Otto Vollnhals

UTILIZATION OF A COMMERCIAL LINGUISTIC DATA-BASE SYSTEM FOR ELECTRONIC STORAGE AND AUTOMATED PRODUCTION OF DICTIONARIES

Introduction

In the age of electronic data processing, data-base systems have given rise to new approaches in lexicography. The computer can do much to facilitate and speed the tedious work of editing vocabulary material, to secure greater accuracy and to produce the photocomposition output immediately after editorial completion. This has contributed considerably to invalidate the old truth that dictionaries are outdated at the time of their publication.

In my paper I would like to explain how a modern data-base system developed for linguistic purposes, i.e. the TEAM system of Siemens AG, Munich, which for more than 10 years has been successfully employed for specialized terminological work, can also be used to store, process and finally publish with the aid of electronic photocomposition the complex information contained in an extensive standard-language dictionary.

Let me illustrate the characteristic phases by giving a practical example for:

(a) compiling the dictionary data-base;
(b) processing options;
(c) output options;
(d) some further applications.

Example

The first essential step before data entry is careful consideration and determination of the input format. This means breaking down the individual dictionary entries into separate information categories, to make them independently accessible in the data pool. These categories represent the basic schedule for data entry and electronic data storage.

At the moment our system comprises 98 information categories for each dictionary entry to be stored. A schedule which would be suitable for a relatively large bilingual dictionary has about 50 such categories.

Later on we will see that these 50 categories are not needed for each entry. Consequently, there is plenty of room for individual extensions (for example, if authors or publishers want to input or store additional data, such as source, valency, reliability identifiers etc., information which normally does not appear in the printed book, but which is important when using and managing the dictionary data-base).

After having set up the dictionary data-base, which is, of course, of paramount importance for the successful outcome of all subsequent stages, we enter the fascinating phase of computer-aided lexicography which starts at the moment the data are stored electronically. Authors and editors employing the computer and appropriate programs have an enormous potential at their disposal for rendering their dictionary work more efficient than ever before.

It would be a great mistake, however, to think of the computer as a potential simulator of a dictionary editor's behaviour. The intellectual tasks entailed in dictionary-making, such as researching the meaning of terms in different languages, will remain the province of man now and for some time to come. But there are tasks that can be conveniently delegated to the machine, and it can do much to speed and facilitate work.

Some of these tasks are:

(a) Alphabetic sorting. By using computers the amount of time saved can be enormous. (Recently I had to sort one of our large technical dictionaries containing some 160,000 entries French-English ... and the machine did it during my lunch-break!)

(b) Updating. The computer-encoded dictionary allows regular revision and alteration as new evidence comes to light. And, what is even more important, right up to the start of typesetting the author can modify entries stored in the computer as he pleases, with virtually no restrictions. He is also free to introduce any number of new and exceptionally up-to-date terms without upsetting either the alphabetical order or the printing format. In a matter of minutes the computer integrates all changes and additions into the work and suitably rearranges the surrounding material.

(c) Error search. Some programs are also excellent tools for finding errors. In this application they far outmatch the performance of human beings, not only by their enormous speed, but also by their uncompromising precision.

(d) Other checks. The computer permits numerous additional checks which could not be performed with the same precision by using traditional methods, e.g. reference checks, statistics, uniformity checks, coordination of defining vocabulary, etc. (in a former version of Wahrig's DEUTSCHES WÖRTERBUCH the word Tasteninstrument was not listed as a headword, although it appeared several times in definitions).

(e) Selection options. On the basis of numerous criteria, terms can be selected from the data-base in any quantity and order, and with practically any combination of elements. For instance, selection can be made by requesting:

- a particular source,
- a particular subject field or fields,
- a certain part of speech,
- a certain usage level,
  ... etc., etc.

All selection criteria can be applied separately or in combination. Selection programs are very useful as well for producing condensed versions of larger dictionaries.

(f) String manipulation. In German, the spelling of words, especially of scientific terms, often changes. (The official German spelling of Jod, for example, was changed to Iod some time ago. It is no problem at all for the computer to sort out those strings containing Jod and to perform the required replacements. Of course, revisions like this cannot be left solely to the computer. But while subsequent checking by human editors is still required, the overall time savings are significant. In our example, the computer would, of course, change the verb jodeln to iodeln as well!)

To summarize the advantages of this option, let me say that although the computer cannot carry out exacting analytical tasks, it can handle huge amounts of data in a short time and can make them more manageable and more convenient to use.

As far as the various output possibilities are concerned, I would like to confine myself to the photocomposition options, because this aspect is the most interesting in our context.

We can assume that modern photocomposition hardly sets any limits to typographical layout. The printing format can be varied in any way prescribed by the authors or publishers. The sequence and arrangement of the terms and supplementary information in the dictionary entries may be changed by altering several parameters. The program automatically formats lines, columns and pages, carries out pagination, and generates guide words or running heads, and initials at the beginning of a new letter of the alphabet.

You will probably have noticed that the stage of traditional typesetting has been skipped. The typesetting work is performed by the CRT photocomposition equipment, which is capable of printing thousands of lines per minute. This means that even for bigger volumes, typesetting can normally be done in one day. Equally important is the fact that this kind of typesetting prevents new typographical errors being generated. In addition, different products can be created from the same data by using other parameters.

Special requirements with regard to symbols and other special characters can be easily satisfied.

A few more applications

The detailed input format lends itself to generation of various other products which can be derived from the stored vocabulary without difficulties by using appropriate programs.

Examples:

- reverse dictionaries,
- etymological dictionaries,
- rhyme dictionaries,
- pronunciation dictionaries,
- dialect dictionaries,
- synonym dictionaries,
- specialized dictionaries,
- dictionaries of idiomatic expressions,

... etc., etc.

Nor are these the only conceivable options. It would even be possible to satisfy a linguist's demand for dictionaries sorted according to pronunciation to facilitate the finding of words, even if their spelling is not known.

As you can see, a dictionary data-base can offer additional advantages far beyond automated sorting and typesetting processes. Generally speaking, the main advantage of a dictionary data-base is the multiple use of stored entries in many different ways.

On the one hand, the data-base may be used with the aid of selection programs etc. to create several different printed products, e.g. different versions in different formats or for different countries, for instance a German-English dictionary for the German market giving phonetic symbols indicating English pronunciation, for the Anglo-American market with German pronunciation symbols, and for the international market with German and English phonetic symbols. If vocabulary storage is clearly structured, it is possible, without much additional effort, to derive abridgements of various sorts from a large dictionary data-base for the scholar, the student, the tourist, etc.

On the other hand, there are quite a number of other possibilities of using such a dictionary data-base, none of which are necessarily connected with the publication of books. Information may be stored in the entries that enables semantic, morphological or syntactic analyses to be performed (for example, all verbs could be given an indicator to show whether they take a dative or an accusative object). The term computer-aided analysis is now commonly used to describe such linguistic activities.

I do not have to stress that other media, such as microfilm, microfiche, screen dialog systems etc., are further possible applications for a data-base of this kind.

A computer-stored dictionary might eventually play an important part in one of the integrated information systems of the future.

Conclusion

I have tried to give a brief outline of the enormous opportunities a data-base system can offer. I am certain you will share my opinon that computer-aided lexicography of the type provided by the TEAM system represents a very useful interface between man and machine.

One of the aims of my paper was to show that new dictionary editions can be produced more rapidly and more economically in this way and that the use of computers makes checking much easier and more efficient, thus assisting in the production of more reliable dictionaries.

Perhaps my comments will make a small contribution to encourage all those involved in the making of a dictionary to avail themselves of such a system.

References

Vollnhals, O. (1982) "Technical dictionaries retrieved from a database" META, Journal des traducteurs 27, 2: 157-166

Vollnhals, O. (1982) "Fachwörterbücher aus Datenbank" Börsenblatt für den deutschen Buchhandel No. 60: 1663-1666