UDC 81'322
DOI https://doi.org/10.32841/2409-1154.2021.52-1.12

Hromovenko V. V.,
Doctor of Philosophy in the branch of knowledge 03 "Humanities",
majoring in 035 "Philology"
Senior Lecturer at the Department of Foreign Languages of Professional Communication
International Humanitarian University

MAIN TYPES OF DATABASES IN LINGUISTIC RESEARCH OF THE XXI CENTURY: FEATURES AND FUNCTIONAL PURPOSE

Summary. The article characterizes the main directions of using linguistic databases (LDB) with consistent demarcation of full-text and actual linguistic databases in modern linguistics. For the successful realization of the goal two main tasks have been solved: 1) acquaintance with full-text LDB and determining the specifics of their content and functioning; 2) delineation of theoretical and applied principles of factual LDB (2011–2021). The relevance of the article is motivated by the lack of a thorough analysis of the experience of creating and operating linguistic databases in domestic and foreign linguistics. The object of research is LDB as a set of systematized linguistic data; subject – specific tested LDB. The main methods are the method of critical analysis and descriptive-analytical method. The methodological basis of the article is the main provisions of applied linguistics and corpus linguistics. The database in modern linguistics is the most effective technology for a compact representation of the set of parameters of linguistic units, the convenience and speed of processing the necessary data to achieve a specific research goal. The factual linguistic database is recognized as the most adequate tool for preserving lexical, phraseological, terminological, morphological and syntactic, stylistic units with the reflection of their characteristics. The use of database technology in linguistics contributes to the convenient presentation of materials and their integration into a single structure, as well as improving the efficiency of working with them, opening new perspectives for further research based on full-text databases or factual databases. We see the prospect of work in the development of a linguistic database of political neologisms in Ukrainian and English.

Key words: database, linguistic database, corpus of texts, factual database, applied linguistics.

Statement of the problem and links with important scientific and practical knowledge. The effective solution of theoretical and applied problems in linguistics of the XXI century requires an appeal to computer technologies in general and the use of database technology, which allows you to quickly obtain lexicographic information from various sources. Thus, one of the productive areas of applied linguistics is the development of linguistic databases, the use of which allows solving a number of new problems and expands the possibilities for solving traditional problems.

Analysis of recent research and publications on the topic. The experience of using databases or data banks for the analysis of linguistic phenomena shows that their creation is mostly not the ultimate goal of the researcher (see the works of M. Bihdai [1], A. Galieva [2], Yu. Kalymon [3], N. Lototska [4], etc.). Databases are created for further multiple analysis of information contained in them, and the format of storing information in the form of databases provides a fundamental opportunity, first, to verify the findings of other researchers, and, secondly, the possibility of further study collected in the database primary empirical information.

The urgency of the study is motivated by the lack of a thorough analysis of the experience of creating and operating linguistic databases in domestic and foreign linguistics.

The purpose of the article: to characterize the main directions of use of linguistic databases with consistent demarcation of full-text and actual linguistic databases in modern linguistics. Successful implementation of the goal involves solving the following tasks: 1) to get acquainted with full-text LDB and determine the specifics of their content and functioning; 2) to outline the theoretical and applied principles of factual LDB, created in 2011-2021.

The object of study is the linguistic database (LDB) as a set of systematized linguistic data. The subject of the work are specific tested LDB. The purpose and objectives of the work necessitated the use of the method of critical analysis and descriptive-analytical method.

The methodological basis of the article is the main provisions of applied linguistics (Ye. Karpilovska [5]) and corpus linguistics (O. Buhakov, T. Hriaznukhina, V. Shyrokov [6]).

Presentation of the main material of the study. Today in applied linguistics there are two main types of LDB: full-text LDB and actually LDB (see the works of N. Mishankina [7] and L. Rychkova [8], etc.).

Full-text LDB are positioned as documentary databases in which complete texts are presented. These include digital libraries, text collections and text corpora.

One of the first was the French base Frantext. Note that this is not a case, but the system of its operation allows the researcher to form his "working case" taking into account a number of parameters (author, date, genre, size, etc.). Work on the base began in 1957 as part of the preparation of the 16-volume "Thesaurus of the French Language" (TLFI), but over time, replenishment and development of facilities for the operation of the database became an independent task. Frantext received financial support, and an abortion of the French National Research Center (CNRS) of 30-50 people worked on it for half a century. The context is the basis of word usage from the texts of the XVI-XX centuries and continues to replenish today. In recent years, work has been carried out to deepen the "historical perspective" of Frantext: it has added databases of the Old French (IX-XIII centuries) and Middle French (XIV-XV centuries) periods, and anyone can use these databases for free. More than half of the database texts are provided with morpho-syntactic markup.

External access to Frantext has been open since 1992 for corporate users (libraries, universities, etc.) and is paid. Free access is provided to the bibliographic database and to the electronic version of the "French Thesaurus".

The advantages of Frantext include its size and long history of formation. However, these advantages are the source of his problems. The formats and systems of operation developed in the 1960's and 1970's are now very outdated and do not meet the capabilities of modern technology and the demands of researchers. Upgrading Frantext, in particular its conversion to XML standard and markup according to TEI recommendations, is a difficult task. However, the current system of exploitation of Frantext allows the solution of many linguistic and literary problems and is widely used by French language researchers around the world.

The first databases appeared in Britain in the 60's of the twentieth century (Brown University Corpus and Lancaster/Oslo-Bergen Corpus (LOB)) but one of the most famous and popular English language corpora is the British National Corpus (BNC). This database was created by the joint efforts of several British universities and publishers, as well as the British Library for the period 1991–1994.

Characteristics of the corpus: a) includes written and oral texts in British English of various genres and functional styles; b) the volume is more than 100,000,000 word usages; c) fragmentary type: texts of more than 45,000 words, presented in fragments, which avoids the influence of the individual style of a particular author on the results; d) equipped with morphological marking: each word form is characterized by belonging to a part of speech, category in the framework of a part of speech and form of word change.

BNC data are widely used in the compilation of dictionaries, grammars and textbooks in English, in linguistic research, in the work of artificial intelligence, as well as in the practice of teaching English.

The Ukrainian National Linguistic Corpus (UNLC) has been established at the Ukrainian Lingua-Information Fund of National Academy of Sciences of Ukraine to study the Ukrainian language and compile modern dictionaries. UNLC is being developed under the leadership of Academician of National Academy of Sciences of Ukraine V. Shyrokov [6, p. 103]. By the order of the Cabinet of Ministers of Ukraine (dated February 11, 2004 №73) the National Dictionary Database of the Ukrainian Lingua-Information Fund of National Academy of Sciences of Ukraine was included in the state register of scientific objects of national heritage. The main directions of UNLC are: 1) provision of textual information according to certain criteria; 2) creation of input streams of linguistic information for various research systems; 3) integration of various linguistic-software means of word processing in a single environment.

There are two functional subsystems in the UNLC system:

1) bibliographic, i.e. electronic library as a collection of digital resources, which is the basis for the development of any corpus; the subsystem serves as a tool for collecting, storing, modeling and using natural language information in digital form;

2) full-text, i.e. full-text search, the user enters a search phrase, sets the maximum desired number of words between search engines and selects additional parameters of full-text search, namely: a) taking into account the order of words; b) search in a certain subset of objects; c) use of the lemmatization procedure; d) use of a synonymous lexicographic database; e) use of certain synonymous series; e) choice of grammatical parameters for each word included in the search fragment. The result of a full-text search is a list of bibliographic descriptions. But unlike a bibliography search, the user has direct access to each localization of the search fragment in the text, i.e. to all contexts that contain the search fragment. By selecting the source, the user can view the contexts (for convenience, the search snippet is highlighted in red).

We also focus on the temporary storage (so-called 'basket') features: the user can select the objects that interest him from various search queries and save the data image for further work in other sessions. Using such a tool, the researcher can select sources of a particular author, style or genre and work only with this part of the body; this procedure makes it possible to distinguish from the general corpus its own sub-corpus, focused on solving personal problems. According to A. Luchyk and I. Ostapova, the functions of the basket allow to generate virtual 'subcorpora' in real time and within a single linguistic corpus, which simulate certain linguistic effects, providing each researcher with opportunities to realize his own ideas about standardity and balance [9, p. 35–36].

The UNLC system is multifaceted and can have many different applications: as a source base of linguistic information for the creation of a fundamental academic multi-volume lexicographic system "Dictionary of the Ukrainian language"; implementation of linguistic research in order to identify new linguistic phenomena and formalize existing ones; provides tools for grammatical and semantic marking of texts; convenient environment for statistical processing of textual information.

One of the main corpora of texts in Russian is the Russian National Corpus. Characteristics of the corpus: a) contains historical, literary, dialectal, written, oral, modern, translated texts covering a period of 200 years, so it is best suited for the study of short (several decades) and medium language changes; b) conditionally divided into two parts – modern (texts created in the period from 1951 to the present) and diachronic (is about 53 million words and combines texts of the XVIII century, XIX century and 1st half of XX century); c) equipped with a significant number of markings: lexical, morphological, syntactic, lexical-semantic and a number of other specialized markings; d) a feature of the corpus is poetic markup, which allows you to search for poetic texts with different parameters.

Includes the following subcorpora:

1) the deeply annotated corpus, in which a complete morphological and syntactic structure (dependence tree) is constructed for each sentence;

2) the parallel Russian-English corpus of texts, in which you can find all the translations for a particular Russian or English word or phrase;

3) the body of dialect texts, including the recording of dialect speech of different regions of Russia while preserving their grammatical specificity; special search is provided taking into account dialect morphology;

4) the corpus of poetic texts, in which it is possible to search not only for lexical and grammatical, but also for specific verse features (search for a certain combination in sonnets, epigrams, poems written by amphibrach, with a certain type of rhyme, etc.);

5) the educational corpus of the Russian language – a corpus with removed homonymy, the layout of which is focused on the school program of the Russian language;

6) the corpus of oral speech includes deciphering of tape recordings of public and private oral speech, as well as transcripts of films of the 2000s.

A distinctive feature of the Czech National Corpus is a) the possibility to obtain all examples of uses together with the contexts in which the word-form occurs, the frequency with which the word-form enters the corpus; b) the morphological analyzer, which allows for morphological and contextual analysis.

Actually, LDB are factual information systems and contain structured information about linguistic units of various kinds. Ye. Karpilovska clarifies that the models for a computer are modern, declarative and procedural knowledge about them to become factual information for the development of singers in computer linguistics [5, p. 34], that is why this type of LDB is also called factual and is classified according to the level of language description into phonetic, morphological, lexical, terminological, phraseological and syntactic [10, p. 30].

Over the last 10 years (2011–2021) the following theoretical and applied principles of factual LDB in linguistics have been updated:

1) electronic dictionary of paronyms, the construction of which involves the following tasks: interpretation at the grapheme level of the main characteristics of paronyms – "sound similarity" (in relation to the written form of language); formation of LDB phonetic paronyms of the Ukrainian language; formation of LDB quasi-paronyms of the Ukrainian language; integration of LDB paronyms with the explanatory dictionary of the Ukrainian language in the general lexicographic system of the Ukrainian Lingua-Information Fund of National Academy of Sciences of Ukraine [11];

2) creation of the database of Ukrainian particles was implemented using software products Microsoft Word and Microsoft Access 2010. The database is a table of 38 fields, which correlates with the relevant information about the particles, and more than 200 rows (records). For convenience of work the form from 9 tabs each of which reflects essential parameters of particles, and the built-in procedures of automatic data processing is constructed. Each tab contains the corresponding text fields and controls [12];

3) the LDB of the special vocabulary of the Belarusian language contains not only lexical units, but also various kinds of linguistic information about them (about part of speech; about lexical and semantic variants having a "special" meaning; about derivatives related to the area of special use, etc.) [13];

4) compiling an electronic dictionary taking into account modern lexicographic methods and technologies to create a detailed description of Taras Shevchenko's language and optimizing the work of researchers with texts of his creative heritage, which involves the following tasks: compiling a database of fixed language units with their grammatical and quantitative characteristics; creation of a user-friendly interface that could search, sort and statistically process the information collected in the database according to the needs of researchers [14];

5) elaboration of tasks of infological [15] and datalogical [16] stages of LDB design "Concept *human* in the phraseology of East Steppe Ukrainian dialects". At the infological stage: a) the corpus of areal phraseological units (APhU) of East Steppe Ukrainian dialects with the archisema 'human' was formed; b) the classification types of APhU by axiological, ideographic, structural parameters are determined; c) the list of culture codes and intercode transitions of the analyzed concept is developed; d) the dictionary article of APhU is constructed; e) the LDB design table is concluded. During the data stage, the LDB was created on the basis of Microsoft Office Access using a single data table, which is an alphabetical dictionary of the analyzed APhU with an indication of belonging to a particular phraseosemantic group or subgroup; queries responsible for sampling data from the table for certain parameters; forms that appear as pages in the menu; macros that provide navigation actions in the menu sections of the LDB. The main menu is presented in the sections "Areal phraseological units", "Codes of culture", "Axiological characteristics", which fully corresponds to the structural components of the concept *human* in the phraseology of East Steppe Ukrainian dialects;

6) design of the database "Syntactic phraseology in the Ukrainian language", which provides infographic (selection and structuring of relevant linguistic and/or linguistic data, i.e. selection of models of syntactic phraseology in the Ukrainian language and establishing their classification features) and datalogical (choice of database management system) stages. The main field in the form mode is the field "Structural scheme of the sentence", which corresponds to the record of the sentence model. The rest of the selected features of syntactic phraseology are grouped by the author in eight tabs: 1) structure (type of sentence by structure; partial language status of the core component; number of word forms within the core component; variants of structural scheme; presence of distributors); 2) semantics (typical semantics of a sentence; additional semantic nuances; semantic type; degree of fusion of components; presence/absence of lexical restrictions on filling the position of a variable component; figurative models); 3) syntactic paradigm reflects the component composition of the syntactic paradigm of the described sentence model (grammatical modifications of time, method, phase, modal and other transformations of the basic sentence model); 4) pragmatics (pragmatic function of syntactic phraseology; pragmatic status of the speaker; pragmatic status of the addressee); 5) statistics (calculation of indicators of association MI,MI^3, MI log Freq, Dice and gmean); 6) semantic-paradigmatic properties (syntactic synonyms; syntactic homonyms); 7) dictionary information (data given in authoritative dictionaries); 8) examples of use in artistic, journalistic and conversational styles [17];

7) concept, structure and content of LDB negative-evaluative vocabulary developed for linguistic research in the field of linguistic expertise [10];

8) an electronic terminographic product in the form of a terminological database "Classification parameters of phraseological units" with a consistent implementation of the infological and datalogical stages. The DB is a special terminological dictionary (monolingual: Ukrainian), which has 113 terms for designating the types of phraseological units, covers 17 phraseological classifications and was created to store information, optimize and intensify system research on fundamental questions of phraseology [18];

9) elaboration of the technology of compiling dictionaries using databases on personal computers: construction of an abstract model of this dictionary; generation of structure and application software: lexicographic database; filling the lexicographic database; converting the lexicographic database to the appropriate final form [19].

Conclusions and prospects. The database in modern linguistics is the most effective technology for a compact representation of the set of parameters of linguistic units, the convenience and speed of processing the necessary data to achieve a specific research goal. The factual linguistic database is recognized as the most adequate tool for preserving lexical, phraseological, terminological, morphological and syntactic, stylistic units with the reflection of their characteristics. The use of database technology in linguistics contributes not only to the convenient presentation of materials and their unification into a single structure, but also to an increase in the efficiency of working with them, and also opens up new prospects for further research on the basis of created full-text databases or using factual databases.

We see the prospect of research in the development of a linguistic database of political neologisms in Ukrainian and English translations.

References:

1. Бігдай М.О. Ідеографічні параметри дієслівної лексики української мови : автореф. дис. ... канд. філол. наук : 10.02.01 «Українська мова». Чернівці, 2020. 20 с.
2. Галиева А.М. Представление глаголов физиологического действия и состояния в лексикографической базе данных татарских глаголов. *Ученые записки Казанского университета. Серия «Гуманитарные науки»*. 2018. Т. 160. Кн. 5. С. 1219–1234.
3. Калимон Ю.О. Структурно-інформаційна модель словника мови новел Василя Стефаника : автореф. дис. ... канд. філол. наук : 10.02.01 «Українська мова». Чернівці, 2020. 20 с.
4. Лотоцька Н.Я. Ідіолект Романа Іваничука: корпуснобазований та лінгвокогнітивний підходи : дис. … докт. філософії : 035 «Філологія» / Національний університет «Львівська політехніка». Львів, 2021. 324 с.
5. Карпіловська Є.А. Вступ до прикладної лінгвістики: комп'ютерна лінгвістика. Донецьк : ТОВ «Юго-Восток, ЛТД», 2006. 188 с.
6. Корпусна лінгвістика / В.А. Широков, О.В. Бугаков, Т.О. Грязнухіна та ін. Київ : Довіра, 2005. 471 с.
7. Мишанкина Н.А. Базы данных в лингвистических исследованиях. *Вопросы лексикографии*. 2013. № 1 (3). С. 25–33.
8. Рычкова Л.В. Полнотекстовые базы данных как материал для лингвистических исследований. *Язык. Образование. Компьютер*. Гродно, 2010. С. 112–115.
9. Лучик А., Остапова І. Синтагматична параметризація еквівалентів слова у парадигмі корпусної лінгвістики. *Human. Computer. Comunication (20-22 September, Lviv)*. 2017. С. 33–37.
10. Кочергина К.С. Лингвистическая база данных отрицательно-оценочной лексики: концепция, структура, наполнение. *Вестник Томского государственного университета*. 2019. № 446. С. 29–40.
11. Грязнухіна Т., Любченко Т. Електронні словники паронімів та їх використання в системах автоматичної обробки тексту. *Комп'ютерна лінгвістика: сучасне та майбутнє* : матеріали Міжнародної науково-практичної конференції (м. Київ, 23–24 лютого 2012 р.). Київ : КНЛУ, 2012. С. 15.
12. Загнітко А., Ситар Г., Данилюк І. Структура і модель бази даних «Українські частки та їхні еквіваленти». *Комп'ютерна лінгвістика: сучасне та майбутнє* : матеріали Міжнародної науково-практичної конференції (м. Київ, 23–24 лютого 2012 р.). Київ : КНЛУ, 2012. С. 21–22.
13. Мицкевич О.С. Лингвистическая база данных (ЛБД) специальной лексики белорусского языка с точки зрения потенциальных пользователей. *Прикладная лингвистика в науке и образовании* : сборник трудов VI Международной научной конференции (Санкт-Петербург, 5–7 апреля 2012 г.). Санкт-Петербург : ООО «Книжный дом», 2012. С. 203–206.
14. Дарчук Н.П., Лангенбах М.О. Електронний словник як дослідницька база даних (на прикладі електронного словника мови Тараса Шевченка). *Мовні і концептуальні картини світу*. 2014. Вип. 50 (1). С. 442–447.
15. Гарбера І. Інфологічний етап проектування лінгвістичної бази даних «Концепт «Людина» у фразеології східностепових українських говірок». *Славистика*. Београд, 2017. № 1–2. S. 112–118.
16. Гарбера І. Лінгвістична база даних «Концепт людина у фразеології східностепових українських говірок»: структура та функції. *Лінгвістичні студії/Linguistic Studies*. Вінниця, 2019. Вип. 37. С. 123–130.
17. Ситар Г. Синтаксичні фразеологізми в розрізі конструкційної граматики. Вінниця : ТОВ «Нілан-ЛТД», 2017. 458 с.
18. Краснобаева-Черная Ж.В. Терминологический банк данных «Классификационные параметры фразеологических единиц» как электронный терминографический продукт: опыт проектирования. *Вопросы лексикографии*. 2020. № 18. С. 117–132.
19. Кульчицький І., Костирко В. Укладання словників за технологією лексикографічної бази даних. *Вісник Державного університету «Львівська політехніка»*. 2000. № 402 : Проблеми української термінології. С. 119–122.

Sources:

1. FRANTEXT. URL: http://www.atilf.fr (дата звернення: 19.12.2021).
2. British National Corpus. URL: https://www.english-corpora.org/bnc/ (дата звернення: 19.12.2021).
3. Український національний лінгвістичний корпус Українського мовно-інформаційного фонду НАН України. URL: http://lcorp.ulif.org.ua/virt_unlc/ (дата звернення: 19.12.2021).
4. Национальный корпус русского языка. URL: http://www.ruscorpora.ru/ (дата звернення: 19.12.2021).
5. Český národní korpus. URL: https://korpus.cz/ (дата звернення: 19.12.2021).

Громовенко В. Основні типи баз даних у лінгвістичних дослідженнях XXI століття: особливості й функційне призначення

Анотація. У статті схарактеризовано основні напрями використання лінгвістичних баз даних (ЛБД) з послідовним розмежуванням повнотекстових і власне лінгвістичних баз даних у сучасному мовознавстві. Для успішної реалізації мети вирішено два основні завдання: 1) ознайомлення з повнотекстовими ЛБД та визначення специфіки їх наповнення та функціювання; 2) окреслення теоретико-прикладних засад фактографічних ЛБД (2011–2021 рр.). Актуальність статті мотивована відсутністю ґрунтовного аналізу досвіду створення й функціювання лінгвістичних баз даних у вітчизняній і закордонній лінгвістиці. Об'єктом дослідження постає ЛБД як сукупність систематизованих лінгвістичних даних; предметом – конкретні апробовані ЛБД. Основними методами постають метод критичного аналізу й дескриптивно-аналітичний метод. Методологічну базу статті становлять основні положення прикладної лінгвістики та корпусної лінгвістики. База даних у сучасному мовознавстві постає найефективнішою технологією для компактної репрезентації сукупності параметрів мовознавчих одиниць, зручності й швидкості опрацювання потрібних даних для реалізації конкретної дослідницької мети. Фактографічну лінгвістичну базу даних визнано найбільш адекватним інструментом збереження лексичних, фразеологічних, термінологічних, морфологічних і синтаксичних, стилістичних одиниць із відображенням їх характеристик. Використання технології БД в лінгвістиці сприяє зручному представленню матеріалів і їх об'єднанню в єдину структуру, а також підвищенню ефективності роботи з ними, відкриваючи нові перспективи для подальших досліджень на основі створених повнотекстових баз даних або з використанням фактографічних баз даних. Перспективу роботи бачимо в розробці лінгвістичної бази даних політичних неологізмів в українській та англійській мовах.

Ключові слова: база даних, лінгвістична база даних, корпус текстів, фактографічна база даних, прикладна лінгвістика.