INSERTION-DELETION NETWORKS

A. ROSS ECKLER
Morristown, New Jersey

Word networks - collections of words of a common length that can be joined by single-letter substitutions, such as the sequence BANKER-BANTER-BATTER-BETTER-SETTER-SETTEE-SETTLE linking BANKER and SETTLE - were studied in some detail in the May and August 1973 issues of **Word Ways**. Non-repeating sequences from word networks, commonly called word ladders or doublets, have a long history; Lewis Carroll, among others, challenged readers to join two specified words by a ladder. However, word networks have one serious flaw: they do not allow one to incorporate words of different lengths. This can be rectified by the concept of an insertion-deletion network, in which a word of n letters is linked to a word of n-1 letters if the first can be converted to the second by the deletion of a single letter. (In the reverse direction, the linkage is called the insertion of a letter in a word to form another word.) Extremely complex insertion-deletion networks containing many loops and branches can be constructed; by examining them, one can easily ascertain whether or not one word can be reached from another by successive insertions and deletions. Further, one can calculate the minimum number of such steps required.

Naturally, the details of the insertion-deletion network depend strongly upon the dictionary chosen. To facilitate comparison with the word network study, I have used boldface words listed in the main section of the 1964-73 edition of the New Merriam-Webster Pocket Dictionary. Hyphenated words, suffixes, prefixes and abbreviations are omitted, as are words listed in groups without further definition (under anti-, un-, over-; re-, self-, sub-, super- and un-). A and I are the only single-letter words allowed. Note in particular that most plurals, past tenses and participles are not listed. It was necessary to keep the vocabulary small because all the work was done by hand; it would, of course, be of considerable interest to program a computer to analyze the characteristics of insertion-deletion networks corresponding to larger dictionaries (such as the Official Scrabble Players Dictionary, or the second edition of the unabridged Merriam-Webster Dictionary).

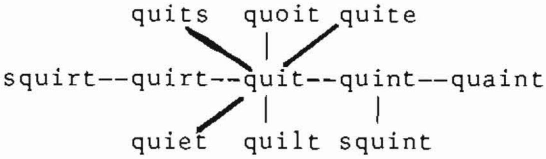
The main network consists of approximately 2800 words. The longest words found in the network have eight letters:

wrangler-wangler-angler-
masterly-mastery-master-aster-caster-
leathern-leather-lather-lathe-
leathery- leather-lather-lathe-
heathery-heather-heater-hater-hate-

splutter-sputter-putter-utter-butter-butte-butt-but-

In each case, various side branches or loops exist, but the only connections to the main network are through the right-hand words. ANGLER connects to the main network by two quite distinct paths, MANGLER-MANGER-MANGE-MANE-MAN and ANGER-RANGER-RANGE-RANG.

The main network contains words using every letter of the alphabet but Q. The largest network of Q-words is:



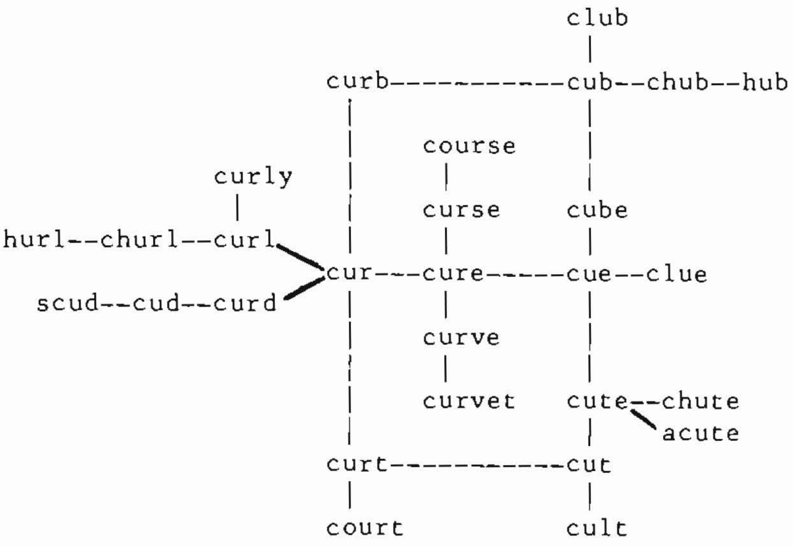
The main network contains many of the most commonly-used English words. In fact, using Kucera and Francis's million-work corpus as the criterion of word commonality, the first thirty words can all be joined: the, of, and, to, a, in, that, is, was, he, for, it, with, as, his, on, be, at, by, I, this, had, not, are, but, from, or, have, an, they. The first missing word is WHICH (31); the next four, HIM (42), WHO (46), ONLY (62), and OTHER (63). It seems quite likely that one could write a connected discourse exclusively of words in the network, although many common words such as WATER, MOTHER and LIGHT are not present.

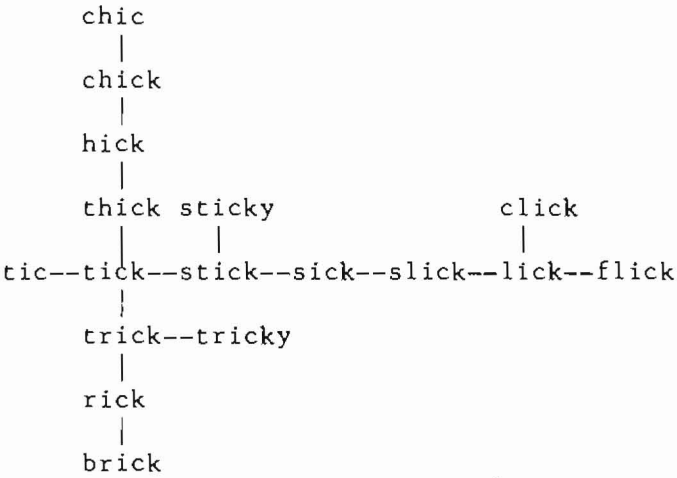
The main insertion-deletion network is larger than any of the individual word networks discussed in May and August 1973; however, it is only about two-thirds as large as the combination of the largest 3-letter, 4-letter, 5-letter, 6-letter, 7-letter and 8-letter word networks. In other words, letter-substitutions are measurably easier to carry out in English than insertions and deletions are. In the 3-letter word network discussed in May 1973, 529 of the 541 Pocket Merriam-Webster words were joined; in contrast, the main insertion-deletion network contains only 416. For the record, the missing 125 are listed below; starred ones connect with no other Pocket Merriam-Webster word:

alp* arc auk* awl bib* bud* bum cab cob cry*
cub cud cue cur cut dab DDT* deb den dew
dig* dip dub ebb* egg* eke* elf elk* emu* err*
few fey* fez* fib* fig* fix* flu fob* gab gem
get gig* gnu* gum gun* guy* gym* gyp* haw him
hob hog* hub* hum icy* irk ivy* jag* jar jaw*
jet Jew* jib jig* job* jog jot joy* jug* keg*
ken key* lab leg* lei* mix* mob* mud* mug mum*
nab* neb* nib* nix* nth* nub nun nut oaf oak
odd* ohm* opt* ova peg* pep pub* pyx* rev* rib
rip rub sex* six* sky* sub tab taw tic tip
tub vex* via vim* vow who why woo yak yaw
yap zed* zen* zip* zoo

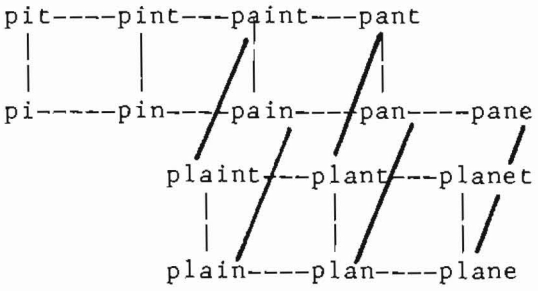
Five of these words are contained in the largest-known insertion-deletion network unconnected to the main network; in all, it has 28 words. The next largest insertion-deletion network has 17 words; both are given on the next page.

The densest part of the insertion-deletion main network is probably at the word AT, which has 19 words connected to it: a, eat,





oat, ant, apt, aft, act, art, ate, bat, cat, fat, hat, mat, pat, rat, sat, tat and vat. PA participates in eight different loops of length four (the minimum loop size): pa-pal-peal-pea-, pa-pas-peas-pea-, pa-pat-peat-pea-, pa-par-pear-pea-, pa-spa-spat-pat-,pa-pan-pant-pat-, pa-pas-past-pat-, and pa-pad-pard-par-pa. An additional conglomeration of closed loops of length four is found in the vicinity:



Note that there is a cube of connections between PAN and PLAINT, and nearly a cube (lacking only PANET) between PAN and PLANET. If S is added to the words forming the cube, one has a hypercube (in four dimensions) joining PAN with PLAINTS.

In the August 1971 **Word Ways**, Dave Silverman introduced the concepts of hospitable and charitable words. A hospitable word is one which admits of the insertion of a letter in any position to form a new word; a charitable word admits of the deletion of a letter in any position to form a new word. To avoid generating the same deleted word twice, let us disallow words with internal doubled letters. Furthermore, to avoid ambiguity, let us require that each inserted letter be different from its neighbors. What are the longest charitable and hospitable words in the Pocket Dictionary network? The longest hospitable word appears to have only three letters, as in PAT: SPAT, PEAT, PANT, PATE or LAD: GLAD, LOAD, LAND, LADY. The longest charitable word, however, has four, as in SEAT: SEA, SET, SAT, EAT or PEAR: PEA, PER, PAR, EAR. No Pocket Dictionary word is simultaneously hospitable and charitable; the only three-letter charitable words, MAY and PAY, are inhospitable.

Let us define the distance between two words in the insertion-deletion network to be the minimum number of words needed to go from one to the other. For example, the distance between HAUNT and HUE is six: HAUNT-HUNT-HUN-HUNG-HUG-HUGE-HUE. The span of the insertion-deletion network is defined as the maximum distance between any pair of words in the network. This is not easy to calculate by hand in a large and complex network having many alternative routes between pairs of words. The largest distance found in the Pocket Merriam-Webster insertion-deletion network is 34, between DUD and MISERLY. Both words, at the end of long branches, allow alternative paths only between the two asterisked words:

dud-dude-due-dune-dun-dung-dug-drug-rug-rung-run*-runt-rut-rout-
out-pout-pot-poet-pet-pert-per-pier-pie-pice-ice*-mice-mince-mine-
miner-minter-miter-mister-miser-misery-miserly

It seems likely that the span of the network is 34, unless some **Word Ways** reader can shorten this path.

In addition to the actual distance between two words in the insertion-deletion network, one can define an idealized distance which forms a lower bound: it is the value of the distance if any combination of letters is allowed to be a "word" in the network. To calculate the idealized distance between two words, cancel the maximum subsequence of letters common to both words and in the same order, and count the number of letters remaining. For example, the idealized distance between MasTerLy and MenTaL is eight, obtained after cancelling the sequence MTL. The idealized distance between two words with no letters in common is equal to the sum of their letters.

Of course, in many cases the distance and idealized distance are identical: ON-ION-IN, STAIR-STIR-SIR-SIRE-SIREN, SKIP-SKI-

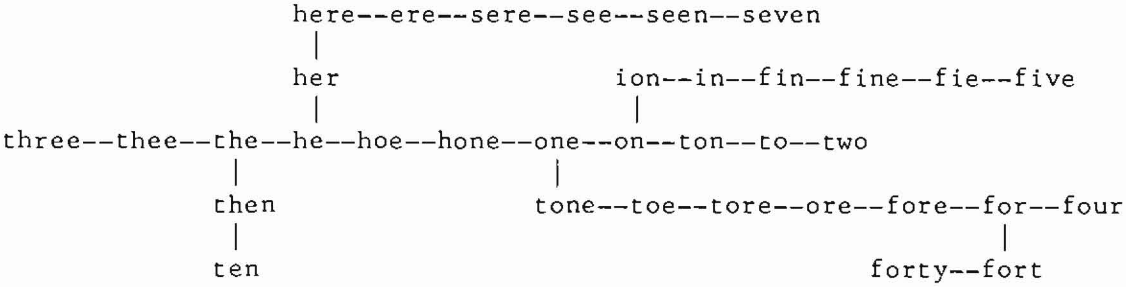
SK1T-S1T-IT-P1T-P1TH-P1TCH. What is the greatest distance for which one can find a pair of words having the same distance and idealized distance? The trick is to find two words as long as possible with no common letters; the answer appears to be 14, achieved by either of the following:

heathery-heather-heater-hater-hate-ate-at-a-an-wan-wain-win-wing-swing-sowing
leathery-leather-lather-lathe-late-ate-at-a-an-wan-wain-win-wing-swing-sowing

A distance of 15 equal to the idealized distance is impossible to achieve, for one of the words must be at least eight letters long, and it is easy to check that none of the six listed earlier qualify. (Of course, if the Pocket Webster had the word SNOW1NG this task would be easy.)

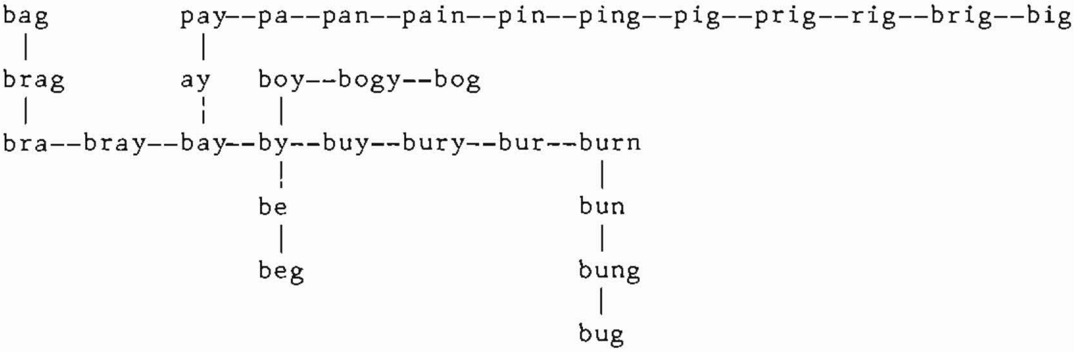
When the distance exceeds the idealized distance, the difference is always a multiple of two. This reflects the fact that additional letters must be both inserted and deleted, an operation requiring two steps.

Among the cardinals, only ONE, TWO, THREE, FOUR, FIVE, SEVEN, TEN and FORTY are in the main insertion-deletion network. The subnetwork joining these eight cardinals is an interesting one:


here--ere--sere--see--seen--seven
her
ion--in--fin--fine--fie--five
three--thee--the--he--hoe--hone--one--on--ton--to--two
then
tone--toe--tore--ore--fore--for--four
ten
forty--fort

Note that the distances between ONE and TWO, ONE and THREE, THREE and TEN, and FOUR and FORTY are all equal to the idealized distances, and the distances between ONE and FOUR and ONE and FIVE exceed the idealized distances by only two.

lt is surprisingly difficult to deal with vowel substitutions in insertion-deletion networks; l illustrate with a subnetwork joining BAG, BEG, BIG, BOG and BUG:


bag
pay--pa--pan--pain--pin--ping--pig--prig--rig--brig--big
brag
ay
boy--bogy--bog
bra--bray--bay--by--buy--bury--bur--burn
be
bun
beg
bung
bug