324	12	152
m	338	70	245	186	z	63	15	76	51	13	74
						Total 26216 letters, 6554 words				Total 1170	

The results are shown in the scatter diagram below. This time, both axes have log scales.