Alonań DOROŻYŃSKA[1]

Opiekun naukowy: Władimir SZYROKOW[2]

DOI: https://doi.org/10.53052/9788366249868.02

# CYFRYZACJA SŁOWNIKA TERMINOLOGICZNEGO

**Streszczenie:** Artykuł przedstawia zakres zadań, podejść i etapów przekształcenia słownika papierowego w produkt internetowy, na podstawie opracowania „Słownik ukraińskiej terminologii biologicznej" [1]. Praca ze słownikami przetłumaczonymi na komputerowe formaty tekstowe jest bardzo nieefektywna i wymaga ich konwersji na formaty leksykograficznych baz danych.

**Słowa kluczowe:** leksykografia komputerowa, system leksykograficzny, parsowanie, XML, baza danych, przestrzeń cyfrowa, web-web.

# DIGITALIZATION OF TERMINOLOGICAL DICTIONARY

**Summary:** The article outlines the range of tasks, approaches and stages of transforming a paper dictionary into an online product based on the study "Dictionary of Ukrainian biological terminology" [1]. Working with dictionaries translated into computer text formats is very inefficient and requires their conversion into lexicographic database formats.

**Keywords:** computer lexicography, lexicographic system, parsing, XML, database, digital space, website.

## 1. Introduction

One of the tasks of modern computer lexicography is to create digital dictionaries, among which are multilingual terminological dictionaries. Till now, most of these paper works in Ukrainian terminology have not digital analogues, which causes the urgent task of their digitization. Today, quite a lot of tools have been created to automate certain stages of terminological work, but universal solutions to the main problems of digital terminography have not yet been developed. This is especially true of the digital reception of traditional terminological heritage, especially multilingual. Among all the dictionary diversity, the Dictionary of Ukrainian

[1] Fundacja Języka Ukraińskiego i Informacji NASU, młodszy pracownik naukowy, alonochkatkachyk@gmail.com
[2] Fundacja Języka Ukraińskiego i Informacji NASU, Akademik Narodowej Akademii Nauk Ukrainy, vshirokov48@gmail.com

Biological Terminology was selected for digitization [1] (according to the authors, this dictionary is the first lexicographical work of a new generation in Ukrainian studies, covering the most common biological terminology in Ukrainian, Russian and English). The offered terminographic work embraces normative general scientific and widely used terminology of biological sciences, fixed in modern encyclopedic, general language and special dictionaries, in scientific, popular science, educational and informative-abstract literature.

Our approach is to gradually convert the dictionary text into a website. Basic technological stages are as follows:

Paper Dictionary
↓
The structure of the L-system of the Dictionary (1)
↓
Marking the text of the Dictionary with XML tags
according to the structure of its L-system (XML-text)
↓
Conversion of XML-text of the Dictionary into a lexicographic database
↓
ULP website[3]

This process, in our opinion, contains steps that can be applied to other dictionaries, so we believe that this sequence is an effective and universal way to translate traditional dictionaries into digital format.

## 2. The steps

Let's consider the individual steps that make up the transformation process.

### 2.1. Paper book

If from the very beginning the text of the Dictionary is accessible in electronic format, then work with it. In particular, the text of the DUBT [2] we had from the beginning in the form of a PDF file. For convenience, this file has been converted to Word doc in order to do some text conversion. Among these transformations we note the following: the disclosure of abbreviations of a certain type (for example: п. загрудни́нна → пульсація загрудни́нна and the like.; replacement of stressed letters with a combination of two characters: «літера#». The following replacements were made: а́—а#, е́—е#, и́—и#, і́—і#, о́—о#, у́—у#, ї́—ї#, я́—я#, є́—є#, ю́—ю#, ы́—ы# (for Cyrillic); ý—y# (for Latin). All dictionary articles were processed in this way.

### 2. 2 The structure of the lexicographic system (L-system) of DUBT

Following the theory of lexicographic systems [2], the structure of the L-system of SMS is presented in the form (Fig. 1):

[3] ULP – Ukrainian linguistic portal

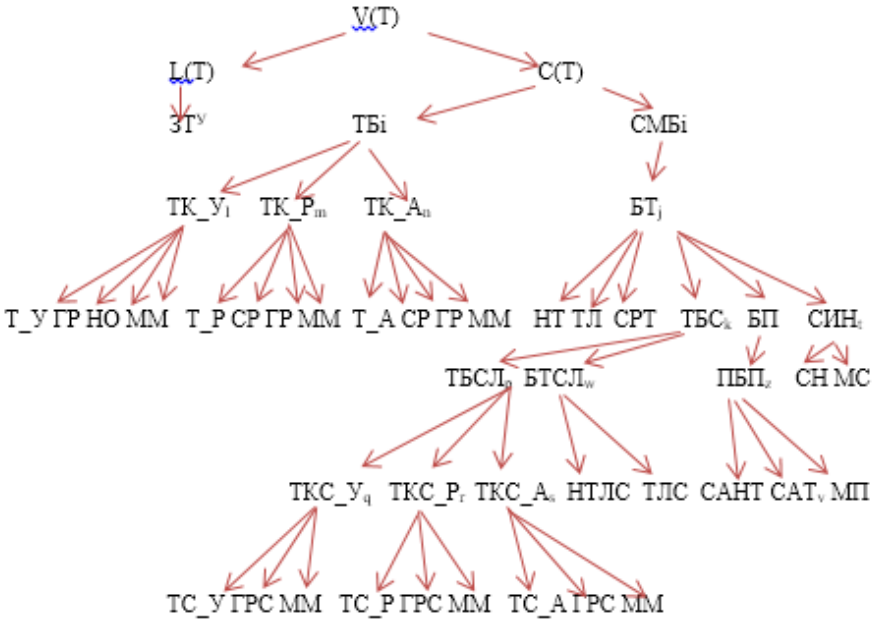


*Figure 1. General scheme of the dictionary article of DUBT*

The following notation is used in this scheme:

СС — dictionary article
ЗТ_У — the title term is Ukrainian
**ТБi — terminological block**
**ТК_Уl — terminological complex ukr.**
Т_У — Ukrainian term
ГР— grammatical trailer
НО — homonym number
ММ — language marker (ukr)
**ТК_Рm — terminological complex rus.**
Т_Р — Russian term
ГР — grammatical trailer
ММ — language marker (rus.)
СР — semantic trailer
**ТК_Аn — terminological complex eng.**
Т_А — English term
ГР — grammatical trailer
ММ — language marker (eng.)
СР — semantic trailer
**СМБi — semantic block**
**БТj — interpretation block**
НТ **— interpretation number**
ТЛ **— interpretation**
СРТ — semantic trailer to ТЛ
**БТСk — block of terminol. Phrases**
САТv **— recipient**
**ТБСЛр — terminol. block of phrases**
ТСК_Уq — terminological complex of phrases ukr.
ТС_У — terminological phrase ukr.
ГРС — grammatical remark of the phrase
ММ — language marker (укр)
ТСК_Рr — terminological complex of phrases rus.
ТС_Р — terminological phrase rus.
ММ — language marker (rus.)
ТСК_Аs — terminological complex of phrases eng.
ТС_А — terminological phrase eng.
ММ — language marker (eng.)
**БТСЛw — block of interp. of phrases**
ТЛС — interpretation
НТЛС — number ТЛС
**СИНt — synonymous block**
СН — synonym
МС — synonym marker (син.)
**БП — link block**
**ПБПz — subblock**
САНТ **— sender**
МП — link marker (*див.*)

### 2.3. Marking DUBT text with XML tags according to its structure L-systems (XML-text DUBT )

The next step is to automatically convert the dictionary text into an XML document, which makes it possible to explain all the structural elements we have identified and the relationships between them. To ensure automatic marking of the dictionary text with XML tags, a program was developed that separates the elements of the text structure according to the structure of the L-system. Polygraphic features of text identification of the L-system are used, namely: the boundaries of the dictionary article (paragraphs), various special symbols that distinguish structural elements, positional characteristics, changes in language, fonts, case, etc.

XML dictionary article schema (CC)

```
<СС> Словникова стаття
    <ЗТ_У>заголовний термін український</ЗТ_У>
            <ТБ номер=i> Термінологічний блок
            <ТК_У номер=l> укр. термінологічний комплекс
                <Т_У> Термін український</Т_У>
                <НО> Номер омоніма</НО>
                <ГР> Граматична ремарка</ГР>
                <ММ> укр.</ММ>
            </ТК_У >
            <ТК_Р номер= m> рос. термінологічний комплекс
                <Т_Р> Російський термін</Т_Р>
                <СР> Семантична ремарка</СР>
                <ГР> Граматична ремарка</ГР>
                <ММ> рос.</ММ>
            </ТК_Р>
            <ТК_А номер=n> англ. термінологічний комплекс
                <Т_А> Термін англійський</Т_А>
                <СР> Семантична ремарка</СР>
                <ГР> Граматична ремарка</ГР>
                <ММ> англ.</ММ>
            </ТК_А >
            </ТБ >
            <СМБ номер=i >
            <БТ номер=j> Блок тлумачення
                <ТЛ> Тлумачення </ТЛ>
                <НТ> Тлумачення </НТ>
                <СРТ> Семантична ремарка </СРТ>
                       <СИН номер=t> Синонімічний блок
                              <Т_У> термін</Т_У>
                              <ТС_У> термін</ТС_У>
                              <МС> Син. </МС>
                       </СИН >
<БТС номер=k> Блок термінологічних словосполучень
          <ТБСЛ номер =p> Термінологічний блок словосполучення
                <ТКС_У номер =q> Український термінологічний комплекс словосполучення
                <ТС_У> Термологічне словосполучення</ТС_У>
                <ГРС> Граматична ремарка</ГРС>
                <ММ> Маркер мови</ММ>
            </ТКС_У>
```

```
            <ТКС_Р номер =r> Російський термінологічний
        комплекс словосполучення
                <ТС_Р> Термологічне словосполучення</ТС_Р>
                <ГРС> Граматична ремарка</ГРС>
                <ММ> Маркер мови</ММ>
            </ТКС_Р>
            <ТКС_А номер =s> Англійський термінологічний
        комплекс словосполучення
                <ТС_А> Термологічне словосполучення</ТС_А>
                <ГРС> Граматична ремарка</ГРС>
                <ММ> Маркер мови</ММ>
            </ТКС_А>
        </ТБСЛ>
        <БТСЛ номер =w> Блок тлумачення словосполучення
                <ТЛС> Тлумачення до словосполучення</ТЛС>
                <НТЛС> Номер тлумачення до сл.</НТЛС>
         </БТСЛ>
    </БТС>
<БП> Блок посилань
    <ПБП номер = z> Підблок посилань
           <САНТ> адресант</САНТ>
           <САТ номер=v> адресат </САТ>
           <МП> маркер посилань <МП>
    </ПБП>
</БП>
</БТ>
</СМБ>
</СС>
```

**Example**

```
<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>
<ТБ номер='1'>
<ЗТ> новонаро#джений </ЗТ>
    <ТК_У номер='1'>
           <Ту> новонаро#джений </Ту>
           <ГР> прикм. </ГР>
           <ММ> укр. </ММ>
    </ТК_У>
    <ТК_Р номер='1' >
           <Тр> новорождённый </Тр>
           <ММ> рос. </ММ>
    </ТК_Р >
    <ТК_А номер='1' >
           <Та> neonatus </Та>
           <ММ> англ. </ММ>
    </ТК _А>
    <ТК_А номер='2' >
           <Та> neonate </Та>
           <ММ> англ. </ММ>
    </ТК_А>
</ТБ>
<ТБ номер='2'>
```

```
<тест_ТБ> 2. ім., -ого (рос. новорождённый, англ. newborn,
infant)</тест_ТБ>
    <ТК_У номер='1'>
          <Ту> новонаро#джений </Ту>
          <ГР> ім. </ГР>
          <ГР> -ого </ГР>
          <ММ> укр. </ММ>
    </ТК_У>
    <ТК номер='1'_Р>
          <Тр> новорождённый </Тр>
          <ММ> рос. </ММ>
    </ТК _Р>
    <ТК_А номер='1' >
          <Та> newborn </Та>
          <ММ> англ. </ММ>
    </ТК_А>
    <ТК_А номер='2' >
          <Та> infant </Та>
          <ММ> англ. </ММ>
    </ТК_А>
</ТБ>
<БТ номер='1'>
    <ТЛ> який недавно або тільки що народився; </ТЛ>
</БТ номер='1'>
<БТ номер='2'>
    <ТЛ> людина, яка недавно народилася. </ТЛ>
</БТ номер='2'>
</СС>
```

The next step is to create a database and create a website.

## 3. Discussion

After going through a number of stages, we achieved many advantages: it allowed us to present the dictionary in a modern way, rather than linking it to the old format; data conversion has allowed us to record many errors and inconsistencies ;. proper XML will facilitate the implementation of search functionality and it will also simplify the creation of direct cross-references on the website.

## REFERENCES

1. Dictionary of Ukrainian biological terminology. - K .: KMM, 2012. – 746 p.
2. Linguistic and information studies: works of the Ukrainian language and information fund of the National Academy of Sciences of Ukraine: in 5 volumes / V. A. Shirokov et al. Vol. 1: Scientific paradigm and basic language and information structures. Kyiv. Ukrainian Language and Information Fund of the National Academy of Sciences of Ukraine. 2018. 271 p.