

Yevhen Kupriianov¹¹NTU "KhPI", Kharkiv, Ukraine, Ukraine, eugeniokupriianov@gmail.com,

ORCID iD: 0000-0002-0801-1789

DEVELOPING SOFTWARE FOR COMPILING ELECTRONIC INFLECTIONAL DICTIONARY OF THE SPANISH LANGUAGE

The paper focuses on the technology of creating software for compiling an electronic inflectional Spanish dictionary using the framework of the L-systems theory by V.A. Shirokov. On its basis all Spanish words were classified into respective paradigmatic types, groups and classes is made, a formal model of the dictionary is built, and the structure of the database and the interface of the virtual lexicographic laboratory to work with the dictionary database are determined. The interface offers a number of functions, including adding, editing, and deleting words and paradigmatic classes. The developed database structure and data editing software tools contribute to the efficient organization of the process of creating a word-based Spanish dictionary. The created database can be successfully used in the study of inflection processes and phenomena.

L-SYSTEM, LEXICOGRAPHIC DATABASE, VIRTUAL LEXICOGRAPHIC LABORATORY, FORMAL MODEL

Купріянов Є. Розробка програмного забезпечення для укладання електронного словозмінного словника іспанської мови. Стаття присвячена технології створення програмного забезпечення для укладання електронного словозмінного словника іспанської мови з використанням апарату теорії Л-систем В.А. Широкова. На її основі вироблено словозмінну класифікацію іспанських мовних одиниць за парадигматичними типами, групами і класами, побудовано формальну модель словника, а також визначено структуру бази даних та інтерфейс віртуальної лексикографічної лабораторії для роботи з базою даних словника. Інтерфейс пропонує низку функцій, зокрема додавання, редагування та вилучення слів та парадигматичних класів. Розроблена структура бази даних та програмні засоби редагування даних сприяють ефективній організації процесу створення словозмінного словника іспанської мови. Створена база даних може успішно використовуватись при дослідженні словозмінних процесів і явищ.

Л-СИСТЕМА, ЛЕКСИКОГРАФІЧНА БАЗА ДАНИХ, ВІРТУАЛЬНА ЛЕКСИКОГРАФІЧНА ЛАБОРАТОРІЯ, ФОРМАЛЬНА МОДЕЛЬ

Introduction

The modern period of computer linguistics development is marked by a shift in the research paradigm caused by factors unrelated to linguistics tasks, namely the rapid development of intellectual information and communication technologies, as well as natural language's rapid acquisition of technological status. These factors necessitate the development of appropriate linguistic resources that cover the widest possible range of language material and linguistic phenomena, which, in turn, requires the development of theoretical and linguistic basics for an integral description of the language system, oriented to use in computer linguistics and lexicography, as well as digital text information processing systems (machine translation, data and knowledge mining, conceptual and ontology-based).

These challenges present extremely important tasks for computer linguistics, particularly the creation of a universal system of digital lexicographic resources. The proceedings of the Ukrainian Language and Information Foundation of the National Academy of Sciences of Ukraine have made a substantial contribution to the problem of building integrated lexicographic objects based on formal models for Ukrainian and some other languages – the monographs "Information Theory of Lexicographic Systems", "Phenomenology of L-Systems", "Elements

of Lexicography", "Computational Lexicography" (V. Shyrovkov), "Linguistic and Technological Bases of Explanatory Lexicography" (V. Shyrovkov, N. Zaika, et al.), "Grammatical Systems" (V. Shyrovkov, I. Shevchenko, T. Liubchenko, K. Shyrovkov), 5-volume set "Linguistic and Information Studies" (V. Shyrovkov, et al.). These publications served as the scientific foundation for the creation of Ukraine's National Dictionary Base, the country's only linguistic property designated as a national treasure.

1. Related Works

As it is stated in [1], for information systems to function properly, the language is to be represented as a formal model. Prof. V. Shyrovkov his team mates [2, 3] discuss in detail the problems of formal modeling of the inflectional system of a language and its representation in software tools, such as virtual lexicographic laboratories and electronic grammar dictionaries. The language system theory proposed by the researcher has been effectively applied to modeling the word change systems of Ukrainian, German, and other languages.

Many works [4-9] are focused on developing a database for different inflectional dictionaries. Among them are:

– **Database of Old Icelandic Inflections (DOI)** is a project that aims to describe the patterns of inflection in Old Icelandic through computer modeling, currently

underway at the Arni Magnússon Institute of Icelandic Studies (SÁM) at the University of Iceland. The inflection models form the basis of the database, and all headwords, regardless of which category they belong to, are assigned to one of them. Each noun inflection pattern consists of two main characteristics: stem type and case endings. The declined forms are entered manually on separate lines according to the inflectional structure.

– **Grammar Dictionary of the Polish Language** aims to provide the most complete description of the Polish language conjugation: a complete morphological characteristics and basic syntactic characteristics of Polish words; for each lexeme, all its declined forms are given with the meanings of all morphological categories (categories according to which the word is declined). In addition, the values of some syntactic features are given: gender for nouns, type for verbs, and required case for prepositions. Due to the large amount of data, the dictionary works using relational databases, which are a means of storing elements of the word-changing model. To build all the conjugated forms based on the dictionary data, other means, such as other tools, would be needed. The declined form in the developed model consists of a prefix, stem, ending, and suffix, which are controlled by several model objects. Each of these parts can be empty. However, since the mapping from endings to forms is universal, this complexity does not affect the process of adding new tokens or checking an existing description. These tasks can be successfully performed on a limited set of basic inflectional forms that are built from stem and ending only.

– **UniMorph** includes 23 meaning parameters and over 212 features. Meaning parameters are morphological categories such as person, number, tense, and mood. Each of them represents a coherent semantic space in inflectional morphology. They include: Participle, animation, aspect, case, comparison, definiteness, deixis, evidentiality, finiteness, gender, information structure, interrogative, mood, number, part of speech, person, polarity, politeness, switching, tense, valence, and voice. These dimensions contain a different number of features, from 2 for definiteness to 39 for case. Features represent the finest differences in meaning that are possible within a given dimension.

2. Method

All Spanish words are grouped into different paradigmatic types, i.e., sets of words with the same grammatical function and inflected according to the same sets of word-invariant parameters. Each paradigmatic type can cover several lexical and grammatical classes, i.e., sets of words united by common grammatical features. In turn, grammatical classes are divided into paradigmatic groups, i.e. groups of words representing a certain type of word-change paradigm (e.g. regular or irregular paradigm for verbs). Then, each paradigmatic group falls into paradigmatic

classes, which are a set of words that use the same set of endings (quasi-flexions) during paradigm generation.

The inflection system of the Spanish language distinguishes the following paradigmatic types: nouns and adjectives T_1 , verbs T_2 , personal and reflexive pronouns T_3 , articles T_4 , and the irreducible or zero T_0 , which includes the uninflected Spanish words, namely adverbs, prepositions, conjunctions, interjections. Each of these paradigmatic types is characterized by a certain set of inflection parameters.

The inflectional change of a certain word is traditionally called its inflectional paradigm, i.e., the set of word forms in all possible grammatical meanings. More formally, a word change is a correspondence (operator) P , according to which each word x corresponds to its word-changing paradigm $[x]$. Formally, this definition can be presented as follows:

$$x = [x]; [x] = \{x^1, x^2, \dots, x^N\}, x^i = c(x) * f(x^i),$$

where x is a word, P is the conjugation operator, $[x] = \{x^1, x^2, \dots, x^N\}$ is the complete inflectional paradigm of the word x ; $x^i, i = 1, 2, \dots, N$, are the components of the paradigm; $c(x)$ is the quasi-base (the unchangeable part of the lexeme x); $f(x^i)$ is the quasi-flexion of the component x^i . In this case, both the quasi-stem $c(x)$ and the quasi-flexion $f(x)$ can take on the following values:

– $c(x) = 0$ if the word has a suppletive form, such as the verb *ser* (to be) in the second and third person singular present tense of the active voice: *eres, es*; the verb *caber* in the first person singular present tense: *quiero*;

– $c(x) = x$ if x is an uninflected word, e.g., all pluralia tantum nouns (*pantalones, gafas, cómicas, esposas*, etc.), all prepositions (*a, por, de, con, bajo*), neuter pronouns (*alguno, ello, lo*), conjunctions (*que, y, como*); some nouns, e.g., those denoting days of the week (*lunes, martes, miércoles*, etc.), etc. etc.), words of Greek origin: *análisis, artritis, crisis*;

– $f(x) = x$ when all forms are complementary, as, for example, with the verb *ir* (to go) in all forms of the present tense: *voy, vas, va, vamos, vais, van*; the verb *ser* in the indefinite past tense: *era, eras, era, éramos, erais, eran*; the plural form of the masculine definite article: *el – los*.

Quasi-stem is a common term for an unchangeable component of a word that can include not just the derivational basis, but also a specific part of it (in some circumstances, only one letter), as well as the entire word. A quasi-flexion is a component of a word that varies during paradigm building. It can include:

– usual suffixes: *comer – como, comes, come; comieras, comiera, comiüramos*, etc.;

– a part of the stem: *plegar – pliego, pliegas, pliega, plegamos, plegáis, pliegan*;

– the whole word form: *orden – yrdenes; ir – voy, vas, va, vamos, vais, van*;

– ending: *gato – gatos, mesa – mesas*.

Here are examples of quasi-stem lengths in the table for the paradigm of the nouns *mano* and *orden*, as well as the verbs *bailar*, *tener*, and *ser*.

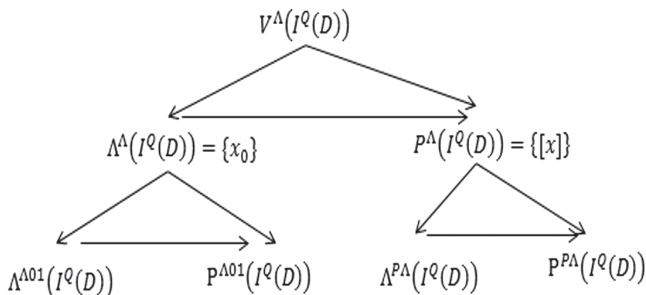
Table 1
Quasi-stems and their lengths

x	$c(x)$	$f(x)$	$x - f(x)$
<i>mano</i>			
<i>mano</i> (sing.)	<i>man-</i>	<i>-o</i>	3
<i>manos</i> (pl.)	<i>man-</i>	<i>-os</i>	3
<i>orden</i>			
<i>orden</i> (sing.)	0	<i>orden</i>	5
<i>órdenes</i> (pl.)	0	<i>órdenes</i>	5
<i>bailar</i>			
<i>bailo</i> (1 pers. sing.)	<i>bail-</i>	<i>-o</i>	4
<i>bailas</i> (2 pers. sing.)	<i>bail-</i>	<i>-as</i>	4
<i>baila</i> (3 pers. sing.)	<i>bail-</i>	<i>-a</i>	4
<i>tener</i>			
<i>tengo</i> (1 pers. sing.)	<i>t-</i>	<i>-engo</i>	4
<i>tienes</i> (2 pers. sing.)	<i>t-</i>	<i>-ienes</i>	4
<i>tiene</i> (3 pers. sing.)	<i>t-</i>	<i>-iene</i>	4
<i>ser</i>			
<i>soy</i> (1 pers. sing.)	0	<i>soy</i>	2
<i>eres</i> (2 pers. sing.)	0	<i>eres</i>	3
<i>es</i> (3 pers. sing.)	0	<i>es</i>	3

Moving on, let us now consider the issue of modeling the grammatical system of the Spanish language represented by word change. Formally, the inflection system can be interpreted as a L-system of a grammatical type (or grammatical system) where the lexicographic effect of inflection is induced:

$$LS^{GRAM} = \{I^Q(D), \Lambda^\Lambda(I^Q(D)), P^\Lambda(I^Q(D)), F', C', H'\},$$

where LS^{GRAM} is L-system of grammatical type; $\Lambda^\Lambda(I^Q(D)) = \{x_0\}$ is a set of words in lemma form, $P^\Lambda(I^Q(D)) = \{[x_0]\}$ is a set of paradigms, F' is an operator that establishes the relationship between a unit and its lemma form, C' is an operator that establishes the relation between a unit and its grammatical content (i.e., a paradigm), H' is an operator that correlates the lemma form of a word with a paradigm. In turn, the elements of the lexicographic system of the grammatical type $\Lambda^\Lambda(I^Q(D))$ and $P^\Lambda(I^Q(D))$ can be decomposed using the recursive reduction mechanism $RR\downarrow[V(I^Q(D))]$, as shown in Fig. 1.


Fig. 1. L-system decomposition using recursive reduction

As a result of the recursive reduction, the set of descriptions $V^\Lambda(I^Q(D))$ is decomposed into smaller information elements containing relevant information about the described units of the Spanish language. The left part $\Lambda^\Lambda(I^Q(D))$ of the grammatical description includes a set of parameters $\Lambda^{\Lambda 01}(I^Q(D))$, that determine the place of the unit in the word-variable system of the Spanish language according to the classification described in [10], as well as $P^{\Lambda 01}(I^Q(D))$, i.e., the Spanish units in the canonical form that correspond to the parameters $\Lambda^{\Lambda 01}(I^Q(D))$. The right side of the grammatical system $P^{P^\Lambda}(I^Q(D))$, contains both all the word forms that make up the paradigm $P^{P^\Lambda}(I^Q(D))$ and the parameters representing the set of grammatical values $\Lambda^{P^\Lambda}(I^Q(D)) \equiv \Omega$. The components of the set of grammatical meanings of the Spanish language $\Omega = \{\Omega^1, \Omega^2, \Omega^3, \Omega^4, \Omega^5, \Omega^6, \Omega^7, \Omega^8, \Omega^9\}$ are:

$\Omega^1 = \{\omega_1^1, \omega_2^1, \omega_3^1, \omega_4^1\} \equiv \{m., f., m. y f., m. o f.\}$ is a set of grammatical meanings of the gender category, where m is masculine, f is feminine, $m y f$ is common, and $m o f$ is indefinite;

$\Omega^2 = \{\omega_1^2, \omega_2^2\} \equiv \{\textit{singular}, \textit{plural}\}$ denotes a set of grammatical meanings of the number category (*singular*, *plural*);

$\Omega^3 = \{\omega_1^3, \omega_2^3, \omega_3^3, \omega_4^3\} \equiv \{\emptyset, \textit{presente}, \textit{pretérito perfecto simple}, \textit{pretérito imperfecto}, \textit{futuro simple}\}$ is a set of grammatical meanings of the tense category (*present*, *imperfect past*, *simple perfect past*, *simple future*);

$\Omega^4 = \{\omega_1^4, \omega_2^4, \omega_3^4, \omega_4^4\} \equiv \{1a \textit{ pers.}, 2a \textit{ pers.}, 2a \textit{ pers.-voseo}, 3a \textit{ pers.}\}$ designates a set of grammatical meanings of the person category (*first*, *second*, *third*);

$\Omega^5 = \{\omega_1^5, \omega_2^5, \omega_3^5, \omega_4^5\} \equiv \{\textit{indicativo}, \textit{condicional}, \textit{subjuntivo}, \textit{imperativo}\}$ is a set of grammatical meanings of the mood category (*active*, *conditional*, *subjunctive*, and *imperative*);

$\Omega^6 = \{\omega_1^6, \omega_2^6, \omega_3^6, \omega_4^6\} \equiv \{\textit{nominativo}, \textit{dativo}, \textit{acusativo}, \textit{preposicional}\}$ means a set of grammatical values of the case category (*nominative*, *dative*, *accusative*, *local*).

$\Omega^7 = \textit{infinitivo}$, $\Omega^8 = \textit{gerundio}$, $\Omega^9 = \textit{participio}$ are non-finite forms of the verb: infinitive, gerund, and participle.

Word forms are defined by complexes of grammatical meanings that together make up the grammatical state of a linguistic unit. These combinations of grammatical meanings, or members of sets, define a word form. An example of a fragment of the paradigm for the verb *hablar* (to speak) is shown in Table 2, and the parameters characterizing its grammatical state are indicated.

Table 2
Parameter values for the grammatical state of the verb *hablar*

Parameter chain	Parameter values in chain	Word form
$\{\omega_1^4, \omega_1^2, \omega_1^3, \omega_1^5\}$	$\{1a \textit{ pers.}, \textit{singular}, \textit{presente}, \textit{indicativo}\}$	<i>hablo</i>
$\{\omega_2^4, \omega_1^2, \omega_1^3, \omega_1^5\}$	$\{2a \textit{ pers.}, \textit{singular}, \textit{presente}, \textit{indicativo}\}$	<i>hablas / hablás</i>

Parameter chain	Parameter values in chain	Word form
$\{\langle\omega_3^4, \omega_1^2, \omega_1^3, \omega_1^5\rangle\}$	$\{\langle\langle 3^a \text{ pers., singular, presente, indicativo}\rangle\rangle\}$	<i>habla</i>
$\{\langle\omega_1^4, \omega_2^2, \omega_1^3, \omega_1^5\rangle\}$	$\{\langle\langle 1^a \text{ pers., plural, presente, indicativo}\rangle\rangle\}$	<i>hablamos</i>
$\{\langle\omega_2^4, \omega_2^2, \omega_1^3, \omega_1^5\rangle\}$	$\{\langle\langle 2^a \text{ pers., plural, presente, indicativo}\rangle\rangle\}$	<i>habláis</i>
$\{\langle\omega_3^4, \omega_2^2, \omega_1^3, \omega_1^5\rangle\}$	$\{\langle\langle 3^a \text{ pers., plural, presente, indicativo}\rangle\rangle\}$	<i>hablan</i>

3. Dictionary Database

The Spanish inflection dictionary database (as shown in Fig. 2) was developed using the conceptual model discussed above. Thus, information regarding paradigmatic kinds may be found in the “par_types” table:

- ID_partyp (paradigmatic type identifier),
- com (paradigmatic type, name),
- ac (comment, word change parameters).

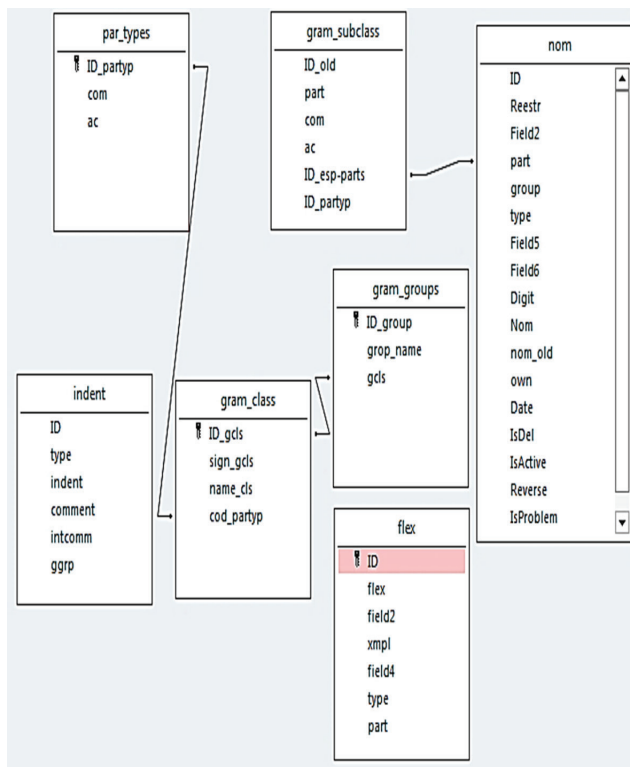


Fig. 2. Database scheme of the Spanish language inflectional dictionary

Parameters and identifiers of grammatical classes in the table “gram_class” include:

- ID_gcls (grammatical class identifier),
- sign_gcls (grammatical class signature according to the conceptual classification),
- name_cls (name of the grammatical class),
- cod_partyp (code of the paradigmatic type).

Table “gram_groups”, which contains information about grammatical groups and includes the following fields:

- ID_group (grammar group identifier),

- grop_name (grammar group name),
- gcls (the code of the grammar class to which the grammar group belongs).

Table “indent” includes the fields that the system uses to determine quasi-stems by paradigmatic classes:

- type (paradigmatic class number),
- indent (the number of letters to be cut off from the end of the word to obtain a quasi-stem),
- comment is the conventional name of the paradigmatic class in the conceptual model.

Table “flex” table contains sets of quasi-flexions for each case word, organized by the following parameters:

- Reestr (headword),
- Field2 (homonymy number),
- Part (grammatical class code),
- Group (grammatical group code),
- Type (paradigmatic class number).

4. Virtual Lexicographic Laboratory to handle Spanish Inflectional Dictionary

To work with the electronic dictionary database, a virtual lexicographic laboratory (VLL) was developed (Fig. 3), the functionality of which currently allows:

- viewing the grammar dictionary wordlist and the full paradigm of each word,
- search for words in the register, as well as display them in forward or reverse order,
- add, delete words from the word list,
- add, delete and edit paradigm classes,
- adding, deleting and editing quasi-flexions in paradigmatic classes,
- filtering the register by the following features (or a combination of them): part of speech, paradigmatic class, homonyms, etc.

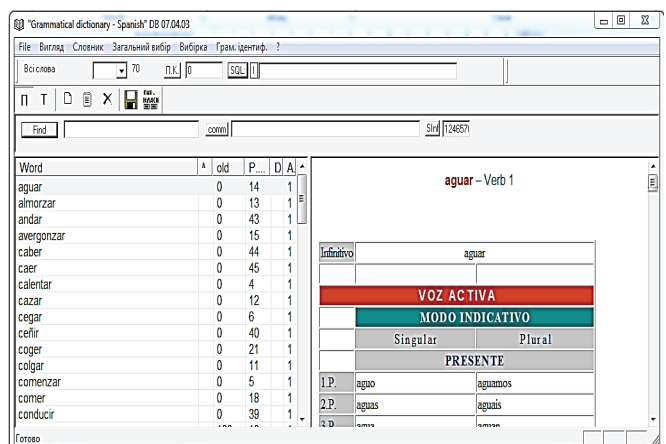


Fig. 3. VLL main window

The main window of the program is divided into three zones: the functional area, the register area, and the lexicographic information area. The functional area consists of the following sub-areas: the general menu, editing tools, tools for executing SQL queries, and the word search interface. The main menu is intended for selecting

the display modes of the dictionary, forming a selection according to certain parameters, and performing operations with the dictionary database file. The word list area represents the word list, indicating the number of the paradigmatic type, grammatical group and class, and paradigmatic class. In case of uninflected word, the number of paradigmatic class isn't given. Lexicographic information area is intended to display information on the word change of the word selected from the list (full inflection paradigm).

Paradigm classes, including inflections, are added using the dialog box (Fig. 4), which is opened by clicking the "Paradigmatic Classes" button on the editing panel. The left part of the window displays the numbers of the paradigmatic type, grammatical class, grammatical group, and paradigmatic class. In addition, technical parameters are displayed, such as the length of the quasi-stem of the word to which the quasi-reflection is to be attached (as shown in Table 2). The right part of the window contains a list of all quasi-flexions belonging to the paradigmatic class selected in the left part. The main functions are: searching for a particular paradigmatic class, creating a new paradigmatic class, and adding and removing flexions.

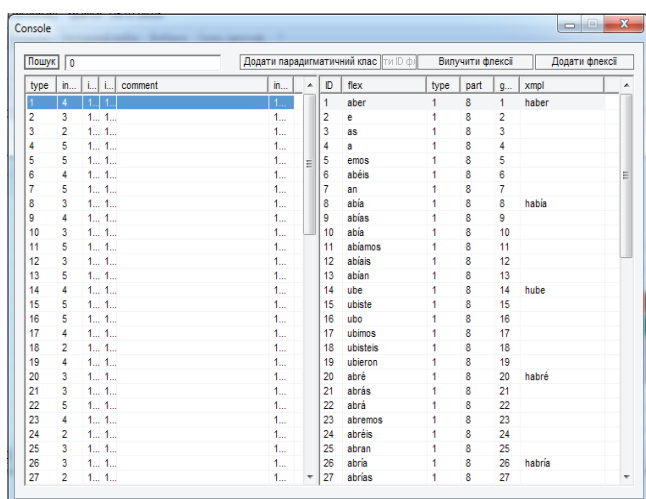


Fig. 4. The window for editing paradigmatic classes

To create a new paradigmatic class, click on the "Add paradigmatic class" button, where you need to enter the numbers of the corresponding paradigmatic type, grammatical class, grammatical group, and paradigmatic class.

Conclusions

The software (VLL) allows compilers to create, edit, and update electronic dictionaries for any language that

has a developed word change system. The inflectional dictionaries created by the VLL tools provide the ability to automatically generate a paradigm for any word, display the whole inflection system together with all its paradigmatic types, groups and classes, and provide grammatical characteristics for any word form included in the paradigm. The developed database structure and software tools for data editing contribute to the efficient organization of the process of creating a word-by-word dictionary of the Spanish language. The created database can be successfully used in the study of word change processes and phenomena.

References

- [1] Shyrokov V. Computer lexicography. – K.: Naukova Dumka, 2011. – 351 p.
- [2] Shyrokov et al. Computational linguistics studies. – V. 2.: Grammar systems. – K.: ULIF NASU, 2018. – 300 p.
- [3] Shyrokov et al. Computational linguistics studies. – V. 3.: Explanatory lexicography. – B. 2: System semantics of explanatory dictionaries – K.: ULIF NASU, 2018. – 250 p.
- [4] Johannsson E., Ingimundarson F. Describing inflectional patterns of nouns in Old Icelandic // CEUR. – 2022. – V. 3232. – P. 260-268.
- [5] Chrzaszcz P. Enrichment of inflection dictionaries: automatic extraction of semantic labels from encyclopedic definitions // 9th International Workshop on Natural Language Processing and Cognitive Science. – 2012. – P. 106-119.
- [6] Wolinski M. A Relational model of Polish inflection in grammatical dictionary of Polish // Human Language Technology. Challenges of the Information Society. – 2009. – №. 5603. – P. 1-11.
- [7] Štěpánková B., Mikulova M., Hajič J. The MorfFlex Dictionary of Czech as a Source of Linguistic Data // Euralex XIX: Congress of European Association of Lexicography. – 2021. – P. 387-391.
- [8] Arista J. M. Old English morphological inflection generation with UniMorph. Assessment with a relational database and training guidelines // Procesamiento del Lenguaje Natural. – 2022. – №. 68. – P. 61-70.
- [9] Wolinski M., Kieras W. The online version of grammatical dictionary of Polish // Język Polski. – 2017. – 97(1). – P. 2589-2594.
- [10] Kupriianov Ye. Lexicographic system of the Spanish Language: Phenomenology of Integral Description. K.: Naukova Dumka, 2018. – 254 p.

The article was delivered to editorial staff on the 30.03.2023