COGNITIVE STUDIES | ÉTUDES COGNITIVES, 16: 143–152
Warsaw 2016

DOI: 10.11649/cs.2016.013

GRYGORII CHETVERIKOV[A], **IRYNA VECHIRSKA**[B], & **OLEKSII PUZIK**[C]

Kharkov National University of Radioelectronics, Ukraine
[A]chetvergg@gmail.com ; [B]ira_se@list.ru ; [C]as308@mail.ru

TECHNICAL FEATURES OF THE ARCHITECTURE OF AN ELECTRONIC TRILINGUAL DICTIONARY

Abstract

This article is devoted to the development of the software system used to create an English-Russian-Ukrainian terminological dictionary. Scanned and recognized documents in MSWord format were the input data for the dictionary. Issues which appeared during the parsing of the input data are analyzed and solutions using regular expressions are identified. This article also describes the scheme of the dictionary's lexicographical database, and its classes of models, views and view models.

In addition, a detailed description of the software system from a user's perspective is included, the prospects for the usage of the dictionary are discussed, and the methods used during the development of the system are described.

The software system is built using the design pattern Model-View-View-Model. Through the use of this pattern, internal logic is separated from user interface, thus changes made in different parts of the software may be independent. The developed software system allows users to edit, to fill, and thus to create new thematic transferable electronic dictionaries. The main advantage of the system is the equality of languages, i.e. each user can decide which language is to be major.

Keywords: algebra of finite predicates; database; lexicography; lexical unit; MVVM pattern; parsing; software system

1 Introduction

At first glance, the creation of an electronic dictionary may not appear so difficult. In reality, it is a time-consuming, multistep procedure which sometimes requires innovative methods. Problems can arise related to the representation, handling and storing of data. The data model which is used in the database should take into account aspects of structure (i.e. descriptions of types and data structures in database), manipulation aspects, ways of changing the state of the data, ways of retrieving data from the database, and integrity aspects.

One of the main reasons it is so challenging to find solutions to these problems is the complex structure of the linguistic material. This leads, in turn, to the complexity of the data model organization. Some of the issues were solved for bilingual dictionaries (Широков, 2011). However, these solutions cannot be applied when building a trilingual dictionary because collisions occur, related to the inadequate representation of results caused by an avalanche-like increase in translated equivalents belonging to different semantic rubrics.

In this article, the challenges mentioned above are overcome using the theory of lexicographical systems (Широков, 2011; Рабулець, Широков, & Якименко, 2004) and the algebra of finite predicates (AFP) (Бондаренко & Шабанов-Кушнаренко, 2006, 2012).

Furthermore, it is assumed that the electronic dictionary is an open system. Therefore, the system built should provide the possibility of changing content without human intervention, and of responding accordingly. It is necessary to provide different levels of access to different users.

2 Method and difficulties faced

It is relatively simple to represent the internal structure of a bilingual dictionary: in general, each term has a corresponding translated equivalent. A dictionary consists of a set of such equivalents. This case is an example of a one-to-many relationship. The problem of defining relations between translated equivalents arises in trilingual dictionaries, especially if all three languages are to be treated equally. The number of relations increases proportionally to the number of languages. This case is an example of a many-to-many relationship. The usual solution for this problem is to introduce an additional level of indirection, which enables the change from a one-to-many to a many-to-many relationship. This approach can be used in the creation of not only trilingual dictionaries, but also multilingual ones.

The creation of electronic dictionaries usually consists of the following steps: scanning and recognition of paper texts into electronic form; splitting the electronic text of the dictionary into an array of dictionary articles; the automatic division of the array of articles based on formal criteria. (Рабулец et al., 2004).

Initially, the dictionary was in the form of scanned and recognized MSWord documents. It was not possible to use this data directly because of internal MSWord text representation, incorrectly recognized characters, and issues caused by hyphens, empty lines, different fonts and so on. These issues had some regularity which allowed them to be solved automatically. Also, all accented letters were incorrectly but universally recognized, so they were fixed by simple replacements. Accented letters were marked by "#" after parsing.

All terms from MSWord documents were transformed to Unicode text files. This was possible because all terms were started by a new line and incorrect "new line" characters were identified by regular expressions. Incorrect parentheses were identified by the same method. In general, it is not possible to use regular expressions to determine if parentheses are correct, but it was possible in this case because the dictionary text does not contain complex nested structures of parentheses.

The most difficult part of the work was the recognition of incorrect hyphens. The majority of hyphens were corrected by regular expression, but some had to be corrected manually. The following is an example of a term split by a new line but not processed in text file:

...
1.(моно, не)хроматѐческая □□(
2.моно, не)- хроматѐчна аберàція)
...
1.(-витîк
2.ампǻр-витîк
...

In the first case, the parenthesis from the first line belongs to the second line. In the second case, the parenthesis is a recognition artefact.

The next step was the extraction of topics, semantics (extended description), changeable parts and so on, which were used to fill the internal structures of the dictionary.

The construction principle of the dictionary is alphabetical-nesting. The nest includes the term's word combinations which contain the heading word as one of the elements. If the heading contains several one-word terms, then the verb in the imperfective form is the first for verbs, and the most widespread word is the first for other parts of speech. The part of the heading word which

is common for all word-combinations in a nest is separated by a straight bold line. Derived words take the tilde (∼) character instead. If the common part is not separated, then whole word is used instead of the tilde character. For terminological word-combinations, the corresponding Ukrainian equivalents preserve the word order of Russian word-combinations. The English equivalents use the word order which is natural for texts. (Остапова, 2007)

3 Description of the software structure of the trilingual dictionary

The software system is designed using an MVVM design pattern (Model-View-View-Model). The user interface is separated from the internal logic because of this pattern. Therefore, changes made in different parts of the software may be independent.

Below is a model of a trilingual dictionary. *TrilingualDictionary* — is a class which describes the entity of a three-language dictionary. It consists of many terms and provides create/read/update/delete (CRUD) operations to control the terms. The database scheme for the dictionary is represented in Fig. 1.

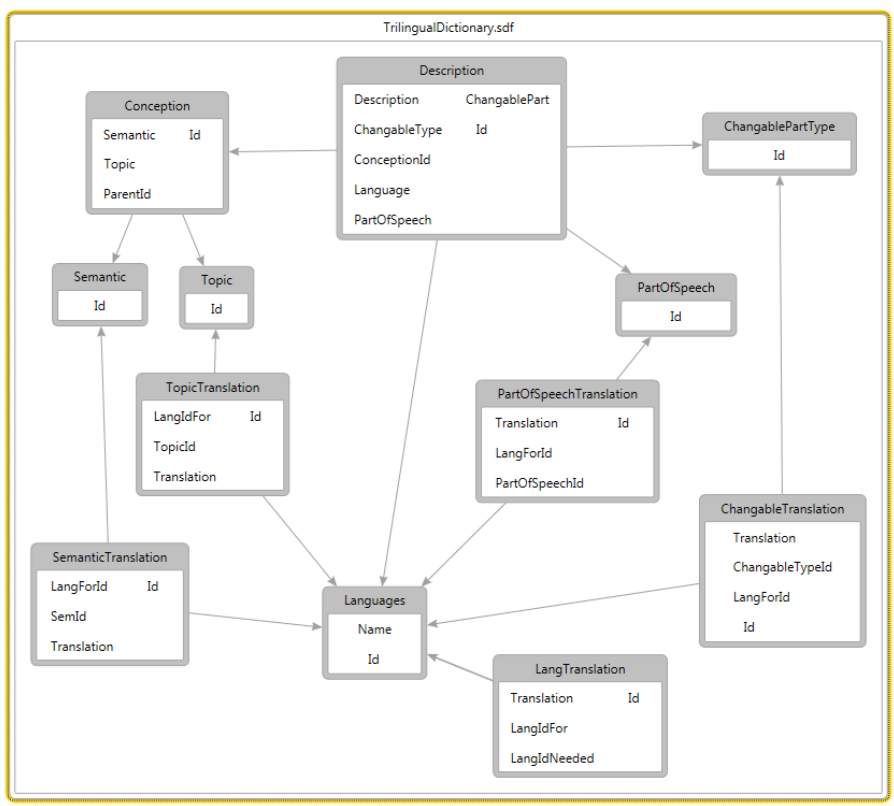

Figure 1: Database scheme

The class *Conception* describes the term itself and contains the following fields:

a) term id (Id);
b) reference to parent term (ParentId);
c) discipline or topic where the term is used (for example мат. — mathematics, рлк — radiolocation, etc.);
d) semantics of the term, especially important for homonymous terms;

e) reference to synonymic term;
f) list of translated equivalents for the term (described in class *ConceptionDescription*) for different languages.

The class *ConceptionDescription* contains a translated equivalent — description of the term. It consists of the following fields:
a) translated equivalent for the term;
b) references to the term's object;
c) changeable part of a term for language used for translation (for now, it is only the genitive case or plural);
d) part of speech, if it is important/required;

Translation classes provide translation of topics, semantics, parts of speech, etc. A diagram of these classes is in Fig. 2.

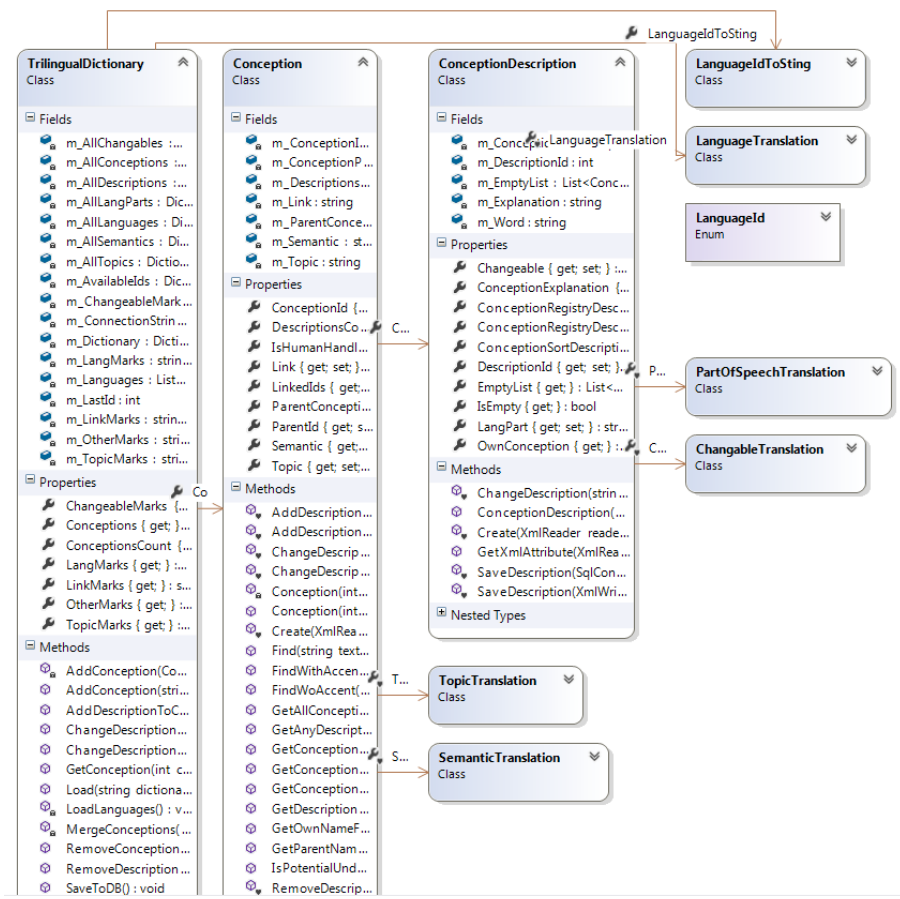

Figure 2: Diagram of core classes for translations

The view model provides a data model and behavior for the view and allows the view to set bindings to the view model. The data model provides data, along with business and validation logic. The view model is responsible for handling the view logic using the data model. It prepares data for binding to the view.

The class *TrilingualDictionaryViewModel* provides fields and methods to display the dictionary inside a view. It contains the following fields:
a) list of alphabets for each language (class *Alphabet*);

b) selected main language.

The class *Alphabet* is a container for words and word-combinations grouped in alphabetical order. It contains the following fields:
a) language id, which alphabet is stored in an object of this class;
b) list of letters of a language (class *Chapter*).

The class *Chapter* is a container for words and word-combinations grouped in alphabetical order in the scope of one alphabet letter, and which do not have a parent term (main or major terms). It contains the following fields:
a) name of a chapter (usually a letter of the alphabet);
b) list of view models for translated equivalents (class *ConceptionDescriptionViewModel*) for main terms.

The class *ConceptionDescriptionViewModel* is a container for words and word-combinations grouped in alphabetical order in the scope of one alphabet letter, and which have a parent term (additional terms). For example, "aberration" — main term, «chromatic aberration» — additional term. This class contains the following fields:
a) object of *ConceptionDescription* class;
b) list of *ConceptionDescriptionViewModel* objects for additional terms.

The class *ConceptionDescriptionEditViewModel* provides fields for editing/adding translated equivalents for each available language. It contains the following fields and methods:
a) translated equivalent for selected language;
b) topic description of a term;
c) semantic description of a term;
d) description of the changeable parts of a term for selected language;
e) part of speech of translated equivalent for selected language if needed;
f) CRUD operations for translated equivalents.

The class *ConceptionViewModel* is designed to represent terms with translations for all available languages. It contains the following fields and methods:
a) list of translated equivalents for all available languages;
b) select operations for translated equivalents for modifications by *ConceptionDescriptionEditViewModel*.

A diagram of view model classes is represented in Fig. 3.

A view is the representation of data of the dictionary taken from the view model. In general, this is the graphical user interface (GUI) of the software system. The view handles events fired by the view model when some values, properties or commands are changed. When the user interacts with elements of the GUI, the view follows corresponding command provided by the view model, thus interaction between the user interface and business logic occurs.

4 User manual of the software system of the trilingual dictionary

The software system of the trilingual dictionary is designed to create, edit, and view dictionary articles and their translated equivalents in three languages: Russian, Ukrainian, English.

Minimal system requirements:
— CPU 1000 MHz;
— RAM 512 MB ;
— More than 100 MB of free space on hard drive;
— OS Windows XP SP3;
— .Net Framework 4 Client Profile.

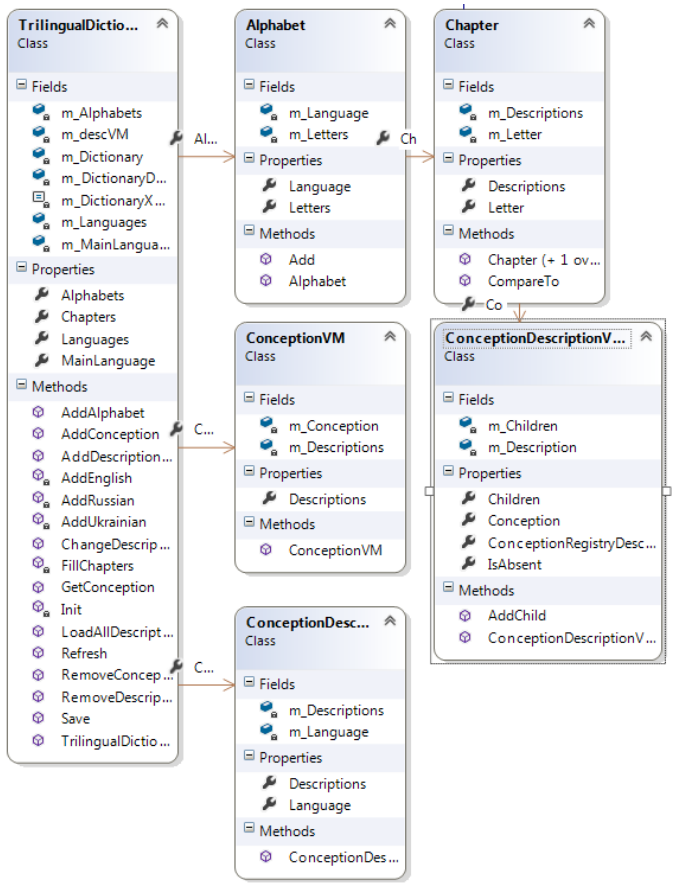

Figure 3: Diagram of view model classes

The user interface is represented in Fig. 4.

The left-hand side of the window is a tree of terms which comprise the dictionary. The dropdown menu “Основной язык” (Main Language) allows the user to select the main language, translated equivalents for which will be found. These will be displayed on left pane. The field “Поиск” (Search) allows the user to enter the part of a word which the needs to search. The character “#” is used as an accent character, and searches may be performed with taking accents into account or not. After the “Найти” (Find) button is pressed, a search of a term in the left pane is performed.

When the term is selected in the tree in left pane, the middle pane will show all translated equivalents, grouped by all available languages.

The right pane is designed to edit terms. To edit an existing term, it should be selected on the left pane. It is necessary to select the required language in the dropdown “Выберите язык” (Select language) to create/edit/remove translation equivalents.

The field “Введите описание” (Enter description) displays the translated equivalent, in the selected language, for the term selected in the left pane. Necessary changes should be made in this field.

The field “Тема” (Topic) should be filled with a conventional label for a topic or scientific discipline if needed (for example мат., math., физ., etc.)

The field “Семантика” (Semantic) should be filled with a semantic description (for example word “акт” may have two meanings “act” and “document”)

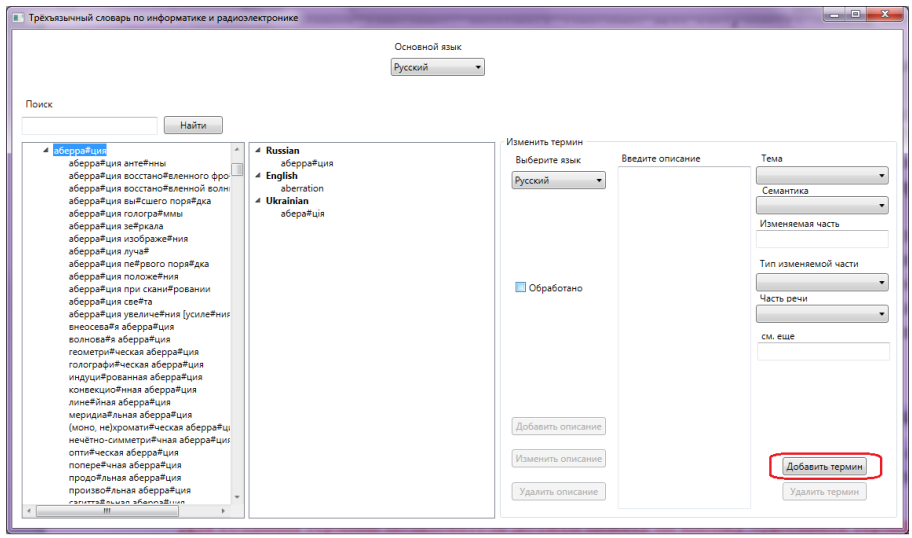

Figure 4: Main window of the dictionary

The field “Изменяемая часть” (Changeable part) should be filled with either the whole word, if it is needed to show accent, or phoneme changes (for example for word “вісь” (axis) this field will contain word “осі” (axes)), or ending (possible some other changeable part) of a word (for example for word “дейтрон” (deuteron) this field will contain “-на” in Ukrainian, for the word “нелінійність” (nonlinearity) it will be filled with “-ності” in Ukrainian).

The field “Тип изменяемой части” (Type of changeable part) should be filled with a type of changeable part (for example genitive case — “род.”, plural — “мн.”, “pl.”).

The field “Часть речи” (Part of speech) should be filled with a part of speech if needed.

The field “см. ещё” (see more) may be filled with reference to another word.

If it is necessary to change an existing term, then one of buttons on the left should be pressed after editing is finished. The buttons will be enabled depending on the existence of translated equivalents before editing.

The button “Добавить описание” (Add description) adds a translated equivalent in the selected language for the selected term. The button is enabled if a translated equivalent is absent in the selected language.

The button “Изменить описание” (Changed description) changes a translated equivalent in the selected language for the selected term. The button is enabled if a translated equivalent exists in the selected language.

The button “Удалить описание” (Remove description) removes a translated equivalent in the selected language for the selected term. The button is enabled if a translated equivalent exists in the selected language.

To create a term, users have to press the button “Добавить термин” (Add term), which is outlined in red in Fig. 4. A window for adding terms will be opened after this. It is necessary to enter translated equivalents for the selected languages.

Fig. 5 represents the process of adding translated equivalents in English. Russian and Ukrainian already have translated equivalents added.

A parent term may be selected from the dropdown menu above. The term “аберрация” (aberration) is selected for the example in Fig. 6.

After all the necessary data is entered, it is necessary to press the button “Готово” (Ready) and then the new term will be added to the database.

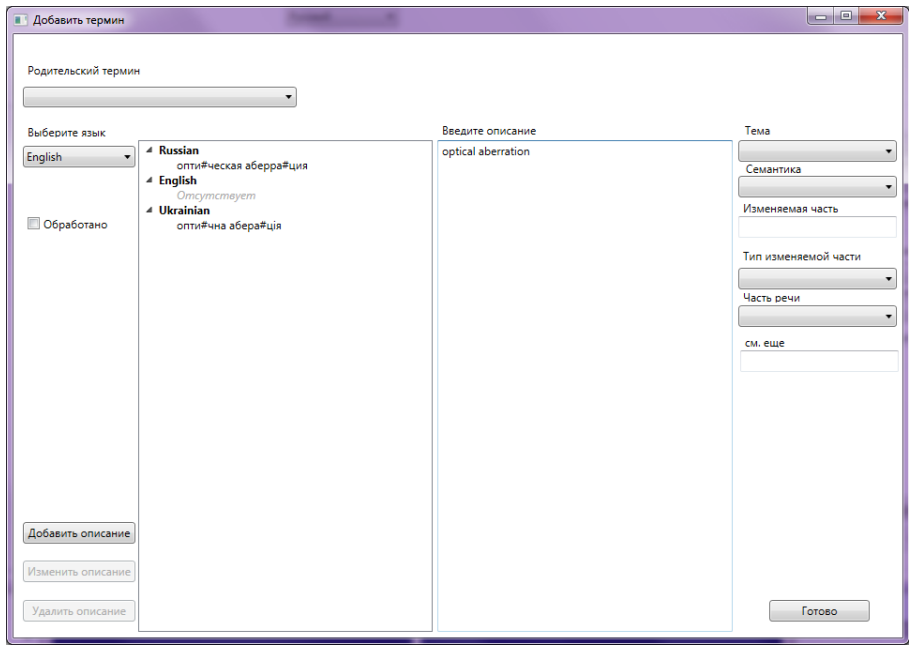

Figure 5: Adding a translated equivalent for a new term

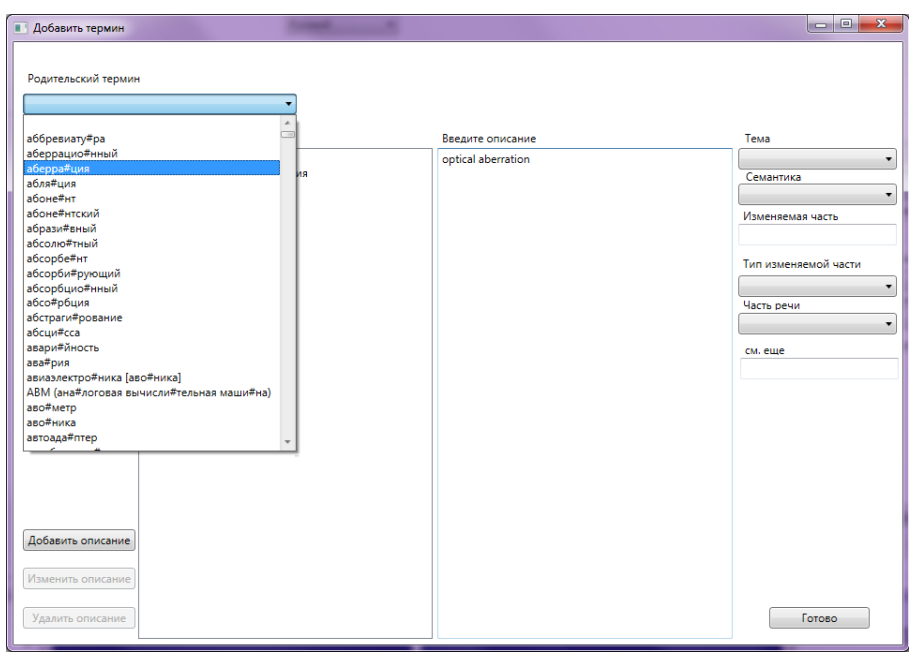

Figure 6: Selection of a parent term

5 Conclusion

The information above describes how the software system of the English-Russian-Ukrainian terminological dictionary was created. Some problems which have appeared in earlier dictionaries

were successfully avoided. (Широков, 2005, 2011). The automation of data input correction using regular expressions was achieved. Furthermore, the application of mathematical apparatus from the theory of lexicographical systems and the algebra of finite predicates (Широков, 2005; Вечирская, 2011) allowed redundancy to be avoided in solutions. A conception of semantic states of language units was used to formalize language information. (Широков, 2005). It was necessary to describe the problem formally using elements of the algebra of finite predicates, such as linear logical transformations, as well as using methods of building relational networks (Четвериков, 2001; Вечірська, 2009; Бондаренко , 2011) to build an adequate data model. Obviously, the approach used to create a trilingual dictionary may also be used for dictionaries with a greater number of languages.

The article describes in detail the structure of the software system of a trilingual dictionary. It presents the corresponding database scheme, describes the classes of entities and view model classes. The article also describes the user manual for the software system.

The developed software system enables the editing, filling, and creation of new thematic electronic dictionaries. The dictionary is easy to use for administrator and end-user alike. The indubitable advantage of the system is that each of the offered languages are equal from beginning, and the user chooses the major language by himself or herself in each case.

References

Бондаренко, М. Ф. & Шабанов-Кушнаренко, Ю. П. (2006). *Теория интеллекта: Учеб.* Харьков: Изд-во СМИТ.

Бондаренко, М. Ф. & Шабанов-Кушнаренко, Ю. П. (2012). *Мозгоподобные структуры: Справочное пособие* (Vol. 1). Київ: Наукова думка.

Бондаренко, М. Ф., Коноплянко, З. Д., & Четвериков, Г. Г. (2011). Концепції уніфікації інформаційно-інтелектуальних технологій в системах мовлення. *Бионика интеллекта: Науч.-техн. журнал*, (3(77)), 150–156.

Вечирская, И. Д. (2011). Разработка трехязычного терминологического словаря на основе алгебры конечных предикатов. *Бионика интеллекта: Науч.-техн. журнал*, (2(76)), 109–113.

Вечірська, І. Д. (2009). Дослідження розмірності предметного простору в задачах моделювання об'єктів у вигляді реляційних мереж. *Бионика интеллекта: Науч.-техн. журнал*, (2(71)), 31–35.

Остапова, И. В. (2007). Лексикографическая структура этимологических словарей и их представление в цифровой среде. *Прикладная лингвистика и лингвистические технологии: Сборник научных трудов*, 236–245.

Рабулець, О. Г., Широков, В. А., & Якименко, К. М. (2004). *Дієслово в лексикографічній системі.* Київ: Довіра.

Четвериков, Г. Г. (2001). Формалізація принципів побудови універсальних k-значних структур мовних систем штучного інтелекту. *Доповіді НАН України*, (1(41)), 76–79.

Широков, В. А. (2005). *Елементи лексикографії.* Київ: Видавництво "Довіра".

Широков, В. А. (2011). *Комп'ютерна лексикографія.* Київ: науково виробниче підприємство «Видавництво «Наукова думка» НАН України».

References (Transliteration)

Bondarenko, M. F. & Shabanov-Kushnarenko, I. P. (2006). *Teoriia intellekta: Ucheb.* Khar'kov: Izd-vo SMIT.

Bondarenko, M. F. & Shabanov-Kushnarenko, IṖ. (2011). *Mozgopodobnye struktury: Spravochnoe posobie* (Vol. 1). Kyïv: Naukova dumka.

Bondarenko, M. F., Konoplianko, Z. D., & Chetverykov, H. H. (2011). Kontseptsiï unifikatsiï informatsiĭno-intelektual'nykh tekhnolohiĭ v systemakh movlennia. *Bionika intellekta: Nauch.-tekhn. Zhurnal*, (3(77)), 150–156.

Chetverykov, H. H. (2001). Formalizatsiia pryntsypiv pobudovy universal'nykh k-znachnykh struktur movnykh system shtuchnoho intelektu. *Dopovidi NAN Ukraïny*, (1(41)), 76–79.

Ostapova, I. V. (2007). Leksikograficheskaia struktura ėtimologicheskikh slovareĭ i ikh predstavlenie v tsifrovoĭ srede. *Prikladnaia lingvistika i lingvisticheskie tekhnologii : Sbornik nauchnykh trudov*, 236–245.

Rabulets', O. H., Shyrokov, V. A., & IAkymenko, K. M. (2004). *Diieslovo v leksykohrafichniĭ systemi*. Kyïv: Dovira.

Shyrokov, V. A. (2005). *Elementy leksykohrafiï*. Kyïv: Vydavnytstvo "Dovira".

Shyrokov, V. A. (2011). *Komp'iuterna leksykohrafiia*. Kyïv: naukovo vyrobnyche pidpriiemstvo «Vydavnytstvo «Naukova dumka» NAN Ukraïny».

Vechirs'ka, I. D. (2009). Doslidzhennia rozmirnosti predmetnoho prostoru v zadachakh modeliuvannia ob'iektiv u vyhliadi reliatsiĭnykh merezh. *Bionika intellekta: Nauch.-tekhn. Zhurnal*, (2(71)), 31–35.

Vechirskaia, I. D. (2011). Razrabotka trekhiazychnogo terminologicheskogo slovaria na osnove algebry konechnykh predikatov. *Bionika intellekta: Nauch.-tekhn. Zhurnal*, (2(76)), 109–113.

Acknowledgment

This work was supported by a core funding for statutory activities from the Ministry of Education and Science of Ukraine.

The authors declare that they have no competing interests.

The authors' contribution was as follows: concept of the study Gryrorii Chetverikov; data analyses: Yrina Vechirska, Oleksii Puzik; the writing: Oleksii Puzik, Yrina Vechirska, Gryrorii Chetverikov

Publisher: Institute of Slavic Studies, Polish Academy of Sciences, University of Silesia & The Slavic Foundation