Automated system for the creation and replenishment of users' electronic lexicographical resources

N.V. Borysova, K.V. Melnyk

National Technical University «Kharkiv Polytechnic Institute», Kharkiv, Ukraine
*Corresponding author. E-mail: borysova.n.v@gmail.com, karina.v.melnyk@gmail.com

Paper received 12.12.18; Revised 17.12.18; Accepted for publication 19.12.18.

https://doi.org/10.31174/SEND-NT2018-186VI22-13

Abstract. This article proposes a solution to improve the efficiency of automated generation of electronic lexicographical resources based on strongly-structured electronic information arrays processing. The developed automated information system for lexicographical resources creation and replenishment have been described is this article. Several supporting subsystems of developed automated system have been characterized. The effectiveness of the information system has been evaluated.

***Keywords**: Automated systems for natural language processing, electronic lexicographical resources, strongly-structured electronic information arrays*

Introduction. The problem of automated lexicographical resources creating needs to be solved, because, firstly, using special automatic lexicographical resources in programs for natural language texts automated and automatic processing have greatly increased the efficiency and quality of such processing; and secondly, in the lexicographical resources the conceptual model of a certain subject area is reflected because they contain concepts, connections between them, definitions of these concepts. The current level of information technology development provides an opportunity to solve this problem only partially by developing and using specialized information systems based on a variety of approaches and methods. As sources of fulfilling the lexicographical resources such systems use natural language texts. But many researchers ignore such sources of information as already created, existing lexicographical resources of different types. Therefore, the problem of developing an information system for the automated creation of electronic lexicographical resources based on the analysis of existing ones can be considered relevant.

Review of the literature. At the moment there are quite a lot of tools and services for the creation and replenishment of electronic lexicographical resources. In the framework of this work, some of them, created in Ukraine and working with the Ukrainian language, were analyzed.

Integrated lexicographic system made by Ukrainian Lingua-Information Fund, NAS of Ukraine, based on the theory of lexicographic systems, consists of three subsystems:

1) a computer library that combines the functions of the electronic catalog, the database and facilities for the processing of generalized storage objects, that is, heterogeneous information presented in the machine form: books, drawings, audio, video, graphic information, databases etc. [1];

2) an automated lexical file system consisting of a texts database (library); Segment bases (microcontexts) obtained from these texts; set of algorithms by which these segments (microcontexts) are extracted; set of all wordforms of the segment base (microcontexts); the set of all complete paradigms for all wordforms of the segments base (microcontexts); dictionary with the necessary information retrieval functions; search-bibliographic block; a block of statistics and control over the new words inputting to the dictionary [2];

3) vocabulary subsystem which is convenient functional environment for working with texts and dictionaries in Ukrainian. [3].

All subsystems described above are united into a single system with the help of the integration program shell «Lexicograph» which provides cross-navigation across all subsystems. The described system creates the preconditions for complex automation of lexicographical activity, from the stage of vocabulary card indexes formation and vocabulary structure design and ending with the stages of the automated typing, pages layout making and dictionaries replication [3].

The system of the multilingual dictionaries creation named PolyDic ML v.3.0 implements the approach «from a computer dictionary to paper one», that is why it is a flexible system by which computer encyclopedic and linguistic dictionaries of different types can be made. The PolyDic ML v.3.0 system consists of two modules and software applications: the main software module – the editor for making and editing dictionaries (PolyDic ML Editor); a module for viewing and working with dictionaries (PolyDic ML Viewer) as well as application programs (including PolyDic ML Localizer – allows the user to locate or edit the system interface in a particular language) [4, 5].

The automated system for managing the integral dictionaries ASVIS, created at the V.M. Glushkov Institute of Cybernetics, NAS of Ukraine, is a program implementation of the integral dictionary concept developed by the Institute. Today, within the framework of this system, a subsystem SIFORS for the dictionaries formation is created, which is oriented on the management of terminological databases [6].

The adaptive linguistic system ALISA, created at the Institute of Applied Informatics, NAS of Ukraine, is a natural language linguistic processor oriented to a number of functions, in particular, automated creation of dictionaries, thesauri, phraseological, terminological databases [6].

The system of support for multilingual terminology dictionaries SLOVO, created at Lviv Polytechnic University, can be used to prepare dictionaries for publication [6].

The complex for the creation of dictionaries provides the development and support of an electronic user dic-

tionary in any chosen domain area [7]. The complex consists of electronic dictionary shell and tools for filling (replenishment) the new articles to the dictionary. The shell implements the wordform normalization using a morphological analyzer; the searching for an article in a normalized wordform (lexeme) and displaying in the window of the article found.

Thus, the reviewed computer systems for the formation of lexicographical resources implement the functions of creation, replenishment and use of lexicographical resources, but they do not provide the user with the opportunity to analyze existing electronic dictionaries and automatically extract the necessary information from them.

The purpose of this article is to solve the problem of automated creation and replenishment of users' electronic lexicographical resources based on strongly structured electronic information arrays processing

Materials and methods. The main function of the developed information system is automated extraction the necessary material from the existing electronic lexicographical resources, based on the lists of dictionary markers. Dictionary markers are a lexicographic abbreviated designation, which given in the dictionary article and contains lexical, grammatical, stylistic and other features of the lexeme [1].

Proceeding from the aforementioned and general requirements for modern information systems [8], the architectural structure of the information system under development should correspond to the following basic principles:

- compliance with current and future goals, as well as functional strategic objectives of the information system;
- information system universality;
- providing user-accessible structuring of data and a sufficient depth of their description;
- providing the required search operativeness and performing analytical-synthetic queries;
- flexibility and the ability to develop and increase the functions and resources of the information system according to the evolution of the using sphere and objectives of its use;
- providing of remote authorized users' access for information system using based on modern GUI;
- realization of technological functions inherent to such information systems (providing integrity, consistency, minimizing data redundancy, data protecting from users' incompetent actions and the possibility of data recovery).

To create the information system for the creation and replenishment of users' electronic lexicographical resources a system approach was chosen, which consists in the complex study of the object as a whole with the representation of its parts as purpose-oriented systems and the study of these systems and the relations between them. In the system approach, an object is considered as a set of interconnected elements of one complex dynamic system, which is in a state of constant changes under the influence of many internal and external factors associated with the processes of transforming input resources on the output. The system approach is based on the following principles: the absolute priority of the ultimate goal, unity, connectivity, modular construction, hierarchy, functionality, development, decentralization, taking into account uncertainties and randomness in the system[8].

Significant features of the system approach are: simultaneous coverage of designing a large number of tasks; maximum typification and standardization of solutions; multi-dimensional representation of the structure of the information system as a system consisting of several classes of elements, and their relative autonomous development; key role of databases; local implementation and increase of functional tasks [8].

Since in the system approach, as already noted above, an object is considered as a set of interconnected elements of one complex dynamic system, the information system can be considered a set of functional subsystems and relationships between them. Functional subsystem is the information system part, highlighted by the functional features commonality. Information system functional decomposition determines the allocation of subsystems, i.e., for what scope it is intended and which main goals, tasks and functions performs. Depending on the complexity of the object, the number of functional subsystems may be different [9].

For allocation the information system functional subsystems the following requirements must be met:

- the tasks that make up the subsystem should not interfere with each other;
- the tasks solved in subsystems, should be closely related to each other in the information plan, that is, when solving them should use a single input information, and the results of solving some tasks should be used to solve the other;
- the results of the decision must have a single consumer [9].

For allocation the information system functional subsystems, their parameters must be determined: the purpose of the subsystem's functioning, the type of resources, and the features of the indicators that are calculated in the subsystem.

The functional subsystems exploitation requires the availability of appropriate resources that creating by the information system's supporting subsystems: mathematical, algorithmic, informational, software, organizational, methodological, technical, linguistic, legal, ergonomic. Let us consider in more detail some of them for the developed information system for the automated creation and replenishment of users' electronic lexicographical resources.

The mathematical support of the information system is a collection of mathematical methods and models used in the information system [8]. Models of lexicographical units' identification by markers in the texts of existing lexicographical resources are used as a mathematical support of the developed information system. Mathematical models were developed using the apparatus of finite predicate algebra and the method of comparative identification. As external identifiers, it is proposed to take $x_1 \div x_n$, which determine the presence of a particular marker in the dictionary article. These identifiers acquire the values *yes* – if the marker is present and *no* – if it is not present. The fields of change of these variables can be formally written in the form of the following equations: $x_1^{yes} \vee x_1^{no} = 1$, $x_2^{yes} \vee x_2^{no} = 1$, $x_3^{yes} \vee x_3^{no} = 1$,…, $x_n^{yes} \vee x_n^{no} = 1$. These identifiers are sufficient to identify the lexicographical units in the electronic lexicographical resources. Consequently, if the set of lexicographical

units from electronic dictionaries is denoted by $T = \{t_i\}$, and the set of identifiers chosen by us through X, then we can enter the predicate $P(t_i, X)$, which accepts the value of 1 in the presence of a identifier or 0 – in the opposite case $P(t_i, X) = P(t_i, (x_1, x_2, x_3, \ldots, x_n))$. That is, the predicate $P(t_i, X)$ implements the recognition of the lexicographic unit in the dictionary text.

On the basis of this model, the algorithm of the process carried out in the information system was constructed. This algorithm is an algorithmic support of the developed information system. The algorithm for the process of creating a user's lexicographical resource is presented in Figure 1.

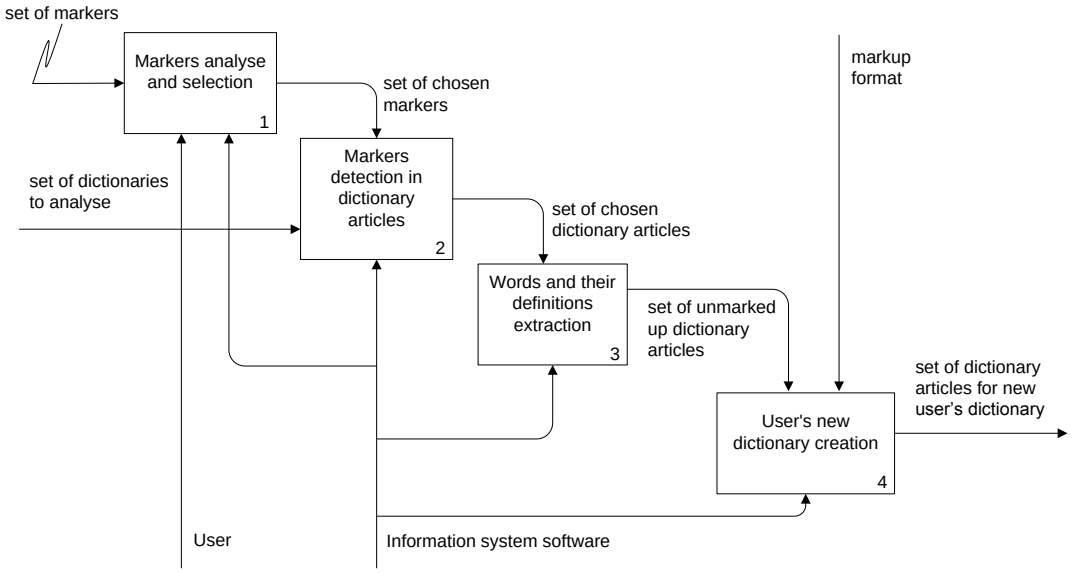

Figure 1. – The process of user's automated electronic lexicographical resource creation

At first, the user from the default markers set in the system selects only the ones he needs. Then he downloads electronic dictionaries that he wants to analyze. The system searches for markers in dictionary articles downloaded for the analysis and extracts dictionary articles that have the markers. If the dictionary article consists of the several word definitions, system extracts just one of them that contains the marker. Vocabulary articles selected in this way or their parts are converted into vocabulary articles using a given set of markup rules, from which the new user dictionary is actually formed.

In addition to create their own dictionary based on the lists of markers, the user of the information system can:

- search for words;
- view information about wordsfrom the dictionary, such as their definitions and information about the marker by which the word was found;
- view other dictionaries available on the system;
- add dictionary articles manually through a special window for dictionary replenishment;
- edit dictionary articles manually through a special window for dictionary articles editing;
- delete dictionary articles and dictionaries;
- create, view, edit, delete markers and lists of markers;
- view information about markers, their meanings;
- search for markers.

The information system architecture includes databases, data processing tools, information resources access tools, user's work organization tools, administration tools and data transaction tools. Information system software is a set of separate components. Each component implements a set of closely interconnected tasks, which ensures the implementation of the necessary set of operations over the data and the sequence of their implementation. The information system functioning is provided by the program, which is a set of software tools implementing the storage and processing environment, the data access interface and the shell for data processing. The data storage environment is the Microsoft SQL Server database management system. The exchange interface is implemented using the Apache web server. The shell for data processing is a program that implements the basic functions of the data management system. The program works through a web-interface. The using of the information system by the submitted architecture provides optimal organization of the user's work with the information resources of the information system.

Thus, an information system for the automated creation and replenishment of users' electronic lexicographical resources is developed as a tool and service of creating and replenishing an individual or corporate electronic dictionary that can be used by the user at his own discretion.

Discussion the results

According to the intergovernmental standard for information in librarianship and publishing business for the evaluation of the efficiency search and extraction of lexicographic units are used the coefficient of accuracy *Precision*, the coefficient of completeness *Recall*, the coefficient of noise *Fallout* and the coefficient of extraction error *Error*, which are determined by the following formulas:

$$Precision = \frac{a}{a+b},\ Recall = \frac{a}{a+c},\ Fallout = \frac{b}{a+b},\ Error = \frac{b+c}{(a+b+c+d)}.$$

where a – number of correctly extracted lexicographic units;

b – number of incorrectly extracted lexicographic units;
c – number of incorrectly unextracted lexicographic units;
d – number of correctly unextracted lexicographic units.

In total, about 700 electronic dictionaries have been analyzed according to different lists of markers; more than 5000 experiments were conducted. These values of coefficients are obtained: *Recall* = 0.97; *Precision* = 0.98; *Fallout* = 0.02; *Error* = 0.01. Since comparing the results with the results of other similar systems is not possible in the absence of systems with the same functionality, the results were compared with the results of systems with a similar functionality. The comparison showed a greater efficiency of the developed information system for solving the problem, since the values of the coefficients found in literary sources, containing the descriptions of other systems, vary within: for the *Recall* coefficient ranging from 0.79 to 0.86; for the *Precision* coefficient ranging from 0.83 to 0.95.

In addition, the values of the *Precision* and *Recall* coefficients for the developed information system are close to the 1 that is the highest value that these coefficients can take, while the *Fallout* and *Error* coefficients are quite low, which also proves the efficiency of the developed information system.

Conclusions. Thus, information system of automated creation and replenishment of user's lexicographic resources is a system that provides satisfaction of the user's information needs in the lexicographical information processing, as well as lexicographical processing of information. The purpose of the information system is realized through its functions: automated creation of various purposes user's lexicographic resources, automated collection, processing, storage of lexicographic information, information support of users etc.

ЛИТЕРАТУРА

1. Широков В. А. Елементи лексикографії / В. А. Широков. – К. : Довіра, 2005. – 304 с.
2. Широков В. А. Інформаційна теорія лексикографічних систем / В. А. Широков. – К. : Довіра, 1998. – 331 с.
3. Широков В. А. Комп'ютерна лексикографія / В. А. Широков. – К. : Наукова думка, 2011. – 352 с.
4. Кінаш Р. Система для укладання комп'ютерних версій словників PolyDic ML 3.0: функції та засоби редактора / Р. Кінаш, Р. Мисак, Ю. Каличак, О. Мельник // Проблеми української термінології: Збірник наукових праць учасників 11-ї Міжнародної наукової конференції. – Львів : Національний університет «Львівська політехніка», 2010. – С. 38-42
5. Мисак Р. Комп'ютерні словники: класифікація та укладання / Р. Мисак // Проблеми української термінології: Збірник наукових праць учасників 10-ї Міжнародної наукової конференції. – Львів : Національний університет «Львівська політехніка», 2008. – С. 52-55.
6. Дубічинський В.В. Українська лексикографія: історія, сучасність та комп'ютерні технології: Навчальний посібник / В. В. Дубічинський. – Харків : НТУ «ХПІ», 2004. – 203 с.
7. Хахалин Г. К. Комплекс по разработке индивидуальных и/или корпоративных электронных толковых словарей / Г. К. Хахалин, Н. К. Богданов, С. В. Платонов: [Электронный ресурс]. – Режим доступа: www.raai.org/about/persons/khakhalin/pages/kogmod2000.doc
8. Береза А.М. Основи створення інформаційних систем: Навчальний посібник / А. М. Береза. – 2-ге видання, перероблене і доповнене. – К.: КНЕУ, 2011. – 205 с.
9. Борисова Н.В. Інформаційна система автоматизованого формування лексикографічних ресурсів / Н. В. Борисова, І. С. Ямшанов // Проблеми інформаційних технологій. – 2014. – № 1(015). – С.193-199

REFERENCES

1. Shirokov V. A. Elements of lexicography / V. A. Shirokov. – K. : Dovira, 2005. – 304 p.
2. Shirokov V. A. Information theory of lexicographical systems / V. A. Shirokov. – K. : Dovira, 1998. – 331 p.
3. Shirokov V. A. Computer lexicography/ V. A. Shirokov. – K. : Naukova dumka, 2011. – 352 p.
4. Kinash R. System for creating computer versions of dictionaries PolyDic ML 3.0: functions and features of editor / R. Kinash, R. Mysak, Iu. Kalychak, O. Melnyk // Problems of Ukrainian terminology: Collection of scientific works of the participants of the 11th International Scientific Conference. – Lviv: National University «Lviv Polytechnic», 2010.–P.38-42
5. Mysak R. Computer Dictionaries: Classification and Creation / R. Mysak // Problems of Ukrainian terminology: Collection of scientific works of the participants of the 10th International Scientific Conference. – Lviv : National University «Lviv Polytechnic», 2008. – P. 52-55.
6. Dubichynskiy V.V. Ukrainian lexicography: history, modernity and computer technologies: textbook / V. V. Dubichynskiy. – Kharkiv : NTU «KhPI», 2004. – 203 p.
7. Khakhalin G. K. Complex for the development of individual and / or corporate electronic dictionaries / G. K. Khakhalin, N. K. Bogdanov, S. V. Platonov: [Electronic resource]. – Access mode: www.raai.org/about/persons/khakhalin/pages/kogmod2000.doc
8. Bereza A. M. Basics of information systems creation: textbook / A. M. Bereza. – 2nd edition, revised and supplemented. – K.: KNEU, 2011. – 205 p.
9. Borysova N. V. Information system for automated generation of lexicographical resources / N. V. Borysova, I. S. Yamshanov // The problems of information technologies. – 2014. – № 1(015). – P. 193-199