Leeds Studies in English

Article:

F. G. Cassidy, 'Dialectology and the Electronic Drudge', *Leeds Studies in English*, n.s. 2 (1968), 135-43

Permanent URL:
https://ludos.leeds.ac.uk:443/R/-?func=dbin-jump-full&object_id=134435&silo_library=GEN01

Leeds Studies in English
School of English
University of Leeds
http://www.leeds.ac.uk/lse

DIALECTOLOGY AND THE ELECTRONIC DRUDGE

By F. G. CASSIDY

Far as the idea of machinery may be from the thoughts of most students of dialect—much as our first reaction may be *against* getting involved with the highly technical subject of electronics—it is probably wise not to reject anything out of hand. Dialect study in any form requires a high degree of detailed, exacting work which one has to undergo not for its own sake but because it has hitherto been largely unavoidable. If computers can really reduce this burden without imposing, in revenge, some other subtle form of discomfort—if they can do within a reasonable time tasks which we have boggled at undertaking at all because of the cost in years—perhaps we would do well to hear what can be said for and against their use.[1]

The first thing to be said is that the journalistic approach to the computer is thoroughly misleading. If there is any "mystery" or "miracle" in it, that is only so for those who avoid learning about it. After all, it is a human invention: a high-grade machine, but no more than a machine. Without human intelligence it would not exist; without human intelligence it cannot be made to work. It has no mind of its own: even when it seems captious or recalcitrant, this is due to mechanical failure or to human mistakes in the programming or handling. It is natural for us to think anthropomorphically, to call the computer a "brain" or a "drudge," to imagine "gremlins" or "bugs" louting about inside it and tampering with the works. A certain folklore has already sprung up around it—the computer of the cartoons, which makes manlike mistakes, plays shrewd tricks, flashes lights, whirs secretly, and is probably planning something diabolical. No one has seen or painted this better than Artzybasheff. Some genuine fear seems to exist, no doubt inspired by the science-fictioneers, that the computer will displace humanity and "take the world over." All very exciting—it raises for our day the half-pleasurable *frisson* that the twenties felt over Capek's robots, or that *Erewhon* roused a century ago. It might be good, just now, to medicine ourselves with Swift's astringent scepticism by reading the Academy of Lagado.[2]

What is the computer really good for that scholars can use? Its

highest virtue is speed, speed that we find it hard to conceive of. Human time is measured in years, hours, at best seconds; computer time is measured in micro-seconds—millionths of a second, and even nano-seconds—billionths of a second. While we are scratching our heads, the computer can search through a hundred thousand items and find the ones we want. Though the job it performs is utterly mechanical,[3] it does it at a speed so prodigious that the investment of human time to use it is more than made up. The more routine the job, the more of the same kind of thing there is to be done, the more human time the computer can save—and this may mean years of our irreplaceable lives. Mental arithmetic, the abacus, the calculating machine: the computer lifts all such operations to a new level through its superior speed; and as to dealing with alphabetic rather than numerical problems, it has no mechanical competitor.

One must add that though the computer is liable to "fatigue" (another anthropomorphic term, often applied to metals and other inanimata) it is less likely to make mistakes from this cause than are human beings. Probably this is because it is harder to shut off the human brain from interfering factors. The things the computer is asked to do at any one time, even when quite complex, are strictly limited. It can do those things only, because they are all it is programmed for, and short of breakdown it cannot escape its programme. The human brain has many potential programmes stored within it which frequently interfere with each other. Probably no human brain is ever programmed completely, either, to do just one set of operations. (Only a well organized human brain can work out a good computer programme.) The brain and the computer are both subject to the effects of temperature, humidity, vibration, physical shock, and so on, but the flesh is heir to many other ills that the computer knows not of. The rare "great thinkers" are people who have disciplined themselves to shut off all kinds of interfering factors which need never bother the computer; we say such men have "great powers of concentration." It has been a preoccupation of certain religions to find ways of releasing thought from bodily influences. The psychedelic[4] or "mind-revealing" drugs, though taken most often by sensationalists, are seriously experimented with by some scientists in this very hope of freeing our "mental" powers from the trammels of the body.[5] The computer is pure mechanism; if we are too, we are more complex and less pure by far. The stuff that we are made on, our "nervous systems," pays for higher sensitivity by a loss in reliability when the level of complexity is raised or when more speed is demanded. Fatigue affects us sooner. We can do anything a computer can do if given time enough. It can do the more

mechanical activities much faster than we can with less danger of error—but only at our volition and through our thinking.

Next to speed, and as a by-product of it, the computer can outdo human beings at manipulating masses of material. A conventional file of cards or slips in drawers in cabinets, as soon as one begins to compare or combine, requires much time-consuming footwork and fingerwork. The slips have to be gathered (and returned to the proper places afterwards), they have to be spread out on tables or sorted into piles. It is all too easy to mislay something. Instances have to be counted up, percentages and statistics, if one goes that far, must be laboriously worked out. In the end one accepts limitations simply because the possible gain from further labour seems hardly worth the candle.

This is the point at which a computer programme can lift the entire operation to a new level. If every item that one might want to recover is properly labelled or delimited when put into the file, any kind of sorting, compilation, combination, or comparison becomes easy. The computer adds nothing to the data, but it will present the data in whole or in part as one desires, making it possible to dig out unexpected correlations from below the obvious surface. The very fact that it is easy to compare invites comparisons which one might never attempt otherwise. Once all the mathematically possible comparisons have been made and examined—an exhaustive study of the data—one can feel that, short of the discovery of new evidence, there is nothing more to be done fruitfully with the subject; it may be laid aside, clearing the way for other investigations.

Dialect study concerns itself either with a closed corpus, if the subject is a problem in the past, or an open corpus if it concerns living language. With the former we have traditionally studied texts one by one, made generalizations on the basis of the most striking features, then returned to correct the details when possible with new evidence or sharper interpretations. In early volumes of the EETS the information about dialect, when any was given at all, was necessarily less than satisfactory. But that was over a century ago. The amount and extent of dialect material given today in newly edited texts is highly variable though on the whole somewhat better. A great step forward was taken in preparation for the *Middle English Dictionary* by generalizing from a series of pretty well localized texts and setting down the major isoglossal lines.[6] These have been corrected in detail since, and Professor McIntosh and his helpers promise to refine on the method, to examine almost double the number of documents and take fuller account of the graphemic basis for phonological analyses.[7] McIntosh's concept of the "fit technique" is interesting and becomes possible to

apply as one gets into the narrower aspects of what is, essentially, dialect geography practised on texts rather than on living speech. Indeed, beyond the study of individual texts progress can best be made by putting more and more texts together. With computer aid one can carry this process to its conclusion: a larger, even a total, corpus can be examined, and examined in the highest possible degree of detail.

The corpus of ME, though closed, is probably too huge to consider tackling as a whole at present, but in the field of OE we have a corpus of exhaustible size. All or almost all that survives from that period is known and could be put together into a single file for total examination. Indeed, the concording of the OE poetic corpus is nearing completion,[8] and work on the prose has begun.[9] It should be noted that accurate editions must still be made, and must be available for reference and consultation, if a concordance is to be of any value—and this requires the work of scholars before the computer begins its part. Further, since whatever one may want a computer to do must be written into the programme, everything must be foreseen. The programmer must therefore himself either know the subject (in this case OE) or work closely with one who does and who can set forth these desiderata.

Within the complete corpus of surviving OE what data might we find it valuable to examine? What have we never or only partially examined before? Scanners at present in operation can read on to electronic tapes material typed in "alphanumeric" characters—a special fount which to the eye appears slightly peculiar but is quite as easily read as any other. If the OE corpus were typed out so and scanned, with the right kind of programming one could retrieve it part by part at will or in various combinations of the parts.[10] One could ask for example to have every character printed out by itself or in context and to have a count made—how many *a*'s, *b*'s, *c*'s, and so on, in what context each appears, and with what frequency. One could then ask to have comparisons made of, say, the use of *c* vs. *k*, *þ* vs. *ð*, of *a* vs. *æ* vs. *ea*. In the end one would have a definitive account of the graphics of any single text, or of the entire corpus, be able to establish the graphemes, the phonemes, correlate these and features of morphology, lexicon, etc. with the geographic source of the data, and so establish dialect distinctions. Correlation with the type of text (ecclesiastical, literary, translated, original, standardized, personal, etc.) would shed further light on stylistic questions. At the very least, one would have at last a full description of OE—as full a one as the surviving corpus permits—which would facilitate comparison with other stages of English or with other languages, and would permit a definitive job of dictionary-making.

For both the linguist and the literary man it would be valuable to have a frequency-count of the letters of OE, a graphotaxis and letter-group count, a phoneme count, a count of consonant-vowel sequences ("canonical forms") within words, a word-count, a count of recurrent word-groups of any kind (some of which would be casual, others syntactic, others formulaic, and so on). With the dictionary one might give exhaustive lists of prefixes and suffixes (inflexional and word-forming). Compounding could be examined in connexion with spacing; accentuation and other prosodic features could be more fully correlated. Emendations accepted and rejected in the past could be reassessed in the light of probabilities now better known. It would obviously be valuable to put together *all* existing evidence on, say, the palatalization of *g* as indicated by *gi, ge, gie* spellings—especially if this could be presented in something like chronological order and with due regard to the various kinds of sources. Once the data are stored it makes no difference to the computer how we want them manipulated and presented. But this must all be anticipated and made a part of the programme.

The four OE dialect areas traditionally recognized are set apart by depending heavily on the few texts which can be best localized in time and place by non-linguistic evidence, then extrapolating from these to the less certain. By computer methods these basic texts could not only be re-examined, comparing their chief features and perhaps revising past conclusions about them, but the less obviously characteristic features could be scrutinized and probably found to furnish additional evidence. How far the computer can carry us towards greater certainty in these matters cannot be safely predicted, but it is reasonable to suppose that the kind of exhaustive examination which it makes possible would squeeze out a few drops of evidence and (changing the metaphor) somewhat thin out the accumulated underbrush of theories and inconclusive interpretations.

The OE dialects, no less than the ME, could be approached by the methods of linguistic geography. Computers, using a "plotter", are at present able to make maps, to set out on paper all the data one wants displayed. They can not only draw the map itself but assign symbols to the various relevant items as programmed and print this "legend" on the map. Thus a complete set of maps can be drawn of any area for which the data are at hand. Possibly even better than that: using a "scope," computers can present the same maps on a television-like screen, superimpose maps for features one wants to compare and between which one may suspect that some relationship exists, expand the scale of some section of the map for better observa-

tion of details, and perhaps delete the commonest feature so that the less common may stand out more clearly. The same data that are used to compose a picture map on the scope can be drawn out on paper in lasting form: using the scope to sort through the possible maps, one might then choose the most significant for the plotter to produce, or simply photograph the scope, as is quite commonly done. Let me repeat that the computer makes no interpretations and draws no conclusions all by itself. The scholar who knows what conclusions may be latent in the data must anticipate and provide for every kind of "output."

The kind of computer methods I have been describing are now in use. For large jobs of data-processing they have already rendered the punch-card and punch tape archaic. Paper tape is especially difficult to correct—cards less so, since a new card can be punched and substituted for the old. But scanner type includes a deletion symbol which will blot out any other character; or, it may be provided that a horizontal line drawn through the type will tell the scanner to ignore it. Thus nothing has to be rubbed out or repunched and substituted—one deletes and goes straight on: the scanner simply skips over the delenda. (Insertions are more difficult to make.)

The scope, too, has a deletion device which is even better, since it can be programmed to remove unwanted letters from the screen at the viewer's command, leaving the remainder clean. For the process of editing—deleting, correcting, adding—the scope should be especially valuable. Data stored on tapes by the scanner may be called on to the screen; the editor reads these and decides which to retain. The rest he dismisses (though they are not lost; if he has an afterthought, any part may be recalled immediately). What he wants to add further is then typed on the keyboard, appears on the screen, and may be placed where he wishes. Thus he may compose his treatment and see it in corrected form. He then pushes a button, and this final output is sent to another tape. From this it may be retrieved and printed out as desired.

At present one is forced to use scanner type to store the file with data, and printouts are all in capitals. But the time is very close when scanners will be able to read conventional type, no longer requiring clearly printed input to be retyped and proof-read. Some typewriters can be coupled to a computer to print out programmes, but this process is less rapid than the usual printout process. Further, printouts are now available (at a price) in upper and lower case characters. This will make for more conventional and readable outputs, even without going from the printout machine to linotype. In working with OE, obviously, certain extra letters and abbreviation symbols have to be provided for, at least in the final stage; though for earlier stages existing punctuation

symbols and others not present in OE (Z, V, Q, %, =, etc.) may be substituted, to be translated ultimately into the proper OE symbols. In working with ME texts, or in using a phonetic alphabet with MnE data, the same device of substitution and translation may be applied—has indeed already been applied.[11]

In broad-scaled investigations of living language, notably the dialect atlases, the computer can be a new ally. It will not do the planning or the field work; the data still have to be collected painstakingly by direct interview of speakers. Some mechanical aids are possible here—the tape-recorder, especially. And if the acoustical engineers succeed in producing the sound-translator, it may ultimately be possible to convert a tape record (or, for that matter, the informant's voice directly) into some visual form—even, conceivably, a conventional phonetic transcription, though this is not ideal, since it segments before the eye something which is not segmented to the ear. The visual form, whatever becomes possible, will make human analysis feasible by converting time to linear extension, so that we can analyze slowly what is said much faster.

When such new aids become available, the computer can be programmed to measure the linear record and analyze it in whatever way the analyst desires. It is conceivable that, ultimately, anyone's words spoken into a machine could be immediately scrutinized and classified, much like a fingerprint—an individual vocal signature. And not merely the voice but, beyond that, language itself. Methods of discourse analysis now being worked out by generative-transformational grammarians are eminently fitted for computer processing. Human language, indeed, is so multiplex in its interlocking systems and subsystems that only the tremendously high speed-powers and mass-handling-powers of the computer can ever be expected to give a full picture of even one idiolect.

This glimpse may not be of a remote future. Much of what I have described is possible at present. The linguistic surveys and atlases now in progress within the English-speaking world (and elsewhere too, of course) have come just before or at the time of change—the transition to the computer age. The maps of the *Linguistic Atlas of New England,*[12] all laboriously lettered by hand, make a fine display, but one not likely to be repeated. Today, with scope and plotter presenting the same data stored in a computer file, the entire process could be done mechanically. Furthermore, correlations between the language of the informants and their ages, degrees of education, occupations, and other relevant non-linguistic facts, all of which were carefully collected and are presented in the *Handbook* but do not appear on the maps, could be put there—or,

better still, could be tabulated or summarized in some relevant arrangement. At present these data are half raw; interpretations of them—the "results" for which they were presumably gathered—have to be worked out in the old laborious way by the scholar, pencil in hand, counting instances and working out sums and percentages—unless he is so advanced as to use a calculating machine.

The admirably detailed and plentiful collections of Professor Orton's *Survey of English Dialects,* presented in careful lists, has cost countless hours of human labour that we could now relegate to the electronic drudge. Everyone of the phonetic features, including segmental characters, length marks, superscript and subscript diacritics, and conventional alphabetic and numeric abbreviations, once stored in a computer file exactly as they appear in the printed volumes, could be sorted and sifted in any desired way and presented on maps or in tabulations. The presence in any part of the area, or all of it. of any single feature—phonetic, morphologic, syntactic, lexical—could be documented in a very short time for the scholar. Summaries and correlations of various kinds similarly. In short, all the preparation has been done except the putting into computer form. It this last step could be taken, the data would yield their latent riches all the sooner, more fully, and without further drain on scholars' lives.[13]

If I may be permitted a short flight of imagination, I picture an accomplished Scribe in Westminster in 1476. He hears of a dubious sort of fellow, one Caxton, who has been consorting in the Low Countries with these Dutch mechanicks—a harlotry lot who are always stirring up some diabolical brew. Now he has brought a new engine and set it up hard by, with boxes of carved letters, huge screws and plates, pots of foul ink—and is blotting away with it in the pretence of imprinting books. Any right-minded man must watch him askance—give him rope enough to hang himself. At least not precipitate oneself into his delusion but follow the old, safe, and honoured track.

Perhaps the parable is plain enough. We who have come up in the old methods of philological scholarship can read *A Grammarian's Funeral* as Browning undoubtedly meant it: straight. Some younger readers cannot believe that it was not meant ironically and try to read it so. The enclitic *de* is still with us and computers will not themselves solve its mysteries. But perhaps I have shown some ways in which—right now —they could save years of labour which is not scholarship, and turn the saved time back to us for labours which are.

NOTES

[1] I know nothing myself about electronics, computer "hardware" (the machines) or "software" (programmes and programming). This non-technical account has been passed as acceptable technically by Dr R. L. Venezky, who was kind enough to read it. Dr Venezky is the author of the computer programme being used by the *Dictionary of American Regional English*. See his article describing this programme: *American Documentation* 19.1 (Jan., 1968), 71-79.

[2] Journalists may be excused better, perhaps, than some university teachers who offer courses on "Artificial Intelligence and Models in Thinking" in departments of Computer Sciences, or who publish articles with titles like "Can Computers Think?"

[3] I am still referring to electronic computers, not to prototypes using gears, relays, etc.

[4] Literally, "soul-revealing" or "-clarifying." The commonly used phrase "mind-expanding" is not an accurate translation.

[5] See, for example, Dr J. C. Lilly's experiments with the "bio-computer" (the human brain) under influence of LSD. Communications Research Institute, Miami, Florida.

[6] Samuel Moore, Sanford B. Meech, Harold Whitehall, "Middle English Dialect Characteristics and Dialect Boundaries: Preliminary Report . . . ," *Ess. and Stud. in English and Comp. Lit. by Members of the English Dept., Univ. of Michigan*, Ann Arbor (U.M. Press, 1935).

[7] See especially Angus McIntosh, "A New Approach to Middle English Dialectology," *English Studies*, XLIV (1963), 1-11, and M. L. Samuels, "Applications of Middle English Dialectology," *ibid.*, 81-94.

[8] J. B. Bessinger, Jr., "A Computer-based Concordance to the Anglo-Saxon Poetic Records," No. L22 in *Computers and the Humanities* I.5 (May, 1967), p. 187, updated *ibid.* II.2 (Nov., 1967), p. 74.

[9] Drs Richard Venezky and Jon Erickson of the Dept. of English, University of Wisconsin, have concorded, at the present writing, nine of the Vercelli homilies. Though done individually, these are programmed for easy conflation. Nearing completion also is a concordance to the Rushworth 1 Matthew. This work will be gradually extended to other works of the OE prose corpus.

[10] At present, lower case, superscripts, and subscripts (accents, italics, etc.) cannot be scanned as such but must be encoded in a linear sequence.

[11] One recent computer-made concordance which deserves notice is Barnett Kottler and Alan M. Markman, *A Concordance to Five Middle English Poems: Cleanness, Saint Erkenwald, Sir Gawain and the Green Knight, Patience, Pearl*, Pittsburgh (Univ. Pitt. Press, 1966).

[12] Ed. Hans Kurath and Bernard Bloch, *Linguistic Atlas of New England*, 3 vols. in 6 parts, Providence (Brown Univ. Press, 1939-43). Also, Hans Kurath *et al.*, *Handbook of the Linguistic Geography of New England*, Providence (Brown U.P., 1939).

[13] Computer processing is expensive, no question. Further, the rapid development of techniques means that a costly investment in computers may no sooner be made than they are succeeded by even more sophisticated and powerful ones. Nevertheless, the per job cost goes down as use goes up, the older machines can be used for less difficult jobs, and competition among manufacturers steadily reduces the cost of "hardware." An up-to-date computing center with a staff competent in processing natural language problems can advise the scholar whether his investigation will profit by computer handling or not, and can estimate the cost.