Prizes and Pitfalls of Computerized Searching for New Words for Dictionaries

David K. Barnhart

Dictionaries: Journal of the Dictionary Society of North America, Number 7, 1985, pp. 253-260 (Article)

Published by Dictionary Society of North America
DOI: https://doi.org/10.1353/dic.1985.0030

PRIZES AND PITFALLS OF COMPUTERIZED SEARCHING FOR NEW WORDS FOR DICTIONARIES

David K. Barnhart

Picture Sir James Murray seated in his editorial office pondering the worthiness for entry of the word *appendicitis*. This should present no problem for him. For, with hindsight, we can see its obvious importance; after all, it is commonplace today even in school dictionaries. In the last five years of *The New York Times* and the last thirty months of *Time*, seventy articles have at least one mention of *appendicitis*. Surely Sir James would enter it. However, Sir James was not entirely his own man. Only eight years before the Vice Chancellor of Oxford University had sent Sir James a set of "suggestions" to help the editor of the dictionary. These suggestions offered the observation that slang and technical words might be included only if they were used in literature. Sir James foresaw the basic problem: language is unpredictable. Beyond the fact of life obvious to us that language *will* change with or without our approval, there is no way to foretell what inventions or fads will impose upon us the urge to invent new words. Murray's quandary started because he had the suggestion that *appendicitis* was a "crack-jaw medical word." To cure his case of *appendicitis* he consulted the Regius Professor of Medicine at Oxford University. *Appendicitis* was consigned to oblivion. A scant eleven years later Edward VII's coronation was delayed: he had appendicitis. The word was on everyone's tongue. Had Sir James had a crystal ball, he would surely have included *appendicitis* in the *OED*.

He might have been unable to resist the urge to include *aeroplane, DDT, penicillin*, or such words from the 1980s as *AIDS, break dancing, computerist, digital disc, exit poll, free-base, gridlock, the hots, ice dancing, jellies* (the shoes), *kneebreaker, laser disk, megatrend, neoliberal, option mortgage, prenatal brain shunt, quality circle, rockvideo, skunkworks, telemarketing, unbanning, videodisk, worksite, Yuppie*, and *Z particle*. There is no telling what will happen. Dictionary editors must guess what will be important, what new

expressions will stick in the language. "The creative act that doesn't respond to some kind of social need isn't going to get picked up," according to Victor Golla, Professor of Sociology at Georgetown University.

There are several problems in measuring the growth of the English language. One is the sheer magnitude of the written word, not to mention the billions of words that are spoken each day. To grasp the notion of how great the task might be, Clarence Barnhart took all the magazines at hand for May 1979 and searched them for words and meanings not covered in *The World Book Dictionary*—about 250,000 entries. He found 1,000 items in need of updating or inclusion. One might justifiably guess that there is a quantity of new words or meanings somewhat in excess of 12,000 each year. Perhaps that figure is too small; more sampling is bound to reveal new items.

To undertake the search for new words always brings up the fear that in this fishing expedition the good ones are somehow getting away. This editor gave each of six researchers identical copies of *Time* for one week. They searched independently for neologisms. The grand total was over 300; no individual found even half that number. In searching in a computer data base for *skybox*, a term I confidently felt was brand new, I discovered that it had been eluding my attention for ten years.

Basic research for evidence to substantiate judgments found in a dictionary based upon original research has traditionally been accumulated on a search-and-find basis. Each example of an item must be found, marked, typed or photocopied, classified, and filed. This meticulous process when intensively practiced by a small staff can yield about 30-50,000 citations per year with a modest budget of about $40,000. In a paper given just a few years ago I outlined the magnitude of the job of sampling changing English usage on a search-and-find basis ("Citations for the Makers and Users of Dictionaries," 1978). In that paper I reported that *Publishers Weekly* estimated that in the United States over 25,500 new books were produced in 1971; by 1974 that figure had grown to 40,800. *Ulrich*

Periodicals Index records a score of "general and popular" weekly or monthly publications for Australia alone. In a year of *Time* magazines there are 2 1/2 million words of text, in *The New Yorker* 2.6 million words, in the *Manchester Guardian Weekly* 3 million words. Ten such sources could quickly add up to over 210 million words of text to survey. With several dozen books at 200 pages per book and 250 words per page the figure would grow another 5 million words. It is evident that the problem is where to stop, not where to start.

Now this problem has grown to the point that *Publishers Weekly* reported for the year 1983 a total of over 53,000 new titles. To compile the five thousand entries in *The Second Barnhart Dictionary of New English* (1980), which covered a selection of the neologisms from 1973 to 1980, a trained staff collected about 250,000 citations. Each citation was the product of search-and-find. The ratio of citations to entries is a staggering 50:1. Out of 50 citations one dictionary entry was forthcoming. The first issue of *The Barnhart Dictionary Companion*—a quarterly journal of neologisms—appeared in the winter of 1981-82. Over 2,000 words have been recorded in its pages that were not available in the dictionaries of record (*OED* and *W3*) or in the dictionaries of new words. The dramatic difference between the dictionaries of new English and the journal is that for the quarterly *Dictionary Companion* there is no file of 100,000 citations, at least not in the old-fashioned sense of a file. The process of procurement for the citations in the first year and a half of the *Dictionary Companion* was search-and-find. For many years computerists have been perfecting their engine to the point that this dictionary editor could no longer ignore the benefits of their seemingly magical kingdom.

Since late 1982 Lexik House has subscribed to Nexis, a data bank encompassing the full text of articles in over 135 publications, comprising more than 14 billion words of text. It has newspapers, such as *The New York Times, The Washington Post, The Christian Science Monitor*, and *Manchester Guardian Weekly*. It has magazines, such as

Business Week, Fortune, Inc., Newsweek, Sports Illustrated, U.S. News & World Report, Discover, Time, and *The Economist.* It has, also, wire services and special subject newsletters. Access to Nexis has allowed us to multiply the yield of search-and-find procurement dramatically. When the *Dictionary Companion* project began it was supposed that fifty entries per quarter would be a manageable quantity. And it was. With Nexis it is possible to enrich that yield so that the *Dictionary Companion* now provides in excess of 250 items per quarter. With the right questions to the information utility's data bank a researcher can find incredible amounts of data in the twinkling of an eye.

The trick of the trade in finding neologisms or other linguistic data in data banks is to create the right questions. Without skill in composing the right questions a researcher of neologisms could quickly sink his or her project with intolerable indebtedness. Large flexible data bases tend to be very expensive. The first example of *post-punk* that I saw was found by one of my researchers in *The New Yorker* in a column describing jazz groups then playing in New York City. The example was for the adjective use. It was the only example. With Nexis the data revealed the variant spelling *postpunk.* Also delivered by the terminal was enough evidence to substantiate four noun meanings for *postpunk* in addition to the adjective meaning already in hand. Also discovered was *post-punker* and its variant, and *pre-punk* and its variant. Just as valuable was the information that if **pre-punker* or **prepunker* did exist, they were not recorded in the sources available through Nexis. A similar story underlies *greenmail, rap music, Star Wars, yuppie, barn burner, buy-down,* and even *xerox* spelled *zerox.*

One of the more interesting problems uncovered with the power of Nexis is seen in the entry for *close, but no cigar.* Searching for it in Nexis introduces an interesting aspect of the data bank that turns out to be at once vexatious and magnificent. For in searching the problem of the unsearchable arises. Nexis will not search some words: *and, but, with, not, his, or,* and so forth. As a consequence, the following appeared:

close and no cigar. This led to reformulating the question, which revealed: *close means no cigar* and *close but few cigar*.

So blind is Nexis to syntax and semantics that end of paragraph and end of sentence remain undistinguished from middle of the sentence position, largely because punctuation is ignored. In searching for *prayer watch*, the noun phrase, I found: "And the Rev. Frances Lawlor of the Rockford diocese led the group of several hundred in prayer. 'Watch over John Paul your servant . . . . ' " But this blindness to periods is not so troublesome as the leveling of phrasal boundaries. I found when looking for *time bank*, noun phrase: "At the same time bank employees in the public sector . . . ." When looking for *round file*, in the sense of wastepaper basket, I received from the data bank: "Final smoothing can be done (if necessary) by using a half-round file or round abrasive stone."

Another blind spot is the leveling of plurals in *-s* and *-es* and *-ies*. The suffix *-en*, as well as the latinate plurals and other irregularities, is not, however, so leveled. When searching for the acronym MIPS, meaning "million instructions per second," I retrieved every example of *mip* as well. The meanings encountered showed at least one usage for each of the following: millions of instructions per second, missile impact predictor, methods improvement department, magnetic induced polarization, and *Dutch* Society for industrial projects. Furthermore, the plural of *bonus* is not called up, just as other singular forms ending in *-s* are not retrieved with their plural *-es*.

Still another blind spot is the loss of distinction between capital and lower case letters. *TV* and *tv* are not differentiated. The combination of this lack of differentiation with the Nexis blindness to punctuation produces for the researcher, without his asking, *T.V.* and *t.v.* along with *TV* and *tv*. This situation was of tremendous benefit in researching *v.j.* for "video jockey." Without expecting it, the researcher finds *v.j.*, *V.J.*, *vj*, and *VJ*. However, in the case of both *v.j.* and *t.v.* one must remember that *T.V.* may stand for initials, as in the name T. V. Smith, an Illinois politician in the 1930s. And VJ will

appear in the expression VJ Day. Furthermore, Nexis lumps some abbreviations and nouns. *T.V.* and *television* is an example. Thus, one could get *television smith.* Beneficially, this produces both *projection television* and *projection t.v.* in the same search.

Hyphens are not leveled with one word forms, but they are with two word forms. So in the search for *neoliberal* one does not get *neo-liberal.* However, in the search for *gado-gado* one does get *gado gado.* And in searching for *figure of merit* one also gets *figure-of-merit.* The nature of the beast also lumps numerals and numbers spelled out. Thus, one retrieves in a search for *Chicago Seven* "$10-23 in Chicago and $7-$20 in Atlanta . . . and "In Chicago, where we have seven lines of business . . . ." Similar things happen when looking for *Wilmington Ten* and *Charlotte Three.* One can imagine what happens when Wilmington plays Charlotte in a game of baseball. We have seen how the nature of what I have chosen to call leveling can bring forth benefits as well as difficulties.

Computer searching has brought about another benefit. Perhaps one of the more exciting innovations in lexicography to come from Nexis via the *Dictionary Companion* is the designation of frequency in the descriptions of usage. With each search, among the data shown is a designation of the number of articles in which the search item is to be found. The first impulse in reporting this information in the *Dictionary Companion* was to give raw data, numbers straight from the computer terminal evidence. However, with some complicated searches this would be virtually impossible. With an item like *close, but no cigar* there are so many "junk citations" that it is unreasonably expensive to do a precise count. We settled on an interpretive scale that gives a relative indication of frequency. The scale adopted is in three stages: 100 plus "frequent," 10-99 "common," and 0-9 "infrequent."

One problem we encountered was how to modify this system to fit the situation when related terms or meanings of particularly wide difference in our scale fall under the same designation, e.g., common. If one aspect of usage is

represented in eighty articles, it would be misleading to lump it under the same designation as a related one with fifteen articles. In these situations we have introduced modifiers such as "very," "less," or "more," as the case may warrant. When national boundaries suggest special problems we try to indicate usage as we can measure it. Some of the problems that arise in counting we have already touched upon. Remember that *T.V.* can be the initials of T. V. Smith. In matters of frequency the greater problem arises with not knowing whether a term for which we wish to search can be a noun or a verb. Just as plurals of nouns in *-s* are not distinguished in searching from singular forms, so third person singular verb forms in the present tense are not distinguished from uninflected forms. To find the examples of a verb one must remember this in framing questions.

Another frequency headache is the new meaning or part of speech for a very common word. When *cruise missile* is reduced to *cruise* or the computerist's usage of *implementation* is the subject of a search, the pursuit of this data will prove to be intolerably expensive. When searching for the uninflected verb form of *limousine* one would have to search through more than 5,568 articles, probably requiring in the neighborhood of twenty-nine hours. The trick in using a data base for linguistic research, at least in the search for specific items heretofore unrecorded in dictionaries, is knowing or artfully guessing the right questions to ask or whether to ask the questions at all given the restriction built into the computer program governing the system. In spite of all the limitations and difficulties that I have noted in this brief presentation, bear in mind that to search several years of issues of *The New York Times* (from June 1980), *The Washington Post* (from January 1977), *The Christian Science Monitor* (from January 1980), and *Manchester Guardian Weekly* (from January 1981) in some cases in less than ten seconds is, in terms of research just a few years ago, beyond the scope of even the Pentagon's budget or the national debt.

Let me provide another example of what an information utility can do. Data conglomerations are variously referred to

as *data base, database, data bank,* or *databank.* How would one rank these forms as to synonomy and to preference of form? When the question was posed to Nexis it reported the following information:

Total articles using one or more of these forms: *16,719*

databank	244	1.45%	earliest date: 1977
data bank	1640	9.80%	1975
database	3606	21.56%	1958
data base	11232	67.18%	1975

If a dictionary manuscript has space for only one form it must be *data base.*

The grandest reward in using a large data bank in editing neologisms comes from having or seeing an item suspected of being heretofore unrecorded and not having to wait for the researchers to find a sufficient number of examples.

But for all its power and majesty, the computer does not solve Sir James's problem. We may better see the past, but computers still cannot see into the future. The cathode-ray tube is close, but no crystal ball.