https://doi.org/10.23939/sisn2019.01.090

УДК 81′ 33

ПАРСИНГ ТЕКСТУ ТЕРМІНОЛОГІЧНИХ СЛОВНИКІВ

А. В. Дорожинська
Український мовно-інформаційний фонд НАНУ
alonochkatkachyk@gmail.com
ORCID - 0000-0001-6554-6731

Окреслено коло завдань, підходів і етапів розроблення технології парсинга тексту багатомовного тлумачного термінологічного словника. Дослідження проведено для "Словника української біологічної термінології". Серед усього словникового розмаїття цей словник обрано тому, що термінологічні словники надають лексико-семантичну базу для подальшого створення систем інтелектуального опрацювання фахових текстів, у яких подається інформація з тих чи інших предметних галузей. Ця термінографічна праця обіймає нормативну загальнонаукову та широковживану термінологію біологічних наук, зафіксовану в сучасних енциклопедичних, загальномовних та спеціальних словниках, у науковій, науково-популярній, навчальній та інформативно-реферативній літературі. Дослідивши обраний словник, ми зможемо узагальнити модель його лексикографічної системи на інші предметні галузі, що створить передумови для формування цілісного багатогалузевого цифрового лексикографічного простору. Робота зі словниками, переведеними у комп'ютерні текстові формати, є дуже неефективною і потребує конвертування їх у формати лексикографічних баз даних, що є спеціальним завданням, не відомим класичній лексикографії. Це і складає зміст терміна "парсинг словників". У процесі роботи побудовано модель лексикографічної системи, яку покладено в основу XML. Подальша робота із перетворення паперової версії словника на онлайн-систему будується на XML-файлі. Проаналізовано поліграфічне оформлення, організацію і структуру друкованого тексту словника з метою ідентифікації елементів концептуальної моделі Л-системи СУБТ. На основі концептуальної моделі побудовано структуру XML-документа, який пропонується використовувати як посередника між паперовою версією словника та його реалізацією як онлайнової лексикографічної системи. Надалі планується побудова універсальної процедури парсингу з удосконаленням структури XML-документа.

Ключові слова: комп'ютерна лексикографія, лексикографічна система, парсинг, синтаксичне дерево, XML, база даних, цифровий простір.

PARSING THE TEXT OF TERMINOLOGY DICTIONARIES

Dorozhynska A. V.
Ukrainian Lingua-Information Fund of NAS of Ukraine
alonochkatkachyk@gmail.com
ORCID - 0000-0001-6554-6731

The article outlines a range of tasks, approaches and stages of developing parsing technology for text of a multilingual explanatory terminology dictionary. Research was conducted for the "Dictionary of Ukrainian Biological Terminology". Among all the vocabulary diversity, this dictionary was chosen because terminology dictionaries provide a lexical-semantic basis for further creation of systems for the intelligent processing of professional texts, which provide information on specific subject areas. This terminographical work encompasses the normative general scientific and widely used terminology of biological

sciences, recorded in modern encyclopedic, general and special dictionaries, in scientific, popular science, educational and informative literature. After studying the chosen dictionary, the model of its lexicographic system into other subject areas, which will create the preconditions for the formation of an integral multidisciplinary digital lexicographic space will be generalized. Working with dictionaries converted into computer text formats is very inefficient and needs to be converted into lexicographic database formats, which is a special task not known in classical lexicography. This is the meaning of the term "parsing dictionaries". During investigation, a model of the lexicographic system, which is the basis of XML, was constructed. Further work on converting a printed version of the dictionary into an online system is based on an XML file. The polygraphic design, organization and structure of the printed text of the dictionary are analyzed in order to identify the elements of the conceptual model of the L-system of the SUBT. Based on the conceptual model, the structure of an XML document is proposed, which is to be used as an intermediary between the printed version of the dictionary and its implementation as an online lexicographic system. In the future, it is planned to build a universal parsing procedure, by improving the structure of the XML document.

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

Вступ

Одним із завдань комп'ютерної лексикографії є створення електронних словників. Колосальні обсяги накопиченої інформації та висока швидкість надходження нової останнім часом привели до необхідності якісно нових засобів опрацювання даних [1, с. 4–5, 15–16]. Сьогодні створено доволі багато інструментальних засобів для автоматизації окремих етапів термінологічної роботи, але немає універсального рішення для основних задач. Саме тому розгляд технології парсингу представляє особливий інтерес і в теоретичному, і в практичному плані.

Особливо актуалізувалася проблема автоматичного переведення природномовних текстів до структурованих форм, адаптованих до машинного аналізу та інтерпретації. Зазначимо, що зазначене опрацювання базується, передусім, на ідентифікації та інтерпретації лінгвістичних параметрів і характеристик аналізованих текстів. Своєю чергою, вказані параметри та характеристики є предметом систематизованого подання в різних за своїм змістом, характером та побудовою словниках. Отже, створення засобів автоматичного опрацювання текстів передбачає наявність електронних словників, інтегрованих у контури відповідних систем.

Серед усього словникового розмаїття особливе місце посідають термінологічні словники, оскільки саме вони надають лексико-семантичну базу для подальшого створення систем інтелектуального опрацювання фахових текстів, у яких подається інформація з тих чи інших предметних галузей. Тобто, постає завдання переведення термінологічного словникового доробку у цифрову форму, що є не зовсім простим завданням. Адже, як зазначено у [12, с. 6]: "Для цього абсолютно недостатньо просто зісканувати тексти відповідних словників. Хоча й це не така проста задача, як може видаватися на перший погляд. Відомо, що метамова багатьох словників містить цілий ряд нестандартних символів, які погано піддаються ідентифікації. Крім того, існуючі програми оптичного розпізнавання текстів далеко не бездоганні і спричиняють чимало помилок у процесі їхнього застосування. Нарешті, робота зі словниками, переведеними у комп'ютерні текстові формати, є дуже неефективною і потребує конвертування їх у формати лексикографічних баз даних, що є спеціальним завданням, не відомим класичній лексикографії". Останнє завдання, власне, і становить зміст терміна "*парсинг словників*", технологічне опрацювання якого і розглянуто у статті.

Наголосимо на важливості парсингу саме термінологічних словників, "оскільки без них неможливий ані розвиток різних галузей знань, ані сучасне міжнародне спілкування в різноманітних сферах" [2]. Проте вітчизняна лексикографія має ще не дуже багатий досвід

створення електронних термінологічних словників. Серед них згадаємо створені в Українському мовно-інформаційному фонді НАН України електронні термінологічні словники [4–8, 9–11].

Ми досідили Словник української біологічної термінології (СУБД) [8], оскільки він має багату структуру, охоплює найуживанішу біологічну термінологію українською, російською та англійською мовами. Ця термінографічна праця обіймає нормативну загальнонаукову та широковживану термінологію біологічних наук, зафіксовану в сучасних енциклопедичних, загальномовних та спеціальних словниках, у науковій, науково-популярній, навчальній та інформативно-реферативній літературі. Дослідивши обраний словник, ми зможемо узагальнити модель його лексикографічної системи на інші предметні галузі, що створить передумови для формування цілісного багатогалузевого цифрового лексикографічного простору.

Під час аналізу об’єкта дослідження виникають певні питання та завдання.

Серед завдань є етапи дослідження формату даних словника та кодування елементів його тексту. Безперечно, важливим питанням є побудова лексикографічної структури словника та опис технології його парсингу. Цілком природно, що концептуальну базу дослідження становить теорія лексикографічних систем [12, с. 89–136]. Наведемо основні її положення.

Теорія лексикографічних систем

Згідно з [12, с. 89–136], лексикографічна система (Л-система) є інформаційним об’єктом доволі загальної природи, який поєднує в собі ознаки моделі даних, моделі знань та логіко-лінгвістичного числення певного типу. Основними *системотвірними відношеннями* Л-системи є: “суб’єкт-об’єкт” та “форма-зміст”. Основним *системотвірним інваріантом* Л-системи є лексикографічний ефект в інформаційних системах [12, с. 89–102].

За теорією В. Широкова, викладеною у [12], та використовуючи позначення з цієї книги, подамо формальне визначення поняття “лексикографічна система”. Здійснюється це в такий спосіб. Спочатку як результат рецепції суб'єктом S лексикографічного ефекту Q у межах об'єкта D, визначається (індукується) **дискретний клас елементарних інформаційних одиниць** (ЕІО) $I^Q(D)$ об’єкта D відносно вказаного лексикографічного ефекту Q. Наступний крок полягає у побудові $V(I^Q(D))$ – **опису класу $I^Q(D)$**, де поняття “опис” тут ужито приблизно в такому самому сенсі, як при визначенні алгоритмічної інформації за А. Колмогоровим. Вважаємо, що суб'єкт S саме й є чинником (оператором), який здійснює цю побудову:

$$S : I^Q(D) \;\rightarrow\; V(I^Q(D)). \qquad (1)$$

Об’єкт $V(I^Q(D))$ у цій моделі являє собою слово (текст) у певному скінченному алфавіті символів $\mathbf{A} = \{a_1, a_2, \ldots, a_k\}$. Зазначений текст є повним описом класу $I^Q(D)$, який дає змогу однозначно реконструювати зазначений клас та властивості всіх його елементів. Якщо $I^Q(D) = \{X_1, X_2, \ldots, X_n, \ldots\}$, то позначимо символом $V(X_i)$ обмеження $V(I^Q(D))$ на X_i: $V(X_i) = V(I^Q(D))^{\frown}{}_{X_i}$. Об’єднання $V(X_i)$ за всіма X_i становить клас $V(I^Q(D))$:

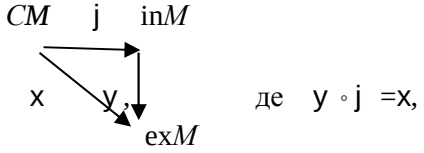
$$V(I^Q(D)) = \dot{E}_i\, V(X_i)\ . \qquad (2)$$

У словниковій інтерпретації клас $I^Q(D) = \{X_1, X_2, \ldots, X_n, \ldots\}$ тлумачиться як клас реєстрових одиниць певного словника; тоді $V(X_i)$ є текстом словникової статті цього словника із реєстровою одиницею X_i.

Одним із основних аспектів у визначенні Л-системи, розглядуваної як інформаційна система певного типу, є поняття її *архітектури*. Слідуючи за [12, с. 116–119], ми використовуємо архітектуру ANSI/X3/SPARK (або просто ANSI/SPARK), яка складається з трьох рівнів представлення даних: концептуального, внутрішнього та зовнішнього, що мають такі інтерпретації. Концептуальна модель (*концептуальний рівень представлення*) предметної галузі – це семіотична, семантична модель, у якій в однозначному, скінченному і несуперечливому вигляді інтегруються уявлення різних фахівців про предметну галузь. У внутрішній моделі (*внутрішньому рівні представлення*) визначаються типи, структури і формати представлення, зберігання та маніпулювання даними, алгоритмічна база та операційно-програмне середовище, в котре „занурюється” концептуальна модель під час її реалізації. Зовнішня модель (зо*внішній рівень*

представлення) відображає погляди кінцевих користувачів (і, отже, прикладних програмістів) на предметну галузь. У ній реалізується комплекс засобів, що дають змогу користувачеві здійснювати дозволені контакти та маніпулювання даними, поданими у внутрішньому рівні. Одній концептуальній моделі може відповідати декілька внутрішніх та зовнішніх.

Отже, елементами архітектури *ARCH* лексикографічної системи вважаємо такі: *ARCH* = {*CM, EXM,INM; Φ,Ψ, Ξ*}, де введено такі позначення: символом *CM* позначено концептуальну модель Л-системи; *EXM* ={ex*M*} – множина її зовнішніх моделей, які відповідають концептуальній моделі *CM*, а *INM* = {in*M*} – відповідна множина її внутрішніх моделей; символами *Φ* = {j }, *Ψ* = {y }, *Ξ* = {x} позначено відображення, які зв'язують *CM*, in*M* та ex*M* у таку комутативну діаграму:



(3)

що забезпечує узгодженість між усіма рівнями представлення даних у Л-системі.

Концептуальна модель *CM* будь-якої Л-системи має певну стандартну будову, яка випливає з інформаційного підходу до її моделювання. Оскільки єдиним джерелом змістового представлення Л-системи є її опис $V(I^Q(D))$, який є словом (текстом) у певному алфавіті **A**, то і єдиним джерелом структури Л-системи можуть виступати певні, інваріантно визначені складники цього тексту, його певні елементи. За теорією лексикографічних систем" [12] позначимо множину цих структурних елементів символом $\beta \circ \beta[V(I^Q(D))] = \{\beta_1, \beta_2, \ldots, \beta_q\}$. Отже, кожний елемент β_i залежить від $X_1, X_2, \ldots, X_n, \ldots$, і набуває певного значення $\beta_j(X_i)$ на кожному елементі X_i з класу $I^Q(D)$. Деякі з елементів $\beta_j(X_i)$ можуть бути порожніми. Множина структурних елементів $\beta_j(X_i)$, j = 1, 2, …, q; i = 1, 2, …, n,… задає фундаментальну структуру Л-системи.

Постає питання: в який спосіб можна побудувати цю структуру для конкретної предметної галузі, що є об'єктом лексикографування? Можливі два способи. Перший полягає у розбудові такої структури за певними "*першими принципами*", тобто з лінгвістичної теорії, яка достатньо повно описує об'єкт лексикографування. Другий спосіб полягає в аналізі тексту вже створеного словника, в якому подано лексикографічний опис певної ділянки мови. Вважають, що словник укладено тоді, коли вже є теорія, настільки повно опрацьована, що відповідний матеріал уже можна подати у словниковій формі. Тоді, аналізуючи текст словника, який у нашій термінології відіграє роль однієї із зовнішніх моделей Л-системи, що є субстратом цього словника, дослідник абстрагує з нього сукупність структурних елементів, яку узагальнює у вигляді структури $\beta_j(X_i)$, j = 1, 2, …, q; i = 1, 2, …, n,… . **Структура β є першою структурою лексикографічної системи**. Після побудови β її можна розглядати вже цілком незалежно від словника, з якого її було абстраговано. Обов'язковими елементами структури β – і це є відмітною особливістю всіх лексикографічних систем – є елементи $\Lambda(I^Q(D))$ та $P(I^Q(D))$, які є носіями, відповідно, форми та змісту елементів класу ЕІО $I^Q(D)$ і, отже, репрезентують відношення "форма-зміст", що є системотвірним відношенням будь-якої Л- системи.

Зі структурних елементів $\beta_j(X_i)$ будують **другу лексикографічну структуру $\sigma[\beta]$**, визначену на β і, отже, на $V(I^Q(D))$. Надалі називатимемо $\sigma[\beta]$ макроструктурою $V(I^Q(D))$; обмеження $\sigma[\beta]$ на $V(x)$: $\sigma[\beta]\frac{1}{2}_{V(x)} \circ \square\sigma(x)$ породжує мікроструктуру $V(x)$. Активне формулювання цього факту полягає у встановленні процедури (оператора, процесу...) σ, яка породжує на β структуру $\sigma[\beta]$:

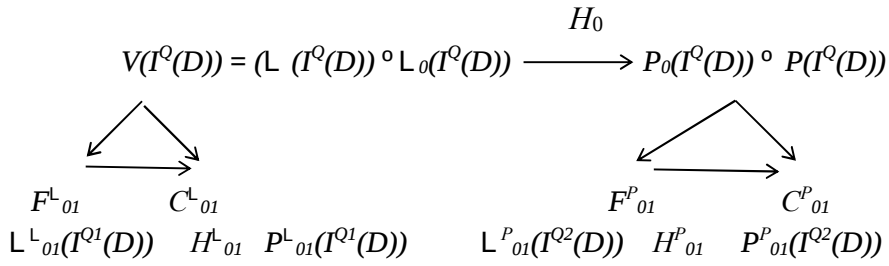
$$\sigma : \beta \rightarrow \sigma[\beta] \qquad (4)$$

Елементи структури $\sigma[\beta]$ представляють закономірності предметної галузі (зокрема, імпліцитні), що є об'єктом лексикографування. Їхнє визначення в кожному конкретному випадку є певним науковим дослідженням, інколи доволі складним та вишуканим, яке залежить від досвіду і навіть мистецтва дослідника.

Тобто, загальну структуру Л-системи подають у такому вигляді:

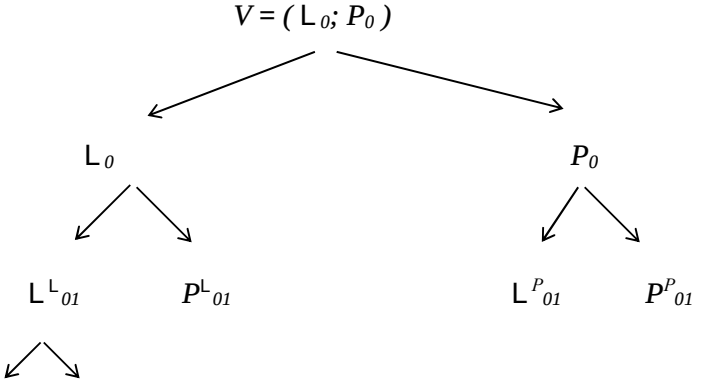
$$\{D, S, Q, I^Q(D), V(I^Q(D)), \mathsf{b}, \mathsf{s}[\mathsf{b}], Red[V(I^Q(D)]; ARCH\}, \quad (5)$$

де всі елементи, крім $Red[V(I^Q(D)]$, визначено вище. Символом $Red[V(I^Q(D)]$ позначено процес так званої рекурсивної редукції лексикографічної системи, суть якого зводиться до наступного. Універсальність явища лексикографічного ефекту дає можливість розглядати $L(I^Q(D))$ і $P(I^Q(D))$ як окремі, автономні елементарні Л-системи, а це уможливлює таку побудову:



(6)

Звернімо увагу на зміну типу лексикографічного ефекту на другому поверсі – замість Q тепер маємо Q_1 й Q_2 відповідно. Отже, приходимо до комплексів об'єктів $I^{Q_1}(D)$ та $I^{Q_2}(D)$. Продовжуючи цей процес, одержуємо рекурсивне розвинення лексикографічної системи $V(I^Q(D))$:



(7)

Цей процес ми називатимемо *рекурсивною редукцією лексикографічної системи*. Він нагадує своєрідний мовно-інформаційний "мікроскоп", що виявляє все тонші деталі структури лексикографічної системи й індукує структуру, схожу на фрактальну.

Моделювання лексикографічної структури та методика парсингу СУБТ

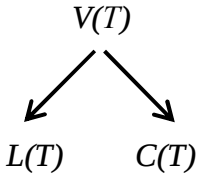
Парсинг словників є частинним випадком загального парсингу будь-яких текстів. Останній полягає у реалізації – автоматичній або автоматизованій – процесу зіставлення природномовному текстові його певної лінгвістичної (лексикографічної) структури. Історично першим було досліджено й певною мірою реалізовано синтаксичний парсинг речень природної мови, тобто зіставлення реченню його синтаксичної структури у вигляді, наприклад, відповідного дерева залежностей.

Узагальнення цієї задачі спричиняє створення систем, які дають змогу в автоматичному або автоматизованому режимі будувати різні формальні лінгвістично або когнітивно осмисленні структури, які можна поставити у відповідність тому чи іншому природномовному текстові.

Теорія лексикографічних систем, в якій введено і формально визначено поняття лексикографічної структури, а також її узагальнення (неелементарні Л-системи, лексикографічні середовища), **дає змогу строго сформулювати парсинг словників.**

Отже, найпершим кроком у парсингу словника є розроблення структури його Л-системи. Розглянемо це на прикладі СУБТ.

Відповідно до теорії лексикографічних систем розкриємо загальну структуру словникової статті СУБТ. Позначимо символом *V(T)* словникову статтю із реєстровою одиницею *T*; ***L(T)*** – її ліву, а ***C(T)*** – праву частину. Тоді справедлива діаграма:



(8)

Схема.1. *Права і ліва частина словникової статті*

де стрілкою позначено відношення вкладення. Своєю чергою: ***L(T)*** – ліва частина – являє собою термінологічний блок, який складається з таких елементів: *У(T)* – українська частина, *Р(T)* – російська частина, *А(T)* – англійська частина, синонім до *Т СИН(T)* (факультативний елемент). Права частина *C(T)* – це семантичний блок. Розглянемо приклад для ілюстрації елементів словникової статті.

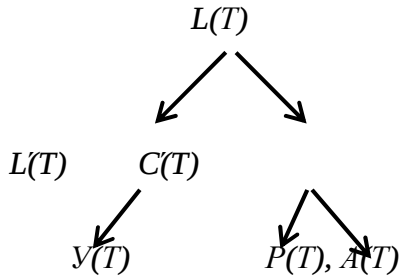
V(T) = **абіоти́ч/ний** (*рос*. абиоти́ческий, *англ.* abiotic), який не належить до живої природи; у якому відсутні життєві процеси Світового океану; **абіотичний фа́ктор** *див*. **фа́ктор: фактори абіоти́чні**. Син. **неорга́нічний**.

L(T) = **абіоти́чний** (*рос*. абиоти́ческий, *англ*.abiotic)

C(T) = який не належить до живої природи; у якому відсутні життєві процеси Світового океану; **абіотичний фа́ктор** *див*. **фа́ктор: фактори абіоти́чні**. Син. **неорга́нічний**.

Елементи *L(T)*: *У(T)* = **абіоти́чний;** *Р(T)* = *рос*. абиоти́ческий; *А(T)*= *англ.* abiotic; *СИН(T)*= Син. **Неорга́нічний**

Застосувавши до *L(T)* принцип рекурсивної редукції, розкладемо її на окремі комплекси, які, своєю чергою, містять терміни і ремарки:



(9)

Схема.2. *Структура лівої частини словникової статті*

Детально опишемо ліву частину ***L(T).*** Вона складається із термінологічних комплексів *(ТК)*. Структура комплексів однакова: комплекс *ТК* складається із терміна і всіх його ремарок. Позначимо: символами *ТК_У, ТК_Р* і *ТК_А* – український, російський та англійський термінологічні комплекси, відповідно.

В **українському термінологічному комплексі** виділяємо український термін *($Т^{У}$)* та граматичну ремарку *(ГР)*. Семантичної ремарки в українському термінологічному комплексі не виявлено. Номер омоніма виділяємо в окремий параметр *(НО)*; якщо омоніма немає, то його не маркують (приклад омонімів № 6).

Російський та англійський термінологічні комплекси містять терміни відповідними мовами *($Т^{Р}$) або ($Т^{А}$)*, граматичну ремарку *(ГР)* та/або семантичну ремарку *(СР)*.

Елементи термінологічного блоку зображено на схемі:

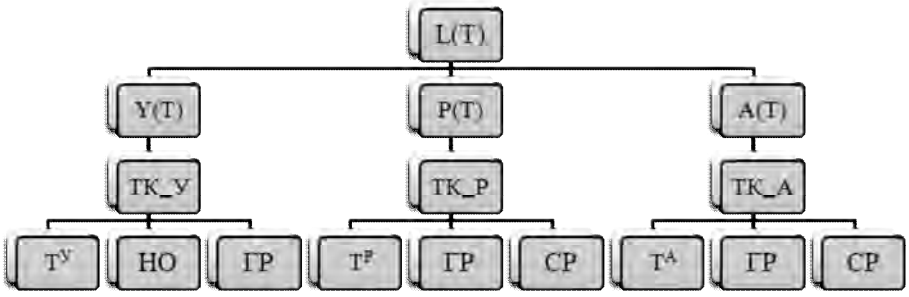


Схема 3. *Представлення структури лівої частини (термінологічний блок)*

Розглянемо *приклад № 2,* в якому виділено елементи термінологічного блоку.

V(T)=**желатин**, -у *і* **желатина** (*рос*. желати́н и желати́на, *англ*. gelatin) – прозорий або жовтуватий білок, отриманий із колагену виварюванням у воді кісток, хрящів, сухожилля тварин. Використовують у мікробіології (як живильне середовище), у медицині – для приготування гліцерину-желатину, а також у кулінарії, хімічному виробництві.

L(T)= **желатин**, -у *і* **желатина** (*рос*. желати́н и желати́на, *англ*. gelatin)

У(Т) = **желатин**, -у *і* **желатина**

ТК_У = **желатин**, -у *і* **желатина**

Т = **желатин**

ГР$_1$= -у

ГР$_2$= **желатина**

Р(Т) = желати́н и желати́на

ТК_Р = желати́н и желати́на

Т= желати́н

ГР= желати́на

А(Т) = gelatin

ТК_А= gelatin

Т= gelatin

Коментар до прикладу № 2

У таких словникових статтях в українському термінологічному комплексі виділяємо термін, граматичний параметр та термін як фонетичний чи морфологічний варіант. Буква "*і*" є параметром, який відділяє заголовне слово від можливих варіантів. Слова "*рос*.", "*англ*." – параметр, що позначає мову.

Розглянемо структуру правої частини *С(Т).*

Семантичний блок *С(Т)* поділяємо на: інтерпретаційний комплекс *KI*, який складається з *і*-ї дефініційної частини D_i (*і*=1, 2, ….N),, блоку термінологічних словосполучень K_i та синонімічного блоку $СИН_i$.

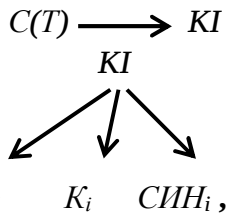


Схема 4. *Структура і-ї інтерпретаційної частини*

Дефініція містить тлумачення і (факультативно) синонім до відповідного термінологічного значення, заданого цим тлумаченням. Блок термінологічних словосполучень (ліва частина блоку словосполучень) K_i містить $УК_i$ – українську частину, $РК_i$ – російську частину, $АК_i$ – англійську частину та (факультативно) синонім до $УК_i$: $СИНУК_i$.. Права частина блоку словосполучень містить тлумачення словосполучення українською мовою. Символічно описана структура набуває такого вигляду:

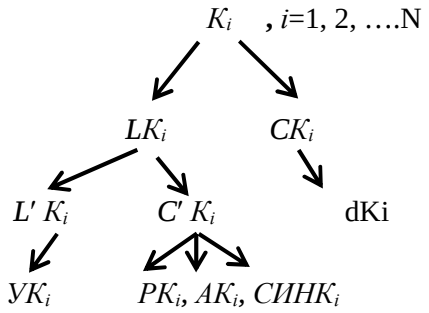


Схема 5. *Складові блоку термінологічних словосполучень*

Розглянемо *приклад № 3* для відображення структури *C(T)*.

V(T)=**непостíйність**, -ості (*рос*. непостоя́нство, *англ*. inconstant) часта зміна своїх поглядів, звичок, ставлення до кого-, чого-небудь; **н**. **соматúчна** (*рос*. непостоя́нство соматúческое, *англ*. somatic inconstancy) коливання кількості хромосом, що спостерігається у хромосомному наборі соматичних клітин багатьох тваринних організмів та людини. Син. **непостíйність хромосóмна**, **анеуплоїдíя соматúчна**.

L(T)= **непостíйність**, -ості (*рос*. непостоя́нство, *англ*. inconstant)

C(T) = часта зміна своїх поглядів, звичок, ставлення до кого-, чого-небудь; **н**. **соматúчна** (*рос*. непостоя́нство соматúческое, *англ*. somatic inconstancy) коливання кількості хромосом, що спостерігається у хромосомному наборі соматичних клітин багатьох тваринних організмів та людини. Син. **непостíйність хромосóмна**, **анеуплоїдíя соматúчна**.

D_i = часта зміна своїх поглядів, звичок, ставлення до кого-, чого-небудь.

K_i = **н**. **соматúчна** (*рос*. непостоя́нство соматúческое, *англ*. somatic inconstancy) коливання кількості хромосом, що спостерігається у хромосомному наборі соматичних клітин багатьох тваринних організмів та людини.

СИНК$_i$ = Син. **непостíйність хромосóмна**, **анеуплоїдíя соматúчна**.

У структурі словникових статей є певні особливості: обмежено подані дієслова в інфінітиві, фонетичні і морфологічні варіанти, синонімія, омонімія. Наприклад, якщо дієслова в інфінітиві в словнику подано обмежено, то для виділення основних блоків та заголовного слова словникову статтю ділимо на дві словникові статті і розкриваємо друге дієслово.

Приклад № 4

репродукувáти, *недок. і док.* (*рос.* репродуци́ровать, *англ.* reproduce) відтворювати, розмножувати рослини і тварин; **-тися** (*рос.* репродуци́роваться, *англ.* reproduce) відтворюватися у процесі розмноження.

Коментар до прикладу № 4

Ця стаття, фактично, складається з двох статей. Перша з них: **репродукувáти**, *недок. і док.* (*рос.* репродуци́ровать, *англ.* reproduce) відтворювати, розмножувати рослини і тварин. Друга: **репродукуватися** (*рос.* репродуци́роваться, *англ.* reproduce) відтворюватися у процесі розмноження. Кожна з цих статей має структуру, описану вище.

Звернімо увагу також на те, що поряд із реєстровим словом у статті можуть бути наведені нормативні фонетичні та морфологічні варіанти у термінологічному блоці; разом вони формують реєстровий ряд словникової статті:

Приклад № 5

одомáшнення, одомáшнювання (*рос*. одомáшнивание; *англ*. domestication) приручення диких тварин та їхнє розведення за вирішального впливу штучного добору для отримання корисної для людини продукції або естетичного задоволення. Син. **доместикáція**.

Коментар до прикладу № 5

У цій словниковій статті реєстровий ряд складається з двох елементів: **одомáшнення** та **одомáшнювання.**

Для виділення ще одного структурного елементу розглянемо терміни-омоніми, тобто слова однакового звучання і написання, але різні за значенням, які подано в реєстрі окремо з цифровими індексами вгорі праворуч.

Приклад № 6

чуб¹, -а (*рос.* хохóл, *англ.* forelock, topknot) жмут шерсті або пір'я на голові деяких тварин.

чуб², -а (*рос*. кисть, *англ*. coma) суцвіття трав'янистих рослин; волоть, китиця.

Коментар до прикладу

У цьому випадку виділяємо дві окремі словникові статті, в яких український термінологічний комплекс має однакове заголовне слово, цифровий індекс виділяємо в окремий структурний елемент *НО* (номер омоніма).

Дефініція ***Di*** містить тлумачення ***di,*** семантичну ремарку *СМРi* та синонім $СИН_i$.

Елементи блоку тлумачень зображено на схемі:

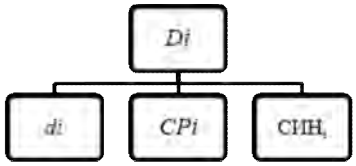


Схема 6*. Представлення структури* ***Di***

Блок термінологічного словосполучення ***Ki*** містить ліву LK_i і праву $СК_i$ частини. Ліва частина містить український комплекс $УК_i$, російську $РК_i$, англійську $АК_i$, частини та синонім $СИНК_i$, права містить інтерпретаційну частину.

Українська, російська та англійська частини, відповідно, складаються з українського (*ТКС_У*), російського (*ТКС_Р*) та англійського (*ТКС_А*) комплексів. Український термінологічний комплекс містить українське термінологічне словосполучення ($ТС^У$) та граматичну ремарку (*ГРС*). Російський термінологічний комплекс містить російське термінологічне словосполучення ($ТС^Р$) та граматичну ремарку (*ГРС*). Англійський термінологічний комплекс містить англійське термінологічне словосполучення ($ТС^А$) та граматичну ремарку (*ГРС*).

Блок інтерпретаційний містить дефініцію і семантичну ремарку.

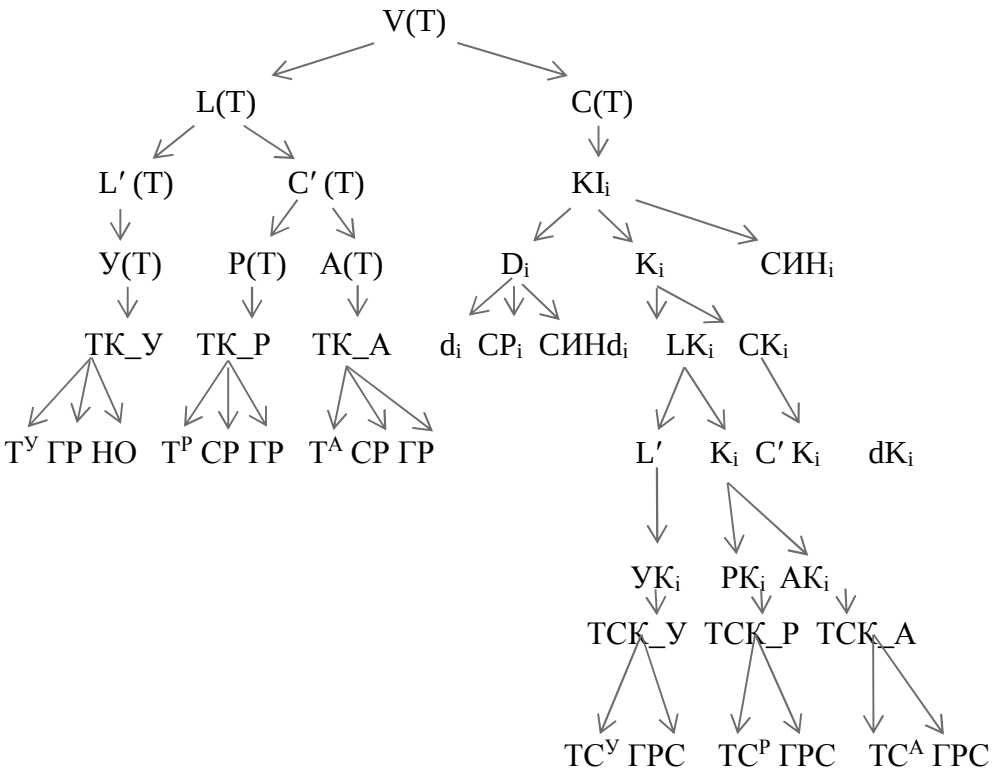


Схема 7*. Загальна схема словникової статті СУБТ*

Технологічні аспекти парсингу Л-систем

У рамках стандартної архітектури Л-систем ASNSI/X3/SPARK задача парсингу словників редукується до створення відповідного програмного забезпечення як елемента внутрішньої моделі, що реалізує парсингові процеси. Втім, проблема розроблення програмного забезпечення, здатного здійснювати парсинг тексту великих за обсягом словників складної структури, досі залишається малодослідженою.

На першому етапі визначаємо формат файла словника. Саме формат визначає технологію оброблення тексту. Якщо текст словника доступний тільки у паперовому вигляді, то першим етапом є переведення його у цифрової форми, що передбачає такі операції, як: сканування словника, розпізнавання тексту та коректура. Але сьогодні більшість словників готують до друку за допомогою комп'ютерних технологій і є доступними у форматі PDF [3] (комунікаційний формат видавничих систем).

Нашим завданням є конвертування тексту словника в XML-документ, що надає можливість експлікувати всі визначені нами структурні елементи та зв'язки між ними. Для автоматичного маркування тексту словника тегами XML було розроблено програму, яка виокремлює елементи структури тексту відповідно до будови Л-системи. Використовують поліграфічні ознаки текстової ідентифікації Л-системи, а саме: межі словникової статті (абзаци), різні спеціальні символи, які відмежовують структурні елементи, позиційні характеристики, зміни мови, шрифтів, регістру літер та ін.

У процесі роботи ми здійснюємо низку перетворень: PDF $\rightarrow$ DOCX $\rightarrow$ XML.

Використовуючи онлайн-конвертори, PDF-файл перетворюємо до формату DOCX. Ми використовуємо DOCX, тому що інструментальні засоби програмування обладнані розвиненими бібліотеками для роботи з цим форматом. Результатом парсингу тексту у форматі DOCX є XML-документ, в якому промарковано всі визначені елементи структури Л-системи СУБТ та зв'язки між ними. Подальша робота щопо організації інструментальних онлайнових систем ґрунтується на XML-документах і може бути виконана автоматично.

Висновки

У цьому дослідженні адаптовано метод лексикографічних систем до лексикографічної системи СУБТ (багатомовного тлумачного термінологічного словника). Проаналізовано поліграфічне оформлення, організацію і структуру друкованого тексту словника з метою ідентифікації елементів концептуальної моделі Л-системи СУБТ. На основі концептуальної моделі будується структура XML-документа, який пропонується використовувати як посередник між паперовою версією словника та його реалізацією як онлайнової лексикографічної системи.

Надалі планується вдосконалення структури XML-документа та побудова універсальної процедури парсингу.

Список літератури

1. Широков В. А. (2018). Эволюция как универсальный естественный закон (Пролегомены к будущей общей теории эволюции). Ч. III. *Бионика интеллекта*. № 1 (90).

2. Olga Karpova. (2009). Lexicography and Terminology: A Worldwide Outlook. Cambridge : Cambridge Scholars Publishing.

3. PDF-конвертер. From: http://pdf2doc.com/.

4. Словник металургійних термінів (грузинсько-російсько-українсько-англо-німецько-французький). (2011). І том. Тбілісі.

5. Словник металургійних термінів (грузинсько-російсько-українсько-англо-німецько-французький). (2011). ІІ том. Тбілісі.

6. Словник металургійних термінів (українсько-грузинсько-російсько-англійсько-німецько-французький). (2014). І том. Київ.

7. Словник металургійних термінів (українсько-грузинсько-російсько-англійсько-німецько-французький.

8. Словник української біологічної термінології. (2012). Київ : КММ.

9. Термінологічний українсько-російсько-англійський словник-довідник зі зварювання. Науково-технічна термінологія. (2013). Київ : Український мовно-інформаційний фонд. Серія: Словники України. [Електронний ресурс. CD]

10. Широков В. А. (Eds.) (2008). Український-російський, Російсько-український словник із зварювання. Київ. [Електронний ресурс. CD].

11. Широков В. А. (Eds.) (2018). Українсько-російсько-англійський словник зі зварювання. Київ. [Електронний ресурс. CD].

12. Широков В.А. *(Eds.)* (2011). Комп'ютерна лексикографія. Київ : Наук. думка.

References

1. Shyrokov V. A. (2018). Evolution as a universal natural law (Prolegomenas to the future general theory of evolution). Part III. Bionics of intelligence. № 1 (90).

2. Olga Karpova. (2009). Lexicography and Terminology: A Worldwide Outlook. Cambridge : Cambridge Scholars Publishing.

3. PDF Converter From: http://pdf2doc.com/.

4. Dictionary of metallurgical terms (Georgian-Russian-Ukrainian-English-German-French). (2011). I tom. Tbilisi.

5. Dictionary of metallurgical terms (Georgian-Russian-Ukrainian-English-German-French). (2011). II tom. Tbilisi.

6. Dictionary of metallurgical terms (Georgian-Russian-Ukrainian-English-German-French). (2014). I tom. Kyiv.

7. Dictionary of metallurgical terms (Georgian-Russian-Ukrainian-English-German-French). (2014). II tom. Kyiv.

8. Dictionary of Ukrainian Biological Terminology. (2012). K. : KMM.

9. Terminological Ukrainian-Russian-English Dictionary-Guide for Welding: Reference Edition; Scientific and technical terminology (2013). Kyiv : Ukrainian Language Information Foundation. Series: Dictionaries of Ukraine. [Electronic resource. CD]

10. Shyrokov V. A. *(Eds.)* (2008). Ukrainian-Russian, Russian-Ukrainian Welding Dictionary. Kyiv. [Electronic resource. CD]

11. Shyrokov V. A. *(Eds.)* (2018). Ukrainian-Russian-English Welding Dictionary. Kyiv. [Electronic resource. CD].

12. Shyrokov V. A. *(Eds.)* (2011). Computer lexicography. Kyiv.: Science opinion.