COGNITIVE STUDIES | ÉTUDES COGNITIVES, 14: 1–11
SOW Publishing House, Warsaw 2014

DOI: 10.11649/cs.2014.001

VOLODYMYR SHYROKOV [1,A], **IRYNA OSTAPOVA**[1,B]
& **KOSTYANTYN YAKYMENKO**[1]

[1]Ukrainian Lingua-Information Fund, National Academy of Sciences of Ukraine, Kyiv, Ukraine
[A]vshirokov48@mail.ru ; [B]irinaostapova@gmail.com

INDEXING THE ETYMOLOGICAL LEXICOGRAPHIC SYSTEMS

Abstract

The main problems and directions for the development of the etymological lexicographic systems in the digital environment are studied. The formal conceptual model of the lexicographic system for fundamental academic Etymological Dictionary of the Ukrainian Language (EDUL) is developed. The lexicographic structure of the EDUL individual elements are developed and described. The EDUL metalanguage was studied and described. The formal model and technology of the EDUL parsing are worked out. That made it possible to convert automatically the EDUL text into the lexicographic database, which corresponds to the conceptual model of the lexicographic system. The conceptual foundations of instrumental tool to form the etymological dictionaries are developed to create the Virtual Lexicographic Laboratory «Etymological Dictionary of the Ukrainian Language», which was implemented with a modern approach to the real lexicographic array of the EDUL. That allowed to form the database of the EDUL multilingual index (about 250 languages) in the automatic mode. This index is a basis of the seventh (final) volume of the EDUL. The possibility of applying the developed models to other etymological dictionaries are studied. The conceptual foundations for integration of the etymological lexicographic systems are discussed.
Keywords: etymological dictionary, index of the lexicographic system, lexicographic databases, virtual lexicographic laboratory.

The index is one of the effective tools to work with the dictionary. An example of the dictionary index are the head entries, graphically isolated from the text data array and arranged in alphabetical order. But this is just one example of the index. With increasing complexity of the entry structure the creation of the effective indexing system becomes a methodological and technological problem that is relevant both for printed dictionaries as well as for digital. This problem is complicated by the uncertainty of the quantity determination and, so to speak, the quality of the indexable lingual objects that meet the lexicographical effects

developing the vocabulary system (Shyrokov, 2011). General formal definition of lexicographical system index is given in the works (Shyrokov, 2011).

Historically, the indexes were specifically designed for phraseological and etymological dictionaries. In the first case, the index is obligatory due to the structural complexity of list units — with an index access to a dictionary entry that is organized for each component of the phraseological unit. For etymological dictionaries index is a tool to establish a genetic (etymological) connection as to the parameter of linguistic origin.

Ideally, every lexicographical parameter of the dictionary should be indexed. In order to implement this requirement, the index entries must be identified with structure formation parameters (invariants) of the lexicographic system, juxtaposed to the dictionary. Obviously, the developed indexation technology can be built only for digital implementations of lexicographic systems.

Similarly to all fundamental lexicographical works, etymological dictionaries are gradually converted into digital format. The "Etymological Dictionary of Russian language" by M. Fasmer is most accessible in the Internet, the first its digital version was published on CD-ROM in 2004. However, all of the digital representations of such lexicographical works are nothing more than species of machine-readable texts, the possibilities for the scalability and the more dynamic indexing are absent.

This paper deals with the Etymological Dictionary of the Ukrainian language (ESYM, 1982–2011) (EDUL, in the printed version — 6 volumes) indexing method and technology, which are offered as universal for any etymological lexicographical system. A formal model of the entry is developed so that the structural elements are divided into two classes: linguistic (descriptive) and structureforming, the latter are singled out by the formal process of linguistic. On the basis of this model is diagram of a computer lexicographical database (LDB) and the dictionary text parsing technology are constructed that allowed to convert the text into LDB automatically. To support a digital version of EDUL a tool set is developed that includes dictionary indexing subsystem as to user generated register.

As a basic structural element of the entry the *etymological class* (indicated as ECL) is defined which is a unit of linear text, that describes certain genetic connections of the head word. For EDUL etymological classes are subdivided as follows: *head words class* (HEAD), a *class of derivatives* (DERIVAT), a *class of Slavic correspondences* (SLAVIA), a *language class* (LANG), a *bibliographic class* (BIBL) and *classes of links* (REF COMP depending on the type of the reference). Each of these classes has unique formal features, due to which it is identified in a linear text entry, and has the individual content and structure. Classes of references and bibliography are considered to be utility. In turn, the classes can be distributed into sub-classes. Minimum obligatory structure of the entry consists of two structural elements: the head word class and one language class. All the other elements of the entry structure are optional (Ostapova, 2009).

Here is an example of the EDUL entry, adequately demonstrating the etymological structure of the text description (the printing marking copies the marking of the original text, class and sub-class separators are singled out in color).

(1)
[**баяти**] «розповідати; ворожити», [*байкати*] «розповідати (писати) байки; балакати», [*бай*] «казкар» Пі, *байка*, *байкар*, *байкарство*, *байківниця* «збірник казок і легенд» Ж, [*байко*] «казкар, брехун» Я, *баєчник* «тс.», [*байла*] «ворожбит» Ж, [*байло*] «тс.» Ж, [*байчар*] «балакун, базіка, пліткар», [*баюн*] «казкар, брехун» Я, [*баян*] «співець» Я, [*байкий*] «говіркий» Ж, *баєчний*, *байбай* (приспів колискової пісні), [*баюбаю*], [*баюлі*] «тс.» Я, [*забаяч*] «заклинач, ворожбит, знахар» Ж; — р. *баять* «говорити», бр. [*баиць*] «говорити, базікати, плести нісенітницю», др. *баяти* «розповідати байки; чарувати», п. *bajać* «розповідати казки; базікати, вигадувати», ч. bájiti «розповідати казки», ст. *báti* «тс.», ст. *baju* «тс.», слц. *báj* «міф», *bájka* «байка», вл. *bać* «розповідати казки, говорити нісенітницю», заст. *bać* «розповідати казки, говорити нісенітницю», нл. *bajaś* «базікати, розповідати казки», болг. *бая* «замовляю, ворожу», м. *бае*, схв. *бајати* «тс.», слн. *bájati* «базікати, заклинати, ворожити, пророкувати», [*bajúlje*] «якісь дитячі пісні», стсл. *баяти* «розповідати байки, замовляти, заворожувати, заклинати»; — псл. bajati; — споріднене з дінд. bhánati «говорить», sabhā «збори», лат. fārī «говорити», гр. φημί (дор. φα¯μί) «говорю», вірм. ban «слово, мова», пн.-фриз. bālen «говорити», дангл. bōian «хвалитися»; з ie. **bhā-*; невірно пов'язується (напр., Sławski I 25; Шанский ЭСРЯ I 2, 65) з лит. *bóti* «звертати увагу», яке походить від п. *dbać* чи бр. *дбаць* і не має приписуваного йому іноді при перекладі значення «питати».— Фасмер—Трубачев I 140; ЭССЯ 1, 138—139; Sł. prasł. I 182; Sadn.—Aitz. VWb. I 115; Bern. I 39; Būga I 584; Fraenkel 53; Mayrhofer II 469—470; Holthausen PBrB 48, 460; Pokorny 105—106.— Пор. **базікати**, **балакати**, **баніт**, **бари**, **барити**, **басні**.

The structure formational blocks of this article are as follows (text blocks are served in angle brackets and, if possible, in the abbreviated form, double index distinguishes subclasses):

ECL_1(**баяти**) ≡ $HEAD$(**баяти**) ≡ <[баяти] «розповідати; ворожити»>;
ECL_2(**баяти**) ≡ $DERIVAT$(**баяти**) ≡ <[байкати] «розповідати (писати) байки; балакати», ..., [забаяч] «заклинач, ворожбит, знахар» Ж>;
ECL_3(**баяти**) ≡ $SLAVIA$(**баяти**) ≡ <р. баять «говорити», ..., стсл. баяти «розповідати байки, замовляти, заворожувати, заклинати»>;
ECL_4(**баяти**) ≡ $LANG_1$(**баяти**) ≡ <псл. bajati>;
ECL_{5-1}(**баяти**) ≡ $LANG_{2-1}$(**баяти**) ≡ <споріднене з дінд. bhánati «говорить», sabhā «збори», лат. fārī «говорити», гр. φημί (дор. φα¯μί) «говорю», вірм. ban «слово, мова», пн.-фриз. bālen «говорити», дангл. bōian «хвалитися»>;
ECL_{5-2}(**баяти**) ≡ $LANG_{2-2}$(**баяти**) ≡ <з ie. *bhā->;
ECL_{5-3}(**баяти**) ≡ $LANG_{2-3}$(**баяти**) ≡ <невірно пов'язується (напр., Sławski I 25; Шанский ЭСРЯ I 2, 65) з лит. bóti «звертати увагу», яке походить від п. dbać чи бр. дбаць і не має приписуваного йому іноді при перекладі значення «питати»>;

ECL_{6-1} (**баяти**) ≡ $BIBL_{1-1}$ (**баяти**) ≡ <Фасмер—Трубачев I 140>;
...
ECL_{6-10} (**баяти**) ≡ $BIBL_{1-10}$ (**баяти**) ≡ <Pokorny 105—106>; ECL_{7-1} ≡ $COMP_{1-1}$ ≡ <**базікати**>
...
$\mathrm{ECL}_{7-6} \equiv \mathrm{COMP}_{1-6} \equiv$ <**басні**>

Thus, the entry V(**баяти**) in its composition has seven etymological classes: head words class, a class of derivatives, a class of Slavic correspondences, two language classes, the bibliographical class and the class of references. The second language class incorporates 3 subclasses, the bibliographical — 10 subclasses, and the references class — 6. The linear sequence of text blocks in the dictionary entry is as follows:

$$V(\textbf{баяти}) \equiv HEAD\{,\ \ \}DERIVAT\{;—\}SLAVIA\{;—\}$$
$$LANG_1\{;—\}LANG_2\{.—\}BIBL\{.—\}COMP.$$

In the curly brackets the sequences of characters that act as delimiters of the structural elements are presented. For the subclasses separation the only one type of separator is used {; }.

To each etymological class except for utility, the etymon structure is assigned, which is a set of word parameters (etymons) with a genetic link established. Eight parameters were defined for EDUL: the linguistic affiliation marker (indicated as PAR_L), the note to the linguistic marker affiliation (PAR_{RL}), a character representation of the etymon in a certain alphabet (PAR_A), an affiliation to the dialect vocabulary (PAR_{DER}), a sign of homonymy (PAR_{OM}), a gloss (PAR_S), a remark (PAR_{REM}), a bibliography (PAR_{BIBL}). All parameters are singled out of the text as to certain formal characteristics. The linguistic affiliation (PAR_L) and character representation (PAR_A). parameters are obligatory. These two parameters provide the uniqueness of each etymon in the text of the entry. For each article at least one etymon — structure (for the head word class) is built.

Let us give the examples of etymon structures of language classes (from the optional parameters only the interpretation is included):

(2)
$LANG_1$(**баяти**) ≡ <псл. bajati>;

№	PAR_L	PAR_A	PAR_S
1	псл.	bajati	

(3)
$LANG_{2-1}$(**баяти**) ≡ <споріднене з дінд. bhánati «говорить», sabhā «збори», лат. fārī «говорити», гр. φημί (дор. φα¯μί) «говорю», вірм. ban «слово, мова», пн.-фриз. bālen «говорити», дангл. bōian «хвалитися»>

№	PAR_L	PAR_A	PAR_S
1	дінд.	bhánati	«говорить»
2	дінд.	sabhā	«збори»
3	лат.	fārī	«говорити»
4	гр.	φημί	«говорю»
5	вірм.	ban	«слово, мова»
6	пн.-фриз.	bālen	«говорити»
7	дангл.	bōian	«хвалитися»

In the printed dictionary text is organized in such a way that the head word has a marking-out and is the first word of the series, which ensures its structural significance by the possibilities of the printed text. In the proposed model structure-forming are obligatory parameters of the etymon structure: the entrance to the dictionary is possible in any language and on any word in the alphabet of the language.

The generalized scheme of the multilingual indexing of the etymological dictionary can be represented as follows:

$$
\begin{array}{c}
Lang\_Reg(S) \\
\downarrow \\
Lang\_Reg\_Ind(S) \\
\swarrow \quad \swarrow \quad \downarrow \quad \downarrow \quad \searrow \quad \searrow \\
Lang\_reg_1 \;\; Lang\_reg_2 \;\; \ldots \;\; Lang\_reg_i \;\; \ldots \;\; Lang\_reg_N
\end{array}
$$

$Lang\_Reg(S)$ — the register of all languages presented in the etymological dictionary of descriptions (also refer to the language of dialects).

$Lang\_Reg\_Ind(S)$ — the language register, chosen for the dictionary indexing. It may coincide with $Lang\_Reg\_Ind(S)$, or be its language subset:
$Lang\_Reg\_Ind(S) \subseteq Lang\_Reg(S)$.

$Lang\_reg_i$ is a subset of the index register, it satisfies the requirement $Lang\_reg_i \cap Lang\_reg_j \equiv \emptyset$, in the case if $j \neq i$, that is the list of languages for each of these subregisters is unique. As a rule, the subregisters include the related languages. Each subregister, in turn, can be subdivided into more specific registers ($Lang\_reg_{i-j}$). The number of the hierarchy levels as well as the register language structure depends on the tasks of the dictionary authors.

Every indexing element ind_el_k ($k = 1, 2, ...K$; K — the element quantity in the corresponding index) is of the following structure:
$ind\_elk \equiv \{e_k, lang(e_k), loc(e_k)\}$, where
ek — etymon,
$lang(e_k)$ — language marker,
$loc(e_k)$ — etymon localization in the dictionary text.

For the EDUL printed index the following form of etymon localization is proposed: the head word of the corresponding dictionary entry, the volume and page number, for example bajati **баяти** I, 356 или bajati *псл.* **баяти** I, 356:

(4)
ind_el ≡ <bajati псл. **баяти** I, 356>, *e* ≡ <bajati>,
lang ≡ псл.
loc ≡ <**баяти** I, 356>;

If the article is placed on multiple pages, the first and last are specified. The volume and the page numbers are tribute to the tradition and the binding to the printed version of the dictionary, the individual volumes of which appeared with a significant time interval. Before converting to a database the text of each article was tied to its printed original, the volume and the page numbers were recorded in the appropriate field of the database. In the EDUL digital version the etymon can be localized with a sequence word number in the etymological class line.

Since each language presented in the dictionary will have a separate register, the following format of printed index (register names in Ukrainian) is suggested:

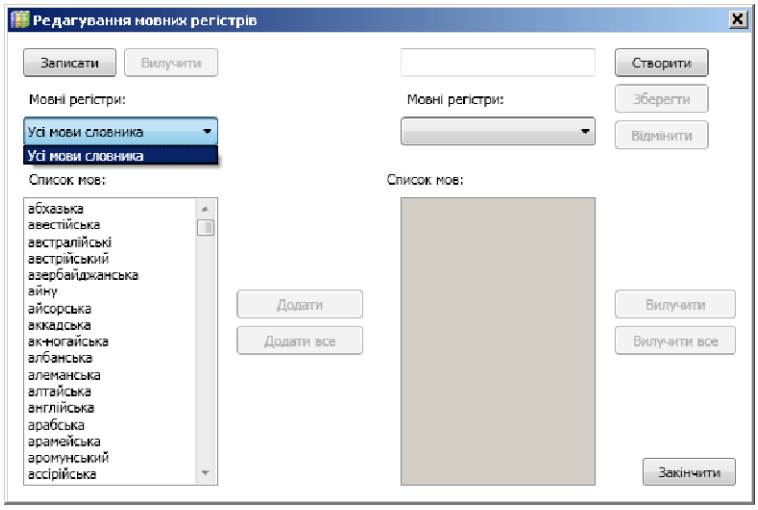
АБХАЗЬКА
...
АВЕСТІЙСЬКА
...
varəsa **во́лос** I, 420–421
varəz **бульва́р** I, 293
varəz- **верста́т** I, 357
АВСТРАЛІЙСЬКІ...
ГРЕЦЬКА
...
κύμβαλον **кімва́л** II 446
...
...**ПРАСЛОВ'ЯНСЬКА**...
...
bajati [**баяти**] I, 356
...
...**ШУМЕРСЬКА**
...
...**ЯКУТСЬКА**
...
ЯПОНСЬКА

The task of the dictionary language index building is entrusted to the software tool system. In developing the tool set of the linguistic indexing the minimum user limit was suggested. The developed tool allows to create a language register with any language set in the dictionary, that is, to combine in one register randomly selected languages, not taking into consideration their kinship. Any number of

such registers can be created, which differ only in name, assigned by the user. You can specify a list of structural elements (etymological classes) that will be indexed.

Fig. 1 shows an interface window, in which registers are formed. The figure fixes the initial state in which neither of the registers are formed.

Figure 1 Linguistic register formation window



The left panel is used to select the already created registers as constant patterns, the right — for the formation of new registers and their editing. The register «Усі мови словника» ("All the dictionary languages") is systemic and contains all the languages recorded in the dictionary. Any formed register, including systemic, can be used for indexing the dictionary.

The user may use several strategies for the new registers formation, for example:

1) to select in the left panel the register «Усі мови словника», to choose from the list of languages a required position and rewrite them in the right pane named list;

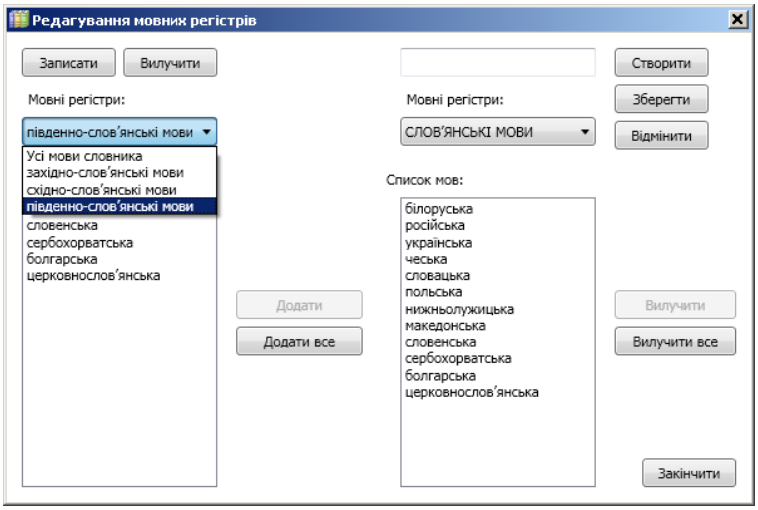
2) to select in the left panel a register with the appropriate language structure, to put it into the right panel, and then edit out not necessary items;

3) to combine several register-templates into one register; in order to do this register series are selected in the left panel and a list of languages is put into the right panel;

4) to select in the right panel the already existing register and to edit it (without changing the name), adding the missing items from the list of left panel.

The register «СЛОВ'ЯНСЬКІ МОВИ» ("Slavic languges") in Fig. 2. is formed in the following manner: «східно-слов'янські мови» ("East Slavic languages"), «західно-слов'янські мови» ("West-Slavic languages"), «південно-слов'янські мови» ("South Slavic languages") have been consistently formed on the basis of the system register, and then the language lists were merged.

Figure 2 The formation of the register «СЛОВ'ЯНСЬКІ МОВИ»



The ability to use the tools developed for the other etymological lexicographical systems primarily exists due to the possibility of a formal model usage. In order to fulfill this task the texts of the dictionary «Этимологический словарь славянских языков: Праславянский лексический фонд» (ESSYa, 1974) are examined. This lexicographical work was chosen for the following reasons. The dictionary appears in the form of small issues in printed format for a long time period, contains the linguistic information of ongoing interest to researchers. The technology developed for the Etymological Dictionary of the Ukrainian language, makes possible the indexing of the dictionary in the process of its creation, without waiting for completion of the work.

The dictionary entry with a head word **bajьka** describes adequately the structure of etymological descriptions:

(5)
***bajtьka**: макед. *бајка* ж. р. 'сказка' (И-С), сербохорв. *bajka* ж. р. 'басня, сказка, fabula' (с XVII в., RJA I, 151), 'сказка, выдумка', 'ворожба, заговоры' (РСА I, 250), словен. *bajka* ж. р. 'сказка' (Plet. I, 10), чеш. *bajka* ж. р. 'басня', 'сказка, выдумка', слвц. *bajka* 'басня, выдумка' (SSJ I, 66), диал. *bajka* ж. р. 'неправда' (Buffa. Dlhá Lúka 130), в.-луж. *bajka* ж. р. 'сказка, басня' (Pfuhl 5), н.-луж. *bajka* 'басня, сказка' (Muka Sł. I, 11), польск. *bajka* ж. р. 'сказка, басня' (Dorosz. I, 301), русск. диал. *байка* ж. р. 'говор', 'сказка' (Опыт 5), ст.-укр. *байка* ж. р. 'выдумка' (Зизаний, Тимченко I, 50), укр. *байка* ж. р. 'басня', 'безделица, шутка, пустяки' (Гринченко I, 21), блр. *байка* ж. р. 'басня, небылица' (Блр.-русск. 115).

Широко распространенное в слав, языках, но, возм., возникшее как параллельное и относительно новое образование с суфф. *-ъka* от основы гл. **bajati* (см.).

Following the logic of the model, four etymological classes are selected:

$EL_1 \equiv$ <***bajьka***>
$ECL_2 \equiv$ <макед. *бајка* ж. р. ‘сказка’ (И-С), ..., блр. *байка* ж. р. ‘басня, небылица’ (Блр.-русск. 115)>
$ECL_3 \equiv$ <Широко ... от основы гл. **bajati* (см.)>
$ECL_4 \equiv$ <*bajati (см.)>

The following distribution by the type of etymological classes can be suggested: head word class ($ECL_1 \equiv HEAD$), Slavic class ($ECL_2 \equiv SLAVIA$), language class ($ECL_3 \equiv LANG$), references class ($ECL_4 \equiv REF$).

As a linear sequence of text blocks this dictionary entry can be represented in the following form:

$$HEAD\{:\}SLAVIA\{passagemarker\}LANG\{(s.)\}REF.$$

The last class is “embedded” in the language class, it can be redefined as follows <s. *bajati>, as the format of the EDUL reference class. There can be presented several classes of the kind, for example, in an article with the header word **bajidlo**:

(6)
***bajidlo**: словен. *bajȗlo* ср. р. ‘колдовство. заговор’ (Plet. I, 10).
Производное с суфф. *-dlo* от основи гл. **бајiti* (см.). Ср. еще **бајadlo* (см.) и ряд слов с основой **ba(d)l-* (см. ниже).

The distribution of the text into the etymological classes in this case is the following:

$HEAD \equiv$ <*bajidlo>
$SLAVIA \equiv$ <словен. *bajȗlo* ср. р. ‘колдовство. заговор’ (Plet. I, 10)>
$LANG \equiv$ <Производное с суфф. *-dlo* от основи гл. **бајiti* (см.). Ср. еще **бајadlo* (см.) и ряд слов с основой **ba(d)l-* (см. ниже)>
$REF_1 \equiv$ <**бајiti* (см.)> $\equiv$ <см. **бајiti*>
$REF_2 \equiv$ <**бајadlo* (см.)> $\equiv$ <см. **бајadlo*>
$REF_3 \equiv$ <**ba(d)l-* (см. ниже)> $\equiv$ <см. **ba(d)l-*>

Let us construct the etymon-structure for the language class of the entry ***bajьka**:

(7)

№	PAR_L	PAR_A	PAR_S
1	макед.	*бајка*	‘сказка’
2	сербохорв.	*бајка*	‘басня, сказка, fabula’
3	словен.	*bajka*	‘сказка’
4	чеш.	*bajka*	‘басня’
5	слвц.	*bajka*	‘басня, выдумка’
6	в.-луж.	*bajka*	‘сказка, басня’

7	н.-луж.	*bajka*	‘басня, сказка’
8	польск.	*bajka*	‘сказка, басня’
9	русск.	*байка*	‘говор’
10	ст.-укр.	*байка*	‘выдумка’
11	укр.	*байка*	‘басня’
12	блр.	*байка*	‘басня, небылица’

While isolating the parameters, we were guided by an algorithm designed for EDUL.

Let us construct the etymon-structure for the article registry-word with a head word ***alovъjь** $HEAD \equiv <$***alovъjь**, ***alovaja** / ***jalovъjь**, ***jalovaja**$>$:

(7)

№	PAR_L	PAR_A
1	праслав.	***alovъjь**
2	праслав.	***alovъjь**
3	праслав.	***alovъjь**
4	праслав.	***alovъjь**

In the etymological dictionaries the linguistic marker of register units is assigned on default, in this case — the language is Proto-Slavic (similar to EDUL is Ukrainian).

On the basis of the above we can make a conclusion that for such lexicographical work (“Etymological Dictionary of the Slavic languages ...”) the model developed for EDUL is quite relevant.

On the basis of the software tools the virtual lexicographical laboratory is built — ВЛЛ «ЕСУМ» (Ostapova & Shyrokov, 2010), which provides the professional lexicographers interaction in the online environment, as well as the ability to use the complete functional of the system in the on-line mode.

The use of the technology for other etymological lexicographical systems (primarily of the Slavic languages) offers the perspective of forming the lexicographical etymological environment. In this approach, the vocabulary systems are regarded as relatively independent entities, and the completeness of the language lexicographical etymological description is created due to their integration. As integrating structure the language code may be used.

References

ESSYa. (1974). *Ėtimologicheskiĭ slovar' slavianskikh iazykov* (Ed. 1). Moskva: Nauka.

ESYM. (1982–2011). *Etymolohichnyĭ slovnyk ukraïnskoï movy. V 7 t.* (Vol. 1–6). Kyïv: Naukova dumka.

Ostapova, I. V., & Shyrokov, V. A. (2010). Virtual'naia leksikograficheskaia laboratoriia dlia tolkovykh slovareĭ. In *Komp'iuternaia lingvistika i intellektual'nye tekhnologii. Po materialam ezhegodnoĭ Mezhdunarodnoĭ konferentsii «Dialog» (Bekasovo, 26–30 maia 2010 g.)* (pp. 363–367). Vyp. 9(16). Moskva: RGGU.

Ostapova, I. V. (2009). Leksikograficheskaia struktura etimologicheskogo slovaria i ego predstavlenie v tsifrovoĭ srede. In *Komp'iuternaia lingvistika i intellektual'nye tekhnologii. Po materialam ezhegodnoĭ Mezhdunar. konf. «Dialog 2009» (Bekasovo, 27–31 maia 2009 g.)* (pp. 359–365). Vyp. 8(15). Moskva: RGGU.

Shyrokov, V. A. (2011). *Komp'iuterna leksykografiia.* Kyïv.

Publisher: Institute of Slavic Studies PAS & University of Silesia in Katowice