SENDING MESSAGES BY TELEPHONE

A. ROSS ECKLER
Morristown, New Jersey

Although the telephone conveys the spoken word with ease, it is ill-suited for the written one needed for pagers or communication with the deaf. The telephone keypad contains only ten alternatives (sometimes twelve), far fewer than the 26 alphabetic letters plus space, which means that a single press of a button is ambiguous. The traditional arrangement ABC/DEF/GHI/JKL/MNO/PRS/TUV/WXY omits Q and Z, but these can be understood to be in their normal positions.

One way to unambiguously convert the keypad choices to letters is to assign two button-presses for each letter, but this is slow to use and requires mastery of a conversion table by the user. It is much simpler to use the alphabet as presented on the telephone keypad and untangle the ambiguities. This task is performed by a computer program developed by Pone-Ex of Lambertville, New Jersey, the brain-child of Bernard Riskin. He has stored some 200,000 words in a computer, sorted by their telephone equivalents: for example, NO and ON appear together, as do SAY, SAW, RAY, RAW and PAW. When the user presses a sequence of buttons on the keypad, the computer retrieves the most likely word and recites it over the phone. If this is not the word wanted (for example, the computer selects ON when NO is desired by the user), the user presses a button requesting the computer to present another word with the same encipherment. This continues until the right word is produced, whereupon the user proceeds to enter the next word of the message.

Of course, even a vocabulary of 200,000 is not sufficient for every user's needs, so Riskin has also devised a method by which new words (for example, proper names) can be added by the user to the database.

How often must the user ask the computer for a second (or third, or even fourth) word before his message is correct? Though it is difficult to give a precise answer, it appears that the user must correct only about one in 25 words. This conclusion is based on the assumption that the most probable word is presented first, and that word frequencies are accurately represented by Kucera and Francis's *Computational Analysis of Present-Day American English* (Brown University Press, 1967), a corpus of approximately one million words taken from various American texts in 1962. The greatest potential for ambiguity resides in words of two to five letters; only occasionally do words of six or more letters cause trouble.

The commonest words which require a second (or third) attempt to retrieve them from the computer are given below. The number gives the Kucera and Francis frequency in a million words of text, and the words in parentheses are the ones which were previously produced by the

computer.

no (on) 2201	seem (seen) 229	lay (law) 139
these (there) 1573	am (an) 228	hot (got) 130
then (them) 1377	soon (room) 199	aid (age) 130
home (good) 547	red (see) 197	note (move) 127
war (was) 464	gone (good) 195	wide (wife) 125
night (might) 411	cut (but) 192	hand (game) 123
saw (say) 352	view (they) 186	blood (alone) 121
York (work) 301	pay (say,saw) 172	sun (run) 112
line (kind) 298	case (care) 162	ball (call) 110
gave (have) 285	fine (find) 161	season (reason) 105
past (part) 281	food (done) 147	walk (wall) 100
boy (any) 242	paid (said) 145	

eight 98, gas 98, Jack 92, base 91, hotel 90
add 88, battle 87, sight 86, shape 85, post 84, bar 82, nine 81, offer
  80
Sam 79, fast 78, die 73, phase 72, rain 70, Rome 70, box 70, Sam 70
aside 67, she'd 67, ends 66, page 66, join 65, wore 65, Tom 63, wind 63,
  cool 62, save 62, soft 61, nose 60, fat 60
lie 59, mine 59, roof 59, tree 59, hole 58, lose 58, stone 58, goods 57,
  truck 57, games 55, win 55, inner 55, runs 55, tall 55, dear 54, band
  53, wet 53, gold 52, sick 51, waves 51, chain 50, proud 50

    As words become less common, the probability that they are dominated
by more-common ones increases. Of the 58848 occurrences of words
appearing 400 to 799 times, 1422 (.024) are dominated; of the 72014
occurrences of words appearing 200 to 399 times, 2216 (.031) are
dominated; of the 82288 occurrences of words appearing 100 to 199 times,
3078 (.037) are dominated; and of the 80673 occurrences of words
appearing 50 to 99 times, 3764 (.047) are dominated.

    The telephone keypad encipherment is a special case of a polyphonic
cipher, discussed in the Feb 1975 and May 1978 issues of **Word Ways**.
Could word correction be reduced by a different allocation of letters to
numbers, preserving alphabetic order? It seems likely that the telephone
allocation is close to optimum. The only possible improvement appears to
be the transfer of O from MNO to PQRS, which introduces four new
ambiguities involving OP and OR with 751 occurrences (out-put, too-top,
foot-fort, word-wood) but eliminates sixteen ambiguities involving NO or
MO with 4071 occurrences (on-no, good-home, boy-any, good-gone, done-
food, alone-blood, needs-offer, any-box, does-ends, too-Tom, some-roof,
stood-stone, homes-goods, homes-inner, stoop-runs, who-win). However,
more drastic changes are counterproductive. The allocation ABC/DEF/GHI
/JKL/MN/OP/QRS/TUVWXYZ creates many TW ambiguities including that-what
1908 and when-them 1789-then 1377, overwhelming any savings from OR, OS,
PR and PS ambiguities. The allocation ABC/DEF/GHI/JKL/MN/OP/QRST/UVWXYZ
removes the TW ambiguities but replaces them with is-it 8756, as-at
5378, to-so 1984 and the-she 2859. In general it is worthwhile
enciphering vowels separately and allowing no more than two consonants
to be enciphered together (not counting rare ones such as J,Q,X,Z); the
only violation of this prescription occurs with PQRS.