Optical Data Storage and Dictionaries

Marcel Lemmens and Herman Wekker

The aim of this paper is to draw the attention of lexicographers and publishers to a new optical data storage medium, the compact disk. What we want to plead for, as practising trainers of translators and teachers, is a careful and critical consideration of the possibilities that this new medium can offer us.

We want to suggest how in our view optical data systems such as CD-ROM and WORM could be most efficiently exploited for lexicographical purposes, what requirements optically stored 'dictionaries' (in the new sense) should meet to be adequate reference tools, and what effect the use of such media may have on the contents and format of such dictionaries. We wish to discuss the concept of such dictionaries rather than the technical details and problems that are no doubt involved in the realization of our ideas.

What is CD—ROM? WHAT is WORM?

CD-ROM (Compact Disk Read Only Memory) is an information storage medium based on the same technology as that of the now very popular audio CD. Both CD-ROM and audio CD are information systems that make use of non-erasable disks and special players which read the digitally stored data from these disks by means of laser beams. The main difference between CD-ROM and WORM (Write Once Read Many times) is that in the case of CD-ROM no information, once stored, can be deleted or changed, whereas in the case of WORM new data can be added (but only once; once added, it cannot be erased or changed). WORM also requires a different and more expensive player. The environment required for these new media is that of a normal personal computer (640K) with a hard disk. The information on the disk can be manipulated by means of the keyboard and shown on the PC screen. An important feature of the CD-ROM and WORM disks is their enormous storage capacity, which is generally estimated at approximately 200,000 typed pages of A4 per disk (more than half a gigabyte of data). CD-ROM and WORM have been used mainly so far for storing written databases such as term banks, bibliographies, medical and legal records, encyclopedias and catalogues of books in print. As far as we know, three English dictionaries are at the moment available on CD-ROM. The largest of the three is THE OXFORD ENGLISH DICTIONARY (OED) (with a total of nearly half a million entries), which is stored on two disks, and costs $1,250 per set. The second is the WEBSTER'S NINTH NEW COLLEGIATE DICTIONARY (W9). This is the first 'talking' CD-ROM, designed for use with the Macintosh microcomputer. Apart from the normal dictionary definitions and illustrations, it provides the pronunciation of words through loudspeakers or headphones. The price of this product is $195. Finally, McGraw-Hill has stored the DICTIONARY OF SCIENTIFIC

AND TECHNICAL TERMS on one disk, together with the CONCISE ENCYCLOPEDIA OF SCIENCE AND TECHNOLOGY. The price is unknown to us. The price of CD-ROM players (also called drives) ranges from $600 to $1000.

We have also heard about two current projects in which Dutch publishers of dictionaries are involved. Wolters has collaborated in a project with some British, French, German, Italian, Spanish and Japanese firms with the purpose of storing a multilingual dictionary database on CD-ROM. Van Dale has had an important share in developing a prototype for a CD-ROM language tool, the so-called MOBIDICK project, which will consist of Van Dale translating dictionaries, a grammar component and a writing aid. The other partners in the development of the prototype were Le Robert and ALP Systems.

Dictionaries on CD-ROM

A complaint that is often heard with respect to traditional printed dictionaries is their lack of space. Publishers use it as an argument against including more entries, better definitions, more examples, etc. This is a problem which can be solved, in our view, by using optical disks as storage devices, since their storage capacity is enormous. We regard this as one of the greatest advantages of the optical media, and they deserve further consideration. It appears that at least 6 or 7 one-volume dictionaries of reasonable size (between 1000 and 1200 pages) can be stored on one single disk. This means that if we were to restrict ourselves to 3 or 4 dictionaries on a disk, this would leave us more than enough space to increase and improve the contents and quality of the dictionaries.

Not only is it possible to expand and change existing dictionaries, it also seems desirable to put more than one type of dictionary (dictionaries which should ideally complement each other) on one disk, for example a combination of one or two monolingual dictionaries with two bilingual ones (L1—L2 and L2—L1).

Another advantage of optically stored dictionaries, as compared with traditional printed dictionaries, is the relatively high speed of access and the ease with which information can be found. It is true that the access speed of CD-ROM is considerably lower than, for example, that of hard disks in PCs; the average maximum access time of CD-ROM is 1—2 seconds. Even so, this is obviously much faster than the speed with which printed dictionaries can be consulted. A related advantage is that cross-referencing can be optimized and that dummy entries can be dispensed with, so that searching becomes more effective and less frustrating. For example, in the LDOCE (2nd ed. 1987) the entry for the word *patella* only contains a cross-reference to *kneecap*^{1} (1), with the abbreviation *med* which stands for 'medical term'. The user is expected to look under *kneecap*^{1} (1) for a definition; in this case the entry for *kneecap*^{1} (1) also contains a reference to an illustration (skeleton). This seems a rather cumbersome way of tracing the meaning of *patella*. The COBUILD is not any better. It also uses the non-medical equivalent *kneecap* in the definition of *patella*, but it gives no explicit cross-reference for users who do not know what a kneecap is. Nor is there a reference to a picture. In principle, CD-ROM can solve this problem by automatically picking up all the relevant information and displaying it on the screen at once. The relevant information might include (near-)synonyms, antonyms, 'false friends', alternative spellings, style variants and illustrations. Information of this kind may be included in the entry itself, but if the entry should become too long, the relevant information can be given in separate entries, but immediately on the same screen, so that it can be easily accessed by means of scrolling, i.e. by moving up or down the page(s).

With the proper software and with extensively indexed dictionary entries, users should be able to define their own wishes and selections as to what they want to see on the screen. This may mean that users should be able to choose between a menu-driven system and a command system. For example, in the case of a command system the experienced user can quickly specify and call up information that he needs. If he should require synonyms or related words, he should be able to study and compare them on the screen. An example of a command system is Medline, with SilverPlatter software (version 1.4). Medline is a bibliography of medical publications with information on title, author(s), source, language, year of publication, and country of origin. Many entries also contain an abstract. By means of a fairly easy set of commands, most of which are self-explanatory, the Medline user can specify about which (kind of) article(s) he wants to have full information. After a FIND prompt one can, for example, ask for articles that have the word *arthritis* in the title (command: arthritis in TI) and/or that have been published in 1984 (command: 1984 in PY) and/or that have been written in English (command: English in LA). It only seems a small step from this application of the medium to dictionaries. Similar retrieval possibilities could enhance the exploitation of data contained in the reference tool on disk. It might, for example, be possible to check whether two particular words collocate or not by keying in those two words as well as a command to make the system search for appropriate examples.

Since the CD-ROM player is equipped with a sophisticated error correction system for the laser beam, the user will hardly ever be disappointed when he wants to access a particular item from the disk, i.e. provided the right commands are given, no typing errors are made, and the item which is wanted is in the dictionary. This means that there will be practically no 'alphabet misses', the time-consuming problem of not searching according to the right alphabetical order in printed dictionaries.

Another advantage we should like to mention, finally, relates to the many compound nouns, phrases and idioms which have always posed problems to dictionary users, because it is often not clear under which headword they can be found. These items can be made directly accessible in a CD-ROM dictionary, if they are indexed. The user need no longer decide which entry has to be consulted to find the meaning of, for example, *It's raining cats and dogs*. Keying in a combination of the truncated words *rain** *cat** and *dog** will take him/her straight to the right place. (This system of truncated forms is also implemented in SilverPlatter).

Apart from the advantages of the new type of dictionary there are also some apparent disadvantages.

The first of these disadvantages seems to be that a work-station consisting of a personal computer with a CD-ROM (or WORM) player is rather static. Although the optical disk itself can be easily removed from the player, the retrieval system as such is not portable. It is also unlikely that a CD-ROM dictionary will soon be able to compete with the printed dictionary as far as portability is concerned. On the other hand, the disk can contain more and bigger dictionaries (one does not carry around WEBSTER'S THIRD INTERNATIONAL very often either). This feature should be particularly attractive for professional dictionary users who normally do their work on PCs anyway and who frequently require a multifunctional language tool in order to do their work properly. For them an optically stored dictionary could be both an integrated medium that can be directly accessed from the work station and an enhanced reference tool.

A second objection could be that electronic dictionaries are far more expensive than printed books. However, given certain circumstances, the optical disk has so much more to offer that it could prove to be good value for money. The individual user can tailor the multi-dictionary tool to meet his/her own needs, so that the speed and accuracy with which specific information can be found in the bulk of data is increased.

The PC screen on which the dictionary data has to be displayed is limited in size; it normally has 24 lines and 80 characters per line. Moreover, part of the screen has to be reserved for menu options and prompts. This means that in the case of long entries or combinations of entries the dictionary user has no general overview of all the components they consist of. If, however, the software provides both a search and a browse mode, the user can not only extract specific details from the dictionary, but can also 'leaf through' his/her database in alphabetical order by means of scrolling. The user should, for example, be able to ask for all the entries from *effective* to *egg*.

Since our ideal dictionary on an optical disk is one that includes both translating dictionaries (L1—L2 and L2—L1) and two monolingual dictionaries, one for L1 and one for L2, cooperation between different publishers is required. This may be a serious problem. Yet the publishers involved may also benefit from such cooperation. First, their dictionaries may penetrate new markets, since dictionaries on CD-ROM can be aimed at the native speakers of both L1 and L2. This may also have spin-off effects on the printed dictionaries offered by publishers. And second, the CD-ROM product requires a new approach to the compilation of lexicographical material. The experiences gained from this new approach may also be beneficial to other lexicographical products of the contributing publishers. Moreover, both points should have positive effects on CD-ROM dictionaries and their price. Close coöperation between publishers should improve the quality of these dictionaries, whereas the retail price of each disk may be lower because of a potentially larger market (two markets: that of the speakers of L1 and of L2).

The target group

It should be obvious that we see dictionaries on optical disks as supplementary tools to traditional printed dictionaries rather than as replacements. There will always be a need for the printed dictionary, particularly among learners and the general public. However, for professional dictionary users like translators, secretaries, correspondents, text writers, and also for language centres and libraries the dictionary on CD-ROM could be a very efficient reference tool. To them the enormously large database and the speed and efficiency with which it can be consulted are extremely important advantages. Furthermore, they do much of the work that involves reference to dictionaries on PCs in an office environment, so that the necessary electronic equipment is already available (which means that there will be no extra costs.). An additional advantage is that these people are used to word processing. They will not easily be scared off by the new medium and its system of accessing data.

Especially for translators it might be interesting to think of a way to enable them to add their own terminology to the database, so that they need no longer go through the hassle of having to fill in separate terminology cards. Some kind of standard blank terminology sheet which is compatible with the dictionary entries should be provided by the software. In the case of CD-ROM the added terms will have to be stored on the hard disk or a floppy. In the case of WORM the new terms can be written on the optical disk. It should be possible to do all this on the PC that one is working with. Features like this make the dictionary an even more powerful language tool. Ideally, the CD-ROM disk should also contain specialist dictionaries for specific user groups, but we realize that in many cases this would not be commercially viable. For some industries or multinationals, however, it might be worth considering.

Conclusion

The aim of this paper has been to draw the attention of lexicographers and publishers to a new data storage medium that could be used for specific lexicographical applications. Printed dictionaries need not be replaced, because they will always remain useful to so many people. At the same time, however, they have certain restrictions like limited space and a strict alphabetical order. Any new medium which could be used to overcome these restrictions deserves further exploration.

The optical disk is such a medium. Like the printed dictionary it has advantages and disadvantages. Many of the disadvantages of the traditional reference book could be solved by using CD. The positive characteristics of the new tools are: an enormous storage capacity, a relatively high speed of access and efficient retrieval of data. The so-called disadvantages of dictionaries on an optical disk disappear in practice. CD-ROM and WORM will normally be used in an environment in which a PC is available, and most users will be professionals who have experience of data processing. This is the environment in which professional dictionary users often work. It is this user group that would greatly appreciate an extensive, flexible and user-friendly dictionary and that would be prepared to pay for a more expensive system that actually turns out to be cost-effective. Admittedly, this market may be relatively small and there are still some problems to be solved, but in the long run publishers could also benefit from the investment.

Cited Dictionaries

COLLINS COBUILD ENGLISH LANGUAGE DICTIONARY (COBUILD). 1987. John Sinclair *et al.* (eds.). London and Glasgow: Collins.

LONGMAN DICTIONARY OF CONTEMPORARY ENGLISH (LDOCE). 1987^2. Della Summers *et al.* (eds.). Harlow and London: Longman.

McGraw-Hill Concise Encyclopedia Of Science And Technology. 1984. Sybil P. Parker (Editor in Chief). New York, etc.: McGraw-Hill, Inc.

McGraw-Hill Dictionary Of Scientific And Technical Terms. 1984^3. Daniel N. Lapedes (Editor in Chief). New York, etc.: McGraw-Hill, Inc.

The Oxford English Dictionary (OED). 1884—1933. James Murray *et al.* (eds.). 12 vols. Oxford: Clarendon P. & Oxford U.P.

Webster's Ninth New Collegiate Dictionary (W9). 1983. F.C. Mish *et al.* (eds.). Springfield MA: Merriam.