

DISCUSSION PAPER
REFLECTIONS ON TECHNOLOGY
IN LEXICOGRAPHY

Richard W. Bailey

*Department of English Language and Literature
The University of Michigan
Ann Arbor, Michigan 48104*

"Properties of the instruments," I. A. Richards reminds us, "enter into the account of the investigation."¹ We should not, I believe, take a narrow definition of *technology* in the title of our section or of *instruments* to mean only those devices mentioned by Professor Venezky in his contribution. Whatever instruments we use are in part conditioned by our previous experiences and by the way we conceive our research. It is no accident that one of the three great nineteenth-century North British lexicographers—Joseph Wright—began life in a mill sorting grades of wool. And if those lexicographers present don't recognize themselves in that apprenticeship, perhaps we might remind ourselves that as his first act as editor of the *New English Dictionary* J. A. H. Murray built at the bottom of his garden at Mill Hill School an iron house with over a thousand pigeonholes built into it.

Doubtless we should not take that iron house in the garden too metaphorically, but it is nevertheless correct, as Richards says, that the outcome of our research will be conditioned by the devices we use in its execution. Pigeonholes impose structure, and dictionary editors continue to see themselves as imposing structure on vagrant word usages that blend into each other imperceptibly. Another metaphor from the general explanation that prefaces the *OED* reveals this conception (and reminds us, by the way, of those "fuzzy features" that Professors Bolinger, Lakoff and Pike have spoken about earlier in the conference): "That vast aggregate of words and phrases which constitutes the Vocabulary of English-speaking men presents, to the mind that endeavours to grasp it as a definite whole, the aspect of one of those nebulous masses familiar to the astronomer, in which a clear and unmistakable nucleus shades off on all sides through zones of decreasing brightness, to a dim marginal film that seems to end no where but to lose itself imperceptibly in the surrounding darkness."² The writer here speaks, of course, of the whole aggregate of English vocabulary as a light that illumines the darkness, but, as we see in almost every entry in the dictionary, the same image informs the conception of word senses blending themselves imperceptibly into one another. And yet the job of the editor is to sort citations into discrete pigeon holes, to enlarge his sorting board if need be until he satisfies himself that the "breaks" between the meanings are revealed.

Such a conception, derived from the technology of pigeonholes, is brought to our attention in Leo Spitzer's memoir of his studies in Romance Philology under Meyer-Lübke, whose conception of linguistic history was akin to that of the founders of English historical lexicography. "We were always looking over our shoulder," Spitzer recalls, "but where was reflected in this teaching my

sensuous, witty, disciplined Frenchman, in his presumably 1000 years of existence? He was left out in the cold while we talked about his language; indeed, French was not the language of the Frenchman, but an agglomeration of unconnected, separate, anecdotic, senseless evolutions."³ Linguistic facts, in our historical dictionaries, are left somehow apart from the life of the community; we have yet to capture in our representations of meaning the continuum from precision to imprecision inherent in our citations and certainly pervasive in the collective whole. The only attempt to present lexical material in this light that I know of is in the sample entry for the word *sonnet* prepared by the late Charles C. Fries as a sample of his *Early Modern English Dictionary*. Fries introduced a separate category for the "imprecise" citations, and, under the label "ambiguous instances," wrote: "In many quotations it is impossible to determine the precise meaning of the word *sonnet*, for the same writer sometimes uses the word in all three of its major senses."⁴ This attempt to introduce the notion of a living and hence fallible idiolect brought severe criticism from those who examined the entry; in his subsequent work on the project Fries abandoned the effort to acknowledge the fuzziness of some illustrations and, like his predecessors, selected only "clear" instances for treatment. Thus the technology of pigeonholes prevailed and the tradition continues that, whatever the imprecision of some speakers, the English language that our dictionaries represent is really a pure light illuminating the darkness: "Life, like a dome of many-coloured glass, / Stains the white radiance of Eternity."

I want to speak further about this matter in a moment, but let me turn to questions of more recent technology in dictionary-making.

First, I would like to endorse what Professor Venezky has said about the potential uses of the computer. He is quite right, I believe, in warning us against expecting too much from machine processing—without human intervention, at least—in the way of semantic analysis.⁵ Certainly we can hardly expect the computer to make the lexicographer an "electronic drudge," though its implementation may do something to affect his wonted humility. Nor can we reasonably expect the computer to execute the direction that Murray issued to the volunteer readers for his dictionary: "Make a quotation for *every* word that strikes you [Oh *machina*] as rare, obsolete, old-fashioned, new, peculiar, or used in a peculiar way."⁶ Instead, the lexicographer, like the sorcerer's apprentice, is likely to flood the editorial office with a tidal wave of slips when all he wanted to do was to draw a little water from the lexicon. In short, I thoroughly agree with Professor Venezky's view that "the evolution of computers into the mundane tasks of lexicography is at or near the saturation point."

The fact that we have reached this point so early results, I think, from our attempt to impose the new technology on the old. That is, we have thought primarily of the computer as a replacement for those now nearly vanished members of the educated leisure class who provided so many of the citation slips upon which the *OED* is based. The time has come to re-examine the role that we have assigned to the computer—or hoped to assign to it—in our dictionary-making. If it has not replaced those selfless volunteers of bygone days, much less will it replace Murray's pigeonholes so delicately imposed on the lexicon. In short, the giant archives of "raw text" promised to us or actually in progress are unlikely to repay the trouble and cost that goes into them if we continue to think in terms of traditional goals.

In the limited time available to me, I would like to sketch out three proposals for re-thinking our lexicological and lexicographical research in light of

the new technology available to us. Some earlier investigation of these ideas is already available in print and I would be glad to have the observations of those present—and the readers of the published proceedings—on these subjects (see References 7–11).

The most modest of these three proposals reflects my own interest in what is now known as the *Michigan Early Modern English Materials*, a scheme that my colleagues—particularly Professors Downer and Robinson—and I have been working on recently. Our intention in this work is to make widely available the citation materials for the English of 1475 to 1700 inherited from the *Oxford English Dictionary* and those added to this basic collection by Professor Fries and his associates in the 1930's. Early in 1973, the first installment of the collection will be available to the scholarly public in two forms: microfiche images (generated directly from computer storage) that will make a portion of our material available to all scholars interested in the period, and a magnetic tape record for those interested in language data processing. In our own planning in Ann Arbor, we look forward to editorial work by means of a computer terminal that will make use of the microfiche "materials" and the power of the central processing unit in assisting the editor in adding his interpretative comments on the evidence.^{12, 13} Such an approach will allow this editorial work to take place at several institutions linked by a computer network, hence saving a single institution from the very practical problems that now make difficult the task of carrying out the long and expensive work of editing a regional or period dictionary. Further, production (and purchase) costs can be reduced by keying the definitions to the citations on the fiche, thus making possible the publication of an historical dictionary in one or two volumes rather than a dozen or more.

But my first proposal only reflects what we are currently undertaking: the creation of a body of machine-processable materials with a few properties that the technology of data-processing machines now makes possible. The corpus is not fixed but subject to easy and frequent addition and alteration. Provision is made for the addition of interpretative comments on the materials which can likewise be readily modified as new evidence or new theoretical pre-occupations arise. But the central property, I suppose, is simply dissemination: lexicography and lexicological research having a common base of data can now take place anywhere, not just in Oxford or Ann Arbor or Springfield or in some other single place to which only a few have access. And the interest of the participants in such research can range from the lifetime given by men like Aitken, Burchfield, Kuhn, and Murison to the briefer commitment typical of contributors to *Notes & Queries*.

The second of my proposals differs from the first more in degree than in kind and derives from my reflections on the *Dictionary of Early Modern English Pronunciation* proposed several years ago by Professor Danielsson of Stockholm University and discussed in some detail at a conference in Bonn in April of this year.⁴ The essential attribute of this scheme, from the point of view of technology, derives from the well-developed science of data management. The proposers of the *DEMEP* anticipate entering the information derived from the works of orthoepists and phoneticians of the period into computer storage, but reports from the conference lead me to believe that the editorial committee has not fully realized that the computer can do more for them than the file of slips that would be traditionally employed in such a project. With careful planning they can anticipate indexing every element of

their transcription rather than just the citation form that happens to occur: not only the consonants and vowels, but also the phonetic environments that may condition sound changes. Given sufficient forethought, they can anticipate a question-answering system that will lend support to work of the kind now familiar in *The Sound Pattern of English* and *English Stress*. Questions that might be submitted include simple ones: At what point does the evidence begin to show the shift from /e/ to /i/ in such words as *tea*? But more complicated ones might easily be possible: Does pre-consonantal and word-final /r/ begin to lose constriction after central vowels sooner than before peripheral ones? In other words, the new technology allows us not only to disseminate lexicological information (as in my first proposal) but also makes it possible for us to anticipate richer results from our collection of data than have emerged from less adaptable and convenient schemes using the old technology.

My third proposal is an outgrowth of what Professor Revard has called the *New Utopian Dictionary of English*, and like him I believe that such a scheme is not really as utopian as it may first appear. In America, at least, the steeply rising costs of higher education are already bringing about a growth of mechanization; as in other labor-intensive industries, when salary costs take up more and more of the total budget, "management" introduces automation. We academics who avoid the fate of technical unemployment will find that the machines will almost certainly have "excess capacity" that will make it possible for us to pursue our uneconomic but amusing research into the English lexicon.

The grand resource that the new technology makes possible will first of all consolidate existing information. Proprietary rights notwithstanding, it is clear that commercial dictionary houses will shortly be obliged to enter into some such central arrangement. Socialized lexicography, in short, is now upon us. If I take the moral of Mr. Burchfield's remarks on Monday correctly, the growth in the volume of printed English and the increase in costs of sampling it will make it nearly impossible for the Oxford University Press to produce another supplement in forty years' time, or even for the G. & C. Merriam Company to give us another unabridged international in the 1980's. The new technology makes it possible for us to work toward the scheme that Professor Revard has outlined: sense-links cross-referencing all dictionary entries, processes of derivation clearly identified, and metaphoric and other word-sense relations abundantly illustrated. It would also contain the detailed information about dialects derived from the work of the Linguistic Atlas projects and from the *Dictionary of American Regional English*. Among these other things, it might also include information about what speakers think they do (or ought to do)—what Labov calls "the audio-monitoring norm"—as well as descriptive statements about what they really do do. Such information would capture the fact that most Americans speak of *drunk drivers* but that a good many of them—the usage panel of the *American Heritage Dictionary*, for instance—believe that they really ought to be saying *drunken drivers* (likewise *iced tea/ice tea, skimmed milk/skim milk*).

Unlike our present lexicological work, the grand resource would have the capacity for sustained and continuous growth through addition (or deletion) of information and, as a practical consequence, would allow many more scholars and interested individuals to participate in this work. Finally, it would give us the capacity to test hypotheses about the connection of language variation and sex, style, class, regional origin, historical growth and change in the vocabulary. It would enable us to view linguistic information from multiple

perspectives: not just in terms of particles—"each word telling its own history," in the familiar slogan—but in terms of waves of growth and change within the diversity of the language community. And to extend this metaphor that I have borrowed from the work of Professor Pike, it would allow us (or future generations like us) to see the whole field of the English language more clearly than is now possible with the old technology that we have too long cherished.

REFERENCES

1. RICHARDS, I. A. 1955. *Speculative Instruments*: 18. Harvest, New York, N.Y.
2. MURRAY, J. A. H. *et al.* 1933. General explanation. *Oxford English Dictionary* 1: xxvii.
3. SPITZER, L. 1948. *Linguistics and Literary History*: 2-3. Princeton Univ. Press, Princeton, N.Y.
4. FRIES, C. C. *et al.* 1932. *The Early Modern English Dictionary*. Publ. Mod. Lang. Assoc. 47: 897.
5. SMITH, R. N. 1972. Interactive Lexical Updating. *Computers and the Humanities* 6: 137-45.
6. MURRAY, J. A. H. *et al.* 1933. Historical introduction. *Oxford English Dictionary* 1: xv.
7. BAILEY, R. W. & J. L. ROBINSON. 1968-69. The University of Michigan Early Modern English Dictionary Project. *Shakespeare Research Opportunities* 4: 120-121. Ann Arbor, Mich.
8. BAILEY, R. W. & J. L. ROBINSON. 1970. Computers and dictionaries. In *Computers and Old English Concordances*. A. Cameron, R. Frank, & J. Leyerle, Eds.: 94-102. Univ. of Toronto Press, Toronto, Canada.
9. BAILEY, R. W. & J. L. ROBINSON. The computer in lexicography. In *Hans Kurath Festschrift*. J. Reidy & H. Scholler, Eds. Forthcoming.
10. BAILEY, R. W. Research dictionaries. *Amer. Speech*. Forthcoming.
11. ROBINSON, J. L. & R. W. BAILEY. 1973. Computer-produced microfilm in lexicography: toward a dictionary of Early Modern English. In *The Computer and Literary Studies*. A. J. Aitken, R. W. Bailey & N. Hamilton-Smith, Eds.: 3-14. Edinburgh Univ. Press, Edinburgh, Scotland.
12. VENEZKY, R. L. 1969. Storage, retrieval, and editing of information for a dictionary. *Amer. Documentation* 19: 71-79.
13. REITMAN, W. A., R. B. ROBERTS, R. W. SAUVAIN, D. D. WHEELER & W. LINN. 1969. Autonote: A personal information storage and retrieval system. *Proc. 24th Nat. Conf. ACM*: 67-76. ACM, New York, N.Y.
14. Information on this project is disseminated through the limited-circulation newsletter, *Studies in Early Modern English*. Correspondence about the project should be directed to Prof. Bror Danielsson, Engelska Institutionen, Stockholm University, Stockholm 50, Sweden.