Reuven Merkin

HISTORICAL DICTIONARIES AND THE COMPUTER - ANOTHER VIEW

The idea of a dictionary on historical principles originated early in the 19th century, but only in the second half of the century did such dictionaries begin to appear: Grimm's DEUTSCHES WÖRTERBUCH (1852-1961), Littré's DICTIONNAIRE DE LA LANGUE FRANÇAISE (1863-73) and the WOORDENBOEK DER NEDERLANDSCHE TAAL (from 1864). Historical lexicography attained to perfection one hundred years ago with the publication of A NEW ENGLISH DICTIONARY ON HISTORICAL PRINCIPLES, later known as the OXFORD ENGLISH DICTIONARY (OED 1884/1933). Besides the general dictionaries such as those mentioned above, there are many period and regional historical dictionaries as the MIDDLE ENGLISH DICTIONARY (from 1925), the TRESOR DE LA LANGUE FRANÇAISE (from 1960), the DICTIONARY OF THE OLDER SCOTTISH TONGUE (from 1931) and the DICTIONARY OF AMERICAN ENGLISH ON HISTORICAL PRINCIPLES (1936-44).

The historical dictionary is therefore the most comprehensive type of scholarly academic dictionary, usually covering a national language with a long recorded history. Each entry in this type of dictionary is the word's biography containing a generous selection of dated quotations extracted from literary and other sources and arranged chronologically. There is a twofold aim in giving quotations: (1) to document a word's existence in different periods and genres, and (2) to give evidence of its meaning, grammatical form and spelling, and the changes which happened in the course of time (Merkin 1983).

As to definitions in a historical dictionary, I agree with A.J. Aitken that "the definitions and descriptive notes, which are also a normal feature of such dictionaries, may be regarded as fulfilling a somewhat secondary purpose, that of signposts or labels to the particular subset of quotations which follows" (Aitken 1971:3). Similarly, I feel that the very detailed and multi-hierarchical semantic subdivision of the entry is of secondary importance in such a dictionary.

Most historical dictionaries give a word's external, or comparative, etymology as well as its internal derivation. I doubt that etymology is essential in a historical dictionary.

The existing historical dictionaries treat combinations as well. For example, of a total of 414,825 entries in the OED 26% are combinations of two kinds: special, or defined (11.5%) and obvious, or undefined (14.5%). Moreover, in the first three volumes of the new OED SUPPLEMENT combinations of both kinds constitute 45% of all entries.

Collocations nevertheless have not yet been dealt with systematically in historical dictionaries. There is a view that a thorough treatment of collocations is outside the scope of a historical dictionary (1) because it is very difficult to identify collocations in older periods; (2) because there is an enormous

quantity of them and (3) because they change in the course of time. I do not share this opinion, because the same argument applies to different meanings and semantic changes, and because I believe that with the computer one can make a thorough search of combinations in a huge lexical archive.

A historical dictionary now even more than before is expected to deal with a word's syntactic features and its stylistic and statistical characteristics as well. The TRESOR DE LA LANGUE FRANÇAISE has illustrated that this can be successfully done.

With a few exceptions, experience shows that making a general historical dictionary may take 100 years, or even longer, and a period dictionary may take some 50 years to complete. Large-scale dictionary projects initiated after 1950 rely more and more on computer-generated lexical archives, each containing many millions of quotations. Can this recent development shorten the very long production time of a historical dictionary, or affect its nature which took definite shape a hundred years ago?

A distinguished lexicographer has recently tried to answer this question. A.J. Aitken, the editor of the DICTIONARY OF THE OLDER SCOTTISH TONGUE, who is the fourth generation in perhaps the most impressive dynasty of historical lexicography (namely James Murray, Henry Bradley and William Alexander Craigie) wrote:

> the time taken to edit older dictionaries based on no more than a few million quotations has always been measured in decades or generations... there is no reason to suppose that the editing of modern technologically aided dictionaries can proceed any more speedily. (Aitken 1971:4)
>
> Of the three main stages in a dictionary's compilation (respectively, of collection, of sorting and of editing) this [the sorting - R.M.] is much the smallest... but whereas the computer can far outstrip the human sorter in simple alphabetic sorting, it has no such advantages in the other processes requiring to be executed at this stage - lemmatization and homograph separation. (1971:7-8)
>
> So efficient are the computers at total excerption and inefficient at selection that they present a strong temptation to accumulate more examples than are strictly necessary. (1971:9)
>
> The real benefit of the computer to this area of scholarship lies not so much in its direct contribution to dictionary-making as such, as in a by-product of this - the computer-readable textual archive... this may turn out to be the most important justification for the extensive employment of computer techniques in historical dictionary projects in progress today. (1971:11)

The OXFORD ENGLISH DICTIONARY - and similarly all other historical dictionaries made in past generations - is based on a file of slips produced manually over 70 years by many hundreds of readers, most of them volunteers, who read through some 16,000 volumes and selected some five million quotations. One of the

severe problems for James Murray and his staff was nevertheless that many entries did not have sufficient documentation, and they had to search for it instead of concentrating on editorial work.

As against this situation, editing a dictionary based on a computer-generated archive faces the problem of an enormous excess of documentation for thousands of entries. Let us take as an example the TRESOR DE LA LANGUE FRANÇAISE: in the course of seven years 1000 texts containing 90 million word-tokens had been processed by a computer. Over 70 million words - those taken from literary texts - had been lemmatized and the number of entries found before homograph separation was 71,415, which is an average frequency of about 1000 occurrences per entry. 79% of all entries occur from 1 to 100 times. Usually there is no special difficulty in selecting out of them the best quotations for the dictionary. The difficulty increases as the frequency of entries becomes higher:

the frequency of	10,490 entries	(14.69%)	is from 101 to 1000
the frequency of	3,809 entries	(5.33%)	is from 1001 to 10,000
the frequency of	651 entries	(0.91%)	is over 10,000
the frequency of	14,950 entries	(20.93%)	is over 100 occurrences

As a matter of curiosity it could be mentioned that the 30 most frequent entries occur 30,602,761 times (44.5% of the whole literary corpus of the TRESOR) - which means that on average each entry occurs more than one million times.

Aitken claims that

> a reasonably skilled lexicographer might hope to work out through no more than 10,000 to 15,000 quotation slips per annum... Thus the editing of a collection of 10 million slips might occupy 1,000 'lexicographer-years' or, to be more realistic, 10 lexicographers 100 years. (Aitken 1971:9)

As a matter of fact, Aitken's opinion is based on his own experience in editing a period-historical dictionary of a medieval language, while the TRESOR DE LA LANGUE FRANÇAISE is a period dictionary of a modern language, which makes a significant difference. Secondly, Aitken talks about "a collection of... slips" - an expression which hardly suits a computer-stored lexical data-base, the treatment of which is quite different from that of a manually produced file of slips. If we accept Aitken's calculation without reservations, it would have been hard to explain how the TRESOR staff managed to edit and publish 10 large quarto volumes of the dictionary in the course of 15 years.

To overcome the difficulty of selecting representative quotations out of high-frequency words found in the lexical data-base, it is essential to ask the computer to classify all occurrences of such words according to combinations, without paying attention in advance to whether the combination is an accidental sequence of consecutive words, a habitual collocation, or an idiomatic expression.

When editing a dictionary entry documented, let us say, by 1000 quotations, it is much easier dealing with 20 or 50 groups of formal combinations each containing an average of 50 or 20 quotations respectively, rather than with all 1000 together, whilst deciding for each quotation which meaning or use of the word it belongs to, and choosing the most representative quotations for that particular meaning; all the more so when editing an entry with 10,000 or 100,000 quotations, which is not unusual in a computer-generated lexical archive.

I am talking about that particular stage of dictionary-making at which all quotations of a given entry to be found in the lexical archive are scanned and a certain percentage of them selected for the dictionary. I am not speaking about the final stage of the precise semantic subdivision and about writing definitions (cf. Aitken 1973). In future however it is possible that a more advanced technique of sorting quotations according to combinations (cf. Choueka, Klein and Neuwitz) is likely to affect the final organization of the entry. I mean that the traditional precise semantic subdivision, which ramifies several times hierarchically, and which is taken for granted in most historical dictionaries, including the TRESOR DE LA LANGUE FRANÇAISE, will be superseded by less precise and less branched semantic divisions, while not only idiomatic expressions but also many habitual collocations, syntactic structures and stylistic uses as well will be given special consideration.

Let us think of an entry for which 50 quotations are selected out of the 500 found in the archive, the different collocations and other habitual uses being considered in the process of selection. There are two possibilities for organizing the entry: one is first to arrange ten groups of collocations and uses each containing two, three, or four quotations, and then to give all other quotations, which are difficult to sort, as a single group. The other possibility is to give all 50 quotations as a single group arranged chronologically, number each quotation, and provide definitions with a list of all the collocations, structures and uses and their quotation number. In any case formal criteria such as collocations and syntactic structures should be considered in selecting quotations no less than purely semantic criteria.

Another original conclusion of Aitken is that "the existence of computer archives would often seem to remove the need to burden library shelves with still larger dictionaries filled with still more detailed information of interest to only a few people" (Aitken 1971:16). To a similar conclusion came Ladislav Zgusta, who believes that in future

> large academic dictionaries will not be published any more. The point is that even the academic dictionaries which consist of ten, twenty, or any number of volumes, do not and cannot present the whole material contained in the archive... then why publish a twenty-volume reduction of the material if a one, two or four-volume reduction could suffice for the first information, which must eventually be followed by the archive search, in any case? (Zgusta 1971: 354-5)

It is not impossible that Zgusta had in mind the excellent

example of the SOED in two large volumes, yet it comprises the essence of the complete OED in twelve volumes, not a reduction of a computer lexical archive.

I am not referring to the possibility that in future books of encyclopaedic scope will no longer be printed, but stored in the memory of a computer. I am talking about the possibility mentioned by Zgusta that in future a printed historical dictionary of two or three volumes "could suffice for the first information" while for any further study the complete lexical archive stored in the computer should be used.

To share this view, I believe, certain conditions must first be fulfilled. A central national, or international, lexical archive must be established, containing tens or hundreds of millions of quotations, stored and operated on-line, and easily, quickly and cheaply accessible by any user anywhere and at any time with his, or her personal terminal.

There are still some technical problems to solve and difficulties to overcome until such a lexical data-base is successfully run, and printed dictionaries still have a few advantages, but it is obvious that viewdata systems, or, as they are called today, videotex, are being developed faster and faster, and it is only a matter of time until such a system is adopted for lexicographic use.

Several full-text data-bases have already been established in different countries. My daily personal experience is with the Responsa Project at Bar-Ilan University (Israel) containing 45 million words, stored on-line, and operated by a sophisticated information-retrieval system (Choueka 1980).

Although technical difficulties seem to be overcome in the course of time, there are still other problems in the way of a computer-stored lexical data-base becoming an efficient alternative to a complete historical dictionary.

(1) The "one, two, or four-volume reduction" mentioned by Zgusta can only serve as an index, while the whole archive is too large to be referred to every time when one needs a dictionary.

(2) A historical dictionary (as opposed to a stylistic dictionary of a single author, for example the GOETHE WÖRTERBUCH (from 1966), or the SLOVAR' JAZYKA PUŠKINA, 1956-61) aims in principle at giving evidence of what is usual and typical in the vocabulary of a particular period or genre. Thus a single quotation, selected from several hundred which testify exactly to the same usage or collocation from that period or genre, is much more useful and economical than the full list, probably in 99% of the cases in which one needs lexical information.

(3) In order that lexical information may be efficiently retrieved from a computer-stored archive, the archive should be organized according to some basic principles of lexicography, even if those principles are not exactly

> the same as in normal printed dictionaries. Three matters should be discussed in this context: homograph separation, polyseme separation and lemmatization.

To illustrate the difficulty of homograph separation in a large computer-generated lexical archive, let me give an example from work on the TRESOR DE LA LANGUE FRANÇAISE. Some 800 literary sources from the 19th and 20th century were processed, containing some 71 million running words. To compile a DICTIONNAIRE DES FREQUENCES this enormous collection was examined quantitatively, and it was found that almost 9.5 million words (13%) belong to word-forms which are considered homographs in French. Since it seemed impossible to read through 9.5 million quotations in order to separate homographs, the editors adopted the following technique: almost half of the words (4,639,591 = 48.91%) were not checked at all, because they seemed in advance to be one-entry words at least in 99% of occurrences. Of the other half (4,847,080 = 51.09%) a sample of some 350,000 words (3.7% of all 9.5 million) was taken and checked thoroughly.

Even if homograph separation is still possible at the stage of preparing the lexical archive, although time-consuming and expensive, separating polysemic words (in a historical dictionary almost every word is polysemic) is possible only within the framework of a selective treatment. I cannot imagine polyseme separation within a lexical archive containing tens of millions of words.

Let us suppose that a thorough semantic separation of polysemes is not essential for the quality of a computer-stored lexical data-base. Still the question of lemmatization seems to be both essential and very complicated, especially in highly inflected languages such as Russian or German, or in Semitic languages like Hebrew and Classical Arabic in which there is an enormous number of homographs - far larger than in any European language - because of peculiar orthography which marks mainly consonants, but only a small number of vowels (Busharia 1979). Imagine that in English common words such as and/end are homographs (which is the case with hundreds of the most common words in Semitic languages).

Let me quote from an as yet unpublished article by Choueka and Lusignan:

> on the one hand lemmatization is one of the most important and crucial steps in any non-trivial text-processing cycle, but on the other hand, no operational, reasonably general, fully automatic and high-quality context-sensitive text lemmatization system nowadays is easily accessible for any natural language.

Since such a lemmatization system does not yet exist, we should restrict our considerations to the existing techniques of automatic lemmatization and, in addition, to some future developments and improvements (Choueka and Lusignan). I believe that in future more sophisticated computer techniques, specially designed for lexicographers, would help them not only at the stage of collecting and sorting data, as is common nowadays, but also at the stage of establishing a lexical archive based

on scholarly principles, using algorithms for homograph separation, collocation retrieval and lemmatization.

In future it will be within the ability of lexicographers and computer-men to lemmatize a gigantic lexical data-base in a reasonably short period of five, ten or fifteen years, which will eventually determine whether such a data-base could be the alternative to a historical dictionary. Perhaps for some languages it might be possible, while for others - not, and then it will be unavoidable to select from the complete archive its best, organize it according to lexicographic principles in order to give enough information to 99% of users, and leave the whole archive for special cases of study.

To the question as to whether it is possible to shorten the very long time needed to produce a historical dictionary based on a computer-generated archive, I would venture to reply in the affirmative, but I believe that the way it will be produced and its general character will be quite different from the ones which have become common in historical lexicography in the last hundred years.

To the question whether it is possible to replace the voluminous academic dictionary by a computer-stored lexical archive, my reply is: I am not sure. Perhaps in some cases the investment in organizing such a data-base according to lexicographic principles might be justified. Generally I am not sure that a computerized lexical data-base can be the alternative to the selective work of the lexicographer.

In future, I believe, every large-scale dictionary project will face the dilemma of whether to invest great efforts and money either in producing a voluminous academic dictionary based on a huge lexical archive, or in organizing the archive itself in a more scholarly way.

I am not convinced that the age of large-scale historical dictionaries is over. We should be careful with forecasts concerning the future of lexicography. R.W. Chapman, Secretary to the Delegates of the Oxford University Press between 1920-1942, the period in which the OED was completed, wrote in 1948: "There is high authority for the view that the day of the comprehensive general dictionary...is over. The ineluctable curse of specialization is branded upon us" (Chapman 1948:16). Chapman died in 1960. Had he lived longer he would have witnessed the enormous prosperity of all branches of lexicography in the last two decades and might have changed his opinion.

References

Aitken, A.J. (1971) "Historical dictionaries and the computer" in The Computer in Literary and Linguistic Research ed. by R.A. Wisbey. Cambridge: U.P.

Aitken, A.J. (1973) "Sense analysis for a historical dictionary" in Lexicography and Dialect Geography. Festgabe for Hans Kurath ed. by H. Scholler and J. Reidy. Wiesbaden: F. Steiner

Busharia, Z. (1979) "Computerized lemmatization of non-vocalized Hebrew texts" in Proceedings of the International Conference on Literary and Linguistic Computing (Israel) ed. by Z. Malachi. Tel Aviv: University

Chapman, R.W. (1948) Lexicography. London: Oxford U.P.

Choueka, Y. (1980) "Computerized full-text retrieval systems and research in the humanities: the Responsa project" Computers in the Humanities 14, 3: 153-169

Choueka, Y. et al. (forthcoming) "Automatic retrieval of frequent idiomatic and collocational expressions in a large corpus"

Choueka, Y. and Lusignan, S. (forthcoming) "Disambiguation by short context"

Merkin, R. (1983) "The historical/academic dictionary" in Lexicography: Principles and Practice ed. by R.R.K. Hartmann. London-New York: Academic Press

Zgusta, L. (1971) Manual of Lexicography. The Hague: Mouton