ELECTRONIC WORD SEARCH PROGRAMS

ANTHONY SEBASTIAN
San Francisco, California

Spelling anxiety is no affliction for writers who use electronic typewriters or personal computer (PC) word processing programs that provide so-called spell checking (see "The Electronic Speller" by Faith Eckler in the May 1988 **Word Ways**). Writers can choose to be notified (via beep) immediately after typing a word incorrectly, or they can invoke global spell checking after completing a manuscript. Most built-in spell checkers both identify a misspelled word and try to "guess" the word intended, offering a candidate list of correctly spelled alternatives from a database ranging in size from 40,000 to over 100,000 words.

Choice Words

Proximity Technology Inc., in collaboration with Merriam-Webster, Inc., has now evolved the spell checker into a true electronic dictionary that supplies definitions as well as spell checking, for about 80,000 words of American English. Other useful information about words for writers is also provided, including hyphenation points, every part-of-speech the word acts as (with appropriate definitions for each), inflected and derivative forms, including derivative phrasal expressions, and occasional usage notes. The dictionary is available as a program called Choice Words, designed for IBM PCs and compatibles. It can be used as a stand-alone reference dictionary, or can be loaded along with the writer's PC word processing program to supplant or supplement the word processor's built-in spell checker.

One feature of Choice Words will be of particular interest to logologists. As a special adaptation of the typical spell checker feature that supplies "guesses" for the word intended by an incorrectly spelled entry, Choice Words will search the dictionary and supply a list of candidate words for an entry word that has one or more letters deliberately unspecified. The user marks the location of each unspecified letter with a ?, called a wild card. With certain limits, the list of candidate words supplied comprises all the dictionary words that satisfy all possible permutations of letters at the locations of the unspecified letters, but only those that contain the specified letters in their original locations. The length of each displayed word equals that of the entry word.

A few examples demonstrate the feature and its logological utility:

(1) A crossword puzzler, seeking a five-letter word ending in M with middle letter L (i.e., --L-M), meaning woody plant tissue, enters ??l?m and receives the following list:

Islam
Salem
xylem

Selecting a likely candidate, **xylem**, simply by positioning a keyboard-controlled pointer on the word and hitting the Enter key, causes the program to look up the word's definition, after a few seconds displaying: noun--woody tissue in higher plants. Entering ???em instead of ??l?m (a different puzzle instance) would have returned the list **Salem, harem, modem, proem, totem, xylem**--again with immediate access to the definitions of the words in the list (e.g., a **proem** is a preliminary comment, or prelude). The entry ?y??m yields **xylem** only.

(2) Another crossword puzzle application: Find a word of the form ---x--e meaning aluminum ore. Entering ???x??e yields **bauxite, dioxide, flexure**. Quick definition check confirms correct choice. Takes the fun out of solving crossword puzzles but facilitates constructing them.

(3) To fill a gap in a tetragram table, a logologist needs a word containing the consecutive letters rstu. Restricting oneself to words of 15 letters or less, it takes less than ten minutes to systematically search the dictionary (?rstu; ?rstu?, ?rstu??..., ??rstu, ??rstu?, ??rstu??..., ???rstu, ???rstu?, ???rstu??...) to find only one result, **understudy**. That's less time than it takes to search back issues of **Word Ways** for the answer.

(4) **Bamboo** is a pattern word coded 123144; find its mates. Using the electronic dictionary, the base word can be thought of as having the form #??#??, where # is a letter to be specified in the entry. We make 26 systematic entries: a??a??, b??b??, c??c??,... z??z??, scanning the returned candidate list of each entry for words ending in doubled letters. We find two mates, **egress** and **excess**; the entire search takes a few minutes.

Many logological activities in addition to those mentioned involve searching the dictionary for words of a specified pattern in which one or more letters are unspecified. Building word networks and ladders of various types, and word squares and diamonds and the like, are additional examples. Such activities likewise would be facilitated by the wild card look-up feature in Choice Words.

Further exploration of the wild card look-up feature reveals one limitation that renders it less than ideally convenient for logologists. If a wild card search turns up more than 12 hits, only the first 12 are displayed, arranged alphabetically by first letter. The remaining hits can be captured, but to do so requires one or more narrowing search requests. Thus, the reader can still be certain of finding all qualifying dictionary words for an entered pattern, but not always with a single search request.

For example, entering ???ke produces a 12-member alphabetically sorted candidate list, the alphabetically last member of which is **quake**. The fact that the list is 12 words long indicates that there may be additional hits beyond quake. To find out, and to get those additional hits, requires nine successive entries, r??ke, s??ke, t??ke, u??ke, v??ke, w??ke, x??ke, y??ke, and z??ke, to ensure that all letters of the alphabet after Q are covered. None of those subsequent entries except s??ke turn up hits; s??ke reveals eight

additional hits, giving a final return of 20 (or 19, if the suffix **-like** is omitted). The entire search is completed in a few minutes.

???ke: -like, alike, awake, awoke, brake, broke, choke, drake, evoke, flake, fluke, quake
s??ke: shake, slake, smoke, snake, spike, spoke, stake, stoke

Despite that inconvenience, the search for groups with more than 12 members sharing common letters is greatly facilitated by comparison with the tedious process of visually scanning a printed dictionary. Hopefully, the program developers can be persuaded to improve the search convenience in a future version of the program.

According to the Choice Words manual, Merriam-Webster supplied all of the lexical data for the dictionary, referred to as Webster's Electronic Dictionary, Concise Edition. A printout of the dictionary's word list is not supplied, and the program offers no option for displaying or printing the entire word list at one time. Beyond stating that the dictionary comprises "almost 80,000 words", little information is provided about its lexical base. What constitutes a "word"? Are there 80,000 stems, or 80,000 combined stems and inflected forms? Proper nouns are included, as well as abbreviations and acronyms, but no information is given as to their number.

A random sampling provides a peek at the lexical database. Among the guide words at 50-page intervals in Merriam-Webster's Ninth New Collegiate Dictionary, exclusive of unhyphenated compounds, the following 28 words were present: ace, achromaticity, archery, biodegradable, cannibalize, cow, crab, diary, electrician, electronics, father, future, gaff, handset, imbecility, immigrant, jolter, local, micrometer, papal, reproach, request, sink, tail, trammel, v, vain, wind. Twenty more were not present: areopagitic, biltong, candida, close-grained, cloven, dichroscope, fashionability, Hansa, micturition, nodus, nomological, panleukopenia, plastron, platyrrhine, psychokinetic, Ptolemaic, schnorrer, scilicet, -sis, taco.

If these results are representative, the electronic dictionary lists about 55 to 60 per cent of the words in the Ninth New Collegiate, favoring words with higher frequency of use. That accords with the book jacket tally of the Ninth's word count.

Additional information about the base of Webster's Electronic Dictionary, Concise Edition would be of considerable interest inasmuch as Choice Word's combined offering of word definitions and wild card searching might popularize the dictionary among PC-using logologists, just as convenience popularized Merriam-Webster's Pocket Dictionary in the era before personal computers. Of course, with concerted effort, the wild card search itself could be used to reveal the dictionary's word list, so the lexical base of the dictionary may emerge out of future research.

WordPerfect

Choice Words is not the only PC program that permits wild card dictionary look-up. While not supplying definitions, the built-in spell checker of a PC word processing program may also offer wild

card searching. Indeed, WordPerfect 5.0, one of the most popular and powerful PC word processing programs, offers a more powerful look-up feature and a larger word list than that of Choice Words (115,000 vs. 80,000 words).

WordPerfect offers two types of wild cards for searching. As in Choice Words, ? specifies "any single letter". In addition, however, * can be used as a wild card to specify "any number of letters" (e.g., *x returns all words ending in X regardless of length: accusatrix, acronyx, adieux,..., hydropneumothorax..., wax..., xerox.

The addition of an "any number of letters" wild card greatly increases the utility of the search capability. For example, the single search request *rstu* accomplishes the same thing as the multiple requests that Choice Words required. Thus, finding all words that contain any permutation of letters is practically instantaneous. (In addition to **understudy**, WordPerfect yielded an additional word with the rstu tetragram, **overstuffed**, which is not contained in the Choice Words dictionary.)

WordPerfect allows use of the two wild cards * and ? in a single specification of a word pattern, further increasing the flexibility of the look-up feature.

In further contrast to Choice Words, WordPerfect does not require multiple searches to capture all hits of a specified pattern when the number of hits exceeds a certain limit. A single search request returns all qualifying words, regardless of number. Indeed, the number of hits can equal the total number of words in the dictionary, a circumstance that obtains if the search request consists of the single character *. When a search turns up more than 24 hits, the display pauses after showing the first 24 words, whereupon the user is prompted to hit any key to view the next block of 24 words. An option is available to allow display of 51 hits between pauses.

The nearly 50 per cent greater number of dictionary words in WordPerfect than in Merriam-Webster's Electronic Dictionary further commends it to logologists. Of the 20 guide words from the Ninth Collegiate not found in Choice Words, half were found in WordPerfect's word list, making the overall success rate close to 75 per cent. Unfortunately, the WordPerfect manual supplies little information about the lexical base of the dictionary. However, since the * wild card conveniently reveals the entire word list, it would not be a daunting research project to characterize the database.

Of potentially great utility to PC-using logologists, WordPerfect allows the user to add words to the main dictionary, either one at a time from the keyboard or en masse from a previously prepared list typed as a WordPerfect document. Thus, the WordPerfect word list can grow in size indefinitely (limited by disk storage capacity). The added words become an integral part of the database, and the wild card look-up feature then searches the expanded word list without discriminating between original and added words. Words can also be deleted from the database. With a little effort,

therefore, a logologist could customize the word list to make it identical to that of, say, a specified standard collegiate dictionary.

The major disadvantage of WordPerfect over Choice Words is cost ($250 to $300 vs. $85 to $99, at discount prices).

The Electronic Oxford English Dictionary

If cost is not limiting, the PC-using logologist might want to own the PC version of the original 12-volume Oxford English Dictionary (OED). Vocabulary entries, definitions, etymologies, quotations, etc., are all displayed--color-coded for easy discrimination -- with the special characters used in the printed text displayed as such or in coded form. The dictionary is available as a read-only laser disk (Tri-Star Publishing, Fort Washington, PA; list price $950). Accessing the disk requires a so-called CD-ROM disk drive reader (price $700+) interfaced with the PC (see review by Edward Mendelson, PC Magazine, Jan 31 1989, p. 219).

The PC-OED comes with a software program that allows complex search strategies, including, for example, searches for all instances of a specified word in quotations, etymologies, and/or definitions, all instances of a specified word in quotations in a certain date range (e.g., 17th century), and wild card searches such as all entry words ending in a specified letter. There are many others.

As I was not able to personally test the PC version of the OED, I cannot describe the capabilities and limitations of its wild card search facility, or compare it with that of Choice Words and WordPerfect, which I did test personally. A full investigation of the logological potential of the PC version of the OED would be of great interest.

Prospects

Logologists who are not computer experts but who use a personal computer can now avail themselves of the computer's power to explore and query large word lists and dictionaries, including the search for words of specified patterns. While for some that prospect trivializes many challenging recreational linguistic activities, for others it promises a world of novel linguistic challenges and recreations.