ПРИКЛАДНА ЛІНГВІСТИКА. ОСВІТНІ КОМП'ЮТЕРНІ ТЕХНОЛОГІЇ

УДК 81'33

ТЛУМАЧНИЙ СЛОВНИК ІСПАНСЬКОЇ МОВИ ЯК ІНСТРУМЕНТ ДЛЯ ЛІНГВІСТИЧНИХ ДОСЛІДЖЕНЬ

Євген КУПРІЯНОВ (Харків, Україна)

e-mail: eugeniokuprianov@gmail.com

КУПРІЯНОВ Євген. ТЛУМАЧНИЙ СЛОВНИК ІСПАНСЬКОЇ МОВИ ЯК ІНСТРУМЕНТ ДЛЯ ЛІНГВІСТИЧНИХ ДОСЛІДЖЕНЬ

У статті висвітлено основні питання створення інструментарію для проведення лінгвістичних досліджень на базі тексту тлумачного словника іспанської мови. Для цього проаналізовано особливості подання словником різних лінгвістичних фактів, виявлені структура словникової статті і особливості його метамови. Розроблено формальну модель DLE 23, схарактеризовано її головні елементи і зв'язки між ними, які мають бути доступні для користувача. Визначено функції інтерфейсу для проведення лінгвістичних досліджень.

Ключові слова: тлумачний словник, комп'ютерна лексикографія, Л-система, іспанська мова, лінгвістичний інструментарій.

KUPRIYANOV Yevhen. EXPLANATORY DICTIONARY OF THE SPANISH LANGUAGE AS A TOOL FOR LINGUISTIC RESEARCHES

The present article covers the main issues of creating tools for conducting linguistic research based on the text of the explanatory dictionary of the Spanish language. For this purpose, the peculiarities of its presentation of various linguistic facts are investigated, the structure of the dictionary entry and the peculiarities of its metalanguage are revealed. A formal model DLE 23 has been developed, its main elements and the main connections between them, which will be available to the user, are described. Interface functions for conducting linguistic research are defined. Based on the theory of L-systems by V.A. Shirokov the formal model of the lexicographic structure of the dictionary DLE 23 (L-system of DLE 23) has been represented as an object containing a set of language units, a set of lexicographic descriptions of the units, the structures comprising lexicographic descriptions (в-structures) and the links by means of which the structures are interrelated with each other (y-links). The formal model served as a basis for creating Virtual lexicographic laboratory VLL DLE 23, in particular linguistic interface. The main functions to be provided by the interface are: 1) conducting researches on different levels of DLE 23 L-systems, and working with separate в-structures and y-links; 2) by applying different y-links to в-structures a comprehensive information of a unit is possible to be got (for example, headword etymology, language of origin, relationships between grammatical and lexical meanings ets.); 3) integrate different linguistic facts in a single information object.

Key words: explanatory dictionary, computer lexicography, L-system, Spanish, linguistic tools.

Постановка проблеми. Великі, здебільшого багатотомні лексикони, містять основну частину національної лексики та фразеології й характеризуються докладним описом лексико-граматичної і лексико-семантичної системи мови. Завдяки великому обсягу, розпрацьованості структури та повноті лексикографічного опису такі словники є носіями величезної кількості імпліцитно заданих лінгвістичних, когнітивних, логічних та інших зв'язків і відношень (переважно неконтрольованих), що перетворює ці великі лексиавкографічні системи на певного роду «речі в собі».

Постає питання про розроблення методології і технології створення такого роду лексикографічних об'єктів, а також дослідження різноманітних ефектів, що явно або неявно функціонують у них. Зауважимо, що від самого початку йдеться про методи комп'ютерної лінгвістики, адже, як наголошується в книзі «Комп'ютерна лексикографія», традиційними методами виконати такі дослідження неможливо вже просто фізично. Отже, найпершою проблемою тут є створення цифрових аналогів відповідних традиційних лексикографічних праць або переведення їх до цифрової форми. Але створення такого інструментарію потребує відповідої теоретичної бази для виявлення, опису та репрезентації відповідних лінгвістичних даних із тексту аналізованого словника.

Аналіз останніх досліджень і публікацій. Проблеми створення лінгвістичного інструментарію на базі тлумачних словників проводили на матеріалі Словника української мови (СУМ 11) і Словника української мови в 20 томах (СУМ 20). Окремі дослідження проведено на словниках російської (Т. П. Любченко) та турецької (К. В. Широков) мов. Зазначені словники було переведено в електронну форму та побудовано для них відповідний інтерфейс, оскільки окремі мовні факти без цього виявити у «паперовій» доволі важко. Детальний опис таких закономірностей, зокрема в СУМ 11 та СУМ 20, детально висвітлено в працях [1–5].

Виділення раніше невирішених раніше частин загальної проблеми. У цьому відношенні вдалося розвинути як теоретичний апарат, так і технологічні засади лексикографічних систем, перетворивши їх на надійний інструмент сучасної лексикографії [1]. Але автоматичне застосування отриманих теоретичних напрацювань до іспанської мови неможливе. Це пояснюється, перш за все, традиціями іспанської лексикографічної школи щодо організації самого словника та подання лінгвістичної інформації.

Мета статті. Виходячи з вищесказаного, мета нашої розвідки – розглянути важливі питання методики виявляння лінгвістичних даних із тексту тлумачного словника іспанської мови (надалі за текстом *DLE 23*). Для цього необхідно: 1) проаналізувати структуру, параметри і зміст *DLE 23*; 2) побудувати формальну модель лексикографічної структури *DLE 23*, використовуючи апарат теорії лексикографічних систем (Л-систем) В. А. Широкова; 3) виокремити та схарактеризувати об'єкти, «індуковані» Л-структурою *DLE 23*; 4) виявити лінгвістичні факти, щом містять виокремлені об'єкти; 5) визначити функції інтерфейсу, що уможливлюють проведення лінгвістичних досліджень.

Виклад основного матеріалу. Словник іспанської мови *DLE 23* є фундаментальною лексикографічною працею, що містить літературну лексику, широко вживану як в Іспанії, так і країнах Латинської Америки. Відповідно його мета – не лише розкрити значення мовної одиниці, а й граматичні, синтаксичні та прагматичні особливості, притаманні їй в тому чи іншому значенні в іспаномовних країнах. Словникову статтю умовно розділено на ліву і праву частини. Як показано на рисунку 1, елементами статті є: 1) ліва частина, яку складають заголовкове слово та інформаційний блок у дужках; 2) права частина, що містить блок тлумачень, блок колокацій та посилання на інші словникові статті.

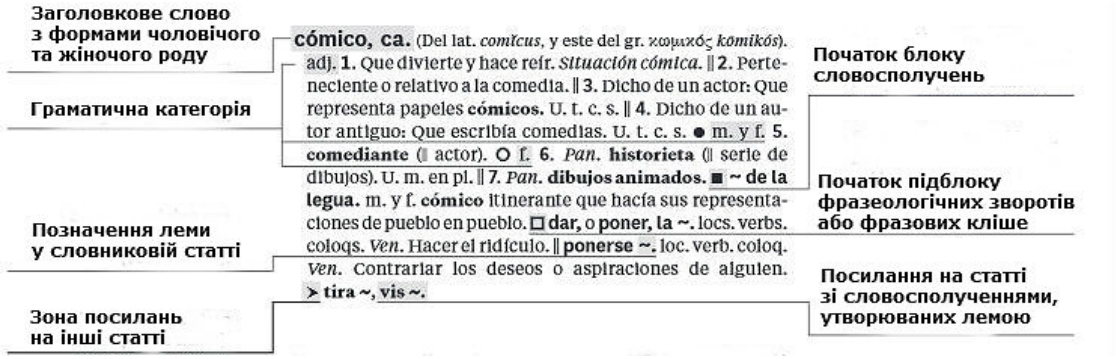

Рис. 1. Приклад словникової статті

Ліву частину складають реєстре слово, реєстровий ряд, інформаційний блок Другим після заголовкового слова іде інформаційний блок що містить дублети етимологічну довідку, орфографічну характеристику та словотвірні особливості:

preterir. (Del lat.praeterire 'pasar adelante'. ♦ Conjug. c.*pedir*. ♦ U. solo las formas cuya desinencia empieza por -*i*).

Права частина розкриває змістові особливості заголовкової частини. Головним та обов'язковим елементом є тлумачення, які групуються відповідно до граматичного значення заголовкового слова. У випадку зі словами, що мають частиномовну варіативність або можуть набувати різних граматичних категорій, тлумачення, що відповідають кожній частині мові (граматичній категорії):

estudiante […] adj. **1.** Que estudia. U. m. c. s. • m. y f. **2.** Persona que cursa estudios en un establecimiento de ensecanza. • m. **3.** Hombre que ayudaba a los actores a estudiar los papeles.

Наприкінці тексту дефініції можуть бути також подані додаткові коментарі у вигляді речень, у яких слова скорочено до першої літери. Від тексту тлумачення відокремлюються крапкою:

gracia. [...] f. **1.** Cualidad o conjunto de cualidades que hacen agradable a la persona o cosa que las tiene. U. t. en sent. fig.

Так, коментар «u. t. en sent. fig.» розшифровується іспанською як «usado tambiйn en sentido figural» (вживають також у переносному значені). Показовим для тлумачного

словника іспанської мови є подання в коментарях додаткових граматичних характеристик слова, наявних у нього в тому чи іншому значенні:

cabryn. [...] adj. [...] || **2.** malson. coloq. Dicho de un hombre: Que padece la infidelidad de su mujer, y en especial si la consiente. U. t. c. s. m.

Так, «U. t. c. s. m.» – usado tambiйn como sustantivo masculino – вказує, що слово також уживають як іменник чоловічого роду.

Текст *Diccionario de la lengua espacola* [6], розглядуваний згідно з нашою методикою, надає необхідний матеріал для побудови лексикографічної структури. У структурі словникової статті виокремлюємо множину реєстрових одиниць $W = \{x\}$, які слугують ідентифікаторами відповідних словникових статей $V(x)$. До складу словникового реєстру входять як саме слова, так і морфеми, певні словосполучення та абревіатури. У кожній словниковій статті $V(x)$ виокремлюється «ліва частина» $L(x)$, яка складається із певних параметрів заголовкового слова, і «права частина» – $P(x)$, у якій подається лексикографічне представлення семантики x.

У випадку тлумачного словника розрізняємо два типи мовних одиниць: одиниці лексичного рівня та словосполучення (до складу яких входить заголовкове слово), яким у мові надається ідіоматичний статус. Тому природньо представити структуру словникової статті $V(x)$ у вигляді об'єднання описів (словникових статей) структурних одиниць обох типів:

$$V(x) \equiv V^{Lex}(x) \cup [\bigcup_{i}^{n(x)} \bigcup_{j}^{m(i)} V_i^{jFras}(x)],$$

де $V^{Lex}(x)$ – лексикографічний опис заголовкового слова x; $V_i^{jFras}(x)$ – опис j-го словосполучення i-го типу; $m(i)$ – кількість словосполучень i-го типа, а $n(x)$ – кількість типів словосполучень у $V(x)$. Кожному лексикографічному – і $V^{Lex}(x)$, і $V_i^{Fras}(x)$ – ставиться у відповідність базова структура: $V \equiv (L_0 ; P_0)$. У випадку $V = V^{Lex}(x)$ в роли L_0 выступає заголовкове слово словникової статті з відповідними параметрами (які далі ми назвивемо параметрами заголовкового слова). Для $V_i^{Fras}(x)$ L_0 – словосполучення в реєстровій словниковій формі плюс параметри заголовкової одиниці. Структура правой части P_0 ідентична для лексеми і словосполучення.

Для побудови формальної моделі лексикографічної структури словника *DLE 23*, враховуючи його особливості, ми спиралися на теорію Л-систем В. А. Широкова [5], а якою будь-який словник можна представити як:

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

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; $в$ – множина структур, виокремлених на $V(I(D))$ шляхом аналізу тексту словника; у[$в$] – окрема структура, породжувана оператором у на $в$; обмеження дії оператора у на $V(x_i)$ породжує мікроструктуру словникової статті у[x_i]; $Red[V(I(D))]$ – механізм рекурсивної редукції, що дає змогу виявити більш тонкі структурні елементи словника. У свою чергу, множину лексикографічних описів кожної одиниці $x_i \in I(D)$, можна розкласти на кілька підмножин (рисунок 1).

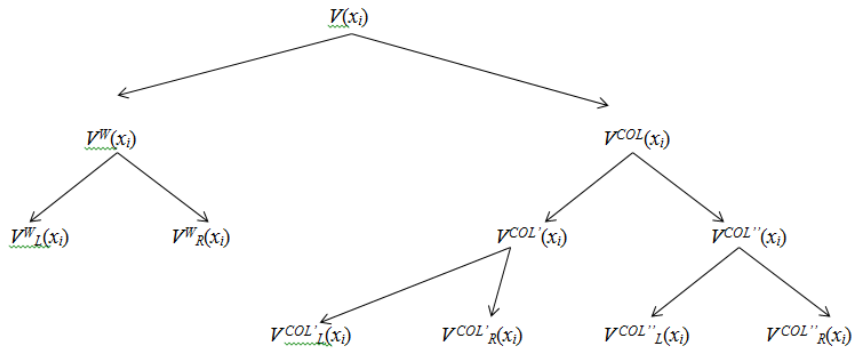

Рис. 2. Множина лексикографічних описів та її підмножини.

Через $V^W(x_i)$ та $V^{COL}(x_i)$ позначено підмножину описів, відповідно, заголовкової одиниці та утворюваних від неї колокацій; $V^W{}_L(x_i)$ та $V^W{}_R(x_i)$ відповідають «лівій» і «правій» частинам.

Показовим для DLE 23 є розділ $V^{COL}(x_i)$ на дві підмножини, що містять опис колокацій: а) типу «іменник + прикметник» $V^{COL'}(x_i)$; б) інших колокацій $V^{COL''}(x_i)$ – віддієслівних, відприслівникових, відприйменникових тощо. «Ліва» та «права» частини також передбачено для $V^{COL'}{}_L(x_i)$ та $V^{COL'}{}_R(x_i)$, $V^{COL''}{}_L(x_i)$ та $V^{COL''}{}_R(x_i)$. Кожна з виокремлених множин $V^W{}_L(x_i)$ та $V^W{}_R(x_i)$; $V^{COL'}{}_L(x_i)$ та $V^{COL'}{}_R(x_i)$; $V^{COL''}{}_L(x_i)$ та $V^{COL''}{}_R(x_i)$ характеризується певним набором *в*-структур, що містять певний елемент лексикографічного опису. До них ми зараховуємо: $в_1(x_i)$ – реєстрове слово x_i, $в_2(x_i)$ – реєстровий ряд, $в_3(x_i)$ – дублети, $в_4(x_i)$ – етимологія, $в_5(x_i)$ – словозмінні особливості, $в_6(x_i)$ – орфографічні особливості, $в_7(x_i)$ – блок тлумачень. Інформаційне наповнення *в*-структур $V^W{}_L(cymico)$ та $V^W{}_R(cymico)$ подано в Таблиці 1 на прикладі статті *cymico*.

cymico, ca. (Del lat. comĭcus, y este del gr. κωμικός kfmikos). adj. **1.** Que divierte y hace reнг. *Situacion comica*. || **2.** Perteneciente o relativo a la comedia. || **3.** Dicho de un actor: Que representa papeles cymicos. U. t. c. s. || **4.** Dicho de un autor antiguo: Que escribna 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** […].

На прикладі статті *cymico* наведемо у Таблиці 1 інформаційне наповнення *в*-структур для виокремлених множинах лексикографічних описів.

Таблиця 1. Об'єкти Л-системи *DLE 23*

$V^W{}_L(cymico)$		$V^W{}_R(cymico)$	
$в_1(x_i)$	cymico	$в_7(x_i)$	adj. **1.** Que divierte y hace reнг. *Situacion comica*. || **2.** Perteneciente o relativo a la comedia. || **3.** Dicho de un actor: Que representa papeles cymicos. U. t. c. s. || **4.** Dicho de un autor antiguo: Que escribna 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**.
$в_2(x_i)$	cymico, cymica		
$в_3(x_i)$	∅		
$в_4(x_i)$	Del lat. comĭcus, y este del gr. κωμικός kfmikos		
$в_5(x_i)$	∅		
$в_6(x_i)$	∅		

Структури $в_j(x_i)$ об'єднуються між собою за допомогою у-зв'язків, утворюючи тим самим більш складні структури. У *DLE 23* наявні зв'язки, що об'єднують: а) структури $в_j(x_i)$ одного типу ($у_0$), тобто лише етимологію, дублети, тлумачення тощо; б) підструктури в межах виокремлених $в_j(x_i)$-структур ($у_1$); в) структури $в_j(x_i)$ за певним типом метаданих, наприклад, тип дефініції ($у_2$); г) структури $в_j(x_i)$, належні до різних рівнів опису: словотвір – семантика, лексика – фразеологія, словоформа – синтаксис, семантика – прагматика тощо ($у_3$).

Використовуючи структури $в_j(x_i)$ і у-зв'язки стає можливим отримання об'єктів вторинної лексикографії (див. таблицю 2). Останні розуміємо як об'єкти, індуковані лексикографічною системою *DLE 23*, внаслідок дії механізму рекурсивної редукції *Red*[*V*(*I*(*D*))]. Наведемо у таблиці 3 характеристики цих об'єктів.

Таблиця 2. Об'єкти Л-системи *DLE 23*

Об'єкти V[*в*]	**Лінгвістині факти, що містять об'єкти V[*в*]**
$V[в_2(x_i)]$	1) Стандартні та / або нестандартні форми жін. роду для іменників та прикметників; 2) Форми жін. роду, у яких під час утворення відбуваються фонетичні зміни;
$V[в_3(x_i)]$	1) Здатність повністю або замінювати слово x_i; 2) Семантичні і лінгвопрагматичні обмеження на використання дублетів;
$V[в_4(x_i)]$	1) Участь інших мов у формуванні лексичного складу іспанської мови; 2) Характер походження мовної одиниці; 3) Фонетичні та / або семантичні зміни; 4) Характер переходу етимона;
$V[в_5(x_i)]$	1) Відхилення від стандартної словозмінної парадигми; 2) Наявність дефектної і подвійної парадигми; 3) Словозмінні особливості слова, що впливають на його семантику;
$V[в_6(x_i)]$	Вплив орфографічних особливостей на семантику слова

$И[в_7^{LEX}(x_i)]$	1) Моносемія або полісемія реєстрового слова; 2) Слова, що містять задану користувачем сему; 3) Слова, що не мають власної семантики (квазісемантичні слова); 4) Вплив певних граматичних категорій (роду, числа) одиниці x_i на її лексичну семантику; 5) Лексичні значення, обмежені: комунікативним наміром; соціальною і предметною сферою; географічним ареалом; частотою уживання.

Для виявлення лінгвістичних фактів, які неможливо побачити «неозброєним» оком, передбачається побудувати програмний інтерфейс, елементи якого надають доступ до відповідних інформаційних зон словника. На кожній користувач може «підключати» різні *в*, а також задавати параметри для у-зв'язків на *в* за рахунок логічних операцій «та», «або», «ні».

Таблиця 2. Лінгвістичні факти, отримувані у[в].

Інформаційна зона	**Тип $в_i(x_i)$**	**Тип у**	**Лінгвістичні факти, що може отримати користувач (у[в])**
Реєстр та реєстровий ряд	$в_1(x_i)$	$у_0$	Мовна одиниця (вводить користувач)
		$у_2$	Тип мовної одиниці (слово, префікс, суфікс)
		$у_2$	Характер походження (питоме, запозичене)
			Омонімічність
			Номер омоніма (вводить користувач)
		$у_3$	Наявність словозмінної парадигми (для дієслів)
	$в_2(x_i)$	$у_0$	Наявність реєстрового ряду для мовної одиниці
Дублети	$в_3(x_i)$	$у_0$	Наявність дублетів
		$у_2$	Дублет із характеристиками
			Дублет без характеристик
Етимони	$в_4(x_i)$	$у_0$	Наявність / відсутність етимона
		$у_2$	Мова походження
			Значення етимону
Словозміна та орфоепія	$в_5(x_i)$	$у_0$	Наявність / відсутність словозміни
	$в_6(x_i)$	$у_0$	Наявність / відсутність орфоепії
Граматичне значення	$в_7(x_i)$	$у_2$	Частина мови
			Граматична категорія
		$у_2$	Варіація граматичного значення
Прагматика		$у_0$	Наявність / відсутність прагматичних характеристик
		$у_2$	Тип прагматичної характеристики
Лексичні значення		$у_0$	Наявність / відсутність дефініції
		$у_1$	Прив'язка до граматичного значення
			Прив'язка до прагматики
		$у_2$	Тип дефініції
		$у_0$	Набір слів, що може містити дефініція
		$у_2$	Повний / неповний збіг
		$у_3$	Формула тлумачення (для похідних слів)
		$у_2$	Тип мовної одиниці (моносемічна, полісемічна)
			Кількість значень (вводить користувач)

Автоматизація лінгвістичних досліджень на базі словника, які потребують доступ до описаних вище $в_j(x_i)$-структур та у-зв'язків, відбуватиметься через інтерфейс віртуальної лексикографічної лабораторії ВЛЛ *DLE 23*. Розроблюваний інтерфейс даватиме змогу:

1) проводити дослідження на рівні окремих підсистем $V^{LEX}(x)$, $V^{COL'}(x_i)$ та $V^{COL''}$ та їхніми $в_j(x_i)$-структурами та у-зв'язками (тобто досліджувати іспанську лексику або сталі словосполучення);

2) за рахунок застосування у-зв'язків до різних *в*-структур можна отримувати різнопланову лінгвістичну інформацію (наприклад, мова та характер походження реєстрової одиниці, залежність лексичної семантики від граматичних характеристик слова);

3) інтегрувати різні мовні факти в єдиному інформаційному об'єкті (наприклад, словозміна – семантика; словотвір – семантика).

Висновки. Тлумачний словник іспанської мови характеризується максимальною повнотою опису граматичних, прагматичних та семантичних характеристик мовних одиниць. Реєстр складають не лише повнозначні слова та словосполучення, а й також службові слова, словотвірні елементи, і абревіатури. Кожна словникова стаття умовно поділяється на ліву та праву частини. Ліва містить характеристики форми (дублети, етимологія, орфографія, орфоепія), а права – граматичні, прагматичні (включаючи синтаксичні) та семантичні характеристики. Показовим для словника є те, що не всі інформаційні елементи в ньому подано наявним чином, тобто не виокремлюються у текстовій структурі спеціальними метамовними маркерами.

Для можливості функціонування *DLE 23* в цифровому середовищі, та, відповідно, побудови інструментарію для проведення лінгвістичних досліджень на цьому словнику, ми розклали всю його текстову структуру на інформаційні елементи, так щоб вони були представлені наявним чином. Ті елементи, що не мають метамовних засобів ідентифікації, можна ідентифікувати їх за місцем розташування серед інших елементів (наприклад, морфологія завжди слідує після етимології, а орфографія – після морфології). Розклад Л-системи на мінімальні інформаційні структури грунтується на дії рекурсивної редукції. На першому етапі за допомогою рекурсивної редукції визначено чотири рівні Л-системи *DLE 23*. Перший представлений заголовковими словами, другий – словосполученнями типу «приметник + іменник», третій – словосполученнями інших типів, а четвертий – відсилковими словосполученнями. Кожний із них можна розглядати як окремиму Л-систему. На другому етапі виявлені Л-системи також зазнали рекурсивної редукції, внаслідок чого вдалося визначити в них набір інформаційних структур (*в*-структур), що представляють елементи лексикографічного опису іспанських мовних одиниць. Кожну Л-система, що входить до складу Л-системи *DLE 23*, складає набір із семи *в*-структур.

У цифровому середовищі лінгвістичний інструментарій *DLE 23* даватиме широкі можливості не лише для доступу та навігації по різних інформаційних елементах, а й для інтеграції різних мовних фактів у єдиному об'єкті. Пропонований інструментарій уможливлює дослідження опису мови не лише на окремому її рівні, а й також на їх стику. За рахунок застосування у-зв'язків до різних *в*-структур можна отримувати різнопланову лінгвістичну інформацію.

БІБЛІОГРАФІЯ

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

2. Широков В. А. Системні ефекти при лексикографічному описі мови // System Analysis and Information Technologies : 16-th International Conference SAIT 2014. Kyiv, Ukraine, May 26–30, 2014: Proceedings. Kyiv, 2014. С. 25.

3. Широков В. А. Язык. Информация. Система. Düsseldorf: Palmarium Academic Publishing, 2017. 280 с.

4. Широков В. А. Інформаційна теорія лексикографічних систем. К.: Довіра, 1998. 331 с.

5. Широков В. А. Системна семантика тлумачних словників // Акцентологія. Етимологія. Семантика. До 75-річчя академіка НАН України В. Г. Скляренка. К. : Наукова думка, 2012. С. 487–510.

6. Diccionario de la lengua espacola: 23Є ed. Madrid: S.L.U. ESPASA LIBROS, 2014. 2432 p.

REFERENCES

1. Shyrokov V. A. (2011) Kompiuterna leksykohrafiia [Computer lexicography]. Kyiv: Naukova dumka.

2. Shyrokov V. (2014). A. Systemni efekty pry leksykohrafichnomu opysi movy // System Analysis and Information Technologies : 16-th International Conference SAIT 2014. [System effects in the lexicographic description of language] Kyiv, Ukraine, May 26–30, 2014: Proceedings. P. 25.

3. Shirokov V. A. (2017) Jazyk. Informacija. Sistema. [Language. Information. System]. Düsseldorf: Palmarium Academic Publishing.

4. Shyrokov V. A. (1998) Informatsiina teoriia leksykohrafichnykh system. [Information theory of lexicographic systems] Kyiv: Dovira.

5. Shyrokov V. A. (2012) Systemna semantyka tlumachnykh slovnykiv // Aktsentolohiia. Etymolohiia. Semantyka. Do 75-richchia akademika NAN Ukrainy V. H. Skliarenka. [System semantics in explanatory dictionaries] P. 487–510.

6. Diccionario de la lengua española: 23ª ed. Madrid: S.L.U. ESPASA LIBROS, 2014. 2432 p.

ВІДОМОСТІ ПРО АВТОРА

Євген Купріянов – кандидат філологічних наук, доцент кафедри інтелектуальних комп'ютерних систем Національного технічного університету «Харківський політехнічний інститут».

Наукові інтереси: прикладна лінгвістика, комп'ютерна лексикографія, науково-технічний переклад.

INFORMATION ABOUT THE AUTHOR

Yevhen Kupriianov – Candidate in Philology, Associate Professor at the Intelligent Computer Systems Department, National Technical University "Kharkiv Polytechnic Institute".

Research interests: applied linguistics, computer lexicography, translation of scientific and technical information.

УДК: 004.436.4:81'373.423

ПРОГРАМНА WSD СИСТЕМА ДЛЯ ВСТАНОВЛЕННЯ ЗНАЧЕНЬ ОМОНІМІВ

Ярослав ШЕВЧУК (Ізмаїл, Україна)

e-mail: yarolegovich@gmail.com

ШЕВЧУК Ярослав. ПРОГРАМНА WSD СИСТЕМА ДЛЯ ВСТАНОВЛЕННЯ ЗНАЧЕНЬ ОМОНІМІВ

В статті представлена розробка програмного забезпечення способу точного вибору значення слова з наявних омонімів з урахуванням семантичного контексту на прикладі англійської мови. Наукова новизна полягає у створенні авторської програми розпізнавання омонімів, складовими якої є проектування математичного визначення; розробка алгоритму; планування архітектури проекту; розробка системи та тестування програмного забезпечення.

Ключові слова: омоніми, семантичний контекст, алгоритм, архітектура проекту.

SHEVCHUK Yaroslav. WSD SOFTWARE FOR HOMONYM MEANING RECOGNITION

Optimization of the cognitive function of language in the architecture of the systems of processing speech phenomena in the process of working with text structures is one of the topical issues of modern computer linguistics, which determines the relevance of the research. The article presents the software for precisely selecting the meaning of a word from existing homonyms, taking into account the semantic context of the Russian and English language. Scientific novelty consists in creation of author's program of recognition of homonyms, the components of which is the design of mathematical definition; development of the algorithm; planning of project architecture; system development and software testing. Methodology of the research is based on the methods of mathematical modeling: methods of Michael Lesk, Dictionary and knowledge-based methods; supervised, semi-supervised and minimally supervised methods. The implementation of the developed program was carried out using the unsupervised method, which is based on the assumption that similar word values are contained in similar contexts, and therefore the meanings of words can be separated from the text by clustering the occurrences of words using a certain degree of context. The architecture of the project with the "unsupervised" vector model algorithm is based on the modular principle (REST interface, Interactor, Feature extractor, WordNet, Classifier, Knowledge base, Database), each of which contains a certain part of the functionality in order to minimize the dependencies between the modules. The modules are fully compatible with the design architecture and encapsulate the implementation details at the package level. To implement the classifier, an algorithm of cosine coefficients was used. The basic implementation of the system is performed in Java programming language. The project class diagram is generated using the CodeIris dependency analysis tool. The server was implemented using Rapidoid technology that runs on asynchronous buffer classes from the java.nio package. Testing has shown that the total execution time of the program's WSD system for setting the value of homonyms is about 400 milliseconds, and virtually everything goes to finding the base vectors.

Key words: homonyms, semantic context, algorithm, project architecture.

Комп'ютерна лінгвістика є відносно молодою науковою галуззю, яка динамічно розвивається і потребує удосконалення програмного забезпечення для обробки мовних явищ, зокрема програмних інструментів вибору точного значення слова. Оптимізація когнітивної функції мови в архітектурі систем обробки мовних явищ у процесі роботи з текстовими структурами є одним з актуальних питань сучасної комп'ютерної лінгвістики. Вилучення інформації з тексту за допомогою комп'ютерних алгоритмів може ускладнитися, якщо в реченні присутні слова-омоніми – різні за значенням, але однакові у написанні одиниці мови. Word Sense Disambiguation (WSD) як проблема комп'ютерної лінгвістики (від англ. *word* – слово; *sense* – смисл; *disambiguation* – усунення конфліктів, неоднозначностей) – означає сукупність операцій, скерованих на віднаходження механізмів вирішення питань лексичної багатозначності в процесі автоматичної обробки текстів; розробку програмного забезпечення лінгвістичної обробки мови для процесів пошукової оптимізації, при перекладі, підвищенні релевантності видачі адекватного значення слова тощо.

Метою дослідження є розробка програмного забезпечення способу точного вибору значення слова з наявних омонімів з урахуванням семантичного контексту на прикладі англійської мови. Досягнення мети передбачає розв'язання наступних завдань: проектування математичного визначення; розробка алгоритму; планування архітектури проекту; розробка системи; тестування програмного забезпечення.

Наукова новизна дослідження полягає у розробці інноваційного програмного забезпечення для встановлення значень омонімів. Архітектура проекту з алгоритмом векторної моделі класу "*unsupervised"* розроблена за модульним принципом (REST інтерфейс, Interactor, Feature extractor, WordNet, Classifier, Knowledge base, Database), кожен з