A microfiche concordance to old english : the first six months of use

by

S. BUTLER
Robarts Library - CANADA

A little less than a year ago, the editors of the Dictionary of Old English received the first copies of *A Microfiche Concordance to Old English*, compiled by Professors Richard Venezky of the University of Delaware and Antonette diPaulo Healey of the University of Toronto, with the aid of a grant from the National Endowment for the Humanities. Professor Healey was responsible for coordinating the Dictionary's share of the work, and for preparing the list of short titles, texts, and bibliography. Professor Venezky was responsible for merging all the individual concordances and producing the fiches.

The basic collections were already in existence, since the Dictionary of Old English maintains a set of computer tapes containing all of the approximately 2,000 individual Old English texts. These tapes were used for generating paper slips for the use of the editors in writing dictionary entries. For the microfiche concordance, we proofread slips, corrected and re-concorded each text, copied the concordances to tapes and sent the tapes to Delaware where Professor Venezky merged the concordances into one large concordance of all Old English, and formatted that concordance for printing directly onto fiches.

The concordance is far too large to be practical to print in book form. There are 412 fiches, 397 of the concordance proper (equivalent to 126,876 pages of citations), 12 of frequency lists (both alphabetical and in descending order of frequency), and 3 containing the lists of texts, editions, and short titles used (this material is also provided in book form as a convenience for users).

The concordance gives all occurrences of every word in Old English, except for 197 spellings of common function words for which only frequency counts are given. Each word is cited with a full sentence of context, which may be as little as one word (as in glossaries) of as much as 10 or 15 lines of text. The concordance documents 2,994,750 occurrences of Old English words, of which 1,660,134 appear with citations under main entries and 1,334,616 are high frequency functions words (such as conjunctions, prepositions, and pronouns). Omitting full citations for these high-frequency words reduced the bulk of the material by about half.

Unfortunately, in eliminating citations for these stopwords, a few homographs were also omitted. The conjunction *ac* 'but' has the homograph *ac* 'oak'. The few oaks are omitted in the 9,358 occurrences of the character-string *ac*. The 12,369 instances of *for* are almost all the preposition--but among them are two nouns (*for*='journey', 'pig') and one verb, part of *faran*, 'to go'. There are 20,206 instances of *is*, part of the verb 'to be'--but among them are mixed some occurrences of *is* 'ice'.

The concordance is unlemmatized--that is, each spelling or character-string has its own list of citations. This means that the main entries are not gathered together under dictionary headwords, and that homographs are not separated into their discrete meanings. Therefore users must check all variant spellings and inflections of a particular word to find all its contexts and must discriminate between homographs for themselves.

For example, with a word like *cirice*, 'church', one must find citations under nearly 100 different spellings. The word can begin with *c* or *ch*. The first vowel can be *e, ea, i, ie,* or *y*. The second vowel can be omitted, *e, ea, i, ie,* or *y*. The second *c* can be *c, cc,* or *ch*. Inflectional endings possible for each of these spellings include *-a, -ae, -an, -ay, -en, -eum, -on, -ian, -ium, -un, -aen, -e, -es, -um, -ean, -a, -ena*. Most of these spellings are predictable (at any rate, when one looks for possible variants of expected spellings on the alphabetical frequency lists, most of them look as if they are probably the 'church' word). But a few of them are both unexpected and homographs with some other word. An example is *cyri*, where the other instances mean 'choice', usually spelled *cyre*, or *cyrre*, where the other instances are part of the verb *cyrran* 'to turn'.

The only attempt we made to group different spellings was in applying "respelling rules" to the unlemmatized headwords. We respelled them (for alphabetization only; the text forms are still printed exactly as they appear in the Old English citations). We included only variant spellings that we believe are not lexically significant. One type of variation is in equivalent letter forms. Thus *u* and *v* are alphabetized together as if both are *u; e* is alphabetized as *æ; j* as *i; k* as *c; ð* as *þ*; and initial and final *th* as *þ*. A second type of rule treats double final consonants as if they are single for alphabetization. The third rule is that we ignore a slash mark (/) following a word. The slash mark is a signal that the manuscript should be consulted by the reader; it marks a deviation of the text from the manuscript (such as an editor's emendation) or an unreadable spot in the manuscript (such as a hole, burn, or stain). Our respelling rules reduce the number of headwords slightly, and mean that a user does not have to travel around the alphabet to find these few non-significant variants. For example, under the word *þol* are gathered the spellings *thol, ðoll,* and *pol*. And under the headword *cyn* are gathered the spellings *cyn, cyn/, cynn, cynn/, kyn, kyn/,* and *kynn*.

We have now been using the concordance for several months, and have found it even more valuable than we had expected. One advantage is that it allows us to write entries before all slips are filed. We have filed all our computer-matched forms, most of A through E, virtually all of C and D, and many complete individual texts (such as AElfric's homilies and the poetry). We are continuing to file while starting entry-writing with C and D words. Much of our work in entry-writing is easier with the conventional paper slips. Paper is easy to annotate, sort, reorganize, and mark the section of the citation to be used in the entry (dictionary entries will, of course, cite only the part of the sentence relevant to the word being defined). But the microfiche concordance allows us to be sure that we have every slip we should. And some tasks, such as skimming through looking for collocations, are more easily performed with the concordance.

An example of this use is the word *ciricbelle*, 'church-bell'. It occurs only three times in Old English, in two passages in the Leechdom. In each case, the word appears in a context which seems unlikely to a modern reader. The medical recipes are for "demonsickness", and the medicine is to be drunk "from a church-bell". When we skim through the approximately sixty entries for 'bell', however, we find the collocation 'church' plus 'bell'. Clearly 'church-bell' should be defined with both compound and collocation.

We make much use of the alphabetical frequency list. It is essentially a compressed alphabetical list of all character-strings appearing as headwords in the concordance. It gives us quick information on relative frequency of spellings. We are not choosing dictionary headwords simply on the basis of frequency, but is is an important consideration. And it is far easier to find possible related spellings in an alphabetical list without citations intervening, and without having to move through so many pages and fiches.

Another very useful tool is the reverse spelling list which we received the first copies of only a few weeks ago. We made such a list because syntacticians and linguists asked for it. It is a frequency list alphabetized from the last rather than the first letter, right-justified to align the identical letters, but with the words spelled in normal order. Our main use for it so far has been in finding compounds involving words we are studying. It is easy to find all compounds with a common first element, such as *candelbora, candelleoht, candelmaesse, candelstaef.* But without the reverse spelling list it is difficult to anticipate compounds with a common second element such as *daegcandel, friþcandel, heofoncandel,* and *woruldcandel.*

We have found the microfiche concordance so valuable that we have decided to complement this concordance by compiling a supplementary concordance of all the words omitted in the first one. It will include only the 197 spellings that were "stopped"in the earlier concordance. The expense of preparing this one for duplication will be much smaller than that of the first microfiche concordance since the programming has already been done. It is necessary to re-concord each text in order to print the words stopped last time and to suppress those printed last time. But the very complicated problems of merging so many texts have been solved.

We had hoped to begin this complementary concordance early this summer. Unfortunately, we were not able to begin the text-by-text concording until this October. There was a problem which involved our handling of the "respelled" forms of words in the first concordance.

The respellings worked perfectly for the words for which citations were printed. But when we did our first trial run of the "reverse concordance", the respelled forms of the stopwords did not appear. They were in neither the first nor the second concordance.

What appears to have happened is that the stopwords were counted and not concorded on a "character-string" basis. The process was to first concord (eliminating the stopwords), then to respell (bringing those specific spelling variants together), and then to write the tape. This meant that there would have been some words for which there would be both a frequency count and some citations. There would be a frequency count for *ac.* But instances of *ac/, acc,* and *ak* would not have been stopped because the respelling was performed later, and would therefore be printed with citations. However, the headword spelling (which was respelled if necessary) would still be *ac*. The two types

of *ac* headwords apparently could not co-exist (each word was to have one or the other type of entry), so they were combined.

Our programmers had to revise the program so that it now first respells, then concords and prints onto tape. Thus all possible spellings of *ac (ac, ac/, acc, acc/, ak, ak/, akk, akk/, ack, ack/, akc, akc/)* will be grouped together under the headword *ac*.

The concording is now well underway and should be completed in January. Sometime next spring we will have two complementary concordances which together give citations for all words in all surviving Old English texts.