ЛЕКСИКОГРАФІЯ

УДК 811’374.81

Є. В. Купріянов

ЛІНГВІСТИЧНИЙ СТАН ІСПАНСЬКИХ МОВНИХ ОДИНИЦЬ ТА ЙОГО ПАРАМЕТРИЗАЦІЯ В ЛЕКСИКОГРАФІЧНІЙ БАЗІ ДАНИХ

Стаття присвячена проблемі створення лексикографічної бази для цифрової версії тлумачного іспанського словника. Проаналізовано лексичні та граматичні особливості мовних одиниць, висвітлених в іспанському тлумачному словнику, що характеризують її лінгвістичний стан. Як результат, визначено формальну модель лінгвістичного стану, виокремлено параметри його опису в лексикографічній базі даних та схарактеризовано їх зміст.

Ключові слова: *комп’ютерна лексикографія, лексикографічна база даних, лінгвістичний стан, граматична семантика, лексична семантика.*

Куприянов Е. В. Лингвистическое состояние испанских языковых единиц и их параметризация в лексикографической базе данных. *Статья посвящена проблеме создания лексикографической базы данных для цифровой версии испанского толкового словаря. Выявлены лексические и грамматические особенности языковых единиц, отраженных в толковом словаре, которые характеризуют ее лингвистическое состояние. В результате определена формальная модель лингвистического состояния, определены его параметры описания в лексикографической базе данных и охарактеризовано их содержание.*

Ключевые слова: *компьютерная лексикография, лексикографическая база данных, лингвистическое состояние, грамматическая семантика, лексическая семантика.*

Kuprijanov Ye. V. Linguistic state of Spanish lexical units and its parametrization in lexicographical database. *The present article is devoted to the creation of the lexicographic database for digital version of Spanish explanatory dictionary. The lexical and grammatical features of language units, being represented in Spanish explanatory dictionary and characterizing their linguistic state, were revealed. Based on the theory of linguistic states, proposed by V. A Shyrokov, the formal model of linguistic states for Spanish lexical units was derived. The model developed takes into account the parameters, by which the linguistic state will be described in the database. The contents of these parameters correspond to the information blocks of the dictionary in consideration. In prospect, the formal model is suggested to be implemented in lexicographic database which is designed to serve as a main part of electronic dictionaries and natural-language processing systems. The access, searching and indexing of the database contents are expected to be assured by these parameters. As for the theoretical value, the research is expected to be helpful in finding the universal way of formal modeling the linguistic features of the units of European languages.*

Key words: *computational lexicography, lexicographical databases, Lexical-grammar state, grammatical semantics, lexical semantics.*

Одним із головних напрямів сучасної національної комп’ютерної лексикографії є створення цифрових лінгвістичних ресурсів на основі паперових фундаментальних лексиконів, із метою їх своєчасного оновлення та підтримки, і призначені як для користувачів, так і для словникарів. У межах

http://dx.doi.org/10.5281/zenodo.48568

зазначеного напряму Український мовно-інформаційний фонд розробив такі електронні словникові проекти: «Інтегрована лексикографічна система «Словники України»», віртуальна лексикографічна лабораторія «Сучасний словник української мови», віртуальна багатомовна лексикографічна система «Mondilex», а також цифрові версії галузевих словників «Російсько-українсько-англійський словник зі зварювання», «Російсько-українсько-англійський словник з механіки», «Російсько-українсько-англійський словник з радіоелектроніки» та інші.

Одним із головних етапів створення цифрових версій словників є розробляння відповідних формальних моделей, що відображають усі можливі граматичні та лексичні особливості мовних одиниць. Сукупність таких особливостей будемо називати лінгвістичним станом. Зауважимо, що мовна одиниця може мати кілька станів. Поняття стану мовної одиниці впровадив А. М. Колмогоров [1], а подальше його обґрунтування з'явилося в працях В. А. Широкова [2, 4, 6] у вигляді теорії лінгвістичних станів. Згідно з нею, лінгвістичний стан трактується як «певна сума ознак граматичної й лексичної семантики», що формує шлях для узагальнення понять граматичного і лексичного значення [5: 71-72]. На цей момент теорію лінгвістичних станів застосовано під час розробляння цифрових версій названих словників. Проте, у зв'язку із подальшою співпрацею нашої держави з європейськими країнами, зокрема Іспанією, актуальним постає питання адаптації зазначеної теорії для створення лексикографічних ресурсів для іспанської мови.

Мета нашої розвідки – розробити принципи параметризації лінгвістичних станів іспанських лексем у лексикографічній базі даних, яка слугуватиме основою для створення цифрової версії тлумачного словника іспанської мови. Для цього необхідно: 1) виокремити чинники, що впливають на лінгвістичний стан іспанських лексем і які враховано під час укладання тлумачного словника іспанської мови; 2) визначити структурні елементи словникової статті, що відповідають за семантизацію реєстрових одиниць; 3) розробити формальну модель опису лінгвістичного стану іспанських мовних одиниць; 4) схарактеризувати параметри, що визначають лінгвістичний стан іспанських лексем і встановити відношення між ними, які мають бути враховані під час розробляння лексикографічної бази даних.

Об'єктом нашого дослідження є лінгвістичні стани мовних одиниць паперового тлумачного словника іспанської мови (Diccionario de la lengua española, 23 ed.), а предметом – формальне моделювання лінгвістичного стану іспанських лексем та його параметризація в лексикографічній базі даних. За приклад візьмемо словникові статті з іменником, оскільки його лексикографічний опис набагато складніший порівняно з іншими мовними одиницями. Трудність опису полягає в тому, що його семантика в іспанській мові залежить від кількох чинників:

1) частиномовне варіювання: іспанські іменники можуть належати одночасно до групи як повнозначних слів (прикметник, прислівник), так і неповнозначних слів (сполучник, вигук), перебуваючи в тій самій граматичній формі;

2) залежність лексичного значення від граматичного. Так, наприклад, слово ***bien***: а) у формі іменника однини та множини має такі лексичні значення ‘добро’, ‘благо’, ‘власність’, ‘товари’; б) прислівника – ‘дуже’, ‘доволі’, ‘належним чином’, ‘з радістю’, ‘із задоволенням’; в) прикметника – ‘заможний’; г) сполучника ‘або ... або’ [7: 306];

3) вплив окремих категорій тієї самої частини мови, до якої належить слово. Так, лексема ***cómico*** у формі чоловічого та середнього родів позначає особу (актора або акторку), що грає комічні ролі. Але, крім цього, у формі жіночого роду аналізоване слово також позначає ‘мультфільм’ та ‘комікс’ [7: 581].

Описані вище чинники стали визначальними для укладачів під час розроблення структури тлумачних зон словникової статті. Крім лінгвістичних характеристик, також ураховувалися прагматичні (наміри мовця, його ставлення до предмета мовлення). Загальний вигляд словникової статті представлено на Рисунку 1. Її умовно поділено на формальну та інтерпретаційну частини, що характеризуються певним набором лексикографічних параметрів та їх змістом. Ліву частину складають заголовкове слово та взятий у дужки блок етимологічних і граматичних характеристик, а праву частину – три блоки: блок тлумачень, блок прикметникових словосполучень і фразеологічних висловів та блок посилань. Зважаючи на мету і завдання нашого дослідження, ми будемо розглядати праву частину, зокрема блок тлумачень.

У свою чергу цей блок складають окремі групи тлумачень, кількість яких залежить від кількості частин мови, у формі яких може виступати лема. Для кожної частини мови подають відповідну групу тлумачень. Для відокремлення груп і зон тлумачень використовують: 1) дві паралельні вертикальні лінії («‖») для відокремлення зон тлумачень, що належать до того самого граматичного класу або граматичної категорії; 2) зони тлумачень, що належать до різних граматичних категорій («●»); 3) у межах того чи того блоку можуть також утворюватися блоки, що групують зони тлумачень для окремих граматичних підкатегорій («○»).

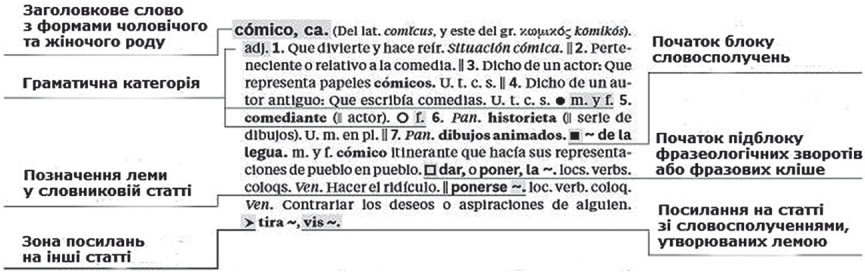

Рис. 1. Приклад словникової статті

Першим та обов’язковим параметром для кожної зони тлумачення є вказівка на частиномовну приналежність слова, а після неї – граматична категорія (наприклад, для іменників це може бути однина або множина, для дієслів – перехідність або неперехідність), група ремарок для розкриття інших особливостей уживання заголовкового слова (стилістика, прагматика,

ареальний статус тощо), тлумачення, ілюстративні приклади та додатковий коментар, що містить супровідну інформацію про граматичні, стилістичні та прагматичні особливості вживання слова в певному значенні. Додатковий коментар може також супроводжуватися ілюстративними прикладами. Обов'язковими для зони тлумачення є частиномовна приналежність і тлумачення, а решта – факультативними.

Як приклад, проаналізуємо лінгвістичні стани лексеми ***cómico***, словникова стаття якої представлена на Рисунку 2. Блок тлумачень починається одразу після зони заголовкового слова та закінчується перед символом «■». Досліджуваний блок містить дві групи тлумачень: перша належить до прикметника (adj.), а друга – до іменника (m. y f.). У першій групі чотири зони тлумачень (1-4), а у другій – дві (5-7). У групі іменника виділяється окрема підгрупа, що відповідає категорії жіночого роду.

cómico, ca. (Del lat. *comĭcus*, y este del gr. *κωμικός kōmikós*). adj. **1.** Que divierte y hace reír. *Situación cómica*. ‖ **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.** ■

Рис. 2. Блок тлумачень словникової статті ***cómico***

Для визначення структури майбутньої лексикографічної бази даних тлумачного словника іспанської мови необхідно розробити формальну модель лінгвістичного стану, що враховує всі особливості граматичної та лексичної семантики мовних одиниць. Розробляння формальної моделі базується на теорії лінгвістичних станів, описаної в [5]. Відправним положенням будемо вважати існування відповідності між мовною одиницею та її станом:

$$s : X \to s(X),$$

де X – іспанська лексема; s – відповідність між X та $s(X)$ – формальним об'єктом, що становить зміст лексеми X. Для будь-якої одиниці X лінгвістичні стани утворюють скінчену, але необмежену множину $\{s(X)\}$. Існує оператор G, що діє на множинності лінгвістичних станів та інтерпретується як оператор граматичної семантики. Зазначимо, що граматична семантика в нашому випадку включає граматичний клас (іменник, прикметник, дієслово) та, у деяких випадках, граматичну категорію (рід, число; перехідність, неперехідність). Отже, для слова ***cómico*** оператор G матиме такі значення: g_1 = «adj.» (прикметник), g_2 = «m. y f.» (іменник спільного роду) та g_3 = «f.» (іменник жіночого роду). Математичну дію оператора G можна виразити як: $Gs_i(X) = g_i s_i(X)$, де g_i – граматичне значення мовної одиниці X, а $s_i(X)$, i = 1, 2... – її зміст (лексична семантика), що відповідає значенню g_i. оператора G. Але, зважаючи на специфіку іспанських лексем, кожному значенню g_i оператора G можуть відповідати одне або кілька $s_i(X)$. Тоді множинність усіх лінгвістичних станів $S(X)$, описуваних у словниковій статті, матиме такий вигляд:

$$S(X) = \sum_{i=1}^{m} \sum_{j=1}^{k} \alpha_{ij} g_i s_j(X),$$

де, i – індекс значення g_i оператора G, який репрезентує граматичну семантику лексичної одиниці X, а j – індекс лексичного значення $s(X)$. У формулу додатково введено коефіцієнт α_{ij}, що характеризує, наприклад, поширеність лексеми в певному лінгвістичному стані в заданій множинності контекстів. Його обчислення є об'єктом окремого дослідження. Цей коефіцієнт передбачено для випадків, коли розроблювану лексикографічну базу даних використовують у системах опрацювання природномовних текстів, зокрема системах машинного перекладу. На прикладі словникової статті ***cómico*** у таблиці 1 наведено елементи множини $S(X)$.

Таблиця 1

Елементи множини S(***cómico***)

i	***j***	g_i	s_j***(cómico)***
1	1	adj.	Que divierte y hace reír. *Situación cómica*.
1	2	adj.	Perteneciente orelativo a la comedia.
1	3	adj.	Dicho de un actor: Que representa papeles cómicos. U. t. c. s.
1	4	adj.	Dicho de un autor antiguo: Que escribía comedias. U. t. c. s.
2	1	m. y f.	**comediante** (‖ actor).
2	2	f.	*Pan*. **historieta** (‖ serie de dibujos). U. m. en pl.
2	3	f.	*Pan*. **dibujos animados**.

Обидва елементи g_i та $s_j(X)$ передбачено візуалізувати в лексикографічній базі даних у вигляді параметрів, умовне позначення та характеристику їх змісту подано нижче. Для g_i передбачено такий набір параметрів опису граматичної семантики для кожної мовної одиниці $s_j(X)$:

1) $<GR^1\ (LS_i)>$ – граматична ремарка першого рівня, що вказує на граматичний клас, відповідний до i-того лексичного значення мовної одиниці. Індекс i позначає номер лексичного значення у лексикографічній базі даних;

2) $<GR^2\ (LS_i)>$ – граматична ремарка другого рівня, що вказує на граматичну категорію, відповідну до i-того лексичного значення мовної одиниці.

Змістове наповнення $s_j(X)$ уключає параметри, що характеризують мовну одиницю з погляду лексичної семантики і прагматики, але необов'язково всі вони наявні в словникових статтях досліджуваного словника:

1) $<Bl\_Rem(LS_i)>$ – блок, що розкриває лінгвопрагматичні особливості вживання слова в лексичному значенні LS_i, і складається з таких параметрів, поданих у тлумачному словнику у вигляді ремарок:

а) $<Rem\_style(LS_i)>$ – віднесеність слова, ужитого в значенні LS_i, до різновиду мови або мовлення: «coloq.» (розмовний), «pop.» (народний), «poet.» (поетичний);

б) $<Rem\_lang\_level(LS_i)>$ – особливості функціонування слова у значенні LS_i у межах тієї чи тієї соціальної групи, що позначаються такими ремарками: «vulg.» (вульгарне), «*jerg.*» (жаргонне), «cult.» (літературне), «infant.» (дитяче);

в) $<Rem\_area(LS_i)>$ – віднесеність слова в значенні LS_i до предметної галузі, наприклад «*Mat.*» (математика), «*Psicol.*» (психологія), «*Quím.*» (хімія), «*Telec.*» (телекомунікації);

г) <*Rem_geogr*(*LS*$_i$)> – поширення слова в значенні *LS*$_i$ на певній території Іспанії («*And.*», «*Ar.*», «*Cast.*», «*Cat.*» тощо) та за її межами («*Arg.*», «*Bol.*», «*C. Rica*», «*Ec.*» тощо);

г) <*Rem_chron_usage*(*LS*$_i$)> – діахронічний аспект уживання слова в тому чи тому значенні *LS*$_i$: «ant.» (давно застаріле), якщо вживання слова не фіксується після 1500 р.; «desus.» (архаїзм), якщо слово вживали в період 1500–1900 рр., «p. us.» (рідко), якщо слово стали рідко вживати після 1900 р;

д) <*Rem_pragm*(*LS*$_i$)> – прагматичні особливості вживання мовної одиниці в значенні (*LS*$_i$), які позначаються ремарками «despect.» (презирливо), «irón.» (іронічно);

е) <*Rem_aesthet*(*LS*$_i$)> – культурно-естетичний аспект уживання слова в значенні (*LS*$_i$): «malson.» (лайливо), «euf.» (евфемізм);

2) <*Bl*(*LS*$_i$)> – блок лексико-семантичного значення мовної одиниці, який містить чотири параметри: <*S*$_i$> – тлумачення; <*Syn*(*X*)> – синонім до заголовкового слова; <*S*(*Syn*(*X*))> – тлумачення до синоніма, якщо він у свою чергу може мати інші значення; <*Il*(*S*$_i$)> – ілюстративні приклади до тлумачення <*S*$_i$>. Параметр <*Il*(*S*$_i$)> може включати кілька речень, тому більш точним буде такий його вигляд: <*Il*(*S*$_i$)> = {*Il*(*S*$_i$)$_j$, j = 1... N};

3) <*Bl_Comment*(*LS*$_i$)> блок додаткових коментарів, який включає: <*Comment*(*LS*$_i$)> – додатковий коментар щодо граматичних, стилістичних та прагматичних особливостей вживання мовної одиниці в лексичному значенні *LS*$_i$. У статтях тлумачного словника становить речення в скороченому вигляді. Наприклад, запис ‘u. t. c. s.’ означає ‘usado también como sustantivo’, тобто – слово в поданому лексичному значенні вживається також як іменник; а також <*Il*(*LS*$_i$ + *Comment*)> – ілюстративні приклади до коментаря.

Наведемо приклади змістового наповнення розглянутих вище параметрів лінгвістичного стану для статей ***agua*** та ***salvavidas***. Для першого слова взято зону тлумачень № 4: ‘f. **lágrimas** (‖ gotas de la glándula lagrimal). *Se le llenaron los ojos de agua*. U. t. en pl. con el mismo significado que en sing.’; та № 14: ‘f. pl. *Mar.* Estela o camino que ha seguido un buque. *Buscar, ganar, seguir lasaguas de un buque*.’ (7, с. 65). Для другого – № 2: ‘m. U. en aposición para indicar que lo designado por el sustantivo al que se pospone sirve para el salvamento de personas en el agua o para mantenerlas aflote. *Bote salvavidas, chaleco salvavidas.* U. t. en sent. fig. *Crédito salvavidas*.’ (7, с. 1965).

Таблиця 2

Змістове наповнення параметрів лінгвістичних станів

Мовна одиниця *X*	agua	agua	salvavidas
Індекс значення *i**	4	14	2
<*GR*^{1} (*LS*$_i$)>*	f.	f.	m.
<*GR*^{2} (*LS*$_i$)>	ø	pl.	ø
<*Bl_Rem* (*LS*$_i$)>			
<*Rem_style* (*LS*$_i$)>	ø	ø	ø
<*Rem_lang_level* (*LS*$_i$)>	ø	ø	ø

Продовження таблиці 2

Мовна одиниця *X*	agua	agua	salvavidas
<*Rem_area* (*LS*$_i$)>	ø	*Mar.*	ø
<*Rem_geogr* (*LS*$_i$)>	ø	ø	ø
<*Rem_chron_usage* (*LS*$_i$)>	ø	ø	ø
<*Rem_pragm* (*LS*$_i$)>	ø	ø	ø
<*Rem_aesthet* (*LS*$_i$)>	ø	ø	ø
<*Bl*(*LS*$_i$)>			
<*S*$_i$>*(а)	ø	Estela o camino que ha seguido un buque	U. en aposición para indicar que ... mantenerlas aflote
<*Syn* (*X*)>*(б)	lágrimas	ø	ø
<*S* (*Syn*(*X*))>	gotas de la glá ndula lagrimal	ø	ø
<*Il* (*S*$_i$)>	*Se le llenaron los ojos de agua.*	*Buscar, ganar, seguir lasaguas de un buque.*	*Bote salvavidas, chaleco salvavidas.*
<*Bl_Comment* (*LS*$_i$)>			
<*Comment* (*LS*$_i$)>	U. t. en pl. con el mismo significado que en sing.	ø	U. t. en sent. fig.
<*Il*(*LS*$_i$ + *Comment*)>	ø	ø	*Crédito salvavidas.*

Параметри, позначені «*», мають бути обов'язково заповнені в будь-якому випадку. До них зараховуємо індекс лексичного значення, що відповідає номеру зони тлумачення в словнику, і граматична ремарка першого рівня. Позначки «*(а)» та «*(б)» указують на те, що один із параметрів обов'язково має бути заповненим. Система параметрів опису лінгвістичних станів іспанських мовних одиниць у лексикографічній базі показана на Рисунку 3.

Перший – граматичний параметр першого рівня, що у свою чергу утворює групу, яку складають граматичний параметр другого рівня, блок лінгвопрагматичних особливостей лексеми, блок семантичних характеристик та блок додаткових коментарів. Обов'язковими є перший та четвертий параметри, тобто частиномовна належність та лексична семантика (у вигляді тлумачення та вказівки на синонім). Усі інші параметри можуть не мати змісту, але все одно повинні бути присутніми у формальній структурі опису лінгвістичного стану та, відповідно, репрезентованими в лексикографічній

базі даних. Послідовність цих параметрів відповідає порядку слідування інформаційних елементів у статті паперового тлумачного словника. Окремі параметри через їхнє призначення (відображення характеристик певного типу) об'єднуються в блоки (наприклад, параметри, позначені на Рисунку 3 як 3.1 – 3.7, 4.1 – 4.4 та 5.1 і 5.2).

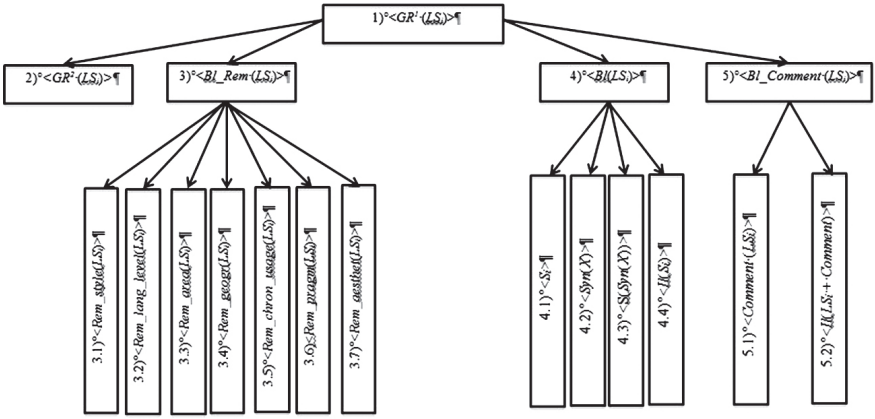

Рис. 3. Система параметрів опису лінгвістичних станів

Отже, ґрунтуючись на проведеному дослідженні структури та змісту статей, а також метамови тлумачного словника іспанської мови «Diccionario de la lengua española», засвідчуємо такі результати і висновки:

– лексичні значення тієї самої іспанської мовної одиниці залежать від її граматичних характеристик. Таке явище отримало назву «лінгвістичний стан», що узагальнює поняття лексичної і граматичної семантики;

– з'ясовано структурні елементи словникової статті, на основі яких виведено формальну модель лінгвістичного стану. Така модель поєднує граматичні та лексичні властивості іспанської мовної одиниці;

– на основі розробленої моделі лінгвістичного стану обрано набір параметрів опису граматичних та лексичних властивостей мовних одиниць, також установлено залежність між цими параметрами;

– доступ, пошук, індексація та опрацювання інформаційного контенту в лексикографічній базі даних будуть забезпечуватися за допомогою цих параметрів.

У перспективі наших подальших досліджень планується винайти універсальну модель лінгвістичних станів, яку можна застосувати для формального опису лексико-граматичних особливостей низки європейських мов (російської, англійської, німецької тощо).

ЛІТЕРАТУРА

1. **Колмогоров А. М.** Три подхода к определению понятия «количества информации» // Теория информации и теория алгоритмов / А. М. Колмогоров. — М. : Наука, 1987. — 304 с. 2. **Лінгвістичні** та технологічні основи тлумачної лексикографії /

В. А. Широков, В. М. Білоноженко, О. В. Булгаков та ін. — К. : Довіра, 2010. — 295 с. 3. **Остапова И. В.**, Широков В. А. Виртуальная лексикографическая лаборатория для толковых словарей [електронний ресурс] / Остапова И. В. и др. – Режим доступу: http://www.dialog-21.ru/digests/ dialog2010/materials/pdf/55.pdf. 4. **Широков В. А.** Елементи лексикографії / В. А. Широков. – К. : Довіра, 2005. — 304 с. 5. **Широков В. А.** Комп'ютерна лексикографія / В. А. Широков. – К. : Наукова думка, 2011. — 351 с. 6. **Широков В. А.** Феноменологія лексикографічних систем/ В. А. Широков. – К. : Наукова думка, 2004. — 326 с. 7. **Diccionario** de la lengua española: 23ª ed. — Madrid: S.L.U. ESPASA LIBROS, 2014. – 2432 p.

Купріянов Євген Валерійович — кандидат філологічних наук, докторант Українського мовно-інформаційного фонду НАН України, Голосіївський проспект, 3, м. Київ-39, 03039, Україна.

Tel.: +38 067-574-85-20
E-mail: cuprijanow.eugen@yandex.ua
http://orcid.org/0000-0002-0801-1789

Kupriianov Yevgen Valeriovych — Candidate of Science in Philology, Associate Professor, Doctoral Candidate, Ukrainian Lingua-Information Fund, NAS of Ukraine, Holosiivskyi av., 3, Kyiv, 03039, Ukraine.