Reports of Enlarged Session of the
Seminar of I.Vekua Institute
of Applied Mathematics
Vol. 19, N1, 2004

THE PROBLEMS OF GEORGIAN LANGUAGE THESAURUS CONSTRUCTION

Veliashvili N., Jibuti M.

I.*Vekua Institute of Applied Mathematics,*
Tbilisi State University

Received in 22.09.04

It is impossible to overestimate the importance of Thesaurus in the language investigation process, and, generally, in the nation's culture. For European Languages appropriate Thesauruses construction have been begun in the 10 – 11 century.
To mirror a language evolution, generally, a Thesaurus construction embraces several phases:

1. The Fixation of lexical fund of nation's cultural heritage – manuscripts.
2. The Fixation of lexical fund of modern literature.
3. Finally, the Fixation of lexical fund of newspapers and journals.

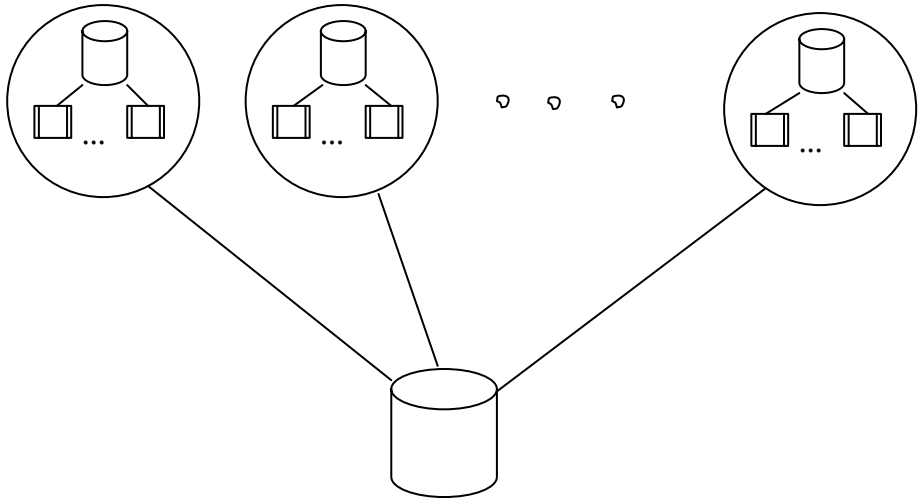
Unfortunately, today Thesaurus of Georgian Language does not exist. However, there was the attempt to construct it. The famous Georgian scientist, academician George Cereteli was at the head of this process. He and his team began with fixation of ancient lexical fund from ancient manuscripts. But after his decease this process was interrupted. What they had time to do is being kept as catalogue of cards in the Georgian Institute of Oriental Studies. This catalogue contains 6 000 000 cards (six million), i.e. six million wordforms, with the indication of source and fragment of text – context – which contains concerned wordform. Also, here we want to note that currently in the catalogue department there is being made a bashful attempt to construct electronic version of this catalogue. But this work is being carried out without any understanding of complexity of this task, understanding of logicality of phases of this process, system view of problem and so on. They have not even IT specialist. So, there is only wish to construct Thesaurus.

Meantime, for Thesaurus construction above topics as well as bringing in this process modern IT technology [1] are absolutely unavoidable. Below we try to show briefly main phases of Thesaurus construction from the IT point of view.

For effective accumulation of language lexical fund it is necessary:

1. To create special programming system using modern IT technologies:
   - Local networks;
   - Database Management Systems;
   - Client/Server methodology;
   - Text recognition and etc.
2. To organize several centers in the different institutes, where lexical fund will be accumulated. Let us name such center Warehousing Center. It is obvious that different Warehousing Centers accumulate different lexical fund from different sources:

manuscripts, books, newspapers and so on. Each Warehousing Center is a Local network with DataBase server and several workstations [2]. On the other hand Warehousing Centers will be combined into one global network for information exchange: Each Warehousing Center will deploy his portion of lexical fund to other Warehousing Centers [3]. So, all Warehousing Centers will have the same lexical fund. This is important for Thesaurus construction – for checking procedures for example.
Global network needs own server to perform management functions:



Thus, in such environment we have two groups of tasks:

1. Tasks for Warehousing Center, i.e. in the framework of a local network. These tasks must be the same for all Warehousing Centers.
2. Tasks for global network for information exchange among Warehousing Centers.

1. Warehousing Center Tasks

Generally, Warehousing Center tasks are dividing into two groups:

1. Lexical Fund accumulation tasks.
2. Lexical Fund using tasks. This group contains such features as lexical fund analysis, searching, statistics and etc. Shortly, we can name this group linguist workplace.

Of course, topics of this paper are accumulation tasks, because we are considering Thesaurus construction.

The Basis of Warehousing Center is DataBase Server. Therefore the first-rate task is:
T0. Designing the architecture of lexical fund database usually, two types of sources are picked out for Thesaurus construction:

1. Manuscripts.
2. Printed, published texts.

In our case, we have two additional sources:

3. Hand-written cards from Cereteli catalogue.
4. Cards DOS-files in the TEXT format.

For entering Hand-written cards and texts into computer we can use only keyboard. Only for some printed texts we can use scanner – generally, it depends on the printed text quality, print's typeface, skittle's size and etc.

So, we picked out four possibilities for information entering:

1. Cards from keyboard. (T2)
2. Cards from DOS-files.
3. Texts from keyboard.
4. Texts from scanner.

Card entering from keyboard is core of accumulation software.

Besides, Accumulation software must contain special program or procedure:

T1. Wordform load into database. This procedure along with wordform storing must perform appropriate checking functions.

This special program must be kernel of accumulation software together with database architecture [4].

As for texts (from keyboard and scanner), it is necessary:

T3. Program, which cuts a text into pieces in the card-format, i.e. wordforms with corresponding contexts. In other words, this program must generate cards from text.

T4. Develop FORMAT of Cards package.

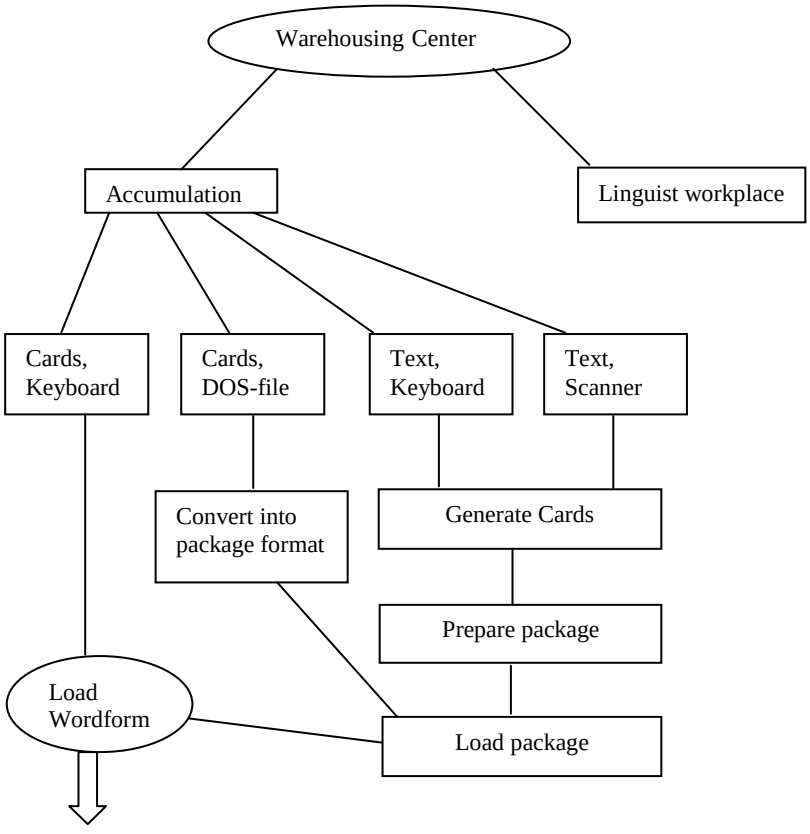
T5. Program, which prepare package from generated cards.

T6. Program, which loads cards package into database and carries out an appropriate checking.

As for cards DOS-files:

T7. They must be converted into developed package format.

And then they will be loaded into database using task T6.



Lexical Fund Database

2. Global network Tasks

The main task on this level, as was said above, is to share each Warehousing Center's lexical fund with other Warehousing Centers. Logically, the easiest way is to copy each Warehousing Center's lexically funds into other Warehousing Centers. But this procedure is not one-off job. This copying procedure must be performed regularly with definite periodicity, say, once per day. This procedure is known as synchronization. It is very hard problem and, therefore, modern database systems offer special tool – replication. Despite the fact that it is not perfect tool, that it is intricate in use, replication is a little in this direction.

Nevertheless, we think that for higher controlling and reliability will be better to construct special soft for Warehousing Center's synchronization. Such soft requires the separate server. Also, it is necessary to take this fact into account during designing the architecture of lexical fund database.

So, the draft list of tasks for Thesaurus construction is:

T0. Designing the architecture of lexical fund database taking into account a synchronization aspect.
T1. Wordform load into database. This procedure along with wordform storing must perform appropriate checking functions.
T2. Wordform input from keyboard.
T3. Cards generation from text.
T4. Development of FORMAT of Cards package.
T5. Package preparation from generated cards.
T6. Package Loading into database.
T7. Cards DOS-files transformation into package format.
T8. Soft for synchronization.

Of course, this list does not contain organizational points, e.g. text correction after scanning.

R E F E R E N C E S

1. Page W., Austin D. and others. Using Oracle 8/8i. Que Corporation, 1999.
2. Riccards G. Principles of Database systems. Addison-Wesley Company, 2001.
3. Date C. Database systems. Addison-Wesley Longman, 2000.
4. Гофман В., Хомененко А. Delphi 6. БХВ-Петербург, 2002.