IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 23, Issue 4, Ser. II (Jul. –Aug. 2021), PP 25-29
www.iosrjournals.org

Architecture of the Georgian Lexicographic System

Manana Khachidze[1], Miranda Makharashvili[2], Vano Kobaidze[3], Nino Bantsuri[4]
(Computer Sciences Department, Iv.Javakhishvili Tbilisi State University, Georgia)

Abstract:
The paper discusses the conceptual model of the development of Georgian lexicographic system, which is presented in the context of information system projection (design). It describes the elements of the system's framework and their functional purpose. The paper suggests the conceptual framework of the lexicographic information system in the form of diagram and its extended view in relation to databases.

***Key Word**: Lexicography; dictionary; information systems.*

Date of Submission: 07-08-2021 Date of Acceptance: 21-08-2021

I. Introduction

Lexicography is one of the oldest sociocultural practices of humanity. It develops with the development of humanity, both in terms of content and it's use. The transformative vision of lexicography emerged in the late 1980s, in which lexicography is presented as a field of social practice and independent science related not only to the analysis and creation of dictionaries, but also to the ability to consider the interests of the user in relation to specific problems (issues) [1]. The establishment of this vision is related to the development of information and communication technologies (ICT).

The development of ICT and its importance in human life has introduced lexicography as part of "Information Society" [2] and has increased availability of dictionaries for the users. At the same time, users' requirements are changing for this type of information systems.

In order to meet the new requirements, it is needed to be done revolutionary changes in almost every aspect of Practical Lexicography - in the demonstration of the final product to the user, in the procedures necessary for the development of this product, as well as in the use of researches and their results [3]. Accordingly, lexicography in terms of practical use has reflected in various information systems. As for the development of information systems – it's based on the rules and methodologies that must be followed in order to develop a fully functioning system [4]. Lexicographic information systems also allow to perform functions relevant to a non-lexical situation, in traditional terms. These situations can be: communication situations - when a communication problem can be solved (e.g. Receive and create text in one language and translate into another); Cognitive situations - when there is a need of receiving knowledge about a subject; Interpretive situations - when it may arise the need of interpretation and understanding of a non-linguistic sign, signal, symbol, etc.

Besides to dictionaries, in the lexicographic information system we consider the existence of various modules related to natural language. These modules are: database of grammatical rules of word formation, database of lexicographic standards, in which there will be describe international standards and additional requirements (that must be taken into account due to the peculiarities of Georgian language), corpus linguistics (language corpus/body), well-known algorithms related to natural language processing and modified and created algorithms for processing Georgian language.

II. Georgian Lexicography

Georgian is one of the oldest and the most morphologically complex languages, which is spoken by only 5 million people worldwide [5] [6]. Georgian is the state language of Georgia and has three unique alphabets (Asomtavruli, Nuskhuri and Mkhedruli [7], [8] two of which are currently used. The oldest Georgian dictionary (in terms of lexicography) is the Georgian-Italian dictionary, which was compiled by Stefano Paolini and Nikifor Irbach (containing 3084 words, published in Rome in 1629). Sulkhan-Saba Orbeliani's "Sitkvis Kona" („სიტყვის კონა“) is an important sample of Georgian lexicography, which was completed by the author in 1713. The number of different types of Georgian dictionaries is about 500, the most of which are presented only in printed way. Although, the most of the dictionaries belong to translation dictionaries type, the collection of Georgian dictionaries includes all existing types of dictionaries (explanatory, orthographic, dialectical, terminological, etc.).

III. The Main Elements of the Lexicographical System Framework

The traditional process of creating any information systems project has a Systems Development Life Cycle (SDLC). SDLC - a complete set of steps which must be followed to ensure definition, development, effectiveness support and modification of information system [9]. In the information system planning one of the approaches is a top-down planning in which “enterprise” data modeling is a key component. As for the database, it is one of the sources of this model. Our "enterprise" for which the information system is created is atypical and therefore "enterprise data" should be established, as well as "enterprise business rules", which is also important for the information system.

Enterprise modeling is the initial stage in the information systems’ planning process, in which the development of a database is one of the main components. During this stage, analysts review current databases and information systems; Analyze the nature of the business field that is the subject of the project; Describe the data required for each information system under consideration.

They determine what data is already available in the database and what new data is proposed to support the new project. After considering the importance of organization of each project, the selected projects are forward to the next phase.

Based on this description, the lexicographic environment should be described as a business area for the lexicographic information system. It is necessary to describe the data that mainly represent the various dictionaries and other related data. Besides dictionaries, lexicographic information system data sources include language corpus, various grammar rules, various natural language processing algorithms, industry standards, and other lexicographic materials. The analysis of this data will form the basis for the development of a conceptual model of the data [10] During the SDLC analysis phase, the analyst creates a detailed data model that identifies all the data that must be managed for this information system. In our case these are basically different types of dictionaries. At the same stage all data categories, all data attributes and "business connections (links)" should also be defined.

- **Data category -** Only high-level categories of data (entities) and major relationships are included at this point. By this we can mean different types of dictionaries as separate essences (main points), as well as texts that will form the body of language included in the system, grammar rules, illustrative texts of words and other illustrations (image, video, audio).
- **Data attributes -** these attributes are different for the different types of dictionaries and must meet international standards (ISO 1951:2007(en) Presentation/representation of entries in dictionaries — Requirements, recommendations and information) [11]. The incomplete list of these attributes: 1. Lemma 2. Sublemma 3. Homonym number 4. Polysemy number 5. Meaning 6. Lexical remark 7. Lexical remark for text production 8. Grammar, word class 9. Grammar, recommended inflexion 10. Grammar, non-recommended inflexion (one or more) 11. Grammar/spelling remark 12. First reference 13. Second reference(s) 14. Collocation(s) 15. Example(s) 16. Word formation(s) 17. Synonym(s) 18. Antonym(s) 19. Synonym remark 20. Proverb(s) 21. Idiom(s) 22. Idiom meaning 23. Internet link 24. Memo field. In the list of attributes it will also be necessary to take into account linguistic peculiarities. These attributes are defined by the dictionary structure that will be used as the primary data source. An important feature of the structure is the disposition of the elements that are interconnected [12] For printed dictionaries the structure (macro and micro) allows you to specify the order for surveying and accessing the dictionary elements. These attributes also define the search route which the user can use to meet a specific lexicographic need regarding a particular dictionary. The dictionary structure can also be used to generate "business rules", search scripts and other rules.
- **"business connections" -** the interdependence between data entities, as well as all the rules that ensure data integrity. In "business connections" we mean the connection between different types of dictionaries, as well as between other objects of the information system, which are represented in the diagram as the main entities. These rules should address the information requirements of the user and should fully represent the informational behavior of the user. They describe and define data and access to information of the users of different rights. The development of rules is based on Functional Theory of Lexicography [13], [14]. This theory focuses on the information requirement of the user in a particular situation, i.e. communicative, cognitive, operational, and interpretive situations: 1. Communication situations - it may require to solve a communication problem; 2. Cognitive situations – it may require specific knowledge; 3. Operational situations – it may require instruction on how to perform a physical or mental action (decision support); 4. Interpretive situations - it may be necessary to interpret and understand different signs, signals, symbols, and others.

A variety of users will be expected in case if we consider the fact that the proposed lexicographic information system is not limited to electronic dictionaries. Accordingly, the requirements will be various as well. As for any business, analysis of operational data is also important for lexicography, based on which it will be possible to optimize the functioning of the system. To achieve this, it is important to include data warehouses as one of the most important components of the information system [15]. The data warehouse will combine information from various internal and external sources and organize it to make accurate and timely decisions in the appropriate format. It provides an informed decision-making process on issues such as: attachment program updates, analysis of user's requirements trends, management of relationship with users, etc.

IV. Architecture of Lexicographical System

Architecture of Lexicographic system is based on one of the popular concepts ARIS (Architecture of Integrated Information Systems - Fig. 1) [16]. The advantage of this model is that it is divided into independent views and descriptions, which allows independent development of individual elements with the method selected for them, without the need to reflect the whole model. This methodology provides a system development lifecycle for business process mapping and optimization.

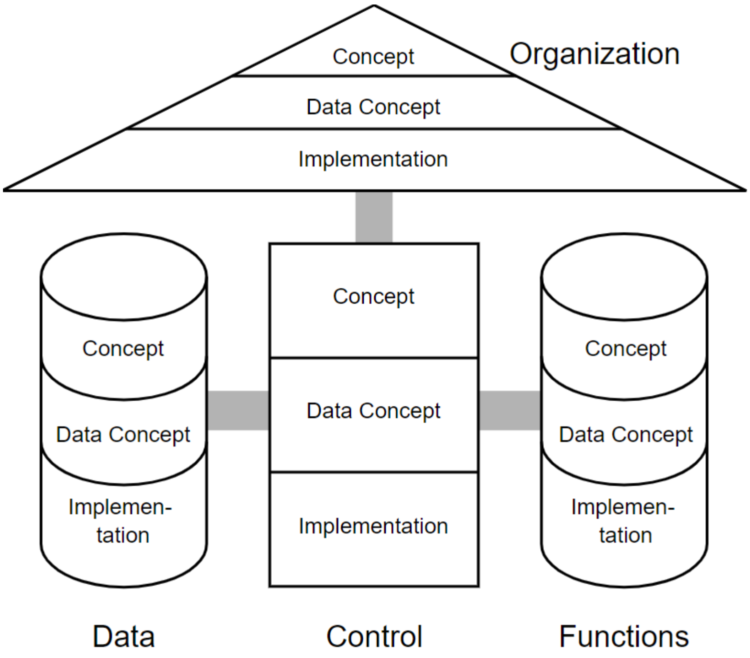

Fig. 1 Model of the ARIS Framework.

ARIS architecture is represented diagrammatically in the form of a house. Each description view of the house is divided into three levels of description: **Concept -** structured representation of business processes through business-friendly models. In this situation, it refers to the relevant rules of "business connections", which we have discussed in the previous section; **Data Concept -** which mainly refers to the IT concept of data processing. Implementation of this concept in descriptive models is related to IT and data types. Due to the typical diversity of data (symbolic, textual, audio, video) that will be placed in a lexicographic system, data concepts are related to the types of database use - structured, semi-structured and unstructured; **Implementation** - IT-technical implementation of the presented parts of the process, which is presented in the form of program code, database systems and various application protocols.

In the terms of ARIS architecture, our lexicographic system's architecture can be represented as following Fig.2.

A block of concepts consists of rules that are thematically divided according to the data source. However, this also refers to rules that are designed to process several different data (dictionary, language corpus, etc.) at the same time.

Database concepts ensure the exchange of data, both between units (entities) united in one database, as well as between different data groups of system's database (dictionary, corpus, etc.).

Implementation IT- the diversity of technical realizations is determined by the study of the requirements of system users, which in addition to the lexicographic requirements is focused on solving communicative, cognitive, operative and interpretive situations/issues of the user.

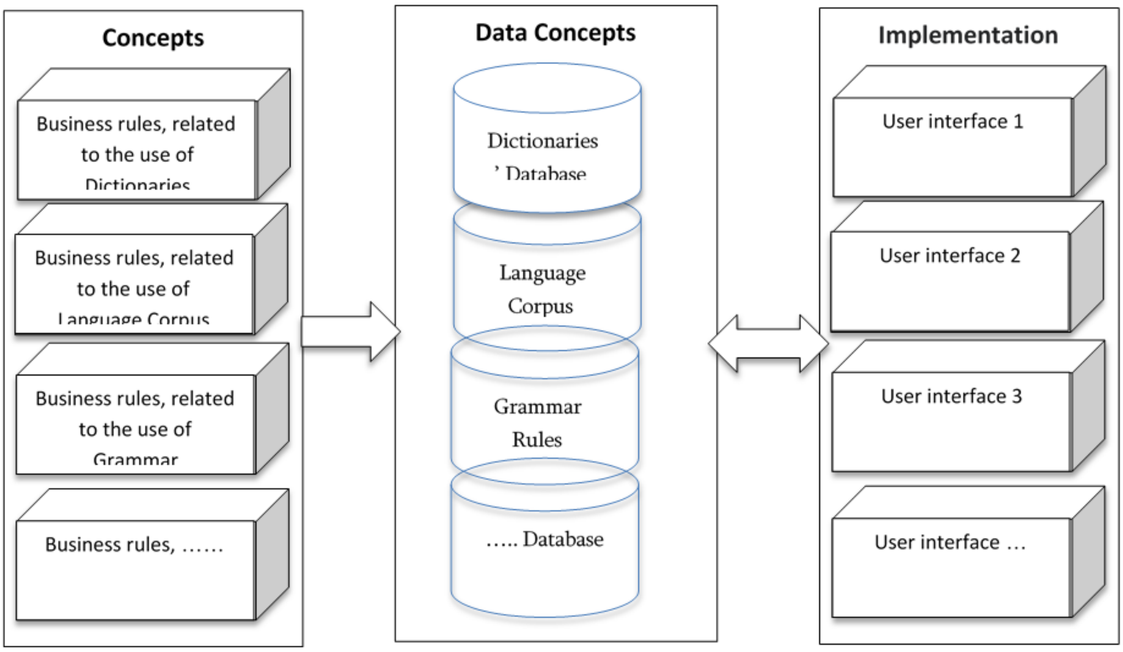

Fig. 2. Architecture of Lexicographical System

A block of concepts consists of rules that are thematically divided according to the data source. However, this also refers to rules that are designed to process several different data (dictionary, language corpus, etc.) at the same time.

Database concepts ensure the exchange of data, both between units (entities) united in one database, as well as between different data groups of system's database (dictionary, corpus, etc.).

Implementation IT- the diversity of technical realizations is determined by the study of the requirements of system users, which in addition to the lexicographic requirements is focused on solving communicative, cognitive, operative and interpretive situations/issues of the user.

It is important to analyze operational data to ensure the life cycle of information systems, which is one of the main functions of data warehouses. In our case it would be more convenient to use data marts (specifically The dependent data mart) - limited volume data warehouses. The data for the data mart is obtained by selecting and summarizing data from the data warehouse or by extracting, converting and uploading separate data from the data source systems. It is considered to be one of the best means for presentation of organizational data to support all users data requests [17].

The most popular in the development of data warehouses is a three-layer data architecture. Each level in this architecture has its own data corresponding to:

1. Operational data – data that is stored in various operating systems, outside the organization (sometimes even on external systems). For the lexicographic system, this data is information about actions and transactions performed by users with different rights. This includes both the traditional use of dictionaries as well as other situational actions; This data is stored for a certain period of time without any further processing;
2. Reconciled data -the type of data that is stored in the enterprise data warehouse and in the operational data mart. Consolidated data is current and detailed and should be the only authoritative source for a decision support program.
3. Derived data - the type of data that is stored in each data mart. Data that is selected, formatted and aggregated for the end-user's decision support programs.

Totally, our lexicographic information system can be shown in the form of the diagram as in Fig. 3.

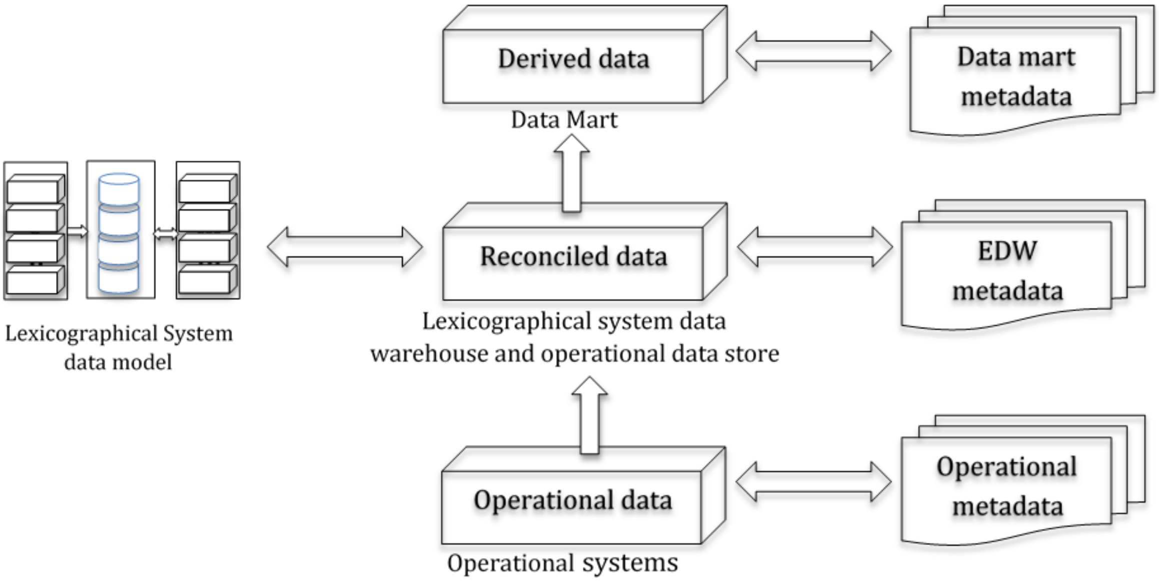

Fig.3 The lexicographic information system general view

V. Conclusion

Developing a model for the lexicographic information system is a difficult task that requires the various activities related to lexicography, as well as reflection of all components involved in this activity in the form of “Industrial model". All these are related with comprehensive practical and theoretical knowledge of lexicographic models and specific characteristics of linguistic structures. Due to the specifics of the field, it will be inevitable to abolish a number of standards and requirements, as well as processing additional value procedures related to information system processing, which will greatly complicate the development process. The design and functional characteristics of lexicographic information system structures are based on specific problems in lexicographic data processing. The database of this system’s models is focused primarily on direct and indirect data search. At the same time, the task of special purpose search is very important for linguistic research. Combining data into one model to provide an effective search engine of different types is a difficult task. Therefore, different search engines have a separate place in the system for the effective solution of various problems. The presented model is implementing and developing in programmatic manner.

References

[1]. Tarp S., "Lexicography in the information age," vol. 17, pp. 170-179, 2007.
[2]. Robins K., Webster F., "Technology and education: progress or control," Critical Social Policy, vol. 5, no. 15, pp. 36-61, 1985.
[3]. Tarp S., "Dictionaries in the Internet Era: Innovation or Business as Usual? (Enrique Alcaraz Memorial Lecture 2014)," licante Journal of English Studies, vol. 27, pp. 233-261, 2014.
[4]. Avison D., Guy F., Information Systems Development: Methodologies, Techniques and Tools, 4. ed, Ed., McGraw-Hill, 2006.
[5]. Hewitt B., Georgian: A Structural Reference Grammar., Amsterdam: John Benjamins., 1995.
[6]. Harris A. C., Georgian Syntax: A Study in Relational Grammar., Cambridge: Cambridge University Press,, 2009.
[7]. Gamkrelidze T., Writing system and the old Georgian script, Tbilisi: TSU, 1989.
[8]. R. Shosted, V. Chikovani, "Standard Georgian," Journal of the International Phonetic Association, vol. 36, no. 2, p. 255–264, 2006.
[9]. Hoffer J. A., George J. F., Valacich J. S., Modern Systems Analysis and Design, vol. Upper Saddle River, 6. ed, Ed., NJ: Prentice Hall, 2014..
[10]. Batini C. , Cer S.i, Navathe S. B., Conceptual Database Design: An Entity-Relationship Approach., Menlo Park, CA: Benjamin/Cummings, 1992.
[11]. "ISO 1951:2007(en)Presentation/representation of entries in dictionaries — Requirements, recommendations and information.".
[12]. Wiegand H. E. Gouws. R., "Macrostructures in Printed Dictionaries ”„" in in R. Gouws, U. Heid, W. Schweickard and E. H. Wiegand (eds.), Dictionaries: An International Encyclopedia of Lexicography, Supplementary Volume HSK 5.4. Berlin/Boston, MA/Ne, 2013.
[13]. Tarp S., Lexicography in the Borderland Between Knowledge and Non-Knowledge., Tübingen: Niemeyer., 2008.
[14]. Fuertes-Olivera P., Tarp S., Theory and Practice of Online Specialised Dictionaries. Lexicography Versus Terminography., Berlin/Boston: De Gruyter, 2014.
[15]. Inmon, W., Strauss D., Neushloss G., DW 2.0: The Architecture for the Next Generation of Data Warehousing., Morgan Kaufmann Series in Data Management Systems, 2008.
[16]. Matthes D., Enterprise Architecture Frameworks Kompendium Über 50 Rahmenwerke für das IT-Management., Berlin: Heidelberg: Springer-Verlag Berlin Heidelberg, 2011.
[17]. Imhoff C., "The Corporate Information Factory," December 1999. [Online]. Available: http://www.informationmanagement.com/issues/19991201/1667-1.html.