Computers and the New OED's new words

John A. Simpson

Several months ago I came across a perfectly ordinary sentence in the *British Medical Journal.* It read like this:

> After this recent personal experience I am now even more firmly convinced that to be awarded the accolade "centre of excellence" a hospital must offer much more than a bright array of scanners, several hundred beds, a handful of professors, and a few dozen medical students (*BMJ* 1985, 2 November).

It is difficult to think of a more ordinary sentence, or one which is less likely to interrupt the safe confidence of lexicographers that their dictionaries contain all there is to be known, and more too. This quotation was in the Oxford Dictionary's files because it is evidence for the expression *centre of excellence,* clearly carded as it was likely to be a new term not in the OXFORD ENGLISH DICTIONARY (OED) or its SUPPLEMENT. That was quite unexceptionable: after some library research, we found that *centre of excellence* dates from around 1971 – and so it was not surprising that it had not got into the SUPPLEMENT. The interesting fact about the quotation is that it includes at least two other lexical items which are not covered by the OED or SUPPLEMENT. *Scanner* appears in the SUPPLEMENT (though *body-scanner* and *brain-scanner* do not), *bed* in the specific sense 'hospital bed' (arguably not a distinct sense) and – remarkably: in fact, astonishingly – *accolade* do not. I should add that their absence from the big dictionaries does not mean that they are necessarily absent from smaller, notionally derivative dictionaries. After a little work we have dated *bed* in this sense to 1881, and *accolade* many years back to 1852, and have now prepared full New OED entries for them. But it serves to remind us that new vocabulary does not always push itself rudely into the language, but often creeps in by stealth when even trained lexicographers are not looking.

It is perhaps appropriate to reiterate a few facts about the New OED project schedule. The project's first dictionary publication is planned to be an integrated version of the OED and SUPPLEMENT (that is, the OED and SUPPLEMENT printed in a single alphabetical sequence). This will be published in conventional book form, and will be followed soon afterwards by a dictionary database available online and, we expect, on compact disc.

The integrated version is scheduled for 1989, and as well as the full text of the OED and SUPPLEMENT, we intend to include several thousand words and meanings which are new to the dictionary. As co-editor of the New OED, my primary responsibility at present is the editing of these new entries. The early letters of the SUPPLEMENT were completed at the beginning of the 1970s, and

fifteen years of linguistic innovation and change since then have brought with them many new terms which deserve a place in the dictionary, such as *asset stripping, breakfast television, cashpoint, cling film,* and *deconstructionism,* to name but a few; and then there are many older lexical items which have only recently come to general prominence, such as *acid rain, artificial intelligence,* and *ayatollah*; in addition, the criteria by which terms were selected changed slightly over the span of the SUPPLEMENT (as, in fact, had been the case with the OED itself), retrospectively causing holes in the earlier coverage. So the introduction of new material allows us to start correcting these anomalies, as well as maintaining editorial practices and traditions unbroken from the SUPPLEMENT through to the New OED.

There are three major areas in which computers are being used to assist the preparation of new-word entries:

(a) selection criteria;
(b) lexicographical research, particularly in the use of online information retrieval systems;

and in terms of the relationship between these new entries and the current dictionary,

(c) the computational integration of new material into the main New OED database.

In the first two categories, it is interesting to examine how traditional methods are used, what computational assistance is currently available, and lastly, what computational support we might profitably use in the next few years. The following discussion relates to the work of supplementing a historical dictionary, and so the priorities may not be the same as for lexicographers working on bilingual dictionaries and monolingual dictionaries of current English.

Selecting material for editing

The traditional methods of selecting which words and senses to edit include working methodically through a large quotation file fed from a general reading programme, noting material entering other dictionaries and publications such as the Barnhart DICTIONARY OF NEW ENGLISH and its COMPANION, and acting on the suggestions of others and on one's own observations. Editorial work is currently centred around a relatively small number of items, and it would not be efficient to scour the whole word-file, so we filter material coming in from the reading programme each day, looking for likely contenders, and also sort sequentially through the first few letters of the file – where the SUPPLEMENT is more out of date and the chance of finding material for drafting is most high. Even so, a short range in B throws up, for example, *bells and whistles, belt-tightening, beltway, bench-test, bench-warmer,* and *benign neglect,* as well as several

other items which will be held back for later. These are some of the traditional techniques which were employed on the OED and the SUPPLEMENT, and to a greater or lesser extent by other dictionaries throughout the world.

If we investigate how computers are being used in selecting material for editing, we see that they are still very secondary aids. The list of current applications is small. There are two small areas worthy of note which are interesting but not particularly significant. The first involves using DIALOG's "expand" function, which lets one look at specified file indexes in which keywords from a file are tabulated alongside frequency of occurrence (Fig. 1). This can sometimes lead one to think that a given collocation may be worth drafting, even though no occurrences have been found by a reading programme. The second is really just chance discovery. For example, while searching for quotations for the expression *career-break* on NEXIS, we discovered that as well as the use we had expected, there were another two senses well-documented enough to merit further editorial attention.

```
File 275:COMPUTER DATABASE 83-86/ISS15
(COPR. 1986 INFORMATION ACCESS COMPANY)
?

      Set  Items  Description
      ---  -----  -----------
?e file

Ref    Items  RT  Index-term
E1         1      FILCOM
E2         1      FILD
E3      7914     *FILE
E4        32   3  FILE ACTIVITY
E5       216   5  FILE DIRECTORIES
E6         1      FILE DIRECTORY
E7        29   1  FILE LOCKING
E8       161   4  FILE MAINTENANCE
E9       811   2  FILE MANAGEMENT
E10      269  17  FILE ORGANIZATION
E11       33   3  FILE REORGANIZATION
E12        1      FILE SECURITY
```

Fig. 1: DIALOG's "expand" function, showing collocations with frequency of occurrence.

The range of computational techniques currently available for selecting material for dictionaries is not impressive. If this facility were given a higher priority, then it would be possible to maintain a type of 'monitor corpus' of the kind discussed by Jeremy Clear (in this volume), from which we might try to identify neologisms against a machine-readable version of the dictionary (though this is not going to be easy). Perhaps we should be thinking of analysing corpora for particular characteristics: frequent collocations (especially of the almost trivial kind which are likely to be overlooked by a reading programme, idiomatic ver-

bal phrases containing high-frequency prepositions and adverbs of the type expensive to work with on the commercial systems); there are many other possibilities. Then there is the question of whether it is expedient to computerize the existing Oxford Dictionary word-files, which, as a collection drawn from heterogenous sources, does not possess the structured controls of, for example, the Brown corpus. I'm not sure that a straight conversion would help the selection procedure at present – at least until we have a comprehensive desk-editing system, which may well be a long-term aim, but is not one which is relevant to the present activity.

Researching dictionary entries

So the selection of terms for editing relies heavily at the moment on traditional methods. But once the words and expressions have been isolated, it is then possible to use computers to help fill out details about their meaning and history.

Traditionally, the research for historical dictionaries has centred around other reference books, library stacks, and expert consultants. The OED needs to illustrate each term's use from its first appearance in English, and for this there is often no substitute for working through rows of source books in search of elusive documentation.

Current computational resources are widely used nowadays by lexicographers, but only in restricted areas. The online information retrieval systems are often geared towards the requirements of other types of user – lexicographers come at these systems indirectly, from the side, trying to extract information for which the system was not really planned. Often, for example, one cannot request frequency reports on variant spellings, since they are indexed under the same form: *advisor* and *adviser*, or *fiber* and *fibre*. In addition, the lexicographer must be very careful that he or she understands the range of sources and meanings over which the frequency is based. It is of no use drawing general conclusions about the frequency of *astrologer* as opposed to *astrologist* unless one realizes that most of the citations for both terms on NEXIS come from American sources. On the other hand, most of the sources on NEXIS are American anyway, so the search actually reveals very little about the use of *astrologist* in Britain.

Both on the SUPPLEMENT TO THE OED and the current new words work, we have used corpora to find not only early examples of words and phrases, but also stock quotations to fill unfortunate gaps in the word-files. This is simply practical lexicography. A typical example might be *conspecificity*: the run of examples from the word-files comes to a close in 1976. This may mean that the term has now been superseded by something else, or it may simply mean that the word has been overlooked by the Dictionary's reading programme. There is no point in trying to find examples for a derivative like this oneself – it is likely to take up much too much time. But the computer can give an answer in seconds. We pro-

bably run searches on NEXIS or DIALOG for about half of the new entries which are being drafted. On balance, the advantages outweigh the disadvantages: the benefit of having extra illustrative quotations for *deconstructionism*, for example, usually outweighs the problems that they often come from a limited range of sources, from a short, recent time-span, and that the quotations used need reverifying. They need reverifying before being cited in the dictionary principally because NEXIS cites by the page on which an article begins, not the page on which the wanted word turns up, and because DIALOG cites later abstracts, not the original sources. In fact, we try not to use too many "agency" quotations, because the sources are rather limited (the *Washington Post* and *New York Times* would otherwise be quoted far more than was acceptable), though these quotations can be useful in trying to understand the meaning of a term, from the way it is used in various contexts. This includes, for instance, the range of collocations in which an adjective can appear – online services can provide information on whether, for example, it is only women who are referred to as *ditsy* in the United States, or whether the adjective is also applied to magazines, pets, etc. And the quotations can be useful in tracing a lexical item's history (though again this is hampered by the generally short and recent date-range of the sources). Occasionally some of the citations will mention information which the lexicographer has not met elsewhere (an alleged inventor's name, or the source of a famous saying), and sometimes they provide a first example: for instance, DIALOG's files have produced our first examples of *paddle* and *menu-driven* in Computing, from 1980 and 1979 respectively.

It may not be easy, however, to find exactly what is required. Suppose one wishes to draft an entry for *crack*, the cocaine derivative. Clearly, all of the examples of the five-letter character string cannot be called up. It is necessary to make a more complex request. A recent search for *crack* in the same context as *cocaine* threw up 1,108 examples, but that failed to solve the matter, because many of these relate to phrases such as "crack down on cocaine traffickers" (Fig. 2). Restricting the search still further, by adding *new drug* to the list of requirements, brought the number of citations down to about 40 (of which some were still false trails). And by this time, many other *real* instances of *crack* as a drug had been passed over by the restrictions that were imposed.

Several other important points are mentioned in Barnhart 1985. And of course the lexicographer has no control over sources (or which parts of the sources are fed into the system – no personal ads, probably) – and which files are deleted from the system, or whether they cover British or American English. Unfortunately, dictionary-compilers are not a big enough pressure group to pay for these features. But any lexicographer using such a system should be aware of its potential shortcomings.

Non-commercial corpora and computer-concordanced texts can be used in much the same way. The point with respect to both is that the lexicographer nowadays really needs a hybrid system which makes use of the broad but slow

Copyright (c) 1981 The Washington Post

January 25, 1981, Sunday, Final Edition

SECTION: Style; G1

LENGTH: 3500 words

HEADLINE: Scapegoats: The Essence of Washington Politics

BYLINE: By Sally Quinn

BODY:

... degree of fighting and carping, until the next crisis arose. This time it was the aftermath of Camp David, when the treaty began to show cracks and Carter began taking criticism for his human rights efforts; the Panama Canal issue came to the forefront, his relations with Congress ...

... spitting Amaretto down the front of a young woman in a bar. There were allegations -- later disproved -- that Jordan sniffed cocaine in the basement of Studio 54. All of these stories were denied. But something was clearly going on. The episodes did have one immediate and very ...

Fig. 2: Unsuccessful search on NEXIS for *crack* (cocaine compound) on keywords *cocaine* and *crack*.

sweep of the traditional methods and the currently limited but fast computational ones.

Additional computer research resources would certainly be useful, if they were developed to a suitable degree. The list of potentially useful applications is, in fact, almost endless – but it relies upon developments in AI and automatic grammatical parsing, etc., and the creation of appropriate knowledge bases.

There is not space here to consider the questions raised by this. The foregoing description has centred around the preparation of new-word entries. This is a small but important aspect of the New OED project. The following section examines the practical matters involved in merging these new entries into the current Dictionary, and leads on to a general description of the present state of development on the project (see also Hulton and Logan 1984, Weiner 1985, Simpson 1987).

Integration of new material into the main New OED database

The keyboarding of the entire text of the OED and SUPPLEMENT was completed in August 1986. The process – including a proofing cycle – took eighteen months, and at its peak involved over 120 keyboarders. The walls of the New OED rooms are lined with photocopies of proof sheets (21,000 OED pages in all, printing out to about 160,000 sheets of proof print, read by a team of fifty freelancers, and collated centrally). The OED and SUPPLEMENT data files have now been passed through a parser to enhance their structural coding, and the separate files are currently being merged automatically into a single alphabetical

sequence and, in cases where that proves impractical, interactively. The merged database will then be proof-read again, before being sent on to the composition system.

The integration of four or five thousand new entries is small fry compared to the logistical problems evident elsewhere on the project. It seemed sensible that this new material should be keyboarded at Oxford, incorporating the parser's coding conventions, so that the entries could then be called individually from data files into the interactive integration system. The new entries have been prepared in the style of SUPPLEMENT entries, so the same interactive integration processes would apply.

However, as the English proverb says – the longest way round is the shortest way home. This system would not include enough checks and balances to ensure that the new material entered the database with uncorrupted coding. Instead, we have decided to follow the course taken by the main OED and SUPPLEMENT text. The new entries are being sent in card-form to International Computaprint

<HW>A'pex,
<PS>*sb.*<HM.2> (*a.*)
<VL>Also APEX.
<ET><OB>Acronym, f. the initial letters of *A*dvance *P*urchase *Ex*cursion.<EB>
<SR>A system whereby airline tickets for scheduled flights may be bought at a reduced rate on certain conditions (usually including payment in advance and a specified interval between outward and return flights); a fare offered on these conditions. <ES> Freq. *attrib.* or as *adj.*
<QP>[1970 *Aviation Week & Space Technol.* 15 June 24/3
<QT>Pan American would also add a new excursion fare, tentatively referred to as an 'advance purchase fare'..at a lower rate than the standard excursion tariff.]
<QN>1971 *Time* 23 Aug. 53/3
<QT>The West Germans argued that..APEX [advance-purchase excursion plan] would only add confusion to..fares.
<QN>1974 *Aviation Week* 28 Oct. 24/2
<QT>The Apex fare, if it is allowed to become effective Nov. 1, will be the lowest of all scheduled fares.
<QN>1976 *Holiday Which?* May 60
<QT>*APEX*, Advance Purchase Excursion Fare. <ES> Available on various routes, using scheduled flights. <ES> Book and pay at least two months in advance.
<QN>*Ibid.*, Johannesburg..<103>356 (Apex).
<QN>1980 *Times* 16 Feb. 11/8
<QT>Travel notes...how season Super-Apex <103>282.
<QN>1985 *Washington Post* 18 Aug. E8/1
<QT>They fly nonstop from New York to Nice for an APEX fare of about $900 round-trip.

Fig. 3: New entry with ICC tagging

Corporation (ICC) at Fort Washington, near Philadelphia, and ICC are keyboarding the text for us (Fig. 3). Once returned to Oxford, the material will be proofread, and the tapes will be run through the SUPPLEMENT parser to convert the tagging to SGML. The criterial path of the project does not allow us to submit these new entries to the automatic integration system, but they will be held in files ready to be called into LEXX, the interactive integration system. These entries will be set in their right place in the database by mid-1988.

Lexicographers seldom stop preparing new entries, even when they are at conferences, and the New OED computer system allows us to compile entries in draft for critical inspection. The tailpiece of this article shows one such entry for *Zürilex*. I should point out that the quotations are adapted from the OED entry for *Mecca*, and that as we found no examples of *Zürilex* in our card file or on the various online computer systems, including the trademark registry, and because we do not normally include specific names of this type, we shall no doubt have to drop the entry.

Zürilex (zue·rileks).[f. Swiss-German dial. *Züri-*, combining form of *Zürich* the name of a major financial centre in Switzerland, as in *Zürisee*, *Züri-Woche*, etc., + *Eura*)*lex* the European Association for Lexicography, established at Exeter in 1983.] The second Euralex International Congress, held at Zurich in 1986. Cf. LEXETER.

1850 BOKER *Anne Boleyn* t. iii, Make to the Zürilex of our hopes, the king, A solemn pilgrimage. 1887 *Times* (weekly ed.) 21 Oct. 9/1 Stratford..is the Zürilex of American pilgrims. 1890 'R. BOLDREWOOD' *Col. Reformer* (1891) 329 He..was..free once more to turn his brow erect and undaunted towards the Zürilex of his dreams. 1986 J. A. SIMPSON (*talk*) I've waited 136 years for Zürilex, and now it's almost over.

(Fig. 4)

References

Cited dictionaries

DICTIONARY OF NEW ENGLISH
C.L. Barnhart et. al., New York: Harper-Row (1973).

OXFORD ENGLISH DICTIONARY (OED)
J.A.H. Murray et al., Oxford: Oxford University Press (1884–1928).

SUPPLEMENT TO THE OXFORD ENGLISH DICTIONARY (SUPPLEMENT)
R.W. Burchfield, Oxford: Oxford University Press (1972–86).

Other literature

Barnhart, David K. (1985), "Prizes and pitfalls of computerized searching for new words for dictionaries", in: *Dictionaries: Journal of the Dictionary Society of North America* 7: 253–60.

Hultin, Neil and Logan, Harry M. (1984), "The *New Oxford English Dictionary* project at Waterloo", in: *Dictionaries* 6: 128, 183–98.

Simpson, John A. (1987), "The New OED project: a year's work in lexicography", in: *University Computing* 9: 2–7.

Weiner, Edmund S.C. (1985), "The *New OED*: problems in the computerization of a dictionary", in: *University Computing* 7, 66–71.