A WORD STRING NETWORK

A. ROSS ECKLER
Morristown, New Jersey

Logologists have been aware of word strings -- overlapping lists of words such as sat, ate, tea, ear, are -- since at least the time of the great English puzzle-constructor Henry Ernest Dudeney. However, no one seems to have realized that word strings can be diagrammed in a network, much as has been done for word ladders (see, for example, the May and August 1973 issues of **Word Ways**, or Chapter 4 of my book Word Recreations, published by Dover in 1979).

However, there are distinct differences between networks based on word ladders, and those based on word strings. Specifically, word string networks allow only those transitions in which a letter is removed from the beginning of a word, and another letter (or, possibly, the same letter) is added to the end to create a new word. Thus, each word in the network may have one or more descendant words which can be reached from it by succesive transformations of this type, one or more ancestor words which can reach it, and one or more words that are inaccessible (you can't get there from here). Some words may have no ancestors; others, no descendants; and sometimes a word can be its own ancestor or descendant, by virtue of being in a loop such as asp, spa, pas.

One can best appreciate the complexities of this network (called a directed network by graph theorists) by looking at an example. The one chosen for this article is the word string network based on the 539 three-letter words listed in boldface type in the 1974 edition of the Merriam-Webster Pocket Dictionary (eliminating from consideration abbreviations such as DDT and TNT). Of the 539 words, 360 are in the main network. This is far too complex to diagram in full, but its essentials are depicted on the next page. All principal thoroughfares are shown, but a number of one-step side streets have been omitted to avoid cluttering the diagram. For example, the word ant is shown as having three immediate ancestors, **ran**, **van** and pan, which in turn have ancestors of their own. However, the complete network should also show **ban**, can, **fan**, **man**, tan and **wan**, none of which has an immediate ancestor. Similarly, the word **boa** has immediate descendants **oat** and oar shown, but oak and oaf, both dead ends, have been omitted. Note also that **has** and **gas** are ancestors of asp in common with **was**; fro, with **pro**; ply and fly, with **sly**; and cry, **fry**, **pry**, try and **wry**, with **dry**.

All routes in the network should be followed from left to right. A word with no ancestors other than one-step ones (as mentioned

above) will have no lines emanating to the left, either directly or on a diagonal, and a word with no descendants will have no lines emanating to the right. The exceptions to this convention are those groups of words forming a closed loop. These are depicted in boldface in a vertical list, with vertical arrows indicating the direction to proceed; when the word at the bottom is reached, one can jump back to the top. There are four such loops: **asp-spa-pas**, **ate-tea-eat**, **era-rap-ape-per**, and **ran-ant-nth-the-her-era**, the latter two having one word, **era**, in common.

To unclutter the network further, a number of terminal subnetworks, printed on the page facing the main network, have been replaced by capitalized words, followed by the number of steps from those words to the end (yes, end is an end). Once one has reached a capitalized word, there is no way back into the main network.

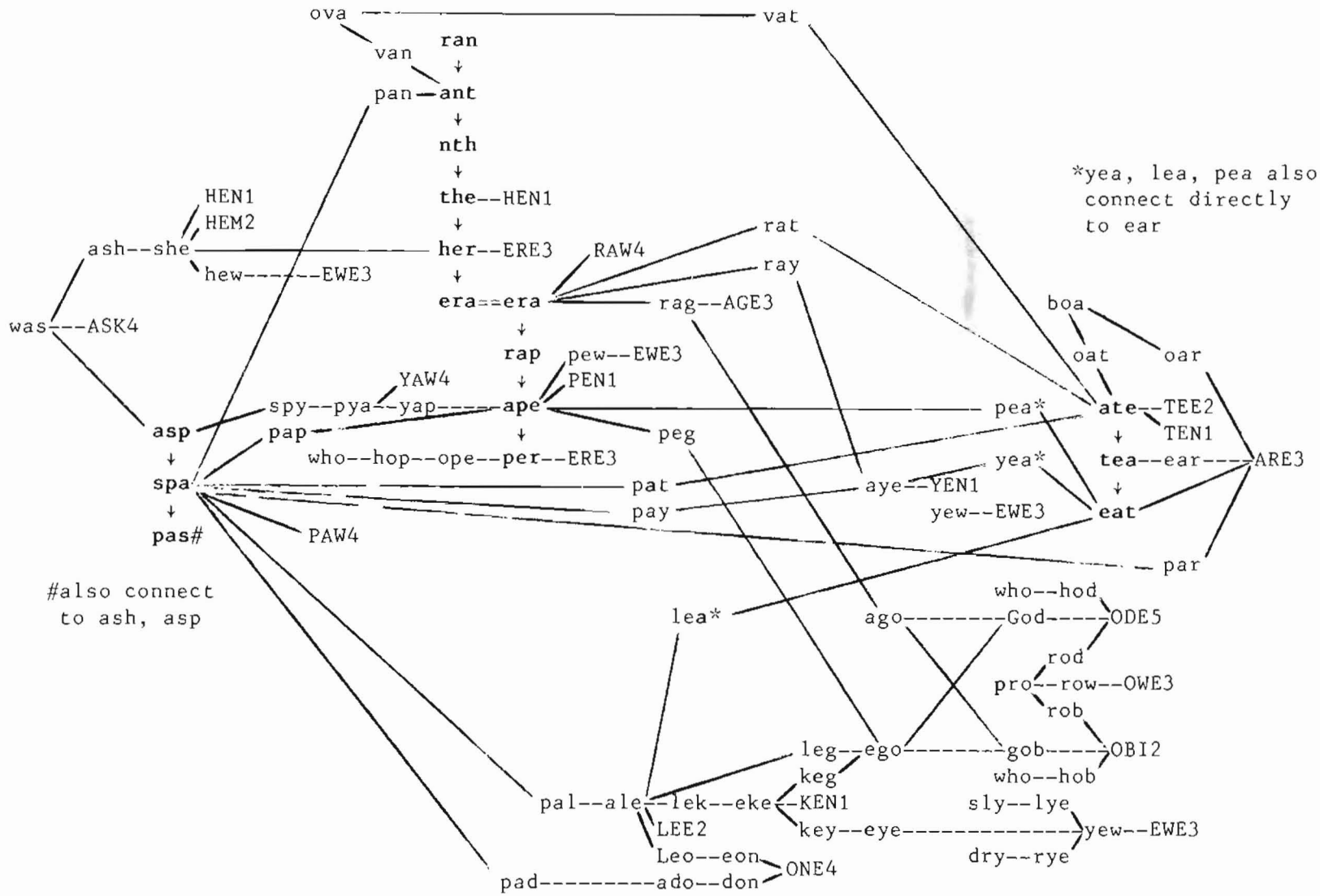
Most of the 179 words not in the main network are isolanos, unconnected to any other ancestor or descendant word. There are only ten groups of two or more words:

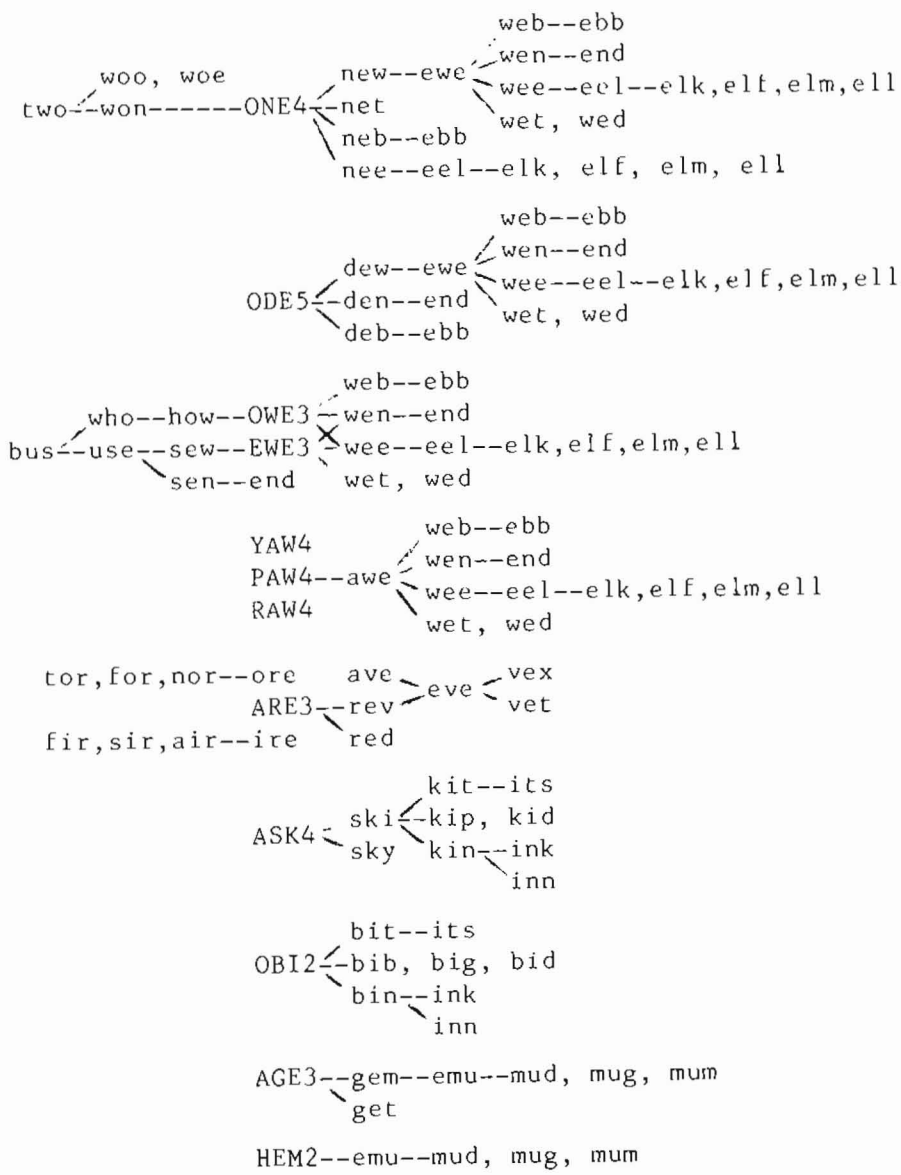
cur bur you-----our>urn out fur	ail,mil,nil,oil--ill,ilk	his--ism
	aim,dim,him,rim,vim--imp	flu--lug
	gnu--nub,nut,nun	sol--old
lax,tax,wax--axe	lac,sac--ace,act	tic--icy,ice

One can determine the span of a directed network just as one can for a standard (two-way streets) network. For each word and a descendant word, there is a minimum-length path of intermediate words joining them; this path is characterized by the positive integer n, one more than the number of intermediate words needed. Different pairs of words in the network generate different values of n; the span of the network is defined as the maximum value possible. In the main network depicted above, the span is 12, as illustrated by the string **ova-van-ant-nth-the-her-era-rag-ago-gob-obi-bin-ink**.

The word having the most descendants is **was** (or **has**, or **gas**) with 158, about 44 per cent of all the words in the main network. The word having the most ancestors is **end** with 195, although **elf** (**ell**, **elm**, **elk**) gives it a good run with 190. **Ebb** is also a popular ending, but **inn** (or **ink**) and **mum** (**mug**, **mud**) are much less richly endowed with ancestors.

Rather than tour the main network as efficiently as possible, one can ask for the longest possible route between a pair of words with no words used more than once. The string of twenty words **was-asp-spa-pan-ant-nth-the-her-era-ray-aye-yea-eat-ate-tea-ear-are-rev-eve-vex** appears to be the best possible; however, twenty can be achieved in other ways as well. If one insists that each letter of the alphabet be used at most once, the string of nine words **spa-pal-ale-leg-ego-gob-obi-bin-ink**, reported in the May 1979 **Word Ways**, is almost certainly the longest.


ova
vat
van
ran
pan
ant
nth
the--HEN1
HEN1
HEM2
ash--she
her--ERE3
hew------EWE3
RAW4
rat
ray
era==era
rag--AGE3
was---ASK4
rap
pew--EWE3
PEN1
YAW4
spy--pya--yap----ape
pea*
asp
pap
peg
who--hop--ope--per--ERE3
yea*
spa
pat
aye--YEN1
pay
yew--EWE3
pas#
PAW4
par
boa
oat
oar
ate--TEE2
TEN1
tea--ear----ARE3
eat
*yea, lea, pea also
connect directly
to ear
#also connect
to ash, asp
lea*
who--hod
ago--------God-----ODE5
rod
pro--row--OWE3
rob
leg--ego--------gob-----OBI2
keg
who--hob
pal--ale--lek--eke--KEN1
sly--lye
LEE2
key--eye----------yew--EWE3
Leo--eon
ONE4
dry--rye
pad---------ado--don



TEE2--eel--elk,elf,elm,ell

LEE2--eel--elk,elf,elm,ell

TEN1--end

HEN1--end

PEN1--end

YEN1--end

KEN1--end

One can conceive of analogous word string networks for four-letter or longer words; however, unless a very much larger dictionary is used, such networks are likely to be extremely fragmentary. One way around this difficulty is to allow less-complete overlaps, such as placing two letters at the end of a four-letter word after removing the first two letters.

Word strings can be written in compact form, such as **ovantheragobink** for the spanning string cited earlier. These "words" can be considered in their own right, in particular as the vocabulary for a game of super-ghost (players take turns adding letters at the front or back of a string of letters, the object being to force one's opponent to complete a "word"). To avoid trivializing the game, one must insist that only "words" from the main network count; otherwise, the person making the second move can nearly always find an isolano to force a win. For example, if the first person picks M, the second adds I, leading only to MIX, MIL, or MID). Further, provision must be made to forbid endless looping of SPA, EAT, APER and ANTHER. There is presumably a set of "safe" letters for the first player to start with to ensure a win, but ascertaining these seems to be a task for a digital computer.

NINETY MILLION NAMES

The business section of the April 14 1991 New York Times contained a short article of considerable interest to well-heeled logologists, genealogists, and tracers of missing persons: the Phone Disc USA Corporation, a small company in Marblehead, Massachusetts, plans to offer in September 1991 the white pages of all 4000 US telephone directories for $1850 on two CD-ROM disks. With the aid of a CD player (a purchase of several hundred additional dollars) the information can be extracted via personal computer. For logologists, this contains a treasure-trove of words not available in dictionaries; according to Social Security records, there exist some 1.1 million different surnames in the United States.

Not quite ready to plunk down two thousand dollars or more? Phone Disc has already sold an earlier version of these disks to about one thousand libraries and bill collection agencies. If you can find a library near you that has this technology, perhaps you could arrange to try it out, finding out the types of questions that the system can answer (and the time it takes to extract answers from 90 million entries). What single-letter surnames exist? What is the last surname alphabetically, excluding those ZZZZ vanity hisses? Does there exist a surname with the bigram FX (probably the most recalcitrant one)? Let Word Ways readers know what you find!