УДК 81’374.3(038)=134.2

Купріянов Є. В.,
кандидат філологічних наук,
докторант Українського мовно-інформаційного фонду НАН України
E-mail: cuprijanow.eugen@yandex.ua

ЛІНГВІСТИЧНИЙ ІНСТРУМЕНТАРІЙ ВІРТУАЛЬНОЇ ЛЕКСИКОГРАФІЧНОЇ ЛАБОРАТОРІЇ ТЛУМАЧНОГО СЛОВНИКА ІСПАНСЬКОЇ МОВИ

У статті висвітлено головні питання щодо створення лінгвістичного інструментарію віртуальної лексикографічної лабораторії тлумачного словника іспанської мови (ВЛЛ DLE 23). Для цього проаналізовано особливості подання словником різних лінгвістичних фактів, виявлено його структуру та особливості метамови. Розроблено формальну модель DLE 23, схарактеризовано її елементи та можливі зв’язки між ними, які мають бути доступними через лінгвістичний інструментарій. Окреслено коло дослідницьких завдань на базі словника, виконання яких має забезпечувати лінгвістичний інструментарій.

Ключові слова: *комп’ютерна лексикографія, віртуальна лексикографічна лабораторія, цифрові середовища, електронні словники.*

Одним із головних завдань сучасної комп’ютерної лексикографії є оновлення та підтримка фундаментальних лексиконів – великих паперових тлумачних словників – у цифровому середовищі. Комп’ютерна лексикографія успішно його розв’язує завдяки поєднанню багатовікового досвіду традиційного словникарства з останніми комп’ютерними технологіями. Результатом поєднання є віртуальні лексикографічні лабораторії (ВЛЛ), тобто системи, що уможливлюють як оперування словниковим матеріалом, так і проведення низки лінгвістичних досліджень.

На сьогодні в Українському мовно-інформаційному фонді розроблено цілу низку ВЛЛ, зокрема для тлумачного словника сучасної української мови (20-томний), граматичного, синонімічного, антонімічного та етимологічного словників. Аналогічні лабораторії створено для тлумачних словників російської та турецької мов. Всі ці системи є інструментальними; вони розташовані на сайті Українського мовно-інформаційного фонду *https://lcorp.ulif.org.ua/*, де функціонують у корпоративному режимі. Не менш важливим є створення ВЛЛ для тлумачного словника іспанської мови (Diccionario de la lengua española. Edición del tricentenario [5], далі за текстом DLE 23) 23 видання. І ця важливість зумовлена потребою подальшого розвитку й апробації виробленої мовно-інформаційним фондом [2–4] теоретичної бази на матеріалі тлумачного словника, побудованого на інших принципах укладання. Крім цього, висвітлені в нашій роботі напрацювання є підґрунтям для розробляння інструментарію, який, у свою чергу, дає користувачеві змогу побудувати власні інструменти для розв’язання лінгвістичних завдань на основі тлумачного словника.

Виходячи зі сказаного, метою цієї роботи є розгляд окремих питань розроблення лінгвістичного інструментарію для ВЛЛ DLE 23, яка сьогодні перебуває на етапі технічної реалізації. Для цього необхідно: 1) виявити лінгвістичні особливості мовних одиниць, репрезентованих у DLE 23; 2) за методологією теорії лексикографічних систем проаналізувати її β-структури та σ-зв’язки між ними; 3) схарактеризувати лінгвістичний інструментарій та виокремити дослідницькі завдання, які він має забезпечувати. Об’єктом розвідки є DLE 23, а предметом – створення лінгвістичного інструментарію, що надає широкі можливості для дослідження граматичних, семантичних, прагматичних та інших особливостей іспанських мовних одиниць. ВЛЛ – цифрове середовище, де словник представлено як «мовно-інформаційний об’єкт, орієнтований на реалізацію комплексного інформаційного опису лексико-граматичних структур певної мови або сукупності мов» [1, с. 359]. На відміну від електронних словників, у тому числі он-лайнових, ВЛЛ пропонує програмний інтерфейс для виконання:

1) *функцій адміністрування доступу*: авторизація та ідентифікація користувачів; додавання та видалення нових користувачів; керування правами доступу (лише читання, читання та редагування словникового матеріалу);

2) *лексикографічних робіт*: редагування словникових статей; створення низки похідних словників на основі тлумачного словника; репрезентація словникових статей у будь-якому форматі;

3) *науково-дослідних робіт*: дослідження на певному рівні мови, представленому в тлумачному словнику (граматика, включаючи словотвір; лексика, зокрема семантика, прагматика); дослідження на стику мовних рівнів: граматики і семантики, словотвору та семантики, семантики і прагматики тощо. У цій статті мова піде про створення інструментарію для робіт, указаних у третьому пункті.

Створення ВЛЛ DLE 23 потребує певної теоретичної бази. З одного боку, її складають наукові здобутки провідних іспанських лексикографів (Х. Касарес, М. А. Ескерра, Л. Ф. Лара, К. Мальдонадо, М. Секо, А. М. Медіна-Геррра та інші), а із другого – теорія опису мовної системи, орієнтованого на застосування в лінгвістичних технологіях – інформації в текстових масивах, систем машинного перекладу, автоматичного аналізу контенту тощо. Перш ніж перейти до розгляду зазначеної проблеми, хочемо наголосити на деяких лінгвістичних особливостях, зафіксованих у DLE 23:

– частиномовне варіювання: іспанські слова можуть одночасно належати до групи як повнозначних слів, так і службових, маючи при цьому ту саму форму;

– залежність лексичного значення від граматичного. Так, наприклад слово *bien* як іменник однини та множини означає «добро», «благо», «власність», «товари»; як прислівника – «дуже», «доволі», «належним чином», «із радістю» і т. д.;

– можливість окремих іспанських слів реалізовувати своє лексичне значення лише за наявності певних граматичних значень. Так, слово *cómico* може перебувати у лексичному значенні «мультфільм» та «комікс» лише у граматичному стані іменника жіночого роду;

– наявність лінгвопрагматичних особливостей вживання слів у певному значенні, зокрема предметна галузь, географічний ареал та застарілість;

– здатність мовних одиниць утворювати синонімічні пари одна з одною (у певному лексичному значенні або певних значеннях).

Ми подали найбільш показові лінгвістичні характеристики, враховані під час розроблення концептуальної моделі *DLE* 23, що встановлює, до яких елементів структури словника має забезпечуватися доступ через розроблюваний інструментарій. Наведемо її узагальнений вигляд:

$$\{I(D), V\big(I(D)\big), \beta, \sigma[\beta], Red[V\big(I(D)\big)]\}$$

де символом D позначено словник *DLE* 23; $I(D) = \{x_i\}$ – множину реєстрових одиниць, представлених у словнику; $V(I(D)) = \{V(x_i)\}$ – множина словникових описів, тобто словникових статей; β – множина структур, виокремлених на $V(I(D))$ шляхом аналізу тексту словника; $\sigma[\beta]$ – окрема структура, породжувана оператором σ на β; обмеження дії оператора σ на $V(x_i)$ породжує мікроструктуру словникової статті $\sigma[x_i]$; $Red[V(I(D))]$ – механізм рекурсивної редукції, що дає змогу виявити більш тонкі структурні елементи словника.

До β-структур ми зараховуємо: $\beta_1(x_i)$ – реєстрове слово x_i, $\beta_2(x_i)$ – реєстровий ряд, $\beta_3(x_i)$ – дублети, $\beta_4(x_i)$ – етимологія, $\beta_5(x_i)$ – словозмінні особливості, $\beta_6(x_i)$ – орфографічні особливості, $\beta_7(x_i)$ – блок тлумачень. Зауважимо, що в межах словникових статей також подано словосполучення і фразеологізми. Але ми надалі розглядатимемо їх як реєстрові одиниці, що мають власну словникову статтю. Зазначимо також, що для певних одиниць зміст окремих β-структур може бути відсутнім, тобто $\beta_j(x_i) = \varnothing$. Але в будь-якому випадку вони присутні на $\{V(x_i)\}$. Наведемо, як приклад, словникові статі *cómico* та *inmaculado*:

cómico, ca. (Del lat. comĭcus, y este del gr. κωμικός kōmikos). adj. **1.** Que divierte y hace reír. *Situacion comica*. || **2.** Perteneciente o relativo a la comedia. || **3.** Dicho de un actor: Que representa papeles cómicos. U. t. c. s. || **4.** Dicho de un autor antiguo: Que escribía comedias. U. t. c. s. ● m. y f. **5. comediante** (|| actor). ○ f. **6.** *Pan*. **historieta** (|| serie de dibujos). U. m. en pl. || **7.** *Pan.* **dibujos animados**.

inmaculado, da. (Del lat. immaculātus. ♦ Escr. con may. inicial en acep. 2). adj. **1.** Que no tiene mancha. ● f. **2.** por antonom. La Virgen María. *La Inmaculada*.

Покажемо для них у Табл. 1 зміст β-структур. Зауважимо, що їх виокремлення ґрунтується на особливостях метамови DLE 23, що чітко встановлює правила репрезентації кожного елемента словникової статті. За основу взято друковану версію, але для більш точного виявлення елементів також використано он-лайнову версію.

Таблиця 1

Приклади змісту $\beta_i(x_i)$-структур

	$V(x_i)$, x_i = **cómico**	$V(x_i)$, x_i = **inmaculado**
$\beta_1(x_i)$	cómico	in?maculado
$\beta_2(x_i)$	cómico, cómica	inmaculado, inmaculada
$\beta_3(x_i)$	∅	∅
$\beta_4(x_i)$	Del lat. comĭcus, y este del gr. κωμικός kōmikos	Del lat. immaculātus
$\beta_5(x_i)$	∅	∅
$\beta_6(x_i)$	∅	Escr. con may. inicial en acep. 2
$\beta_7(x_i)$	adj. **1.** Que divierte y hace reír […].	adj. **1.** Que no tiene mancha […].

Окремі $\beta_j(x_i)$ здатні розкладатися на більш дрібні структури в результаті дії механізму рекурсивної редукції. Зокрема, до таких належать $\beta_3(x_i)$, $\beta_4(x_i)$ та $\beta_7(x_i)$, як це видно у Табл. 2.

Таблиця 2

Підструктури на $\beta_j(x_i)$

$\beta_j(x_i)$	Назва структури	Підструктура	Назва підструктури
$\beta_3(x_i)$	дублети	$\beta_3^{DUP}(x_i)$	форма дублета
		$\beta_3^{CHAR}(x_i)$	характеристика
$\beta_4(x_i)$	етимологія	$\beta_4^{LANG}(x_i)$	мова походження
		$\beta_4^{ETYM}(x_i)$	форма етимона
$\beta_7(x_i)$	блок тлумачення	$\beta_7^{GRAM}(x_i)_k$	граматична характеристика
		$\beta_7^{LEX}(x_i)_{k[p]}$	лексична характеристика
$\beta_7^{LEX}(x_i)_{k[p]}$	лексичний підблок	$\beta_7^{LEX}(x_i)^{PRAGM}$	прагматика
		$\beta_7^{LEX}(x_i)^{SEM}$	тлумачення

Блок тлумачень $\beta_7(x_i)$ може розкладатися на кілька $\beta_7^{GRAM}(x_i)_k$, де k позначає індекс граматичного значення, репрезентованого в DLE 23, та $\beta_7^{LEX}(x_i)_{k[p]}$, де p – індекс лексичного значення (у словнику відображається як номер тлумачення), відповідного до граматичного з індексом k. Лексичний підблок також дає підструктури $\beta_7^{LEX}(x_i)^{PRAGM}$ – прагматичні особливості уживання реєстрового слова в значенні $\beta_7^{LEX}(x_i)^{SEM}$. Наведемо зміст підструктур на прикладі словникової статті *cómico*:

Таблиця 3

Приклади змісту структур та підструктур

Структура	Зміст структури	Підструктура	Зміст підструктури
β_3(*cómico*)	∅	β_3^{DUP}(*cómico*)	∅
		β_3^{CHAR}(*cómico*)	∅
β_4(*cómico*)	Del lat. comĭcus, y este del gr. κωμικός kōmikos	β_4^{LANG}(*cómico*)	lat.
		β_4^{ETYM}(*cómico*)	comĭcus, y este del gr. κωμικός kōmikos
β_7(*cómico*)	adj. **1.** Que divierte y hace reír. *Situacion comica*. || **2.** Perteneciente o relativo a la comedia. || **3.** Dicho de un actor: Que representa papeles cómicos. U. t. c. s. || **4.** Dicho de un autor antiguo: Que escribía comedias. U. t. c. s. ● m. y f. **5. comediante** (|| actor). ○ f. **6.** *Pan.* **historieta** (|| serie de dibujos). U. m. en pl. || **7.** *Pan.* **dibujos animados**.	$\beta_7^{GRAM}(cómico)_1$	adj.
		$\beta_7^{LEX}(cómico)_{1[1]}$	**1.** Que divierte y hace reír. *Situacion comica*.
		$\beta_7^{LEX}(cómico)_{1[2]}$	**2.** Perteneciente o relativo a la comedia.
		$\beta_7^{LEX}(cómico)_{1[3]}$	**3.** Dicho de un actor: Que representa papeles cómicos. U. t. c. s.
		$\beta_7^{LEX}(cómico)_{1[4]}$	**4.** Dicho de un autor antiguo: Que escribía comedias. U. t. c. s.
		$\beta_7^{GRAM}(cómico)_2$	m. y f.
		$\beta_7^{LEX}(cómico)_{2[5]}$	**5. comediante** (|| actor).
		$\beta_7^{GRAM}(cómico)_3$	f.
		$\beta_7^{LEX}(cómico)_{3[6]}$	**6.** *Pan.* **historieta** (|| serie de dibujos). U. m. en pl.
		$\beta_7^{LEX}(cómico)_{3[7]}$	**7.** *Pan.* **dibujos animados**.

Лексичні підблоки розкладаються на прагматичний та семантичний компоненти. Прагматичний компонент у DLE 23 представлено групою ремарок, що характеризують стилістичні, дискурсивні, галузеві, регіональні та частотні особливості уживання заголовкового слова в поданому значенні. Семантичним компонентом є власне тлумачення. Так, у двох останніх лексичних підблоках прагматичний компонент представлено географічною ремаркою *Pan.* (Панама), а семантичний – тлумаченнями у вигляді синонімів: '**historieta** (|| serie de dibujos)' та '**dibujos animados**'. Виокремлені структури $\beta_j(x_i)$, а також їхні підструктури об'єднуються між собою за допомогою σ-зв'язків, утворюючи тим самим більш складні структури $\sigma[\beta_j(x_i)]$. У DLE 23 наявні зв'язки, що об'єднують:

1) σ_0: структури $\beta_j(x_i)$ одного типу. Наприклад, $\sigma_0[\beta_1(x_i)]$ утворює словниковий реєстр, а $\sigma_0[\beta_7(x_i)]$ – масив тлумачень усіх реєстрових одиниць тощо;

2) σ_1: підструктури в межах $\beta_j(x_i)$. Так, $\sigma_1[\beta_7^{GRAM}(x_i)_k, \beta_7^{LEX}(x_i)_{k[p]}]$ об'єднує підблоки лексичного значення мовної одиниці відповідно до граматичного;

3) σ_2: $\beta_j(x_i)$ за певним типом метаданих. Наприклад, за цим відношенням можна отримати усі $\beta_7(x_i)$, що містять лише родо-видові тлумачення.

4) σ_3: відображають відношення між різними $\beta_j(x_i)$, такі як: словотвір – семантика, лексика – фразеологія, словоформа – синтаксис (утворення похідних словосполучень), словоформа – словозміна (для дієслів).

Виходячи з вищесказаного, розроблюваний лінгвістичний інструментарій має забезпечувати користувачеві доступ не лише до структур $\beta_j(x_i)$ та їх підструктур, а й також можливість їх комбінації шляхом активування відповідних σ-зв'язків. Розроблюваний лінгвістичний інструментарій – програмний комплекс, що пропонує користувачеві низку програмних інтерфейсів для виконання на рівні:

1) словникового реєстру (структура $\beta_1(x_i)$), для чого передбачаються такі режими відображення: а) мовний інвентар: лексеми (належні як до питомої, так і запозиченої лексики), службові слова (прийменники, сполучники, частки тощо), словотвірні елементи (суфікси, префікси тощо), абревіатури; б) реєстр слів, що можна відображати за певними параметрами: «ціле слово», «починаються з», «закінчуються», «точний збіг» та «містять»;

2) окремих структур $\beta_j(x_i)$ або підструктур, що відображають певну лінгвістичну інформацію, параметри відображення якої можуть бути: а) загальними, що вказують на відсутність / наявність змісту однієї або кількох $\beta_j(x_i)$; б) специфічні параметри показу змісту для окремої $\beta_j(x_i)$, встановлювані користувачем (наприклад, тип та/або кількість дефініцій, мова походження мовних одиниць);

3) взаємозв'язків між підструктурами $\beta_7(x_i)$, що репрезентують різні мовні рівні у словнику (граматика, словотвір, лексика): а) залежність лексичного значення від граматичного ($\beta_7^{GRAM}(x_i)_k \rightarrow \beta_7^{LEX}(x_i)_{k[p]}$); б) залежність лексичного значення від прагматики ($\beta_7^{LEX}(x_i)^{PRAGM} \rightarrow \beta_7^{LEX}(x_i)^{SEM}$); в) співвідношення словотвору та семантики (морфосемантика);

4) системних відношень між реєстровими одиницями, індукованих структурою тлумачного словника. Наприклад: заголовкове слово – синоніми, заголовкове слово – гіпоніми та інші.

Лінгвістичні інструменти передбачається розподілити по відповідних вкладинках (Табл. 4). На кожній користувач може «підключати» різні β, а також задавати параметри для σ-зв'язків на β за рахунок логічних операцій «та», «або», «ні».

Таблиця 4

Лінгвістичний інструментарій

Вкладинка	$\boldsymbol{\beta}$	**σ**	**Параметри σ[β], що задають через інструментарій**
Реєстр та реєстровий ряд	$\beta_1(x_i)$	σ_0	Мовна одиниця (вводить користувач)
		σ_2	Тип мовної одиниці (слово, префікс, суфікс)
		σ_2	Характер походження (питоме, запозичене)
			Омонімічність
			Номер омоніма (вводить користувач)
		σ_3	Наявність словозмінної парадигми (для дієслів)
	$\beta_2(x_i)$	σ_0	Наявність реєстрового ряду для мовної одиниці
Дублети	$\beta_3(x_i)$	σ_0	Наявність дублетів
		σ_2	Дублет із характеристиками
			Дублет без характеристик
Етимони	$\beta_4(x_i)$	σ_0	Наявність / відсутність етимона
		σ_2	Мова (походження)
			Значення етимону
Словозміна та орфоепія	$\beta_5(x_i)$	σ_0	Наявність / відсутність словозміни
	$\beta_6(x_i)$	σ_0	Наявність / відсутність орфоепії
Граматичне значення	$\beta_7(x_i)$	σ_2	Частина мови
			Граматична категорія
		σ_2	Варіація граматичного значення
Прагматика		σ_0	Наявність / відсутність прагматичних характеристик
		σ_2	Тип прагматичної характеристики
Лексичні значення		σ_0	Наявність / відсутність дефініції
		σ_1	Прив'язка до граматичного значення
			Прив'язка до прагматики
		σ_2	Тип дефініції
		σ_0	Набір слів, що може містити дефініція
		σ_2	Повний / неповний збіг
		σ_3	Формула тлумачення (для похідних слів)
		σ_2	Тип мовної одиниці (моносемічна, полісемічна)
			Кількість значень (вводить користувач)
Словосполучення та фразеологізми	$\beta_1(x_i)$	σ_3	Наявність / відсутність словосполучень
		σ_3	Наявність / відсутність фразеологізмів

Наприклад, якщо користувач досліджує слова із суфіксом *-aje*, що позначають процес та результат, то в цьому випадку набір параметрів матиме такий вигляд: $\sigma_0[\beta_1(x_i)]$ = «aje» AND $\sigma_2[\beta_1(x_i)]$ = «suf.» AND $\sigma_3[\beta_1(x_i), \beta_7(x_i)]$ = «Acción y efecto de + X». Результатом установлення цих параметрів є відображення: *etiquetaje*, *embalaje*, *montaje* і т. д. Можна також відобразити групу слів, суфікс яких позначає процес та результат: $\sigma_2[\beta_1(x_i)]$ = «suf.» AND $\sigma_3[\beta_1(x_i), \beta_7(x_i)]$ = «Acción y efecto de + X». Тоді буде показано слова із суфіксами *-ado* (*lavado*, *peinado*), *-aje* (*etiquetaje*, *embalaje*), *-ión* (*cubrición*, *gestión*).

Отже, великі тлумачні словники в цифровій формі, до яких належить *DLE 23*, пропонують широкі можливості не лише для доступу та навігації по різних структурних елементах, а й для інтеграції різних мовних фактів у єдиному об'єкті. Це досягається за рахунок формалізації структури текстового масиву тлумачного словника у вигляді β-структур та σ-зв'язків для його переведення до комп'ютерної бази даних. Виокремлені в DLE *23* β-структури є тими частинами тексту, до яких буде забезпечено доступ у базі даних через розроблюваний лінгвістичний інструментарій ВЛЛ DLE *23*, а σ установлюють параметри доступу до β та їх комбінації, причому доступ забезпечуватиметься як до текстових, так і метатекстових елементів словника. Саме це визначає перевагу ВЛЛ перед системами, орієнтованими на мову розмічення. Перспективою нашого дослідження є проблеми інтеграції двох компонентів мови, – словника і граматики – в одному цифровому просторі.

Список використаної літератури

1. Остапова И. В. Лексикографическая структура этимологического словаря и его представление в цифровой среде / Остапова И. В. // Компьютерная лингвистика и интеллектуальные технологии: по мат. Междунар. науч. конф. «Диалог 2009». – Вып. 8 (15). – М. : РГГУ, 2009. – С. 359–364.

2. Широков В. А. Комп'ютерна лексикографія / В. А. Широков. – К. : Наукова думка, 2011. – 351 с.

3. Широков В. А. Лінгвістичні та технологічні основи тлумачної лексикографії / В. А. Широковта ін. – К. : Довіра, 2010. – 295 с.

4. Широков К. В. Іменна словозміна у сучасній турецькій мові / К. В. Широков. – К. : Довіра, 2009. – 318 с.

5. Diccionario de la lengua española : ed. 23. – Madrid : Espasa Calpe, 2014. – 2432 p.

Куприянов Е. В. Лингвистический инструментарий виртуальной лексикографической лаборатории толкового словаря испанского языка.

Аннотация

В статье описаны основные проблемы создания лингвистического инструментария виртуальной лексикографической лаборатории толкового словаря испанского языка (ВЛЛ DLE 23). Для этого проанализированы особенности представления в словаре разных лингвистических фактов, выявлено его структуру и особенности метаязыка. Разработано формальную модель DLE 23, охарактеризованы ее элементы и возможные связи между ними, которые должны быть доступными через лингвистический инструментарий.

Ключевые слова: *компьютерная лексикография, виртуальная лексикографическая лаборатория, цифровые среды, электронные словари.*

Kuprianov Ye. V. Linguistic tools for virtual lexicographic laboratory of Spanish explanatory dictionary.

Summary

The present article is devoted to the problems of creating linguistic tools for the virtual lexicographic laboratory of Spanish explanatory dictionary (VLL DLE 23). The goal of the research is to consider some issues related to the development of linguistic tools for VLL DLE 23. The object is VLL DLE 23 under development.

To achieve this goal the dictionary was analyzed for the peculiarities of linguistic facts representation, its structure and metalanguage. On the basis of the dictionary analysis the formal model of DLE 23 was developed and its main components, including their relationships, to be made available via linguistic tools for accessing linguistic information were determined. The range of research activities to be performed by using the linguistic tools was outlined.

Unabridged monolingual dictionaries, among them DLE 23, in digital format are found to be powerful research environment facilitating the navigation and access to their structural elements and integration of language facts in one object. This can be achieved by formalizing the structure of dictionary text in a form of β-structures and σ-links for converting the dictionary into database.

The prospect of our research is related to the problem of integrating two main components of the language, – vocabulary and grammar, – in one digital environment.

Key words: *computer lexicography, virtual lexicographic laboratory, digital environments, electronic dictionaries.*