Traditional Dictionaries and Data Base of Dictionaries

Ülle Viks

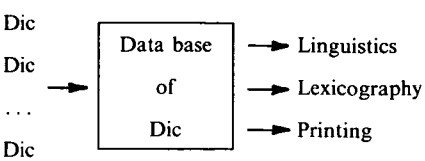
1. My report deals with the automation of lexicographical work in a situation where there are no large corpuses of text nor special computer-oriented dictionaries, but, at the same time, there is abundant lexical and grammatical information available in a large number of traditional dictionaries (TDs), such as normative, monolingual, dialect, bilingual dictionaries etc. These have been compiled manually by high-skilled lexicographers.

The information of TDs must be made accessible to the computer. One possible way to do this is to compile a data base of dictionaries (DBD) incorporating all the information in traditional dictionaries.

The question is what the DBD should be like and how it is to be done. In the Institute of Language and Literature of the Estonian Academy of Sciences work on computer dictionaries began some 10 years ago but up to now we have been working on single dictionaries. Now the preparation of an integrated DBD is under way.

This is meant to ambrace all the major dictionaries of the Estonian language published up to now and in the future, including bilingual dictionaries, some dictionaries of cognitive languages and also some encyclopedias. At present we are working on more than 10 dictionaries. The normative dictionary, the morphological dictionaries, the concise dialect dictionary, the dictionary of synonyms and the SCHOOL-CHILDREN'S ENCYCLOPEDIA are fully available in computer-readable form, while the material of the monolingual dictionary, the Russian-Estonian Dictionary, the English-Estonian Dictionary, the Dictionary of the Votic language and some other dictionaries is now being prepared for the computer. New material is being added constantly.

2. Our aim is a universal DBD that is not oriented towards strictly concrete tasks. At present we can say that the main outputs of the DBD are concerned with three domains of application: 1) linguistics, 2) lexicography and 3) printing industry. The general structure of the system is represented by the following scheme.



The texts of input dictionaries are fed into the computer where they are transformed into component parts of an integral DBD.

From the point of view of linguistics, the DBD serves as an information retrieval system where researchers can obtain the required material, or it serves as a base for linguistic data processing.

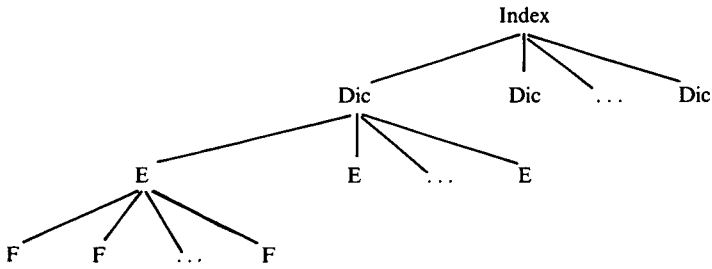
The lexicographic system is meant for compiling new dictionaries from existing ones. The new dictionaries are also linked to the DBD. The printing system is meant for printing dictionaries by means of photocomposition. The greatest advantage is gained here in publishing new computer-compiled dictionaries. In this case the text of a dictionary is set automatically, without human involvement.

The input format of the dictionaries has not been unified. This depends on the concrete purpose of the input. The text of dictionaries fed into the computer for printing is provided with special commands for the photocomposition system. The dictionaries fed into the computer for processing may have special labels marking the arrangement of the components of word-entries, or they can be tabulated. Some dictionaries have no additional symbols in their text.

3. Thus, the DB of TDs is characterized by a considerable heterogeneity of the input material and a wide diversity of application. This determines the basic requirements to the DBD: First, it must be an open system, because new dictionaries are added constantly, and new domains of application may appear: Second, the inner structure of DBD must be independent of the input as well as of the output structures: Third, the DBD must be provided with the possibility to apply new software, because we cannot predict all the tasks to be solved.

At present we are not agreed on the structure of the DBD. Should it be a physically unitary superdictionary or should it consist of many independent parts linked by a complicated cross-reference system?

Here I shall present a possible solution to the problem. To meet the above-mentioned requirements and render the system as flexible as possible we may proceed from the module and hierarchy principles that have been effective in information processing. In our case this means compiling a DBD from many independent but interconnected and hierarchically organized modules.



The middle level in the hierarchy is occupied by individual dictionaries, the number of which increases with the expansion of the DBD. All D-modules are linked by a collective index including all entry words fixed as such in at least one dictionary. Every entry word in this index has references to all the dictionaries where the corresponding word occurs.

The formation of modules on the lower levels is based on the general principles of D structuring. Every dictionary subdivides into word-entries consisting of certain components, e.g. head word, grammatical characteristics, explanations, translation equivalents, stylistic and pragmatic references, geographical distribution, etymological data, etc.

Our aim is to compile a DBD having a logical structure corresponding to the lexicographical reality. I mean the following correspondences:

a) one D-module — a single TD,

b) the logical entry of DB — the entry of D,

c) the field of the DB entry — the fragment of the D entry, representing certain lexical information (phonological, grammatical, semantic, pragmatic, etymological, etc.).

Different types of TDs (e.g. monolingual and frequency dictionaries) may require DBs of a different structure.

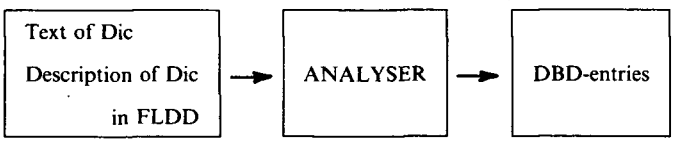
Even if the dictionaries are physically separated in the DBD, they are integrated into one system by the cumulative index and cross-references.

Such a module structure renders the DBD flexible and easy to handle. For example, we can use two-step processing. In order to solve a concrete task it is not obligatory to go through all the DBD. As the first step it is possible to form the most suitable small DB and including only necessary information: we can select the required fields of the entries to compile a working file. This saves on memory capacity and increases operating speed.

Thanks to the module structure of DBD inter-D information processing is no more complex than intra-D processing, while a working file may be based on several D-modules.

And finally, the DBD can be developed step by step for adding or altering single modules, and this will not interfere with the operation of the entire system.

4. Now a few words about the process of turning the ordinary dictionary text into the part of the DBD.



Entry to the DBD is gained through the analyser where the initial text of a dictionary is analysed and segmented according to the fields of DB entries.

The analyser uses the formal language of dictionary description (FLDD). It represents the data description language ELMAMETA (developed at Tallinn Technical University) adapted to the lexicographic material. The FLDD should be able to describe the structure of any dictionary irrespective of the concrete language or the type of the dictionary. The same FLDD should be capable of analyzing the TDs (with the aim of incorporating the following TD into the DBD) as well as synthesizing new dictionaries in the lexicographic component of the system.

The inputs for the analyser are: (a) the text of the input TD, (b) the description of the input TD in terms of FLDD.

This description consists of three components:

a) a description of the lexical means of the FLDD, keyed to the concrete TD, defining the classes of characters, lexemes, key-words and separators;

b) syntactic rules that describe the structure of the TD entries;

c) semantic rules that define the relations between the TD entries and DBD entries.

In the syntactic rules it is important to define the diagnostic features in order to recognize the strucutral elements or fragments of TD entries in the initial text. As diagnostic features marks of punctuation, commands for photoposition system, special distinctive marks in text, abbreviations, some keywords, etc. can be used. With some dictionaries these diagnostic features are sufficient and unambiguous, with others they are not.

5. The automatic analysis of TDs does not always give good results since the TDs, compiled by humans for humans, are, as a rule, formalized unsatisfactorily for the computer. There are three ways of improving the quality of the results:

1) correcting the description of the input TD (in a man-machine dialog),

2) pre-editing of the input text (by adding the separators, marking the components of the D-entries),

3) post-editing of the output entries.