Formal Model of Explanatory Trilingual Terminology Dictionary

Alona Dorozhynska

Ukrainian Lingua-Information Fund of NAS of Ukraine,3, Holosiivskyi avenue, Kyiv, 03039, Ukraine

Abstract

The object of research is the academic trilingual (Ukrainian-Russian-English) "Dictionary of Ukrainian biological terminology" (SUBT) [1]. This dictionary is an authoritative terminographic work that embraces the normative general scientific and widely used and narrowly specialized terminology of biological sciences, recorded in encyclopedic, general language and special dictionaries, in scientific, popular science, educational and informative literature. The dictionary is published in small editions in traditional paper form; it is very popular among users - scientists, graduate students and students of biological specialties.

The digital version of the dictionary in pdf-format was used in the research. The research process included the construction of a formal model of the lexicographic system of SUBT, the analysis of which provides an opportunity to generalize it to other terminological dictionaries. The developed model of the lexicographic system of SUBT was used to create a representation of the text of the Dictionary by means of XML markup language, which was used to convert the pdf-file of SUBT into XML-file, which completely reproduces the formal structure of the lexicographic system.

The presence of an XML file allows you to build a database according to the structure of the dictionary article. The XML file is proposed to be used as an intermediary between the paper version of the dictionary and its implementation as an online lexicographic system. The possibility of building a formal model for dictionaries of this type is considered.

Keywords [1]

Computer lexicography, lexicographic system, parsing, XML, database, digital space, website.

1. Introduction

As you know, the dictionary consists of dictionary articles, is a certain set of them. If the dictionary is built according to the canons of lexicographic science, then it has a certain structure to which all its dictionary articles are subject. Usually the structure of dictionary articles is described in the preface to the dictionary. But between different dictionary entries of the same dictionary there are many connections, relations, reflections, which reflect the content of the subject area that is the object of lexicography, and these relations are usually implicit. However, they are very important for users and those who want to use the dictionary to conduct their own research and create new dictionaries. Partially mentioned structural vocabulary effects are described in the works [2, 4, 6]. General theory of dictionary structures - the theory of lexicographic systems was created by V. A. Shirokov in the 90s of last century; The most complete presentation of this theory and its applications is published in the seven-volume edition "Linguistic and Information Studies", which is available for free on the website of the Ukrainian Linguistic Portal at https://www.ulif.org.ua/publication. In our work we will follow the principles of this theory.

It is important to distinguish between the most formal model and the XML scheme (coding scheme). That is, it is necessary to consider the form and content of lexical information in the abstract, regardless of the requirements and restrictions imposed on its final presentation as a coded or printed object [7, 8]. This process is important, because dictionaries can be coded not only for the purposes of publication in

COLINS-2021: 5th International Conference on Computational Linguistics and Intelligent Systems, April 22–23, 2021, Kharkiv, Ukraine
EMAIL: alonochkatkachyk@gmail.com (A.Dorozhynska)
ORCID: 0000-0001-6554-6731 (A.Dorozhynska)

printed (book) or electronic form (website), but also to create computational lexicons. Therefore, it is very important to develop a model that can later be transformed into a variety of alternative formats [3].

In this paper, we outline the following stages:

1. Lexicographic system (L-system) development of dictionary structure

2. Marking the text of the Dictionary with XML tags according to the structure of its L-system (XML document)

Using the examples of dictionary articles of the selected dictionary, we will demonstrate how the XML schema can be applied to any dictionary article. Due to its generality, we believe that our model can serve as a basis for presenting, combining and extracting information not only from dictionaries of the same type, but also from a wide range of terminological dictionaries [5].

2. The steps

2.1. Lexicographic system development of Dictionary of Ukrainian biological terminology (SUBT)

We introduce the notation: $A \rightarrow B$ it will mean, that $A \supseteq B$. Then, following the theory of lexicographic systems [2], the structure of the L-system of SUBT is presented in the form:

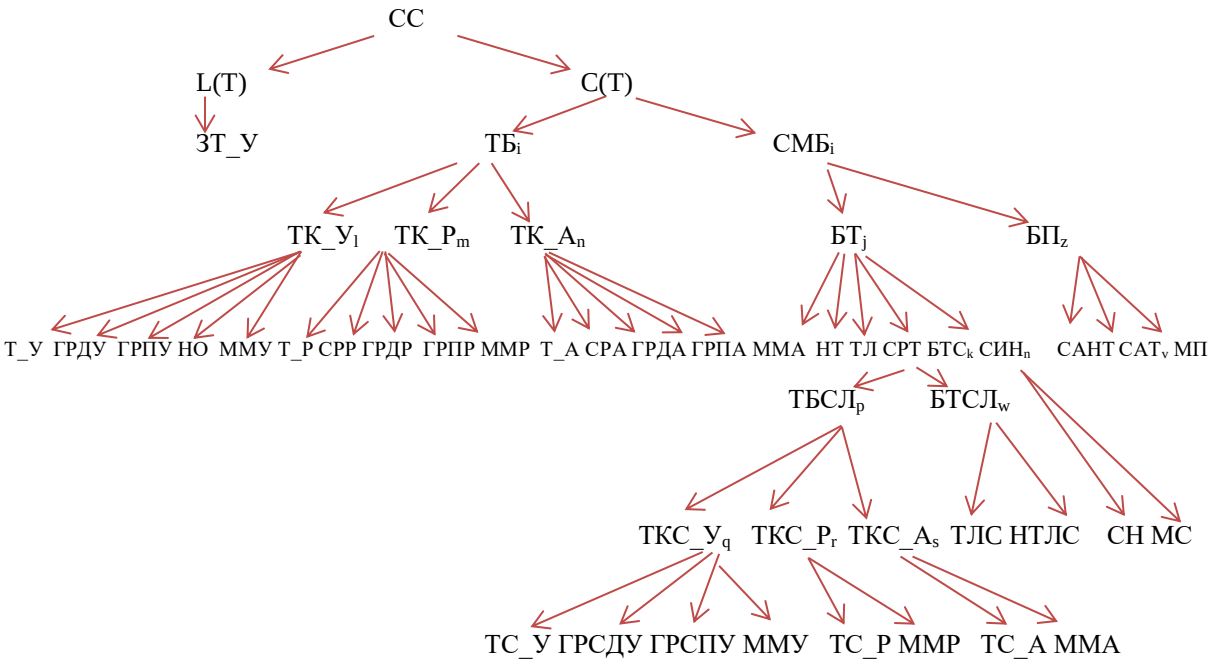

Figure 1. General scheme of the dictionary article of the SUBT

In scheme 1, the structural elements mean the following:

СС — dictionary article text
ЗТ_У — the title term is Ukrainian
ТБ$_i$ — terminological block
ТК_У$_l$ — terminol. complex ukr.

Т_У — the term is Ukrainian
ГРДУ— grammatical remark to the term
ГРПУ — grammatical remark after the term
НО — homonym number
ММУ — language marker (ukr)
ТК_Р$_m$ — terminol. complex of rus.
Т_Р — the term is Russian
ГРДР — grammatical remark to the term
ГРПР— grammatical remark after the term
ММР — language marker (rus)
СРР — semantic trailer
ТК_А$_n$ — terminol. English complex
Т_А — English term
ГРДА — grammatical remark to the term
ГРПА — grammatical remark after the term
ММА — language marker
СРА — semantic trailer
СМБ$_i$ — semantic block
БТ$_j$ — interpretation block
НТ — interpretation number
ТЛ — interpretation
СРТ — semantic trailer to tl.
БТС$_k$ — block of terminological phrases
ТБСЛ$_p$ — terminological block of phrases

ТСК_У$_q$ — terminological complex of phrases ukr.
ТС_У — terminological phrase in the Ukrainian language
ГРСДУ — grammatical remark to the phrase
ГРСПУ — grammatical remark after the phrase
ММУ — language marker (ukr)
ТСК_Р$_r$ — terminological complex of words. Rus.
ТС_Р — terminological phrase in Russian
ММР — language marker (in Russian)
ТСК_А$_s$ — terminological complex of phrases in English.
ТС_А — terminological phrase in English
ММА — language market
БТСЛ$_p$ — block of interpretations of phrases
ТЛС — interpretation
НТС — phrase interpretation number
СИН$_t$ — synonymous block
СН — synonym
МС — synonym marker (Син.)
БП$_z$ — link block
САНТ — sender
САТ$_v$ — recipient (can be several)
МП — link token (*див*.)

Lexical information in dictionary articles can be represented in the form of a tree structure, which largely reflects the natural hierarchical organization of entries in printed dictionaries. Consider the examples in the notation of the scheme:

Example 1.

ву#са, -ів, *мн.*, *одн.* **вус**, -а *(рос.* усы#, *ед.* ус., *англ.* whisker (*у тварин*), tendril (*у рослин*), moustache (*у людини*), antenna (*у комах*)) 1. Загальна назва розміщених біля рота чутливих волосин (вібрисів) у ссавців та щетинкоподібного пір'я у птахів, деякі дотикові утвори у безхребетних тварин; 2. Надземні виткі прикріплювальні пагони у рослин.

Marking	Representation in the SS
ТБ	**ву#са,** -ів, *мн.*, *одн.* **вус**, -а *(рос.* усы#, *ед.* ус*., англ.* whisker (*у тварин*), tendril (*у рослин*), moustache (*у людини*), antenna (*у комах*))
ЗТ	**ву#са**
ТК_У $_1$	**ву#са,** -ів, *мн.*

ММУ	*укр.*
Т_У	**ву#са**
ГРПУ	**-**ів, *мн.*,
ТК_$У_2$	*одн.* **вус**, -а
ММУ	*укр.*
Т_У	**вус**
ГРДУ	*одн.*
ГРПУ	-а
ТК_$Р_1$	*рос.* усы#
ММР	*рос.*
$Т^Р$	усы#
ТК_$Р_2$	*ед.* ус
ММР	*рос.*
Т_Р	ус
ГРДР	*ед.*
ТК_$А_1$	*англ.* whisker (*у тварин*)
ММА	*англ.*
$Т^А$	whisker
СРА	*у тварин*
ТК_$А_2$	tendril (*у рослин*)
ММА	*англ.*
$Т^А$	tendril
СРА	*у рослин*
ТК_$А_3$	moustache (*у людини*)
ММА	*англ.*
$Т^А$	moustache
СРА	*у людини*
ТК_$А_4$	antenna (*у комах*)
ММА	*англ.*
$Т^А$	antenna
СРА	*у комах*
СБ	1. Загальна назва розміщених біля рота чутливих волосин (вібрисів) у ссавців та щетинкоподібного пір'я у птахів, деякі дотикові утвори у безхребетних тварин; 2. Надземні виткі прикріплювальні пагони у рослин.
$БТ_1$	1. Загальна назва розміщених біля рота чутливих волосин (вібрисів) у ссавців та щетинкоподібного пір'я у птахів, деякі дотикові утвори у безхребетних тварин
НТ	1
ТЛ	Загальна назва розміщених біля рота чутливих волосин (вібрисів) у ссавців та щетинкоподібного пір'я у птахів, деякі дотикові утвори у безхребетних тварин
$БТ_2$	2. Надземні виткі прикріплювальні пагони у рослин.
НТ	2
ТЛ	Надземні виткі прикріплювальні пагони у рослин.

In the example, there is only one terminology block and one semantic block, respectively. It was found

that in dictionary articles there are as many terminological blocks as semantic ones and vice versa. There are two Ukrainian complexes in the terminological block. Presented as: **вус#са,** -ів, *мн.*; *одн.* **вус**, -а. In the semantic block there can be some interpretations in an example of them two are found: 1. Загальна назва розміщених біля рота чутливих волосин (вібрисів) у ссавців та щетинкоподібного пір'я у птахів, деякі дотикові утвори у безхребетних тварин; 2. Надземні витки прикріплювальні пагони у рослин.

Example 2.

новонаро#джений 1. *прикм. (рос*. новорождённый, *англ*. neonatus, neonate) який недавно або тільки що народився; 2. *ім.*, -ого *(рос*. новорождённый, *англ*. newborn, infant) людина, яка недавно народилася.

Marking	Representation in the SS
$ТБ_1$	**новонаро#джений** 1. *прикм. (рос*. новорождённый, *англ*. neonatus, neonate)
ЗТ	**новонаро#джений**
ТК_У	**новонаро#джений** 1. *прикм.*
ММУ	*укр.*
Т_У	**новонаро#джений**
ГРПУ	*прикм.*
ТК_Р	*рос*. новорождённый
ММР	*рос.*
$Т^{Р}$	новорождённый
ТК_А	*англ*. neonatus
ММА	*англ.*
$Т^{А}$	neonatus
$СБ_1$	1. який недавно або тільки щонародився;
$БТ_1$	1. який недавно або тільки щонародився;
НТ	1
ТЛ	який недавно або тільки щонародився;
$ТБ_2$	*ім.*, -ого (*рос*. новорождённый, *англ*. newborn, infant)
ТК_У	**новонаро#джений** *ім.*, -ого
ММУ	*укр.*
Т_У	**новонаро#джений**
ГРПУ	*ім.*, -ого
ТК_Р	*рос*. новорождённый
Т_Р	новорождённый
ММР	*рос*.
$ТК\_А_1$	*англ*. newborn
Т_А	newborn
ММА	*англ*.
$ТК\_А_2$	infant
Т_А	infant
ММА	*англ*.
$СБ_2$	2. людина, яка недавно народилася.
$БТ_2$	2. людина, яка недавно народилася.
НТ	2.

ТЛ	людина, яка недавно народилася.

The example reveals two terminological blocks and two semantic ones, respectively. The first terminological block is complete, it consists of Ukrainian, Russian and English complexes. The second block is cut. The terminological block Ukrainian in both complexes has a common Ukrainian term, which is the title word. Semantic blocks, respectively, consist only of interpretations.

Example 3.

ацидофі#льний *(рос.* ацидофи#льный, *англ.* acidophilic) 1. Який має здатність забарвлюватися кислими барвниками; **ацидофі#льні органі#зми** *див.* **органі#зм**: **органі#зми ацидофі#льні.** Син**. кислотолю#бний**; 2. Який росте тільки в кислому середовищі.

Marking	Representation in the SS
ТБ	**ацидофі#льний** *(рос.* ацидофи#льный, *англ.* acidophilic)
ЗТ	**ацидофі#льний**
ТК_У	**ацидофі#льний**
ММУ	*укр.*
Т_У	**ацидофі#льний**
ТК_Р	*рос.* ацидофи#льный
ММР	*рос.*
$Т^{Р}$	ацидофи#льный
ТК_А	*англ.* acidophilic
ММА	*англ.*
$Т^{А}$	acidophilic
СБ	1. Який має здатність забарвлюватися кислими барвниками; **ацидофі#льні органі#зми** *див.* **органі#зм**: **органі#зми ацидофі#льні.** Син**. кислотолю#бний**; 2. Який росте тільки в кислому середовищі.
$БТ_1$	1. Який має здатність забарвлюватися кислими барвниками;
НТ	1
ТЛ	Який має здатність забарвлюватися кислими барвниками;
$БТ_2$	2. Який росте тільки в кислому середовищі.
НТ	**2**
ТЛ	Який росте тільки в кислому середовищі.
СИН	Син**. кислотолю#бний**
СН	**кислотолю#бний**
МС	Син**.**
БП	**ацидофі#льні органі#зми** *див.* **органі#зм**: **органі#зми ацидофі#льні.**
САНТ	**ацидофі#льні органі#зми**
САТ	**органі#зм**
МП	*див.*

The semantic block can be filled with blocks of terminological phrases, synonyms, blocks of references. Terminological blocks can be several, they can be presented in one or two complexes. The examples illustrate some variants of the structure in the notation of the scheme.

2.2. Marking the text of the Dictionary with XML tags according to the structure of its L-system (XML document)

The next stage is the automatic conversion of the lexicographic structure of the dictionary into an XML document. However, it is obvious that the XML file explains and stores all the structural elements we have identified and the relationships between them. This is done using a special software procedure developed by us to automatically mark the text of the dictionary. The marking algorithm is developed based on polygraphic features of text identification of structural elements of the L-system (boundaries of the dictionary article (paragraphs), special symbols, positional characteristics, changes of language, fonts, case of letters, etc.).

XML dictionary article schema (SS)

<СС> *Словникова стаття*
<ЗТ^{У}>*заголовний термін український***</ЗТ^{У}>**
<ТБ номер=p> *Термінологічний блок*
<ТК_У номер=i> *український термінологічний комплекс*
<Т^{У}> *Термін український***</Т^{У}>**
<НО> *Номер омоніма***</НО>**
<ГР> *Граматична ремарка***</ГР>**
<ММУ> *укр.***</ММУ>**
</ТК_У >
<ТК_Р номер= j> *російський термінологічний комплекс*
<Т^{Р}> *Російський термін***</Т^{Р}>**
<СР> *Семантична ремарка***</СР>**
<ГР> *Граматична ремарка***</ГР>**
<ММР> *рос.***</ММР>**
</ТК_Р>
<ТК_А номер=k> *англійський термінологічний комплекс*
<Т^{А}> *Термін англійський***</Т^{А}>**
<СР> *Семантична ремарка***</СР>**
<ГР> *Граматична ремарка***</ГР>**
<ММА> *англ.***</ММА>**
</ТК_А >
</ТБ >
<СМБ номер=p>
<БТ номер=m> *Блок тлумачення*
<ТЛ> *Тлумачення* **</ТЛ>**
<СРТ> *Семантична ремарка* **</СРТ>**
<СИН номер=n> *Синонімічний блок*
<Т^{У}> термін</**Т^{У}**>
<ТС^{У}> термін</**ТС^{У}**>
<МС> Син.**</МС>**
</СИН номер=n >
</БТ >
<БТС номер=l> *Блок термінологічних словосполучень*
<ТБс номер =t> Термінологічний блок словосполучення
<ТКС_У номер =f> *Український термінологічний комплекс словосполучення*

<ТС^{У}> *Термологічне словосполучення* **</ТС^{У}>**
<ГРС> *Граматична ремарка* **</ГРС>**
<ММУ> *Маркер мови* **</ММУ>**
</ТКС_У >
<ТКС_Р номер =g> *Російський термінологічний комплекс словосполучення*
<ТС^{Р}> *Термологічне словосполучення* **</ТС^{Р}>**
<ГРС> *Граматична ремарка* **</ГРС>**
<ММР> *Маркер мови* **</ММР>**
</ТКС_Р >
<ТКС_А номер =h> *Англійський термінологічний комплекс словосполучення*
<ТС^{Р}> *Термологічне словосполучення* **</ТС^{Р}>**
<ГРС> *Граматична ремарка* **</ГРС>**
<ММА> *Маркер мови* **</ММА>**
</ТКС_А>
</ТБс>
<БТсл номер =v> Блок тлумачення словосполучення
<ТЛс> Тлумачення до словосполучення</ТЛс>
</БТсл>
</БТС>
<БП> *Блок посилань* **</БП>**
</СМБ>
</СС>

According to the scheme, all dictionary articles were marked. Consider the labeling by example 1,2.

Example 1

СС> <текст_СС> **ву#са,** -ів, *мн.*, *одн.* **вус**, -а *(рос.* усы#, *ед.* ус., *англ.* whisker (*у тварин*), tendril (*у рослин*), moustache (*у людини*), antenna (*у комах*)) 1. Загальна назва розміщених біля рота чутливих волосин (вібрисів) у ссавців та щетинкоподібного пір'я у птахів, деякі дотикові утвори у безхребетних тварин; 2. Надземні виткі прикріплювальні пагони у рослин. <текст_СС>

<ТБ> <текст_ТБ> **ву#са,** -ів, *мн.*, *одн.* **вус**, -а *(рос.* усы#, *ед.* ус., *англ.* whisker (*у тварин*), tendril (*у рослин*), moustache (*у людини*), antenna (*у комах*)) </текст_ТБ>
<ТК номер='1'_У>
<Т_У> ву#са</Т_У>
<ГРПУ> -ів, *мн.* </ГРПУ>
<ММУ> *укр.* </ММУ>
</ТК_У>
<ТК номер='2'_У>
<Т_У> вус</Т_У>
<ГРПУ> -а</ГРПУ>
<ГРДУ> одн.</ГРДУ>
<ММУ> *укр.* </ММУ>

</ТК_У>

<ТК номер='1'_Р>
<Т_Р> усы#</Т_Р>
<ММР> *рос.* </ММР>

</ТК_Р>
<ТК номер='2'_Р>
<Т_Р> ус</Т_Р>
<ГРДР> ед.</ГРДР>
<ММР> *рос.* </ММР>
</ТК_Р>
<ТК номер='1'_А>
<Т_А> whisker</Т_А>
<СР> у тварин</СР>
</ТК_А>
<ТК номер='2'_А>
<Т_А> tendril</Т_А>
<СР> у рослин</СР>
<ММА> *англ.* </ММА>
</ТК_А>
<ТК номер='3'_А>
<Т_А> moustache</Т_А>
<СР> у людини</СР>
<ММА> *англ.* </ММА>
</ТК_А>
<ТК номер='4'_А>
<Т_А> antenna</Т_А>
<СР> у комах</СР>
<ММА> *англ.* </ММА>
</ТК_А>
</ТБ>
<СМБ>

<БТ номер='1'>
<ТЛ> Загальна назва розміщених біля рота чутливих волосин (вібрисів) у ссавців та щетинкоподібного пір'я у птахів, деякі дотикові утвори у безхребетних тварин; </ТЛ>
</БТ>
<БТ номер='2'>
<ТЛ> Надземні виткі прикріплювальні пагони у рослин.</ТЛ>
</БТ>
</СМБ>
</СС>

Example 2

<СС>
<текст_СС> **новонаро#джений** 1. *прикм. (рос*. новорождённый, *англ.* neonatus, neonate) який недавно або тільки що народився; 2. *ім.*, -ого *(рос*. новорождённый, *англ.* newborn, infant) людина, яка недавно народилася. </текст_СС>
<ЗТ> **новонаро#джений** </ЗТ>
<ТБ номер='1'>
<тест_ТБ> **новонаро#джений** 1. *прикм. (рос*. новорождённый, *англ.* neonatus, neonate) </тест_ТБ>
<ТК_У номер='1'>
<Т_У> **новонаро#джений** </Т_У>
<ГРПУ> *прикм.* </ГРПУ>
<ММУ> *укр.* </ММУ>

```
                    </ТК_У>
                    <ТК_Р номер='1'>
                            <ТР> новорождённый </ТР>
                            <ММР> рос. </ММР>
                    </ТК_Р >
                    <ТК_А номер='1' >
                            <Т_А> neonatus </Т_А>
                            <ММА> англ. </ММА>
                    </ТК _А>
                    <ТК_А номер='2' >
                            <Т_А> neonate </Т_А>
                            <ММА> англ. </ММА>
                    </ТК_А>
    </ТБ>
    <ТБ номер='2'>
    <тест_ТБ> 2. ім., -ого (рос. новорождённый, англ. newborn, infant)</тест_ТБ>
                    <ТК_У номер='1'>
                            <Т__У> новонаро#джений </Т_У>
                            <ГРПУ> ім. </ГРПУ>
                            <ГРПУ> -ого </ГРПУ>
                            <ММУ> укр. </ММУ>
                    </ТК_У>
                    <ТК номер='1'_Р>
                            <Т_Р> новорождённый </Т_Р>
                            <ММР> рос. </ММР>
                    </ТК _Р>
                    <ТК_А номер='1'>
                            <Т_А> newborn </Т_А>
                            <ММА> англ. </ММА>
                    </ТК_А>
                    <ТК_А номер='2'>
                            <Т_А> infant </Т_А>
                            <ММА> англ. </ММА>
                    </ТК_А>
    </ТБ>
    <СМБ номер='1'>

    <БТ номер='1'>
                    <НТ>1</НТ>
                    <ТЛ> який недавно або тільки щонародився; </ТЛ>
    </БТ номер='1'>
    </СМБ>
    <СМБ номер='2'>
    <БТ номер='2'>
                    <НТ>2</НТ>
                    <ТЛ> людина, яка недавно народилася. </ТЛ>
    </БТ номер='2'>
    </СМБ>
</СС>
```

The transition to an XML document is due to the need to define author tag sets and attribute names. Document XML structures can also be nested, providing any level of hierarchy, as long as the rules for

embedding XML documents are followed. XML documents can contain any optional grammar descriptions of the document so that other programs can check its structure. The XML representation of the dictionary obtained in this way makes it possible to form its lexicographic database in automatic mode. This stage will be considered in a separate paper.

3. Discussion

After going through a number of stages, we have achieved many benefits:

1. In the digital world, dictionaries will be given new life as they are presented in a modern way.
2. Working with the content showed many errors that the program highlighted.
3. In the future proper XML will help implement the right search engine on site.
4. Any changes that will need to be made to the site can be made through a modern editing system.

4. Conclusion

Although there are still questions, it has been demonstrated that it is possible to digitize a paper dictionary and save it in XML and on the Internet. The key is to use standard components that can be reused in other projects and have simple data formats that are easy to edit with free tools

5. References

[1] D. M. Grodzinsky, L. O. Simonenko and other. Ukrainian biological terminology Dictionary. – K.: KMM, 2012. – 746 p.

[2] V. A. Shyrokov Computer lexicography: Monograph / Palagin O.V.; Ukrainian Lingua-Information Fund – Kyiv. : Nauk. dumka, 2011. – 351 p.

[3] V. A. Shyrokov etc. Linguistic and information studies: works of the Ukrainian Language and Information Fund NAS of Ukraine: y 5 V. V. 1 : Scientific paradigm and basic language and information structures. Kyiv. Ukrainian Lingua-Information Fund of NAS of Ukraine. 2018. 271 p. URL: https://movoznavstvo.org.ua/files/tom_1_B5_print.pdf. doi: 10.33190/978-966-02-8683-2/8684-9.

[4] V. A. Shyrokov etc. Linguistic and information studies: works of the Ukrainian Language and Information Fund NAS of Ukraine: in 5 V. V. 2 : Grammar systems. Kyiv. Ukrainian Lingua-Information Fund of NAS of Ukraine. 2018. 300 p.

[5] V. A. Shyrokov etc. Linguistic and information studies: works of the Ukrainian Language and Information Fund NAS of Ukraine: in 5 V. V. 5 : Virtualization of linguistic technologies. Kyiv. Ukrainian Lingua-Information Fund of NAS of Ukraine. 2018. 239 p. URL: https://movoznavstvo.org.ua/files/Ling_inf_studio_TOM_5_umif_B5.pdf. doi: 10.33190 /978-966-02-8683-2/8690-0

[6] V. A. Shyrokov Grammatical systems: phenomenological approach / V. A. Shyrokov, T. P. Lyubchenko, I. V. Shevchenko, K. V. Shyrokov. – K. : Nauk. dumka, 2018. – 310 p.

[7] O. Karpova Lexicography and Terminology: A Worldwide Outlook / Olga Karpova, Faina Kartashkova. – Cambridge : Cambridge Scholars Publishing, 2009. – 205 p.

[8] I. Kernerman A multilingual trilogy: Developing three multi-language lexicographic datasets. Electronic Lexicography in the 21st Century: Linking lexical data in the digital age. Proceedings of eLex 2015, 11-13 August 2015. – 372-383p. URL: https://elex.link/elex2015/