# THE GROWTH OF A WORD NETWORK

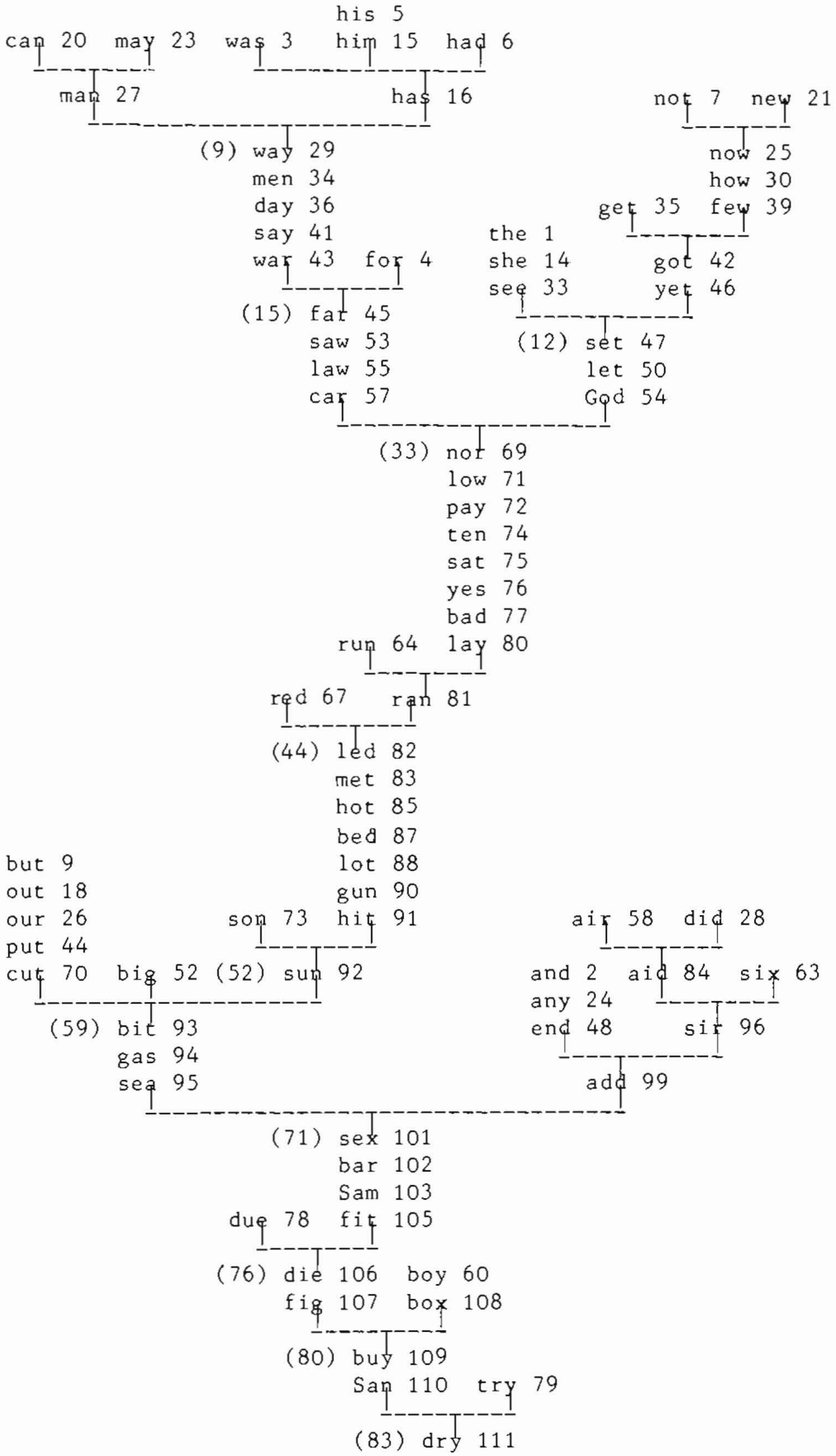
A. ROSS ECKLER
Morristown, New Jersey
LEONARD J. GORDON
Chico, California

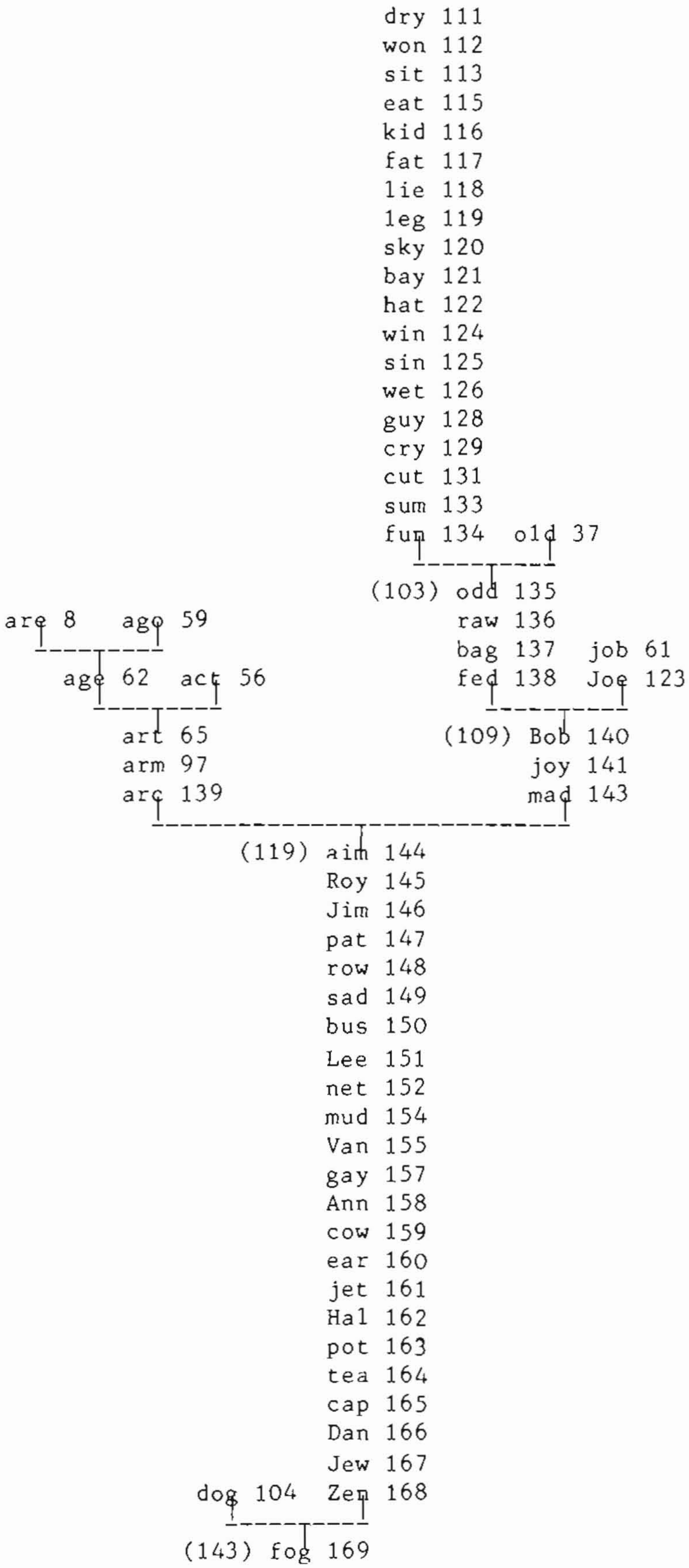
The May 1973 issue of **Word Ways** introduced the concept of a word network, a connected diagram of words of a single length, in which any two words differing in only one letter in the same position (as BRA and ERA, or TON and TEN) are joined by a line. Networks with more than 50 to 100 words are generally too complicated to diagram on a single sheet of paper, and it is difficult to assess their properties.

In particular, it is hard to assess the degree of connectivity of a network--is any pair of words connected by many different routes, or are there isthmuses (such as Panama between North and South America) through which most routes must go? If most of the words are contained in a single main network, with only a few words or small groups of words left out, it is conjectured that no major isthmuses exist. However, it is possible that such isthmuses transitorily form as networks are built up a word at a time, joining formerly isolated groups for the first time.

To evaluate this possibility, this article shows how one can chart the evolution of a word network from a set of unconnected words. The diagram on the next two pages shows how it can be done, using three-letter words as an example. Kucera and Francis's Computational Analysis of Present-Day American English (Brown University Press, 1967) tabulates the observed frequencies of words in a million-word sample of materials printed in the United States in 1962. One can build up a network of words one at a time in decreasing order of frequency, starting with THE, AND, WAS, FOR, and HIS. At first, no words are connected. The initial link occurs between HIS (ranked 5th in the list) and HIM (ranked 15th). When HAS (16) appears, it joins the HIS-HIM fragment with two isolated words, HAD (6) and WAS (3). All this is diagrammed at the top of the tree-like structure on the next page, in which each word is followed by its Kucera-Francis ranking. The only word that might be fairly identified as an isthmus is NOR (69), which unites the 18-word network on the left with the 14-word network on the right by joining FOR in the former with NOT (or NOW) in the latter. (The parenthesized number preceding NOR indicates that the main network now has 33 members, or 48 per cent of all the words thus far sampled: 33/69 = 0.48.) Note that the isthmus begins to disappear with the very next word added, LOW, which joins LAW with NOW (or HOW).

Scanning the diagram, one sees that the main network soon becomes so large that it attracts other word groups to it before they


his 5
can 20
may 23
was 3
him 15
had 6
man 27
has 16
not 7
new 21
(9) way 29
men 34
day 36
say 41
war 43
for 4
now 25
how 30
get 35
few 39
the 1
she 14
see 33
got 42
yet 46
(15) far 45
saw 53
law 55
car 57
(12) set 47
let 50
God 54
(33) nor 69
low 71
pay 72
ten 74
sat 75
yes 76
bad 77
run 64
lay 80
red 67
ran 81
(44) led 82
met 83
hot 85
bed 87
lot 88
gun 90
son 73
hit 91
but 9
out 18
our 26
put 44
cut 70
big 52
(52) sun 92
air 58
did 28
and 2
any 24
end 48
aid 84
six 63
sir 96
(59) bit 93
gas 94
sea 95
add 99
(71) sex 101
bar 102
Sam 103
due 78
fit 105
(76) die 106
fig 107
boy 60
box 108
(80) buy 109
San 110
try 79
(83) dry 111



```
                   dry 111
                   won 112
                   sit 113
                   eat 115
                   kid 116
                   fat 117
                   lie 118
                   leg 119
                   sky 120
                   bay 121
                   hat 122
                   win 124
                   sin 125
                   wet 126
                   guy 128
                   cry 129
                   cut 131
                   sum 133
                   fun 134  old 37
                    ---------
                 (103) odd 135
are 8  ago 59          raw 136
---------              bag 137  job 61
   age 62  act 56      fed 138  Joe 123
   ---------            ---------
      art 65          (109) Bob 140
      arm 97                joy 141
      arc 139               mad 143
      ----------------------------
            (119) aim 144
                  Roy 145
                  Jim 146
                  pat 147
                  row 148
                  sad 149
                  bus 150
                  Lee 151
                  net 152
                  mud 154
                  Van 155
                  gay 157
                  Ann 158
                  cow 159
                  ear 160
                  jet 161
                  Hal 162
                  pot 163
                  tea 164
                  cap 165
                  Dan 166
                  Jew 167
          dog 104 Zen 168
           ---------
       (143) fog 169
```

become very large. There are no hard-to-reach places in three-dimensional word space where a group of words can build up considerably before joining the main network. The greatest potential seems to be in words beginning with a vowel, which have difficulty in joining words beginning with a consonant.

By the time one has sampled 169 Kucera-Franics words (omitting abbreviations such as REV(erend), SEN(ator), and AUG(ust)), the main network has incorporated 85 per cent of the sample, leaving a number of isolated words plus the groups WHO-WHY, HER-PER, ALL-ILL, and TWO-TOO-TOP-TOM. When the next word, TIM 170, is reached, this four-word group is brought into the main network.

Each pair of non-adjacent words in a network can be joined by a series of intermediate words, often in more than one way; however, there is always some path of minimum length. Looking at all word-pairs in the network, the number of steps in the longest of these minimum-length paths is defined to be the span of the network. In the simple network

```
                          had
can-man-may-way-was-has-his-him
    men
```

the span is 7, achieved by CAN (or MEN) to HIM.

It is instructive to see how the span changes as the network evolves. One expects it to increase with network size to some maximum value, probably achieved when the main network first forms out of islands, and then slowly decrease when there are no large islands left to annex and most of the words added to the network merely create alternate paths and short-cuts. However, as the network grows, the span may temporarily increase. This will occur if an island being annexed is moderately large, and the point of annexation is near the edge of the main network, so that a peninsula is formed.

In the evolutionary network diagrammed previously, NOR joins a 14-word island with a span of 9 (THE to FEW) to an 18-word island with a span of 7 (CAN to HIM), to form a 33-word network with a span of 14:

the-she-see-set-get-got-not-nor-for-far-war-was-has-his-him

There are 539 three-letter words in the Merriam-Webster Pocket Dictionary (excluding abbreviations such as DDT and TNT); the span of the main network has been reduced to 11:

ivy-icy-ice-ire-ere-err-ear-bar-bay-say-sky-ski

There are 907 three-letter words in the Official Scrabble Players Dictionary (OSPD). All but six of these--GNU, QUA, EBB, ISM, UGH, and OXY--are in the main network, and the span has been further reduced to 10:

ivy-icy-ice-ace-aye-tye-the-thy-try-fry-fro
ivy-icy-ice-ace-aye-tye-the-thy-try-pry-pro

Some paths of the network are extremely dense; for example, there are 33 words one step away from PAT, and 244 more two steps away.

Probably one-third of the network lies within two steps of the cycle PAT-PAY-SAT-SAY.

How do corresponding networks of four- and five-letter words evolve? Their behavior is very similar to the three-letter one: at first a large number of islets which coalesce into a few large islands, then a grand coalition of the major islands into a main network containing some 40 to 50 per cent of the words in the sample. However, these newly-formed main networks are much larger in size. The four-letter main network is formed when HOLE unites a 71-word island with a 56-word one, followed immediately by LOSE which unites the resultant 128-word island with a 27-word one. The newly-formed main network of 156 words uses 47 per cent of the 331 words sampled to that point.

As the four-letter main network is five times as large as the three-letter one, it is not practicable to show the detailed diagram of its growth here. However, one can capture the flavor of the network growth by noting those words which unite the network with islands. For example, in the three-letter network, WAY joins an island of size 5 containing WAS, to an island of size 3 containing MAY, to form a new island of size 9. By the time this has grown to size 13 by four single-word accretions (not specified below), the word FAR links WAR with the single-word island FOR to form a new island of size 15. The growth of the largest island is thus charted until it becomes so overwhelmingly larger than other islands that it can be fairly termed the main network (at NOR).

5 has (his 2, was 1, had 1)
9 way (was 5, may 3)
15 far (war 13, for 1)
33 NOR (for 18, not 14)
42 ran (man 40, run 1)
44 led (let 42, red 1)
52 sun (run 50, son 1)
59 bit (hit 52, big 1, but 5)
71 sex (see 61, six 9)
76 die (did 74, due 1)
80 buy (but 77, boy 2)
83 dry (day 81, try 1)
103 odd (add 101, old 1)
109 Bob (boy 106, job 2)
119 aim (aid 111, arm 7)
143 fog (fig 141, dog 1)
148 Tim (aim 143, Tom 4)

Note how, as the main network grows, the average size of the annexed islands decreases, and the time between successive annexations increases.

The corresponding list of annexations for the four-letter evolutionary network is given below. By the time the main network attains a size of 344 with FRED, it contains 67 per cent of the 513 sampled words.

4 gave (give 2, have 1)
11 live (five 5, love 1, like 4)
13 move (love 11, more 1)
17 fine (five 13, find 3)
27 game (gave 17, same 9)
31 firm (fire 28, form 2)
37 Mike (like 33, make 3)
44 nine (line 39, none 4)
52 wine (line 48, wide 3)
57 Rome (come 52, role 4)
65 wore (more 58, were 2, work 4)
128 HOLE (home 71, hold 56)
156 LOSE (nose 128, lost 27)
165 fort (form 159, sort 2, foot 3)
176 lake (like 168, late 7)
188 mile (mine 186, milk 1)
196 wave (have 192, wage 3)
199 load (road 197, loan 1)

201 meat (meet 199, mean 1)
204 sale (same 202, salt 1)
222 seed (need 216, seen 5)
257 fail (fall 233, mail 23)
268 fool (foot 258, pool 9)
273 mood (food 270, moon 2)
277 cash (case 273, wash 3)

293 vary (Mary 291, very 1)
298 Jess (less 295, Jews 2)
309 lane (land 306, Jane 2)
320 cure (care 317, curt 2)
341 flew (fled 324, flow 16)
344 Fred (feed 341, free 2)

The networks leading to HOLE and LOSE evolve similarly:

3 held (head 1, help 1)
7 hold (held 4, told 2)
10 read (head 7, real 1, road 1)
13 hear (head 10, year 2)
31 hell (held 17, hall 13)
38 fell (hell 31, feel 6)
51 beat (heat 39, boat 1, best 10)
56 text (test 54, next 1)

3 past (part 1, last 1)
9 lost (last 4, cost 4)
13 park (part 11, dark 1)
16 pass (past 13, mass 2)
18 loss (lost 16, less 1)
21 mark (park 18, Mary 2)
23 post (past 21, poet 1)

There are a number of small islands that join the main network later. The largest two are:

```
                when
what-that-than=then =thin-chin
                they this
                them
```

which joins to SHIP in the main network when CHIP, the 636th word, is sampled, and

```
               died-tied
view-Viet-diet<
               dies=ties=lies-lips
        goes=does=toes
               dogs
```

which joins to LOTS through DOTS, the 784th word in the sample.

At the time the main network is first formed with the aid of HOLE and LOSE, the span reaches a value of 23:

wait-want-went-west-best-beat-heat-head-held-hold-hole-role-rose-
lose-lost-last-fast-fact-pact-part-park-mark-Mary-many

However, the largest known span of 25 is achieved for the first time when VERY is added to the network and it reaches a size of 293:

very-vary-Mary-mark-park-part-pact-pace-pale-male-mile-file-fill-
fall-fail-fair-pair-paid-laid-land-band-bank-back-lack-luck-Lucy

Curiously, the 16-word island annexed by FLEW does not form a peninsula which extends span beyond 25. The span decreases to 14 by the time 3670 words in the OSPD have been sampled, as reported in the February 1989 issue of **Word Ways**.

The five-letter evolutionary network resembles the four-letter one, again scaled up by a factor of five or so. The five-letter main network is formed when BEATS unites a 383-word island with a 233-word one, BLINK unites this 617-word island with a 69-word one, and BLOTS unites this 687-word island with a 27-word one, to form a main network with 715 words, 43 per cent of the 1664 words sampled to that point. The full list of annexations is:

5 lives (gives 2, lines 1, lived 1)
8 loved (lived 6, moved 1)
18 mines (lines 9, minds 4, miles 4)
21 wines (lines 19, wings 1)
28 wives (lives 23, waves 4)
33 lover (loved 29, cover 1, lower 2)
41 males (miles 34, sales 6)
44 talks (tales 42, tasks 1)
47 tanks (talks 44, banks 2)
54 balls (bills 47, calls 6)
58 liver (lived 56, river 1)
60 rider (river 58, wider 1)
70 lever (liver 64, level 1, never 4)
87 files (miles 72, filed 14)
89 Sally (Wally 87, silly 1)
96 maker (makes 94, Baker 1)
98 sadly (sally 96, badly 1)
100 tires (fires 98, times 1)
123 bands (banks 100, bonds 5, hands 17)
126 wired (tired 124, wiped 1)
147 mates (males 127, rates 19)
155 rides (rider 148, rises 1, sides 5)
160 sings (wings 158, songs 1)
172 cares (cared 162, cards 2, cases 7)
175 filly (Billy 172, fully 2)
180 pants (parts 178, wants 1)
187 timed (times 182, aimed 4)
191 belts (bells 189, pelts 1)
210 domes (doses 193, homes 16)
218 merry (marry 213, mercy 1, Jerry 3)
221 modes (moves 219, codes 1)
224 model (modes 221, motel 2)
233 forks (forms 231, folks 1)
239 mails (fails 237, maids 1)
246 molds (holds 239, moods 6)
249 paces (faces 247, paced 1)
252 tunes (tones 249, dunes 1, tubes 1)
284 casts (cases 262, costs 21)
295 noses (loses 290, notes 4)
302 towel (tower 300, vowel 1)
305 baths (Bates 303, paths 1)
331 hairy (Harry 316, dairy 3, hairs 11)
345 Mayer (maker 334, mayor 4, Meyer 3, layer 3)
353 rains (gains 349, reins 3)
357 roofs (roots 355, hoofs 1)
363 sting (stint 360, swing 2)
368 vowed (vowel 365, bowed 2)
372 Willy (Billy 370, Wiley 1)
381 barns (yarns 375, burns 5)
617 BEATS (belts 383, bears 233)
687 BLINK (blank 617, blind 69)
715 BLOTS (plots 687, blows 27)
717 Bruce (truce 715, brute 1)
719 bully (Billy 717, bulky 1)
722 capes (cares 719, caper 2)
726 crush (crash 724, brush 1)
732 disks (risks 730, discs 1)
741 dross (cross 735, dress 5)
744 ducks (bucks 741, ducts 1, decks 1)
757 forts (forms 748, sorts 1, forth 7)
769 groin (grown 760, grain 8)
773 hates (dates 771, Hayes 1)
779 knack (snack 777, knock 1)
789 lends (lands 781, tends 7)
794 poked (poker 792, posed 1)
803 rover (lover 801, Roger 1)
827 stale (stole 814, scale 1, stall 11)
830 tapes (tales 828, types 1)
833 texts (tests 830, Texas 2)
837 vents (tents 835, Venus 1)
845 babes (Bates 842, Babel 2)
850 Barry (marry 845, barre 4)
860 Beame (beams 851, blame 8)

The largest island to join the main network later contains 27 words:

```
fetch-letch
     latch=match=catch=watch=patch=batch=hatch
           march             pitch =bitch =hitch =ditch
           marsh             pinch              Dutch
           harsh             punch=hunch=bunch
                                         bench       Reich
                                         beach=teach=reach
                                                     react
```

This island joins to PEACE in the main network via PEACH.

By the time BEAME (the surname of a former New York City mayor) is reached, the main network still contains less than half of the words sampled (46 per cent).

At the time the main network is formed with the aid of BEATS, BLINK, and BLOTS, the span reaches a probable maximum value of 52:

bored-bowed-vowed-voted-noted-notes-noses-roses-roles-poles-polls-pills-bills-bells-belts-beats-seats-seams-seems-stems-steps-stops-shops-shots-scots-Scott-scout-shout-shoot-shook-shock-stock-stack-slack-black-blank-blink-blind-blond-blood-brood-broad-bread-breed-creed-creek-cheek-cheer-sheer-steer-steel-steal-steam

The span is 48 at the time BEAME is added to the network. As reported in the February 1989 **Word Ways**, the span decreases to 29 for the 8200+ five-letter words in the OSPD.

Here is a summary of the span behavior in the evolutionary network as a function of word length:

| | 3 | 4 | 5 |
|---|---|---|---|
| Span at time of coalition | 14 | 23 | 52 |
| Largest-known value of span | 17 | 25 | 52 |
| Span for OSPD | 10 | 14 | 29 |
| Ratio of largest to OSPD span | 1.70 | 1.79 | I.79 |

As the word length increases, the span at the time of coalition seems to approach the maximum span in value.

Finally, one should note that the Kucera-Francis ranking of words in decreasing order of frequency is to a considerable extent dependent upon the sample of words drawn -- a different million-word sample would change the relative ranking of all but the few dozen commonest words. There is no particular significance in the fact that NOR, HOLE, LOSE, BEATS, BLINK, and BLOTS were the essential links that first formed the main networks; in another sample, other words would assume this role. Nevertheless, it is likely that the statistical features of this study -- the size of the maximum span, the percentage of the sampled words included in the main network when it first forms, etc. -- would be approximately the same if a new sample were taken. In fact, one might even be able to simulate the behavior of evolutionary word networks by taking random samples of various sizes from the OSPD word lists. This is, in fact, the only way available to study evolutionary networks with words of six letters or more, for the Kucera-Francis sample is not large enough even to show the coalition of islands into a main network.