Gert Engel and Bodil N. Madsen

FROM DICTIONARY TO DATA-BASE

Introduction

In a world with increasing information transfer and growing language barriers the small nation is in a particularly difficult situation. It is crucial to procure tools to help overcome the language barriers; at the same time it may be almost impossible to raise the necessary financial means. The fact that the DANSK-FRANSK ORDBOG (Blinkenberg and Høybye), the only major scientifically based reference work of its kind in Denmark, with more than 200,000 entries, is out of print, therefore was bad news, both for the Blinkenberg & Høybye Foundation, now proprietor of the copy-right, and for the publishers, Erhvervsøkonomisk Forlag. A new edition is imperative, but how will it be possible to achieve?

The thought of a photomechanical reprint was soon rejected. Outstanding corrections and some 20,000 new articles would have required a smaller supplementary volume, and it would not have been possible to make amendments in the body of the present work. Nor was an entirely new composition based on manual keying with simultaneous proof-reading of existing articles and insertion of new articles considered a viable solution, because the cost of the typing alone would be in excess of two million Danish Kroner, which is more than 200,000 Dollars.

However, if the revised composition were based on optical reading of the present volumes and manual keying and subsequent insertion of the proofs, the cost could be reduced by more than 60%. Providing that grants from public sources matching those for previous editions were made available, it would be possible to cover production costs for a new edition of the DANSK-FRANSK ORDBOG, based on the work method referred to above. The production itself has been left with the Schultz Printing House at Copenhagen; it was started four months ago, and the optical character recognition is being handled by a KURZWEIL-reader situated with a subcontractor in Sweden.

The reading process turned out to be far more difficult than anticipated. The printing quality of the originals available was not the best possible, therefore the copy-proofs produced for the KURZWEIL-reader had to be enlarged by 1½. The error margin after the training period of the reader was 3% and could not be reduced to the 5 0/00 which had been stipulated in the contract; but when the first 100 pages were ready, it turned out that the majority of errors could be systematized. Typical examples were the apostrophe read as an accent, so that c'est, n'est, s'est were turned into cést, nést, sést. In such cases it was possible to reduce errors by especially designed proof-reading programmes.

This means that the reading process will take twice as long as planned. It is now envisaged that it will be completed by

the beginning of October, 1983. At the same time the 20,000 new articles will have been keyed in and corrected so that it is now possible to make insertions into the existing machine-readable dictionary. Thereafter a tape for photo-composition will be prepared, leaving only the actual printing. Thus, the spring of 1984 should see a revised French-Danish dictionary in the book-shops.

The modern solution: a data-base

But this new edition will only be new for a very brief sales period as calculated by the publisher. Up-dating must be initiated immediately, and in this connection all possibilities should be exploited which can be offered by modern technology. First and foremost, the results of this continuous process should be made accessible to the users immediately after the new information has been added.

With this in view a lexical data-base should be created, with on-line access and also with the possibility of printing out any new dictionary entries, if necessary limited to certain subject areas.

The creation of such a data-base rests on the following conditions:

(1) that we have a structured machine-readable text collection;

(2) that a system is developed or purchased to process these texts;

(3) that the necessary computer facilities are available;

(4) that manpower for the creation of the data-base is available.

It does not seem to be impossible to meet these requirements.

The machine-readable text base was established by producing a tape for photocomposition. The inconvenience that these texts were not structured originally for a data-base has to be accepted. As regards the system and the necessary computer capacity, the Modern Language Department of the Copenhagen School of Economics has agreed that the project can be based on the STATUS II-system previously purchased and further developed in connection with the DANTERM Project which is now available through the computer PRIME (550) of the Copenhagen School of Economics. Further explanation of this project is given below.

As a result of the optical character reading very comprehensive dictionary materials may be obtained in machine-readable form. But this material is not always structured with a view to being used in a data-base. Therefore, the question is whether this structure - without involving major expense - may be made explicit and whether in its existing form it will be sufficient for the desired exploitation as a lexical data-base.

Among the easiest categories of information to structure is

pronunciation, which is indicated by the sign [ ] and comes immediately after the entry headword. The same applies to subject labels, which consist in one of a number of established abbreviations, usually in italics and enclosed by ( ).

Among the slightly more complicated information categories is the headword itself. This is a semi-solid string; another possibility is to make it a prerequisite of the entry word that it must be followed by pronunciation in [ ] and/or grammatical information.

Another complicated information type is grammatical information. It seems reasonable to list all possible abbreviations, which, where necessary, are checked by the programme to be used for the insertion of labels. On the basis of our preliminary investigations we assume that the insertion of labels may be made automatically with an acceptable margin of error.

Preliminary proposal for information categories

The following data-elements are required:

headword ;
pronunciation;
grammatical information;
synonym(s);
subject label;
reference;
translation of headword;
word combinations;
translation of word combinations;
idiomatic constructions.

As the data-base will be used partly for on-line search, partly for new editions of the dictionary and more detailed linguistic investigations, a number of special considerations will have to be taken into account.

As regards searches, all information must appear in one type-face, but at the same time it should be possible to produce dictionaries with different font types. In most cases these are category-dependent so that in the printing it is possible to specify that all information under one label shall appear in a given print.

The character set normally used for data-processing, the ASCII-character set, comprises 96 characters. For a correct reproduction of the information categories required in dictionaries the following characters will have to be added:

the relevant combinations of letters and accents;
phonetic alphabet;
four different type-faces.

The Copenhagen School of Economics has acquired STATUS II, an information and documentation retrieval system with many applications. The system has been developed by AERE (Atomic Energy Research Establishment) at Harwell and is sold and serviced for

PRIME computers by BNF (British Non-Ferreous Metals Technology Centre) at Wantage. The dictionary's information structure may be adjusted to the STATUS II data-base, which is hierarchically designed with chapters, articles and paragraphs. The paragraphs of the individual article may be named and contain the information categories of the dictionary entries. In co-operation with the Computer Centre of the Copenhagen School of Economics we have developed a special Danish version of STATUS with an expanded character set of some 250 characters and an amended sorting procedure.

By using special routines, input and output to and from STATUS are controlled. For each special character they contain both a font-change-control character and an ASCII character. Software has been developed for three types of terminals:

TEKTRONIX, a graphical display unit almost without limitations in character numbers;

TANDBERG, a semi-graphical display unit with 214 different characters;

NEC, a printer terminal capable of printing 137 characters.

If terminals with normal ASCII character set are to be used, special input and output solutions will have to be prepared as regards character types which are not found on the keyboard, and a number of characters will not be able to appear on the visual display unit.

On the basis of an analysis of the implicit dictionary entry structure, a first draft working plan as a basis for programming has been developed (see Appendix). Before the data-base is released for on-line search, special search and display profiles will have to be designed.

The command language and the search procedures in STATUS II may be applied without major prerequisites. By exploiting the macro-facilities of the system, it is possible to develop special commands to facilitate user access to the data-base without requiring the user to have special knowledge of the details of the system.

Exploitation of the data-base

As soon as the planned tasks have been completed, the data-base will be accessible for users of the Copenhagen School of Economics library, in the same way that other facilities of the library are open to the general public. In October 1983 the library will open a lexicographical and terminological information centre which is intended, inter alia, as on-line access to the data-base.

For commercial exploitation the data-base will be placed with the Danish company Schultz whose SIEMENS machine will act as host for the EURONET.

It is likely that the data-base will grow relatively quickly. In the same way the chance of publishing the data-base in the

form of a book will be reduced, but it will be possible - and also commercially interesting - to separate out certain parts which cover special subject areas for special addressee groups.

We are able to ascertain that

hardware resources are available;
the computer system is available;
all software development for the improvement of the system and for communication between the central unit and the peripherals have been completed;
the optical reading of the DANSK-FRANSK ORDBOG is at the stage of completion;
photocomposition and printing of the new edition has been secured.

In addition, we envisage further structural improvements and updating of the total data-base, which may take many years to complete.

Conclusion

At the early stages of planning it became clear that the project could not be completed on a purely commercial basis. Although usually the Blinkenberg & Høybye Foundation receives royalties on the sale of their dictionaries, according to the stipulations of the Foundation this modest income is only to be used for the maintenance and improvement of these works.

A very important contribution to the project has been made by the Modern Language Department of the Copenhagen School of Economics, which has also made the necessary computer resources available and agreed that the development of a dictionary data-base may form part of the research activities of its staff. The cooperation of the Modern Language Department with the Institut National de la Langue Française at Nancy, already established, will be very important in this connection.

Appendix

Working plan for production
of F/D D/F Ordbog data-base