BJÖRN Þ. SVAVARSSON & JÖRGEN PIND

Database Systems for Lexicographic Work

Abstract

At the Institute of Lexicography of the University of Iceland work revolves around very large collections of data in computers. There are mainly two types of databases which are kept in computer storage.

The first one is a relatively simple database containing the whole vocabulary of the main collection of the Institute, along with the age, number of citation, word class, word type, and oldest citation for each word. For maintaining this database there is no need for a very complex or powerful database system, but it must be fast since it contains over 600,000 words.

The second database is much more complex. It contains the lexicographic analysis and is used to construct the dictionary itself. A system like this must be more "intelligent" and more flexible than the first one, but speed is not as important a feature.

This paper describes some of the properties of the database systems we have been working on under MS-DOS and UNIX at the Institute.

1 The Background

The Institute of Lexicography at the University of Iceland was established in 1947 with the major purpose of making a historical dictionary of the Icelandic language, covering the period 1540 to the present. Over the past four decades, major effort has been put into the excerption, building up a collection of some 2.6 million citations. The state of excerption is now such that it is 'complete' for the 16th, 17th and 18th centuries, 'fairly complete' for the 19th century, and 'incomplete but substantial' for the 20th century. The collection encompasses a vocabulary in excess of 600,000 words.

When confronted with such a massive amount of material, the question arises as to what would be the natural way of proceeding with the editing. The first question one would presumably ask is: How *has* it been done? Well, the answer to that question is pretty well known. Traditionally, after the material has been collected, or at least a substantial part of it, the editor sits down with the first few hundred slips from the first box and tries not to think too much about the

Proceedings of NODALIDA 1989, pages 326-341

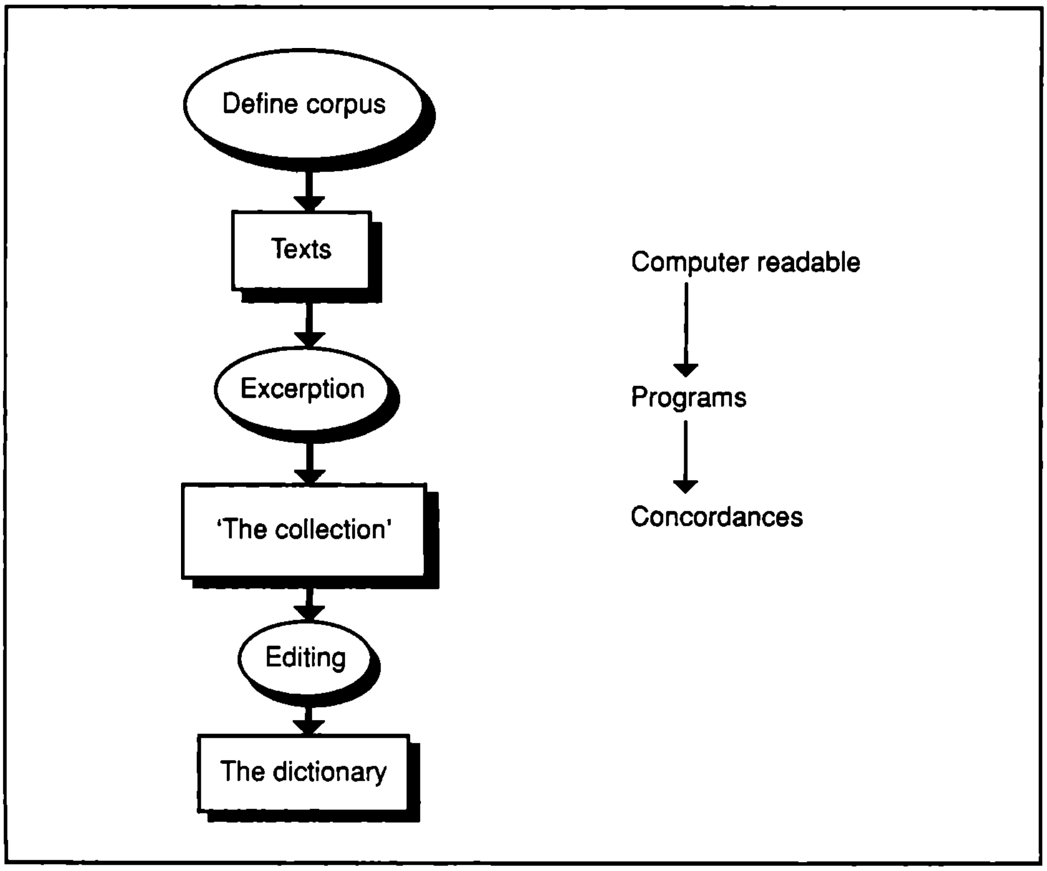

Figure 1: The procedure of making of historical dictionary

2.6 million slips awaiting! As material gathers, it is brought out in installments over a period of decades, sometimes even extending to a century or two. This procedure, almost universally adhered to in the making of historical dictionaries, is shown on figure 1.

We want to emphasize here that the historical dictionary, because of its enormous scope and sheer size, poses some problems which can conveniently be labelled 'order of magnitude effects.' They are not primarily problems of theoretical difficulty (although such problems are also abundant, as discussed in the paper by Jón Hilmar Jónsson in this volume), but simply problems which are related to the enormous bulk of these books. The major aim of this paper will be to discuss some ways in which we have attempted to deal with this problem. We would also like to emphasize that we are dealing with a project which is in some respects unique, since the whole editorial process will be computerized. This is in contrast to projects for computerizing existing dictionaries, e.g. the Oxford English Dictionary (Raymond and Tompa 1988) or the Svenska Akademiens Ordbok (Rydstedt 1988).

In 1983 a pilot editing project along traditional lines was undertaken. A reasonably full dictionary text was made up, containing words from a small subsection of the alphabet. By comparing this sample with a standard dictionary of the Icelandic language (Sigfús Blöndal 1920–1924), it emerged that it would take close to two centuries to finish the editing using traditional methods and present levels of staffing. At the time this seemed to us an unacceptable way of proceeding. There were two reasons for this. On the one hand we felt the time scale to be absolutely unacceptable, on the other there are well-known problems with the traditional approach which we felt we could do something to solve by following another approach to the editing task.

2 A Different Strategy

Based on the experience gained in 1983, a new strategy for the editing of the historical dictionary has gradually been taking shape in the work carried out at the Institute. This strategy stands in rather sharp contrast to earlier methods and approaches. It is characterized by heavy reliance on computational tools. Indeed, we think it would be fair to say that without the existence of such tools this strategy would not be feasible.

The contrast between our new methodology and the traditional one can be captured with two terms borrowed from computer science. The traditional methodology can be characterized as **depth-first**, whereas ours comes much closer to being a **breadth-first** strategy.

A depth-first strategy means that every detail in the editorial process must be resolved to completion as the need arises. There is no chance of delayed commitment, no matter how much the lexicographer so desires! Since the editorial process turns out pages evenly and constantly, any revisions, which may be called for at a later date, may lead to the necessity of redoing the analysis, taking the manuscript apart, as it were, and putting it back together again.

How much more efficient it would be if it were possible to carry out the analysis independently of making up the manuscript. This is a goal not readily achieved, however, since the lexicographic analysis needs to be linked to the sources, i.e. the citations, in some manner. Traditionally, that link has been achieved by interspersing the analysis with quotations from the collection of citations.

It is precisely this reliance on alphabetical ordering, coupled with the enormous wealth of material which has to be accounted for, which gives rise to the aforementioned 'order of magnitude' effects. The following points are worth mentioning:

- There is a time lag of decades between the editing of different parts of the dictionary with many *generations* of lexicographers involved.

- The editor has a very *limited view* of the task at hand since he is labouring in a small corner of the dictionary.

- The alphabetical ordering of the text forces the editor to deal with material which is in no way related.

These problems are acute because of the size of the project. With a single volume dictionary, which takes perhaps 10 years to complete, it is relatively easy to bypass these problems. Not so with the historical dictionary. The research carried out at the University of Waterloo on the computerized Oxford English Dictionary has brought this out (Raymond and Tompa 1988). It turns out, for instance, if the cross-references of the OED are examined, that references to previous letters of the alphabet are much more frequent than those to subsequent letters. Most cross-references are, however, to words starting with the same letter. While the latter is to be expected (many cross-references are probably to closely related words), the former result can only be explained by the individual editor's limited view of the whole project.

A breadth-first strategy, of the sort now being developed at the Institute of Lexicography, proceeds with the editorial work from the top down, by making a number of passes through the citation collection, deepening the analysis at each stage.

There are numerous advantages to such an approach:

- Since the editing is computer-based it can be made available on the computer at an early stage.
- It is possible to deal with coherent parts of the vocabulary at any one time.
- It opens the way for defining significant phases in the work which can be finished in the relatively short time span of 5–10 years.

The first stage in the editorial process, which began in late 1983, involved a complete pass through the main collection, making up a database of the total vocabulary.

This database is of twofold use. In the first place it is of tremendous value for work in linguistics, especially those areas dealing with word-formation and morphology. The database has already been used to investigate the nature of compounding in Icelandic (Kristín Bjarnadóttir 1990). Such research is not only of theoretical importance, but is also relevant in the making of practical language tools on the computer, such as the ubiquitous spelling-checker. Spelling-checkers are usually dictionary-based. Dictionary-based checkers, however, have some limitations in most Germanic languages where the process of compounding is quite active. Some dictionary-based checkers now offer compound word parsing (and one Icelandic checker is completely based on the idea of word parsing). Due to the limited knowledge about compounding, however, it has generally not been possible to state the constraints which compounding obeys, and thus the mechanisms tend to overgeneralize wildly.

Secondly, the database is of great value for further lexicographic work as it, at once, opens up multiple search paths into the collection and frees the editor from the 'tyranny' of the alphabet (since a collection of written dictionary slips permits only one search key!). Now, however, the editor can access the collection on the basis of grammatical category, age of citations, oldest source, etc.

3 The Vocabulary Database

3.1 The Nature of the Database

As already mentioned, the first computational project involved a database covering the whole vocabulary of the main collection of the Institute. This database has 8 fields which can be briefly described as follows:

The word itself. The choice of words was primarily based on the orthographical form.

Word class. This, of course, has traditionally been indicated on the dictionary slips.

Age of oldest citation. The age is established to the nearest third of a century or to a greater time period if it is not possible to uniquely position the citation in this time frame (of 33 years). See also note below on the **source**.

Age of most recent citation. This is coded in the same manner as the **age of oldest citation**.

Number of citations. This is noted exactly for 1 to 5 citations. All words having more than five citations are marked as such with no finer distinctions being made.

Word type. This is the only type of information which cannot be read directly from the citation slips. We felt, however, that it would be advantageous to attempt a rough classification of the vocabulary according to word-type so words are marked as being compounded, affixed, or noncompounded.

Source. Finally, the source for the oldest citation is noted. This is often followed by a note detailing the exact age of the citation, e.g. a specific year. In fact, this field is built up of two subfields: abbreviation for the source and reference of page, exact age, etc.

Word in reverse. Part of the word is kept reversed in this field. This is done for the indexing of word endings.

Actually, while the Vocabulary Database makes up a relatively simple dictionary, it took quite a while to prepare. This was mainly due to the enormous vocabulary in the main collection. While the total number of citations is approximately 2.6 million, the total number of different words contained in the vocabulary database now stands at 608,205. This shows that the number of 'singletons' (words attested by only one citation) is relatively high, as noted by Jón Hilmar Jónsson (this volume).

Work on the database started in October 1983, we reached the 100,000 mark in May the following year, and keyboarding finished on the 7th of March 1986! The keyboarding was done on a Victor 9000 microcomputer using a specially made BASIC program.

After the keyboarding was completed, proofs were read, mostly in 1987. Due to the simple nature of the database, it was possible to perform numerous integrity and consistency checks with specially written programs. This considerably eased the onerous task of proofreading. The database reached its present form at the end of 1988, or five years after work on it was originally started.

Though we did not keep an exact account of the work involved, we would guess that it probably amounted to about 10–12 man-years. This gives some indication of the magnitude of the task of compiling a true dictionary of the material contained in the collections of the Institute.

3.2 The Computer Database

As already mentioned, the Vocabulary Database contains over 600,000 records at the present time. Putting the database online under the MS-DOS operating system (which has been our platform for most of this period) was never considered a viable option, since we felt that this operating system would have considerable difficulty in coping with a database of this size. This was one of the reason for a decision made late in 1987 to change to the Unix operating system.[1]

Our main computers are two IBM 6150 machines (perhaps better known as IBM RT/PC), running the AIX operating system. At present their total disk capacity is 840 Mb. We also have two IBM PS/2 machines running AIX PS/2 with 230 Mb of disk space. The RTs are connected by *Ethernet* and the PS/2s will also shortly be linked to the network. The *Network File System* (NFS) runs on top of the *Ethernet* connecting the machines.

The Vocabulary Database can conveniently be described by the relational database model. It consists of one main table, containing the records for all the words, along with a secondary table which contains information about the sources used to gather the citations for the main collection. These two tables are linked on the source fields as shown in figure 2.

There are a number of relational database managers available for AIX, among the better known are *Oracle* and *Informix*. Due mainly to considerations of cost, we decided to use the *Informix* database system. This is a fully relational system with the SQL query language. Unfortunately, it turned out that *Informix* has a nasty peculiarity[2] in that it does not allow fields with a variable number of characters (VARCHAR). Now this, obviously, makes life pretty difficult for the lexicographer!

For this reason the database grows to approximately 150 Mb when it has been indexed under *Informix*[3]. While the database is quite voluminous, it is also quite fast. It will instantly find any word and search on indexed keys is also fast.

[1] There were, of course, other reasons as well. Unix is well-known for its outstanding collection of tools, its preeminence in dealing with text files, the easy access to e-mail, etc.

[2] When compared with the description of relational database systems given, for example, by Date (1986).

[3] In textform it is about 30 Mb.

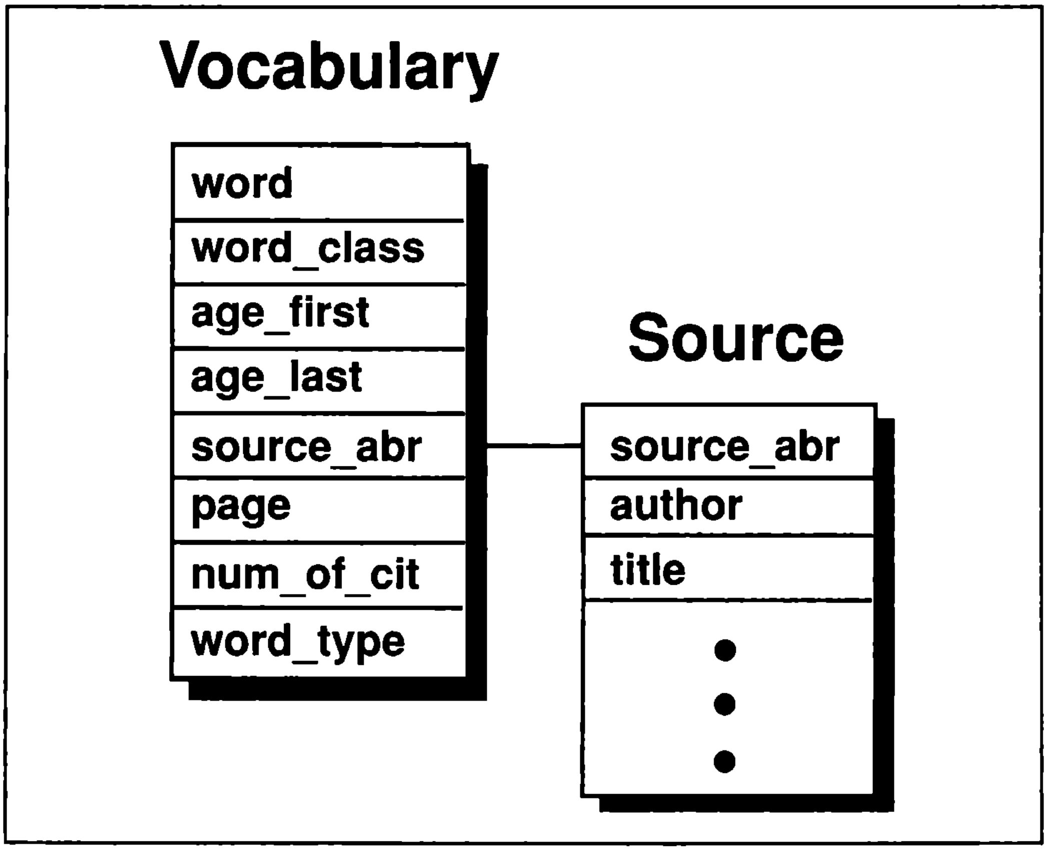

Figure 2: The tables of the Vocabulary Database as they are configured at present. The link on the source fields is shown.

3.3 Searching the Database

We will not particularly elaborate on the existing possibilities for searching the database. It is evident, from the description given so far, that we now have the possibility of searching the collection in ways not possible earlier. The following are examples of queries which have been put to the database by members of staff at the Institute, as well as by other researchers.

1. Find all words which occur for the first time in the works of the 19th century poet Jónas Hallgrímsson.
2. Find all adverbs ending in 'is'.
3. What is the proportion of verbs in the total vocabulary?
4. List all nouns for which there are more than five citations in the collection, with examples attested both in the 16th or 17th centuries and in the 20th century.
5. List all noncompound nouns for which first examples are attested in the first third of the 19th century.

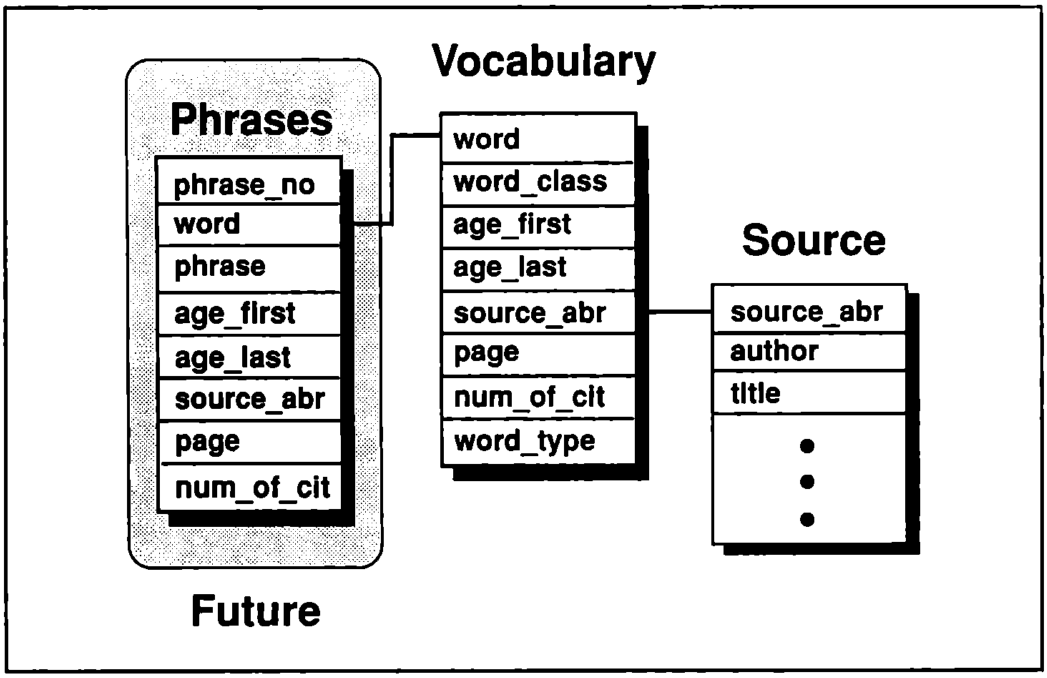

Figure 3: Relationship of tables in the Vocabulary Database if a table of phrases is attached to it.

3.4 Enhancing the Vocabulary Database

We have given some thought to the possibility of enhancing the Vocabulary Database with other kinds of information. As described by Jón Hilmar Jónsson in his paper (this volume), the main thrust of the editorial process concerns the description of verbs. Evidently, quite a lot of material which is of relevance to the verbs is filed under other word classes in the collection. This holds, for example, for phrases and idioms which are often filed under the noun rather than the verb. Some attempts were made by those collecting the citations to file them under all the relevant categories, but it goes without saying that in a task of this magnitude, stretching over decades, it is inevitable that various inconsistencies of practice will arise. While this extension has not been implemented, figure 3 gives an example of how it could be carried out.

4 Editing, Excerption: Computational Approaches

While material was being gathered for the Vocabulary Database, work was also being carried on in various other areas relating to the dictionary project as a whole. Four things in particular stand out:

1. The editorial process was begun in 1983 and gathered momentum during 1984 and 1985.

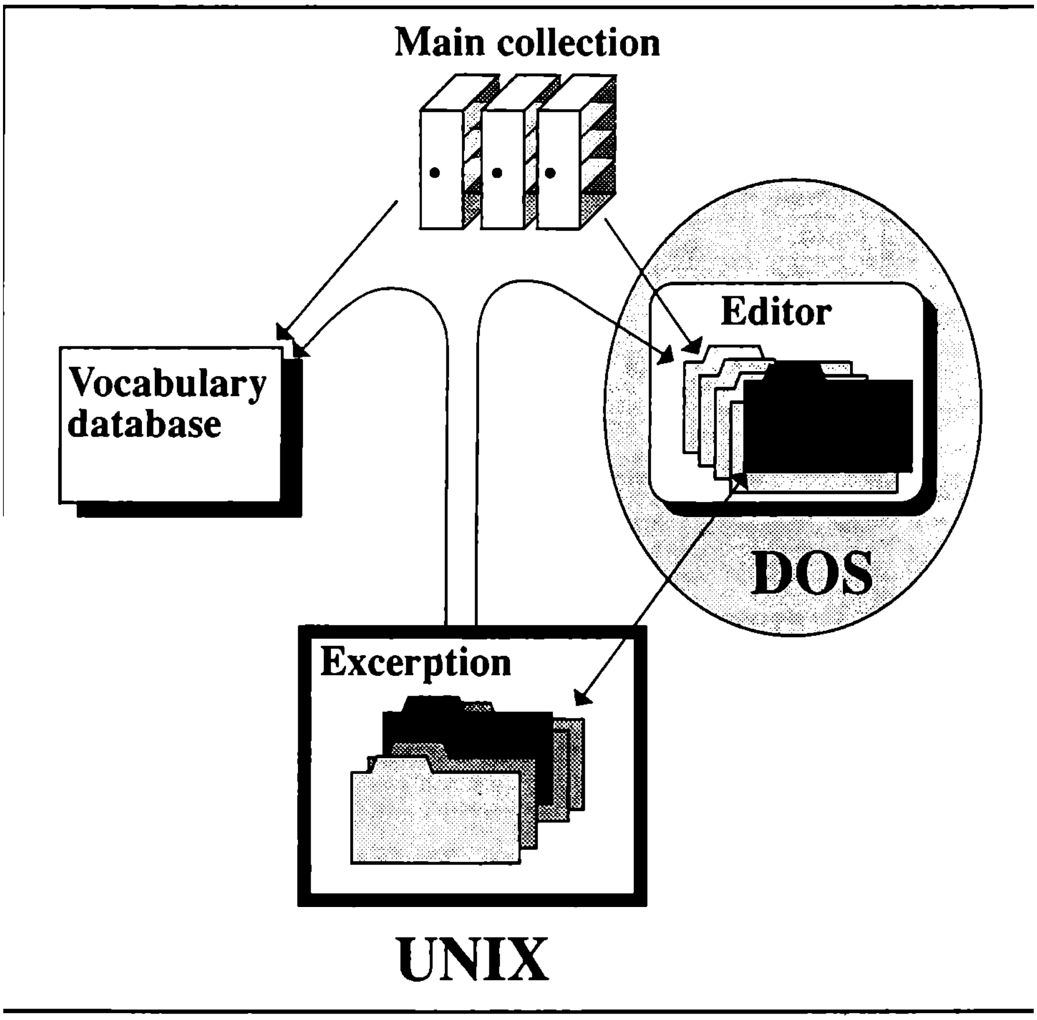

Figure 4: Computerized tasks at the Institute, as divided between MS-DOS and Unix at present.

2. The Institute started a collection af machine-readable texts. These texts can readily be used to augment the collections. It should also be noted that although most of the texts are modern texts from publishers and printers we have also endeavoured to acquire older texts. Some of these have been specially keyboarded at the Institute.

3. Excerpting for the main collection continues. All new citations are now entered directly to the computer.

4. We have adopted the TeX typesetting system for typesetting dictionaries. This is further discussed in the paper by Jörgen Pind (this volume).

Figure 4 shows in a schematic way all the major activities where computers have been brought to bear on the lexicographic work. This figure shows this activity as it is carried out at present under MS-DOS and Unix. We are presently in the process of moving all these tasks to Unix. A number of points are worth discussing in some detail.

The editorial process was carried out using the *Revelation* database system for MS-DOS—which later became *Advanced Revelation* (Cosmos 1987). Revelation is a database system, based on the Pick Operating system which has enjoyed some success in the commercial environment (Rochkind 1985). The main reason for originally choosing Revelation was the absence of any (significant) constraints on the length of individual fields; using, as it does, completely variable length fields. The main characteristics of Revelation are summarized in the following:

The benefits of Advanced Revelation are:

- It is very flexible. It is easy to reorganize datafiles and reconstruct applications.
- It has a user-friendly interface.
- It has variable field lengths. Each field can range from 0 to 65 kilobytes, and predefinition of length is unnecessary.
- It is possible to define as many as 65k fields in each record, and the number of records in one file is only limited by disk space.
- It is possible to have many files open concurrently, and these can be related.
- It has its own procedural programming language (R/BASIC), similar to BASIC, but more structured.
- It has a powerful query language, similar to SQL.

The major disadvantages of Advanced Revelation are:

- It is too slow.
- It is only available on computers running MS-DOS which is a primitive operating system.[4]
- Since the system has only been available on computers running MS-DOS, disk storage is limited.

In spite of these disadvantages we have found Revelation to be a singularly useful product for lexicographic work, and it is with some sadness that we take leave of it now that we have moved over to Unix! The combination of variable length record fields with a very powerful programming language has turned out to be ideal.

[4] As of 1990 it is also available under OS/2.

4.1 The Nitty Gritty of the Editing System

We will now describe the editing system as it was implemented under *Advanced Revelation.*

As J6n Hilmar J6nsson has already described in his paper, the editing of verbs proceeds mainly with reference to the formal, syntactic and morphological behaviour. The editing is based on the citations and proceeds in a number of steps. In effect, it is possible to view the editing as being, to a large extent, a continuation of the excerption in that the citations are provided with grammatical markers. This is in complete contrast with the methodology traditionally employed (Kuhn 1982).

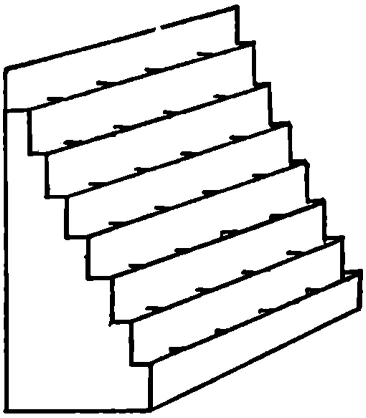

Figure 5: Kuhn's sorting board

Traditionally, the editing proceeds with the editor spreading all the slips attesting the occurences of a particular word out on a table, attempting to sort them manually into semantic categories. Kuhn has a nice illustration of an 'exceedingly useful' sorting board (shown in figure 5) which has been used to assist in the preliminary sorting done for the *Middle English Dictionary.*

It goes without saying that the strategy employed at our Institute is diametrically opposed to Kuhn's approach since the editing process starts out by augmenting the citations with various entries detailing such factors as conjugation, subject, object, meaning, etc.

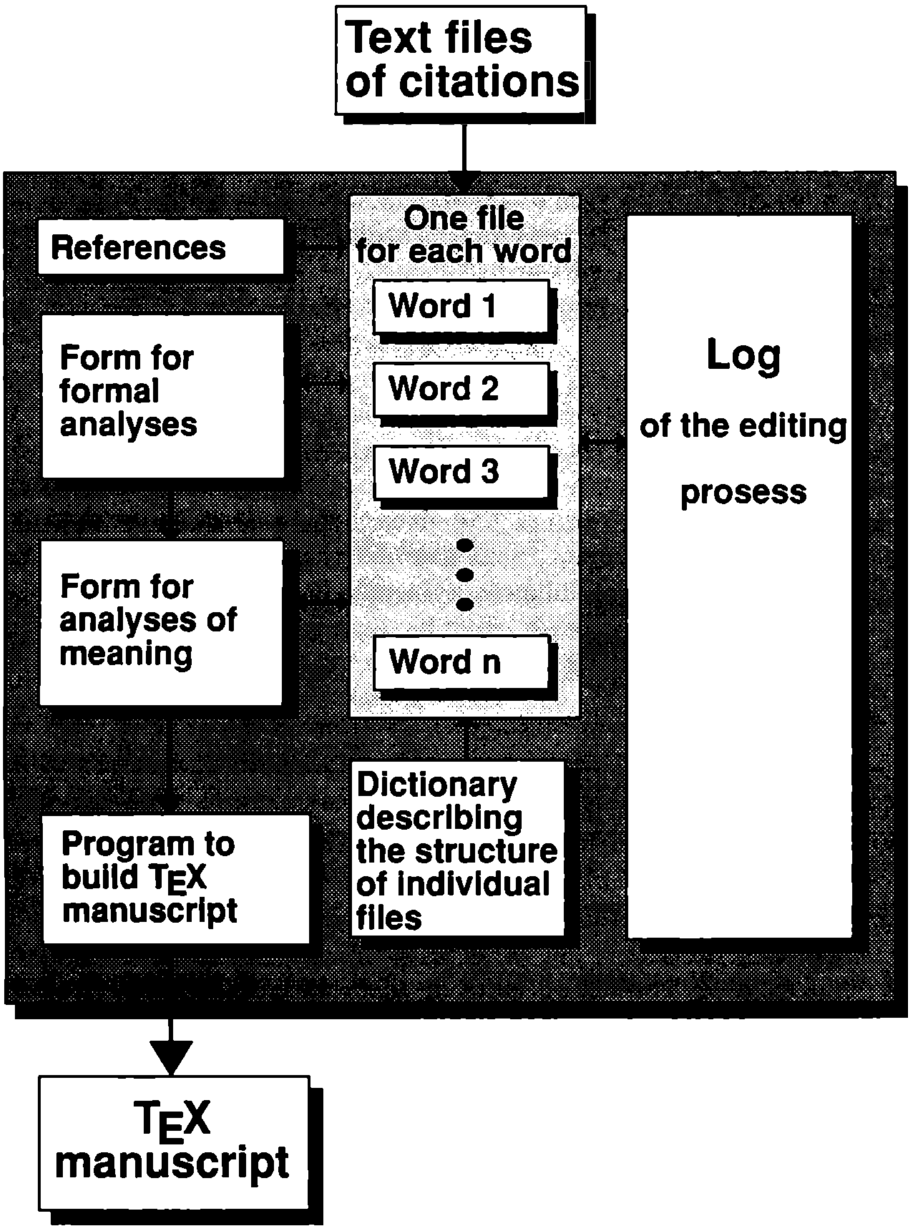
Text files of citations
One file for each word
References
Word 1
Word 2
Word 3
Word n
Form for formal analyses
Form for analyses of meaning
Log of the editing prosess
Dictionary describing the structure of individual files
Program to build TEX manuscript
TEX manuscript

Figure 6: The structure of the Lexicographic Database as implemented under Advanced Revelation.

The method used in the editing is also reflected in the structure of the database program which is used for the editing. Figure 6 shows in a schematic way how the editing takes place. All citations are keyboarded by a secretary into files which are then imported into the database system. One file keeps a log of the editing process, noting which verbs have been analysed, the state of analysis, and other such information of a general nature. Furthermore, it contains fields with detailed information associated with most of the citations. This file functions as a log of the editing process.

The citations for each word are kept in separate files which also contain the analysis. Jón Hilmar Jónsson, in his paper, describes the fields which are entered into these files and the editorial process as such. A special 'dictionary' contains information about the structure of the files. Each time a word is imported into the database the dictionary is consulted regarding the structure of the file which is to be made for the imported word.

Two screen input forms are used for the editorial analysis, one for the formal analysis, the other for the semantic analysis. When the editor starts on a particular phase of the analysis, he or she can use the query mechanism to arrange the citations in any desired order and browse through the collection while entering the analysis, thus dealing with similar citations at each point. After each pass through the citations, it is possible to rearrange the citations using the fields which have been entered. The command given is the `select` command. The following command is, for instance, used to order the verb 'koma' (come) on the fields *form*, followed by *voice*, followed by *conjugation*:

```
select koma by form by mynd by beyging
```

This possibility of ordering the citations prior to the analysis and at each subsequent step, is one of the main attractions of a database system such as *Advanced Revelation* as it ensures consistency of treatment and also speeds up the analysis considerably. Another advantage of Revelation is the fact that the entry forms have been defined in such a way that they carry over default values from the preceding citation. When the citations are sorted such carry-over defaults also serve to ease the data entry task and ensure consistency.

For the person responsible for the maintainance of the system it is very important that changes to the system, especially the entry forms, can be easily achieved. A special forms editor (which, incidentally, goes by the name 'painter') makes it very easy to change the layout of forms. It is also easy to change the structure of the files.

When the analysis of each word is finished the system will output a TEX-coded manuscript. This is done by a special program written in R/BASIC, Revelation's procedural language. This program uses a query command to select the citations which are marked as suitable for the printed dictionary.

To summarize briefly at this point:

- The editorial process augments the citations by attaching entries to the citations.
- These entries make up the sort key which is used to deliver automatically the major structural lines of each article.
- The database's "report writer" has been changed so that instead of producing the normal columnar reports it turns out manuscripts for TEX which can then be directly typeset.

The discerning reader is now probably wondering how we guarantee the integrity of our database, when on the one hand we have a collection of citations, and on the other hand citations contained within the editorial database. If errors are found in the editorial database, where will they be corrected, in the citation database, in the editing database, or in both places?

Obviously, this is a weakness of our approach as implemented in the Advanced Revelation database system of which we are fully aware. This, in fact, brings us right up to the present date as regards the development of our lexicographers "workbench". It is obvious that the computerized citation collection needs to be integrated with the editorial database.

5 Computational Lexicography under Unix

How should this database be implemented under Unix? It is quite clear that a database system like *Informix* is not at all up to this task since it cannot deal with fields having variable length records. Presumably the *Oracle* DBMS would be able to cope and this is already used for at least one large lexicographic database, CELEX in the Netherlands. However, Oracle is an expensive system, and, furthermore, other considerations have lead us to consider a different approach.

Our experience with TEX, which is a completely open system with all source code freely available, has brought home to us the importance of having programs which are available in source code form. If the program does not behave quite as desired one can always change the program code. This lesson has been reinforced with our experience of Unix, where an abundance of excellent software in source code form is also available. For this reason, we will be making a serious attempt to create a system where we can use available source code which is (preferably) available in the public domain (e.g. TEX and GNU) or commercially. The details of the actual implementation are currently being worked out, so here we will limit ourselves to an overall view of the toolset, as we have currently come to view it.

Perhaps the major concern for us is the following: Since the computerized citation collection is still being added to, we need some way of keeping access to the collection open, while at the same time using the citations, or at least a part of them, for the editorial work. Now, by referring to the description given earlier of the editing process as it was implemented in the Revelation database, it can be seen that quite a lot of the information which gets added to the citations in the editorial process is only dependent on the citation itself. As a matter of fact, this holds for 25 out of the 29 features analysed. It would, therefore, seem natural to detach these features from the editorial database and store them along with the citations. This would leave the citation collection intact and still allow us to progress with the editorial process. In this way, the formal analysis would be taken care of.

In this manner the citation collection, thus augmented, could be input directly to a sort routine similar to that illustrated earlier for Advanced Revelation.

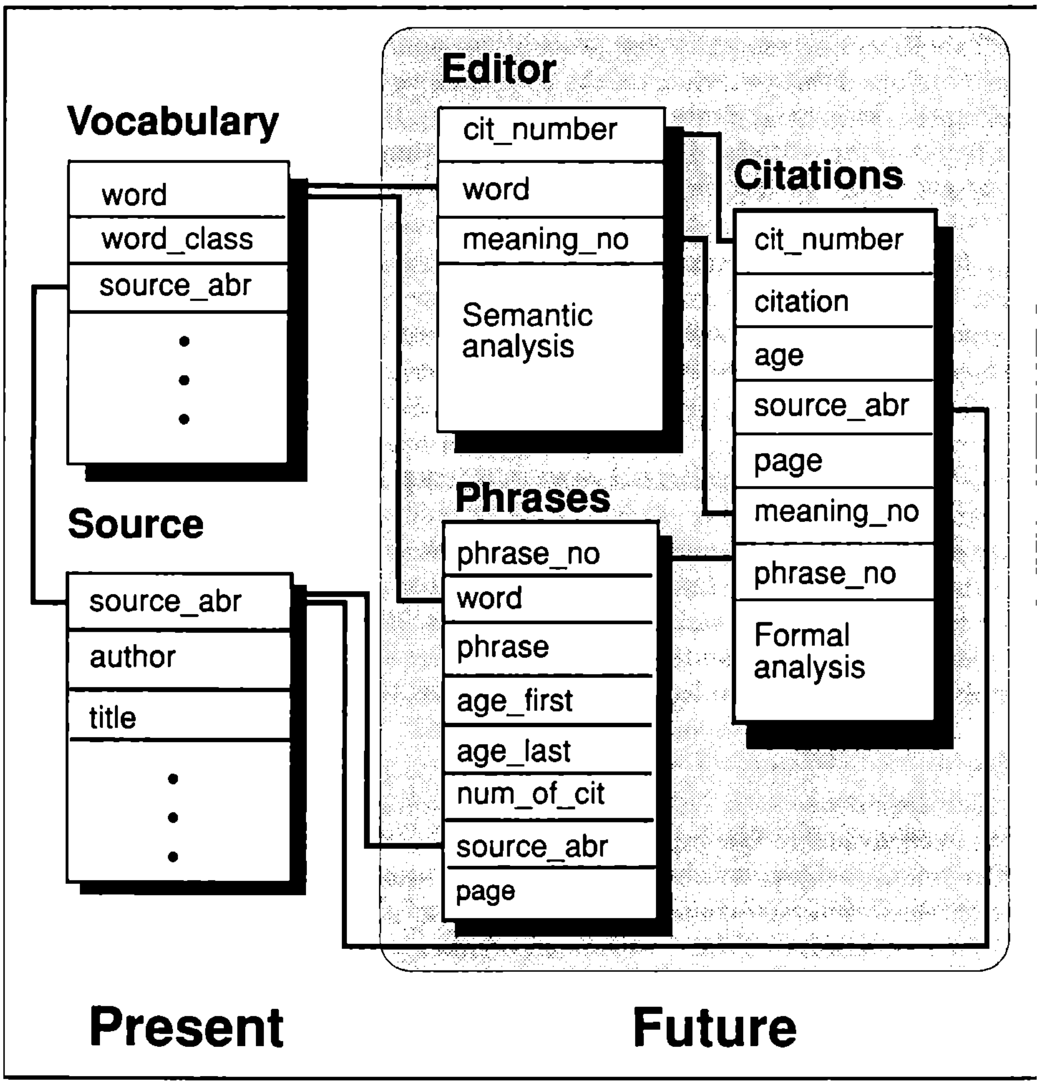

Figure 7: Our present ideas of how the lexicographic database can be implemented under Unix in the future

From this point on, we need to rely on a specifically made editorial program or a collection of programs for further analysis. Such a collection of program needs to be able to handle the following:

- It must keep a journal of the editorial process.
- As the citation collection grows, there must be some mechanism for notifying the editor that further citations have been added to the collection. These need to be incorporated into the editorial database.
- The generation of TEX manuscripts should be automated.

Work on the implemention of such a procedure has just been started under

Unix so it is not possible for us to elaborate in detail as to how the programs will be implemented, but figure 7 shows our present ideas.

References

Kristín Bjarnadóttir. 1990. *Um stofnhlutagreiningu samsettra orða.* BA thesis in preparation.

Sigfús Blöndal. 1920–1924. *Íslensk-dönsk orðabók.* Reykjavík.

Cosmos. 1987. *Advanced Revelation—Documentation.* 4 volumes. Cosmos, Seattle, Washington.

Date, C. J. 1986. *An Introduction to Database Systems.* Addison-Wesley, Reading, Mass.

Jón Hilmar Jónsson. 1989. A Standardized Dictionary of Icelandic Verbs. (This volume).

Jörgen Pind. 1989. Computers, Typesetting, and Lexicography. (This volume).

Kuhn, Sherman. 1982. On the Making of the Middle English Dictionary. *Dictionaries: Journal of the Dictionary Society of North America* 4:14–41.

Raymond, Darrel R., and Frank Wm. Tompa. 1986. Hypertext and the Oxford English Dictionary. *Communications of the ACM* 31:871–879.

Rochkind, Marc. J. 1986. Pick, Coherent and Theos. *Byte* 10(11):231–239.

Rydstedt, Rudolf. 1988. Creating a Lexical Database from a Dictionary. Martin Gellerstam [ed.]. *Studies in Computer-Aided Lexicology.* Data Linguistica, 18. Almqvist & Wiksell, Stockholm.

Institute of Lexicography
University of Iceland
Reykjavík 101
Iceland
bjorn@lexis.hi.is
jorgen@lexis.hi.is