ANOTHER POLYPHONIC CIPHER

A. ROSS ECKLER
Morristown, New Jersey

A polyphonic substitution cipher is one in which several different plaintext letters are enciphered into a single cipher letter or symbol. Perhaps the most simple and well-known example of a polyphonic cipher is the telephone dial, in which the letters ABC are encoded by the number 2, DEF by 3, GHI by 4, JKL by 5, MNO by 6, PRS by 7, TUV by 8, and WXY by 9. Polyphonic ciphers have tended to be shunned by cryptologists because of the inevitable ambiguity encountered in recovering a message (does 117 equal BAR or CAP?). However, if one turns the problem around and asks how one should encode the alphabet to make it as easy as possible to recover a message, then polyphonic ciphers are deserving of study. Since the English language is highly redundant, it is possible to tolerate a considerable amount of ambiguity in decoding.

Obviously, the fewer different symbols that are used in the cipher, the more ambiguity will result. In "A Readable Polyphonic Cipher" in the February 1975 Word Ways, I devised a nine-symbol cipher (conveniently encoded by the digits 1 through 9, with 0 a word-space) in which messages could be easily deciphered. Specifically, I assigned letters so that the commonest bigrams in English-language text all had distinct cipher representations. It turned out that the commonest 30 bigrams could be assigned different number-pairs (DE was the commonest bigram with a number-pair already used by a yet commoner bigram); in fact, of the 81 bigrams accommodated by the number-pairs 11, 12, ... , 98, 99, 57 were among the 70 commonest bigrams in the language.

Can the redundancy of the English language support a polyphonic cipher using fewer than nine symbols? This article demonstrates that one can reduce the number of symbols to six if one is willing to spend a bit more time working out the message. The difficulty of deciphering is increased, but the results of the decipherment do lead to the intended message with a high degree of certainty. This improvement is made possible by using the commonness of words rather than bigrams as the basis of the cipher.

The cipher works on the principle that one should always select the commonest English word corresponding to a given sequence of symbols. Thus, if the pattern 32 can be decoded as either OF or ON, the first word should be selected since this occurs more commonly in the English language. Statistics giving the frequency of occurrence of words in English are readily available in H. Kucera and W. N. Fran-

cis' Computational Analysis of Present-Day American English (Brown University Press, 1967), based on a sample of a million words from English texts first printed in 1961. The commonest-word strategy is tedious to implement without the aid of a computer containing the Kucera and Francis corpus, for one must construct a list of 6^i possibilities for words i letters long -- 36 two-letter word candidates, 216 three-letter word candidates, and so on. (For longer words, one probably ends up listing nearly every word's cipher separately, as ambiguity is rare.) For this reason, this cipher is unlikely to be widely used, but instead should be viewed as an interesting demonstration of the redundancy of the English language.

I used a minimax philosophy to assign letters to symbols -- that is, I tried to make the commonest word dominated by a yet-commoner word in the cipher as rare as possible. For example, I made sure that F and N were enciphered differently in order to avoid having ON (6742 occurrences in Kucera and Francis) be interpreted as OF (36411 occurrences). Likewise, R and S had to be kept distinct (THERE occurs 2724 times, THESE 1573). By a somewhat tedious procedure of trial and error, it was possible to assign letters to symbols in such a way that the commonest word lost was TAKE, the 148th word on the list with 611 occurrences (dominated by MAKE, with 794).

Why did I pick six symbols, instead of seven or four? Short words have fewer alternative codes than long ones, and two-letter codes have the fewest of all. I early decided that the 24 common two-letter words ought to have unique ciphers; this is first possible if one allows a six-symbol cipher with 36 possibilities. It turned out that I had to sacrifice the least common of these 24 words, AM, to avoid even worse three-letter word scrambles (AM is dominated by AT). The polyphonic cipher is given at the left, and the codes for the other 23 two-letter words are given at the right:

		1	2	3	4	5	6
KAMT	1	at	as	an	me	to	my
ZJLIS	2	it	is	in		so	if
NPW	3				we	no	
QBDE	4				be	do	by
XVCOR	5			on		or	of
GHFUY	6		us	up	he	go	

This cipher is not unique; a number of the rare letters can be assigned other symbols if desired without violating the minimax principle.

Three-letter words proved almost as difficult as two-letter words to accommodate; although they have 216 possible codes, the number of very common words is much greater. The following table of three-letter words is typical of the ones to be stored in the computer; the commonest Kucera and Francis word is indicated first, with the second commonest word following in parenthesis. (The most common of these dominated words is YET, with 419 occurrences.) Sampling vagaries make Kucera and Francis unreliable when only a few instances of a

		1	2	3	4	5	6
1	1	mat	mal	man(tap)	mad(ate)	tax(Max)	may(Kay)
	2	ask(aim)	all	tip(tin)	aid(kid)	air(mix)	ash
	3	apt		Ann	and	two	any
	4	met(tea)	ads	men(ten)	add	ado	key
	5	act(art)		top(ton)	are(ace)	too(arc)	try
	6	mum		tun	the(age)	ago	thy
2	1	sat(Sam)	its	saw(law)	sad	jar(lax)	say(lay)
	2	sit(Jim)	ill	sin(lip)	lie(lid)	six(sir)	sly
	3	ink					spy
	4	set(let)	Les	Jew(Zen)	see(led)	sex	leg
	5	lot	Los	low(son)	job(Joe)	sox	joy(Lou)
	6	sum		sun	she(sue)		shy
3	1	pat	was	pan	pad	war(wax)	way(pay)
	2	wit(pit)	nil	win(pip)	pie		pig
	3						
	4	wet(net)	Wes	new(pen)	wed	per	peg
	5	not(pot)	poi	now(won)	nod	nor	pry
	6	put(nut)	pus	pun	nub	who	why
4	1	eat(bat)	bas	Dan	bad(Dad)	bar(ear)	day(bay)
	2	bit(dim)		bin(dip)	did(die)		big(dig)
	3				end		
	4	bet	Del	Ben	bed(bee)		beg
	5	era(dot)		don(bow)	Bob(eve)	box(doc)	boy(dog)
	6	but	bus	dun(bun)	due(eye)	ego(quo)	buy(dug)
5	1	cat(oak)	Cal	can(ran)	cab	car	ray(rag)
	2	via	oil	rip	old(rid)	Rio	rig
	3			own	one(owe)		
	4	vet			red(odd)	Rex	rey
	5	rot	col	row(cow)	rob(rod)	roc	cry(Roy)
	6	out(cut)		run(cup)	vue	our	off(rug)
6	1	fat(hat)	has(gas)	fan(gap)	had	far	gay(hay)
	2	him(hit)	his	gin(hip)	use	fix	fig(fly)
	3		ups				
	4	get(yet)	yes	few(hen)	fed(fee)	her	hey
	5	got(hot)		how	god	for(fox)	you(fog)
	6	gum(hut)	Gus	gun(fun)	hub	fur	guy

word were found, so plausible substitutions have been made in several cases (such as SHY for SHU, or JAR for LAO). The commonest doubly-dominated (or more) three-letter words, which do not appear in the above table, are ARM, FIT, SAN, SEA, TOM, SKY, ICE, RAW, BAG and LEE.

A brief examination of the above table reveals that three-letter words are rather unevenly distributed in it, the most notable lack

being words with second letter encoded by 3 (W, N or P). Originally it was conjectured that the vowels should be uniformly distributed among the cipher symbols to minimize problems of this sort, but it turned out that YOU dominated HOW (a word with 834 occurrences) when U was encoded by 3.

Four-letter words are somewhat easier to keep separate, since they are sorted into 1296 boxes. The brief table below indicates the commonest dominated words, with the dominating word in parenthesis:

take 611 (make)	four 359 (your)	gave 285 (have)
face 371 (have)	kind 313 (mind)	rate 209 (came)
form 370 (from)	York 301 (from)	hard 202 (have)
	five 286 (give)	

Note that 6154 is an especially popular encipherment, with HAVE, FACE, GAVE and HARD all among the 150 commonest four-letter words.

A lengthy check of five-letter words indicated that the first dominated one is WHOSE (with 252 occurrences) which yields to WHOLE. Since five-letter words played very little role in the formation of the cipher (there are only 25 that occur more often than TAKE, the first dominated word of any length), it is of interest to see how many words of five letters had to be examined before the first dominated one was found. WHOSE is, in fact, the 72nd word in the five-letter list. A simple probability calculation (similar to the one used to determine that there is a 50-50 chance that two or more people in a group of 22 will have the same month and day of birth) reveals that there is a probability of 0.72 that the first domination will occur later than the 72nd word if all five-letter ciphers are equally likely; our bad luck is hardly surprising, given the unequal distribution of words noted in the three-letter word table.

Dominations for words of six or more letters have not been checked, but it is likely to take several hundred words to find one.

How likely is it that a randomly-chosen word in English text is dominated by another one? Obviously, this depends on the word length, and it is feasible to calculate only for words of two, three or four letters. Over 98 per cent of all two-letter words (weighted by textual occurrence) were examined; among these, the probability of drawing a dominated one was an infinitesimal 0.004. A similar percentage of three-letter words were examined, leading to a probability of domination of 0.027. It was possible to examine only 84 per cent of all four-letter words, but among these only 0.046 were dominated. From this, one can guess that there is an overall probability of less than 10 per cent that a randomly-chosen word is dominated, with the most likely problems occurring among four-letter words. To test this conjecture, the first sentence in the first chapter of W. Allen Wallis and Harry V. Roberts' Statistics: A New Approach (Free Press, 1956) was enciphered and deciphered: "Statistics is a body of methods for

making wise decisions in the face of uncertainty" translated to "Statistics is a body of methods for making WILD decisions in the HAVE of uncertainty" -- two errors in fifteen words. To deal with situations like these, it is useful to have the computer programmed to deliver not only the most probable word, but the most probable two or three words. In this case, WISE (36 occurrences) is dominated only by WILD (56 occurrences), and FACE (371 occurrences) only by HAVE (3941 occurrences), so the correct message is easily found. However, I concede that some non-statisticians might still prefer WILD to WISE!

A DICTIONARY OF CATCH PHRASES

This is the title of a new book by Eric Partridge, well-known as the solo compiler of earlier dictionaries on word origins, slang, underworld jargon and Shakespearean bawdy words. What is a catch phrase? Alas, Partridge refuses to define it, other than by means of the too-general "a saying that has caught on and pleases the public", hopelessly confusing it with the concepts of cliche, proverbial saying, and famous quotation. Set against Partridge's too-general definition, I offer a too-specific one: a catch phrase is a phrase having a meaning different from that suggested by its words taken at face value, and which is used as a conversational shorthand to respond to a commonly-occurring social situation. Classic examples are: don't hold your breath; big deal; drop dead; I bet you say that to all the girls; I couldn't care less; let's get the show on the road; pardon my French; you're the doctor. This is a great book for browsing through at random; there are oddities and delights on every page. However, Partridge is not always up-to-date on American catch phrases, and the book is not error-free: I note that Pearl Harbor occurred on December 10, 1941 (p.53), "fish -- or cut bail" (p.62), and a reference to the cartoon character Smokey Storer (instead of Stover) (p.63).