Murray T. Wilton

BILINGUAL LEXICOGRAPHY: COMPUTER-AIDED EDITING

Bilingual dictionaries present special difficulties for the lexicographer who is determined to employ the computer to facilitate his editorial work. In a sense, these dictionaries include everything contained in a normal monolingual edition and a good deal more. The single-language definition dictionary is consulted as the authority on orthography, pronunciation and stress-pattern, grammar, level of formality, field of application, definitions, examples, usage and etymology. A bilingual dictionary which purports to be more than a pocket edition will treat all of these with the exception of etymology, which is not normally in the domain of the translator. In addition, it will devote itself to providing accurate translations, which necessarily presuppose an intimate acquaintance with the correct definitions in both languages. Such a dictionary is a far cry from its mediaeval ancestor, the two-language glossary, which was usually a one-way device furnishing equivalent forms for simple words and expressions in the opposite language. The modern bilingual dictionary is usually two-way, each section constituting a complete dictionary in its own right and contrived to cater for a variety of translation requirements. Yet the two sections are inextricably linked by an intricate network of translations and cross-references which guide the consulter and ensure that he does not falter when semantic equivalence fails to overlap smoothly. Since semantic equivalence is the important basic feature of bilingual dictionaries, deviations from the normal pattern will require special treatment. In closely related languages, like French and English, numerous pairs of words of common origin are only slightly, if at all, altered in their modern form (e.g. Eng. *versatile*/ Fr. *versatile*). But the disparate development of two modes of expression in different cultural and historical environments has left a residue of such word pairs whose only similarity is in fact the visual image of the sign. Their definitions are often very remote from each other. It is yet another task of bilingual lexicography to distinguish clearly between the meanings of these deceptive cognates or " faux amis ".

These, then, in very brief outline, are some of the features common to all good bilingual dictionaries. *The Canadian Dictionary* (*/Dictionnaire canadien*)[1] is no exception to these general remarks. First published ten years ago under the editorship of Professor Jean-Paul Vinay at the University of Montreal, it is now undergoing a major revision and updating at the University of Victoria, still under Vinay's supervision. The new editions should see the corpus of the original version increased from 40,000 to about 100,000 entry words. The first edition was specifically tailored for the unique linguistic situation in Canada and takes into account the two main dialects of each of the official languages it represents, namely, European and Canadian French, and British and Canadian English. This, however, is a gross simplification of a complicated dialect situation fraught with all the problems associated with social and official acceptability. But it is sufficient for the purposes of this discussion to mention that a good deal of importance is attached to Canadian content in both languages, thereby adding a further unit of complexity to the material to be presented. Accordingly, in addition to the data common to all bilingual dictionaries, *The Canadian Dictionary* furnishes information on the dialect status of most words and expressions.

In the textual published form of the dictionary a variety of symbols and abbreviations are employed to designate special features, such as dialect type, semantic level, range of formality, " faux amis " and so on. The basic element of the manual editing process is the file card upon which all the data for a given entry word is recorded in textual format. This file card, which is designed to conform to the exigencies of the final typographical layout, is called the " physical fiche " to distinguish it from the file card to be used in computer-aided editing, known as the " software fiche ". It includes, in addition to lexicographical material, details of the source of the words and expressions and of their translations as well as the signature of the autor of the fiche. During manual editing these details may be required for verification purposes.

The first stage in converting the complex textual data to machine-readable form is the recording of the material on the " software fiche ", which has been designed in such a way that it is readily interpreted by keypunch operators and yet remains a comprehensible manual file at the same time, in case manual editing should be necessary. The structure

[1] J-P. Vinay, *et al.* (1962).

of this fiche reflects the conceptual format of the informatic data file, which must be so organized that useful manipulations can be effected rapidly and accurately within the file and across files in either or both languages. The unit of manipulation in each data file is the entry word, the address for the whole file which enables the file itself to be accessed or sorted according to editorial requirements. Interior addresses enable the file to be structured so that data can be " exploded " in order to reveal certain elements which that file contains in common with other files and which the editor wishes to cluster for various reasons. For instance, it will be possible, under this system, to obtain a printout of all words concerned with, say, *information theory*, so that the total list can be submitted for verification to a translator with expertise in this field. This kind of procedure is precluded in manual editing unless the corpus is built up and maintained in subject-fields; even then, it is impossible to ensure that all the words in this area have been successfully gathered, since many will belong to other fields as well; and at some stage of development the fields must be disintegrated in order to position the words in alphabetical order within the corpus.

Interior addresses are to be reserved for key items which the editors want to be accessible for various reasons: grammatical category, pronunciation, subject field, dialect, level of formality, usage, translations, cross-references and usage notes being those envisaged at the moment.

Three phases have been anticipated for this project. The first phase, which is presently being operated, consists in the production of simple " skeleton " word lists containing a few essential features such as grammar, subject field and dialect, but without translations or any other textual material. The purpose of this exercise is to be able to produce lists for verification by spelling and pronunciation consultants and to have check lists of words by subject matter without the necessity for breaking alphabetical dictionary order. The second phase will elaborate on phase one by adding textual material and transforming the entire corpus to computational form. The third phase is a feasibility study of the pros and cons of using automatic typesetting procedures in the final published version.

One of the basic problems studied in the first phase was the necessity of maintaining the entry word as the unique address for its file without requiring special programmes to be written to account for accented words in sorting procedures. The question of homographs of the type *fine* (noun), *fine* (adjective), *fine* (verb), *fine* (adverb) and French *fine* (feminine noun), *fine* (feminine adjective) is resolved by

*

the grammatical function attached to the word which readily distinguishes one file from the other. Where the main difficulty arises is in the orthography of accented French words and capitalized words in both languages. If printouts of the French wordlists were to be correctly spelt with accents included, using a French print train, some means had to be devised for representing the spelling internally without interfering with the address. Any intrusion of special symbols to dessignate accented letters would mean that existing programmes for sorting would have to be abandoned and special programmes written. Two solutions presented themselves to avoid doing this. LEXAUTOM, an automatic dictionary research project at the University of Victoria,[2] represented the entry word twice in two separate fields of each file. The first field contained the entry word without special indications for accents (e.g. *été* is represented internally in the entry field as *ETE*). This procedure enables the entry word to be sorted and accessed with existing programmes. The second field, which is located at the end of the file, is known as the " graphics " field and duplicates the entry word but with the addition of special characters for internal representation of the thirteen accented letters of French. Unaccented words would have this field of their file left blank, although the fixed format used requires that the appropriate space be reserved anyway. The system was devised for a limited corpus of 100 words for which it was adequate. However, in a corpus of 100,000 words it would be somewhat cumbersome; it involves duplication of information on the fiche, is wasteful of space and makes sight checking of input cards rather difficult. A second solution is to have a " code " field immediately following the " entry " field in which special characters indicate that a given letter of the " entry " field is to be treated in a certain way. For instance, to ensure that *pécher* (verb), *pêcher* (verb) and *pêcher* (noun) are distinguished from each other and placed in their correct alphabetical order, or to show that the noun *Italien* is to be capitalized to distinguish the person it denotes from the noun *italien*, which indicates the language, and the adjective *italien*. The code system renders duplication unnecessary and is much less wasteful of storage space since accented words usually contain only one, or at most two or three accents and capitalization involves only the first letter of the word. In sorting procedures the " entry " field provides the cue for positioning while the " code "

[2] See J-P. VINAY, B. KALLIO, *Bilingual Lexicography and the Computer*, University of Victoria (mimeographed), 1971.

field ensures distinction between homographs and correct spelling for printouts.

In the first phase, because of the limited size of files, it is possible for a fixed-length format to be employed without undue storage waste. A typical file will contain the entry word, its grammatical function, subject field or fields and a dialect indication where necessary. Any of these features, located at different but fixed levels in the file, will be accessible. Other non-accessible elements which may be useful in printout lists are cross-references, usage notes, and an indication as to whether inversion of the fiche has been carried out.

The second phase of the project involves the conversion of the entire corpus to machine-readable form. If programme systems for the " skeleton " word lists have been successful, this phase should be an extension of the first, avoiding the necessity to duplicate keypunching of all the material. The adoption of on-line procedures would enable all editorial tasks in the final stages to be accomplished rapidly and automatically, perhaps from remote terminals with video screens. The data file in this phase is clearly much more complex and gives rise to the problem of variability in file size; in textual form, entry word " files " range from one or two lines to several columns or even pages of continuous typographical text. In order to maintain the capability of access to certain elements of the file, these elements must be located at fixed levels. But a fixed format file system is tremendously wasteful of storage space since, in theory at least, every file must have at least the capacity of the largest of them, whereas most files are, in fact, relatively small. Thus the only solution, other than employing an individual programme for each file, is to design a variable length format.

The question of variability is currently being studied by programmers at the Computer Centre of the University of Victoria and I.B.M. (Canada) Ltd. The most likely outcome of this study is a theoretical conception of the data file based on the " catalogue " system devised by the Rand Corporation for the production of the *Random House Dictionary* (monolingual) but suitably adapted in certain respects for the bilingual situation. The Rand system was designed with a structure which could accommodate new elements of variable size and yet remain subject to a single management programme applicable to every file. The individuality of each file is preserved by means of a key, a reproduction in miniature of the contents of the file. This key, in coded form, describes the data classes included in the file and the byte space occupied by them. The key is always interpreted first, prior to the reading of the file itself, and therefore has the total effect of an individual programme for that file. For *The Canadian Dictionary* the catalogue would contain a common structure of ten or eleven levels corresponding to the elements for which an accessible address is required. These discrete levels are fixed, so that the code for their access remains constant; it is only the coding of the variable contents of the file which is flexible.

If it becomes possible to employ a system of this type, given the expense and difficulties of keypunching a vast quantity of complex textual data, it will permit the editors to manipulate their material in various ways without requiring a large research team and many man-hours of work. One particularly time-consuming and laborious task carried out in manual bilingual lexicography is the inversion of the contents of a fiche. This procedure is to ensure that translations of entry words or expressions in the source language are converted to entry words or derivative expressions in the target language. In fact, the circularity of dictionaries is such that enormous quantities of a bilingual work could theoretically be built up by a continuos " back-and-forth " movement of translation equivalents, each translated word giving rise, in turn, to a whole series of different words and expressions. In practice, however, inversion is carried out only to establish the existence of immediate translations in the opposite-language section of the dictionary. Computational procedures would accomplish this far more rapidly and fully than the present manual methods and should ensure that, as far as possible, semantic closure is achieved. Similarly, many other editorial tasks, from simple checking of spelling to complete updating of files, will become relatively elementary operations consuming far less time and effort. But computer-aided bilingual lexicography looks further ahead than the mere acceleration of the editorial process. The automated dictionary will overcome the problem shared by all new editions-the stigma of instant obsolescence at the very moment of birth. For a new edition can now be produced virtually at a moment's notice with the rapid addition of new material and weeding out of outmoded forms. The time required to produce such a revised edition can be reduced from decades to months, or even weeks, and in general there will be a proliferation of useful consultative material which is always up to date.