УДК 81’32:004.65
DOI https://doi.org/10.26661/2414-9594-2020-2-18

ТЕКСТОВА БАЗА ДАНИХ FRANTEXT: СТРУКТУРА, ПАРАМЕТРИ, ПРИНЦИПИ ВИКОРИСТАННЯ

Страшко І. В.
кандидатка філософських наук,
докторантка кафедри прикладної лінгвістики, порівняльного мовознавства та перекладу
Національний педагогічний університет імені М. П. Драгоманова
вул. Тургенєвська, 8/14, Київ, Україна
orcid.org/0000-0001-5137-991X
i.v.strashko@npu.edu.ua

Ключові слова: *корпус текстів, французька мова, інформаційні системи, корпусна лінгвістика, прикладна лінгвістика.*

У статті розглядаються стан, детермінативні характеристики та історія розроблення французької текстової бази даних Frantext. Frantext – це унікальне зібрання французьких і франкомовних синхронно-діахронічних текстових корпусів не тільки за репрезентативністю матеріалу, а й за глибиною представлення різних його аспектів. Ця текстуальна база надає широкий спектр мовної інформації, відображаючи використання французької мови в усій своїй різноманітності. Принципи побудови та збагачення її новими текстами ґрунтуються на балансуванні різних епох та жанрів.

Frantext дає можливість користуватися наявними попередньо визначеними корпусами, створювати власні користувацькі корпуси і здійснювати широке коло пошукових завдань: за словоформою, лемою, граматичними категоріями, семантикою, сполученням лексичних одиниць тощо. Вибираючи корпус, комбінуючи команди і різні типи пошуку, користувач може диверсифікувати його застосування відповідно до своїх наукових інтересів і пріоритетів. Доступні опції експлуатації корпусних ресурсів дають змогу дублювання і видалення корпусу, редагування його назви та опису, створення власного корпусу за автором, датою, літературним жанром, додавання текстів, пошук метаданих для сортування та фільтрування творів.

Серед інших можливостей, запропонованих текстовою базою Frantext, є створення, редагування та перегляд списків слів, які можуть створюватися за допомогою регулярних виразів або CQL і містити окремі слова або їх послідовності.

Сьогодні управління текстовою базою здійснюється за допомогою програмного забезпечення Allegro, компоненти якого – індексатор, робоче середовище і сервер – забезпечують ефективну роботу користувачів із корпусами.

Обсяг, хронологічні межі, текстове наповнення цієї бази свідчать про те, що вона неухильно розвивається, нарощується її ефективність, зростає поле її застосування, збільшується коло користувачів.

Існування такого текстового ресурсу з інструментами для його використання у наукових дослідженнях мови становить інтерес для всіх, хто вивчає і викладає французьку мову як рідну або іноземну.

TEXT DATABASE FRANTEXT: STRUCTURE, PARAMETERS, APPLICATION PRINCIPLES

Strashko I. V.
Ph.D. in Philosophy,
Postdoctoral Student at the Department of Applied Linguistics,
Comparative Linguistics and Translation
National Pedagogical Dragomanov University
Turgenevska str., 8/14, Kyiv, Ukraine
orcid.org/ 0000-0001-5137-991X
i.v.strashko@npu.edu.ua

Key words: *text corpus, French language, information systems, corpus linguistics, applied linguistics.*

The paper focuses on the state, determinative characteristics and development history of the French text database. It is intended to make the French textual database Frantext known to Ukrainian researchers interested in text corpora. Enormous scientists' theoretical and practical activities have culminated in a representative collection of texts belonging to the different fields. In terms of the representativeness of its textual material, and presentation of its various aspects, Frantext is a unique collection of French-language corpora. This textual base provides a wide range of lingual information, reflecting the use of the French language in all its diversity. The principles of constructing and enriching the database with new texts are grounded on balancing different eras and diverse genres.

Frantext allows using predefined corpora, creating user's own corpora and performing a wide range of search tasks, such as: word form, lemmas, grammatical categories, semantics, combination of lexical items and more. By choosing a corpus and combining commands, and different types of queries, as a simple search, an assisted search, and an advanced search, the user can increase the number of applications according to proper scientific interests and priorities. The available options for exploiting corpus resources allow corpus duplication and removal, its title editing and description, making user's own corpus by the author, date, literary genre, etc. Texts adding, metadata search for sorting and filtering texts are also possible.

Among the other possibilities offered by the Frantext textbase is the wordlists' creation, editing and viewing. The last ones can be constituted using regular expressions or CQL and can contain single words or their sequences.

Today, the textbase is operated through software called Allegro, whose components such as an indexer, a runtime environment, and a server ensure that users work efficiently with the corpus.

The volume, chronological framework, textual content of this base indicate that it is steadily developing, its effectiveness is increasing, its fields of application are augmenting, users number is growing.

The existence of such corpora, with tools for their use in scientific and educational research, in order to get acquainted with the structure and functioning of the French language, is of interest to the French-speaking community, and all who study and teach it as a native or foreign language.

Дослідження з корпусної лінгвістики та автоматичної обробки текстів вимагають використання величезних лінгвістичних ресурсів: словників, текстів та корпусів, а також інструментів для управління ними та їх аналізу. Питання якості та доступності корпусів, практик їх застосування залишається ключовим не лише для лінгвістичних досліджень, а й багатьох інших галузей гуманітарного знання. Отже, зберігається необхідність у великих стандартизованих, анотованих та затверджених корпусах. У зв'язку із цим заслуговує на увагу досвід Франції, де протягом кількох десятиліть ведеться інтенсивна робота щодо створення текстових корпусів.

У цій статті мова піде про французьку і франкомовну текстову базу даних Frantext, яка, з одного боку, відзначається належною лінгвістичною якістю, а з іншого – є доступною в мережі Інтернет.

Незважаючи на те що «дослідження мови на базі електронних корпусів сьогодні стали одним з основних дослідних методів у лінгвістиці» [1, с. 154], в українській мовознавчій традиції практично відсутні спеціальні систематичні розвідки, які б стосувалися опису саме французьких корпусів: наявні лише окремі згадування, зокрема В.В. Жуковська [2], аналізуючи наявні національні корпуси, називає й Frantext. Серед французьких дослідників, які досить детально розглядали функціонал цієї текстової бази, доцільно відзначити роботи P. Bernard, J. Dendien, J. Lecomte, E. Martin, É. Petitjean, J. Pierrel.

З огляду на це, **актуальність дослідження** зумовлена відсутністю аналізу функціонування французької текстуальної бази Frantext в українському науковому дискурсі.

Мета – теоретичне дослідження детермінативних параметрів та узагальнення досвіду використання французької бази даних Frantext.

Frantext – це текстова база даних, розроблена в лабораторії комп'ютерного аналізу та обробки французького лексикону ATILF (Analyses et Traitements Informatiques du Lexique français). Варто зазначити, що лабораторія пропонує набір комп'ютеризованих ресурсів, які складаються з текстової та лексикологічної бази, серед яких переважно виділяють власне Frantext [3] і комп'ютеризовану версію «Тезауруса французької мови» (Trésor de la Langue Française informatisé [4], знаного під абревіатурою TLFI), яка базується на «Тезаурусі французької мови» XIX і XX століть (Le Trésor de la Langue Française) у 16 томах, виданому друком за період 1976–1994 років Національним інститутом французької мови. З 2002 року TLFI є у вільному доступі в мережі. Як указують розробники, комп'ютеризована версія містить 100 000 слів з їх історією, 270 000 визначень, 430 000 прикладів. Проте, оскільки написання «Тезауруса….» завершилося у 1994 році, цей ресурс більше не оновлюється і є закритим так «як є» (фр. *en l'état*) [4] (тут і далі переклад мій. – І. С.).

Уже сама ідея створення бази з набору текстів з їх автоматичною обробкою була революційною для лексикографічного ландшафту 70-х років XX сторіччя [5]. Спочатку її основною метою було конституювання «файлів слів» для використання авторами-редакторами під час розроблення словника TLF: стаття в TLF, присвячена певному слову, була наділена його систематичним узгодженням, відібраним за різними критеріями (за алфавітом, граматичними категоріями лівого або правого контекстів) [6]. База використовувалася і на завершальній стадії редагування слова для відбору тексту з прикладів, наведених у TLF. Конституювання файлів слів і вилучення остаточно збережених прикладів надавалися важким неінтерактивним програмним забезпеченням (тип пакетної обробки) з послідовною обробкою корпусу. Після того як словник був завершений, база даних продовжувала розвиватися: вперше доступна у 90-х роках у формі CD-ROM Discotext, вона була розміщена в Інтернеті в 1998 році разом із першою пошуковою системою Stella.

У 1980-х роках лабораторія ATILF створила текстову платформу, яка дала змогу вражаюче підвищити продуктивність завдяки можливості прямого доступу до слів корпусу. У 1985 році вона забезпечила реалізацію проєкту користувальницького інтерфейсу з використанням терміналів Transpac і Minitel. Приблизно 90% із 430 000 зразків, процитованих у TLF / TLFi, було взято з Frantext [7; 8].

Джерельне наповнення текстуальної бази поступово збільшується: за даними, наведеними на сайті Frantext (https://www.frantext.fr), у 1992 році база нараховує 2 345 текстів, доступних на постійній та інтерактивній основі. Шляхом традиційного введення, оптичного зчитування та придбання фотокомпозиційних стрічок творів усіх століть відбувається планомірне зростання бази, і вона перетворюється на один із найбільших доступних франкомовних ресурсів [9].

Е. Мартен, характеризуючи функціонал цієї мономовної текстуальної бази, обстоює ідею, що «Frantext ілюструє, крім фактів мови, факти, проілюстровані мовою» (фр. *Frantext illustre, outre les faits de langue, les faits illustrés par la langue*) [9]. Величезна теоретична і практична діяльність науковців увінчалася репрезентативною колекцією текстів, які належать до галузей науки, мистецтва та техніки. А це, підкреслює Е. Мартен, приблизно 160 мільйонів цитат, отриманих у результаті комп'ютерної обробки чотирьох століть літератури, близько 600 000 друкованих сторінок, 300 000 форм, понад 100 мільйонів слововживань і понад мільярд символів. У цей період у Frantext уже представлено близько 900 письменників (майже 450 видавців). У виборі видань, акцентує дослідниця, укладачі керувалися бажанням записати текст першого видання кожного твору. Оригінальні корпуси були вибрані групою фахівців із назв, відібраних за частотою їх появи в основних бібліографіях. Що стосується власне художніх текстів, які становили орієнтовно 80% від загальної кількості, спостерігається справедливий баланс залежно від розподілу в часі (приблизно 6 мільйонів слів на десятиліття) та представництва жанрів. У кожному періоді

представлені в порядку спадання: роман, театр, поезія, мемуари, листування, подорожі, брошури, ораторське мистецтво. Основними науково-технічними галузями, представленими під жанром «трактат або есе», були такі: державне управління, мистецтво, астрономія, будівництво, біологія, хімія, літературознавство, право, економіка, енергетика, етнологія, історія, інформація, лінгвістика, відпочинок, математика, окультизм, філософія, фізика, політика, психологія, релігії, науки про землю, спорт. Зазначені жанри та галузі призначаються як дескриптори кожному тексту в базі даних, окрім власне бібліографічних описів, які дають змогу скласти відповідний робочий корпус. До цього корпусу, уточнює дослідниця, було додано набір текстів, уведених у Лабораторії лексичного аналізу [9].

У цей час, як, утім, і сьогодні, здійснюється постійне спостереження за текстуальною базою з кількома цілями:

− перевірка якості введення даних: певна частина матеріалів потребує внесення змін, оскільки все ще перебуває у первинному стані початку 1960-х років, що пов'язане із технологією тих часів, коли використовувалася перфорована стрічка;

− контролювання якості видань: хороша якість видань, що розглядаються для збору даних, і належним чином оформлене посилання на них – одна з характерних особливостей Frantext. Оскільки деякі публікації можна вважати застарілими та/або ненадійними, вони є предметом корегування;

− поповнення корпусів: політика формування, наповнення і збагачення бази даних новими текстами ґрунтується на принципах збалансування різних епох та різних жанрів і сприяння конкретним дослідницьким та навчальним завданням [6].

«Робочі корпуси – це те, чим ми хочемо, щоб вони були (фр. *Les corpus de travail sont ce qu'on veut qu'ils soient*), – констатує Е. Мартен [9] і робить висновок, що, користуючись ресурсами Frantext уже на ранньому етапі її розвитку, спеціаліст-літературознавець мав змогу вибрати свій корпус: твір або певні твори письменника, літературну продукцію певного хронологічного періоду, епохи, тексти, що належать до певного жанру чи до конкретної галузі. Спеціаліст-мовник міг працювати із цілим корпусом або з великими групами текстів. В останньому випадку машина потім сама сортувала слово/слова за допомогою функції індексу – для визначення місця розташування слова, за допомогою функції пошуку – для ілюстрації його контекстів і за допомогою функції частотності – для підрахунку його вживання. Система також надавала можливість створення списків словоформ. У разі складених форм списки автоматично формувалися навколо інфінітиву та з попередньо встановленого довідкового словника. Для проведення досліджень у галузі семантики можливості Frantext у цей період менш суттєві, зазначає вона, оскільки пошук місця розташування слова стосується тільки його форми. Однак можна було класифікувати хронологічні списки свідчень уживання слова, у яких описуються умови його використання, характер семантичного зсуву, еволюція, зміна сфери вживання слова. У сфері синтаксису діапазон запитів був більш широкий: можна було знайти приклади множинних конструкцій, вирази, синтаксичні теми. Особливо вона відзначає внесок Frantext у пошук лексичних одиниць із переносним значенням, дослідження нюансів значення слів, жодного свідчення яких не було зафіксовано у традиційних паперових словниках. Навіть за відсутності семантичного сортування, особливо в часто цитованих випадках омографів і багатозначних слів, підкреслює вона, існувала можливість усунути деякі двозначності, зокрема шляхом вибору корпусів, місця слів, усунення збігів. Контексти, надані базою даних, також давали змогу розвинути бібліографію терміну та поняття, яке він охоплює, наприклад шляхом запису цитат авторів, зроблених у більш-менш тісному середовищі [9].

Тобто вже на початковому етапі функціонування Frantext шляхом вибору корпусів і поєднання команд користувач міг збільшити кількість застосувань відповідно до своїх потреб, а саме: проілюструвати прикладами відоме значення слова, підтвердити, засвідчити нове, рідкісне або тільки ймовірне значення, знайти цитату, свідчення вживання, створити файл послідовностей, що функціонують за даною моделлю, датувати форму слова або фразу, вибрати приклади визначень, скласти перелік авторів, ідентифікувати тип мовлення, визначити синтаксичні теми, перелічити орфограми тощо.

Поступово початковий сенс існування Frantext – на службі «Тезаурусу…» – був витіснений бажанням зробити доступним науковому співтовариству вдосконалену колекцію текстових матеріалів з ефективними інструментами пошуку. Спочатку зосереджена на мові XIX і особливо XX століття, база продовжувала розвиватися, розширюючи свій діахронічний простір, виформовуючись як зібрання корпусів: додавалися середньовічні тексти, тексти на середньофранцузькій (приблизно 300 текстів), докласичній та класичній французькій мові [3].

У 2003 році Frantext уже нараховує 3 665 текстів, датованих із 1507 по 1998 рік, і містить близько 80% художніх текстів (у повній версії) та 20% технічних, що представляють основні наукові дисципліни [8].

Дані щорічного зростання кількості текстів, наведені на сайті Frantext (https://www.frantext.fr/), свідчать про те, що база регулярно збагачується: у 2004 році вона налічувала 3 737 текстів, у 2008 році – вже 3 911, 2009 рік показує 3 985 робіт, у 2011 – 4 084 тексти, у 2012 році додалось 164 нових тексти, у 2013 році загальна кількість робіт сягає 4 515, у 2014 році база збільшилася на 16 текстів, 2015 рік дає вже 4 746 текстів, у 2016 році – 5 116 текстів. Таке зростання стало можливим завдяки науковій співпраці з різними науково-дослідними інститутами. Амбітна мета авторів цього проєкту – відобразити використання писемної французької мови в її різноманітності – від літературної класики до кулінарних посібників і мисливських трактатів. Репрезентативна вибірка також включає лінгвістичні праці, підручники з географії, журналістські спогади, сучасні романи чи навіть так звані «звичайні» твори.

До липня 2018 року база Frantext функціонувала з програмним забезпеченням STELLA (фр. Système de Textes en Ligne en Libre Accès). Це система Інтернет-текстів із відкритим доступом, розроблена в лабораторії INaLF (нині ATILF) Жаком Дендієном (Jacques Dendien). Вона працювала під управлінням MULTICS на сервері Міжрегіонального комп'ютерного центру Лотарингії (CIRIL) у Нансі [9].

Програмне забезпечення STELLA було представлене у вигляді набору інструментів (C ++) з такими компонентами:

1) утиліти, які включають сортування, обробку регулярних виразів, переважно з бази даних на основі номенклатури TLF, що дає змогу виконувати операції флексії або лематизації;

2) вебінтерфейс, який дає змогу легко реалізувати користувальницький інтерфейс, функції управління «сеансами користувача», гіпернавігацію між різними програмами, керованими STELLA, незалежно від того, знаходяться вони на одному сервері чи ні;

3) систему управління текстовою базою, яка забезпечує функції зберігання та доступу до інформації [8].

Оскільки програмне забезпечення STELLA, яке було розроблене в кінці 1980-х років, більш не здатне відповідати новим IT-завданням, виникла ідея розробити нову, простішу в обслуговуванні пошукову систему, яку з плином часу можна було б модернізувати. Система отримала назву Allegro. Як зазначають її розробники, на даний момент Allegro працює лише на серверах ATILF, але в перспективі планується запропонувати науковому співтовариству безкоштовний інструмент, що дасть змогу користувачам визначати власні сховища зі своїми файлами та їх метаданими. Порівняно з попередньою ця платформа більш зручна у застосуванні: вона пропонує алгоритми пошуку для оптимізації часу відгуку під час пошуку форм у лексиконі або виконання корпусних запитів [8; 10].

Allegro складається з трьох різних програмних компонентів: індексатора, робочого середовища і сервера. Індексатор приймає дані і метадані як вхідні, реструктурує їх і створює оптимізований вихідний формат, який дає змогу здійснювати ефективний пошук як даних, так і їх структури. Середовище виконання дає змогу визначати корпус, виконувати запити й отримувати результати у вибраному форматі виведення. Сервер інкапсулює індексатор і середовище виконання, щоб забезпечити доступ із вебсервера до всіх їх функцій. Платформа дає змогу використовувати текстові корпуси, що містять будь-яку кількість шарів анотації, а розмір цих корпусів обмежується лише доступною оперативною пам'яттю. Як відзначає Етьєн Петіжан, один з основних розробників Allegro, індексація всієї бази даних Frantext здійснюється приблизно за хвилину [10].

Отже, з 2018 року Frantext має оновлений інтерфейс, нову платформу Allegro, нові функції, включаючи використання регулярних виразів та CQL, збагачений, лематизований і повністю категоризований єдиний корпус текстів зі стандартними функціями Frantext та новими інструментами пошуку і візуалізації.

Сьогоднішній Frantext (https://www.frantext.fr/) дає змогу здійснювати простий і складний пошук щодо словоформ, лем, граматичних категорій, використання регулярних виразів та відображає результати в контексті 700 символів. Нині база включає французькі і франкомовні твори й містить 10% так званих «наукових» та технічних текстів і 90% тих, що вважаються «літературними», об'єднуючи всі жанри: романи, мемуари, автобіографії, щоденники, театр, поезію, есе. Різні елементи, що становлять Frantext, оновлюються кілька разів на рік.

Починаючи з жовтня 2020 року почала функціонувати версія Frantext 20.1, так звана «Агрегація 2021», яка включає тексти в програму агрегування. Станом на жовтень 2020 року Frantext містить 5 469 посилань, або 258 мільйонів слововживань [3].

Сьогодні ознайомитися з усією різноманітністю текстів, наявних у Frantext для навчальних та дослідницьких цілей, можуть науковці, викладачі-дослідники, студенти та наукові співробітники. Доступ до текстуальної бази мають також університети, лабораторії, дослідницькі центри, центри документації, бібліотеки, медіатеки.

Frantext доступний в Інтернеті у трьох версіях:

− повна версія за передплатою (3 665 текстів), у якій можна вести пошук за графічними формами тексту;

− класифікована, де представлено частину доступних матеріалів (1 940 текстів), перегляд якої можливий після індивідуального запиту на підписку;

− демонстративна версія, яка пропонує добірку із сорока текстів, не захищених авторським правом, відкритий доступ та безкоштовне тестування функційних можливостей.

Перші дві доступні за підпискою установи або за персональною підпискою, отриманою від ATILF. Вона дійсна протягом календарного або навчального року. Сума передплати використовується для покриття витрат на обслуговування бази та її збагачення. Індивідуальна передплата дає змогу входити за паролем, тоді як інституційна реалізується через прямий доступ за допомогою розпізнавання IP-адреси. Окрім того, у межах партнерства з Національним синдикатом видавців, за платною підпискою надаються додаткові консультації [3; 7].

Варто підкреслити, що у Frantext існує можливість працювати і з попередньо визначеними корпусами, серед яких:

− старофранцузький корпус, де наведено тексти до 1300 року;

− корпус середньофранцузької мови, куди входить зібрання творів 1300–1549 років;

− докласичний корпус репрезентує збірку текстів 1550–1649 років;

− класичний корпус, який охоплює період 1650–1799 років;

− модерн, скомпонований із текстів 1800–1959 років;

− корпус RL-fr: 1950+ – це довідковий корпус французької лексичної мережі (RL-fr), де зібрані тексти з 1950 року;

− сучасний корпус, у якому представлені твори з 1980 року по теперішній час;

− корпус XX століття репрезентує корпус творів XX століття;

− повний корпус, який включає корпус усіх документів [3].

Попередньо визначені корпуси є загальними для всіх користувачів. Змінити попередньо визначений корпус неможливо, але його можна продублювати. Під час завантаження певного корпусу він стає придатним для використання із застосуванням інструментів Frantext. Можна дублювати і видаляти корпус, додавати тексти, редагувати назву та опис корпусу, шукати метадані корпусу для сортування та фільтрування творів.

Кілька слів про принципи роботи з Frantext. Перший етап – вибір корпусу роботи. Можна вибирати з усіх наявних текстів, одного або кількох авторів, твір або сукупність творів, один або кілька літературних жанрів, певний хронологічний розділ або поєднати кілька критеріїв. Додавання текстів до корпусу здійснюється шляхом вибору метаданих із текстової бази даних Frantext. Тобто можна створити власний корпус за автором, датою, літературним жанром тощо. Особисті корпуси зберігаються в обліковому записі користувача, якщо він пройшов автентифікацію, в іншому разі – у браузері.

Другий етап – власне робота з корпусом. Принагідно відзначимо, що кожен пошуковий запит супроводжує Інтернет-довідка.

Frantext дає змогу виконувати різні типи пошуку: конкорданс, частота, суміжність тощо. Простий пошук дає змогу швидко шукати в текстах основного твору слово чи ряд слів, фразу, декілька варіантів написання в одному або кількох реченнях. Керований – дає можливість виконувати складні пошуки спрощеним способом із використанням випадних меню та поєднанням різної інформації (наприклад, форма + лема). Розширений пошук дає змогу проводити дослідження безпосередньо мовою CQL [3].

Наведемо приблизний, а отже, неповний перелік можливостей системи для здійснення більш складного пошуку: відмінкові форми дієслова, іменника чи прикметника; усічені форми; списки слововживань; вирази з множинним вибором в одному запиті; статистичні характеристики лексичних одиниць; вивчення словникового запасу у реченнях, що містять тільки конкретне слово; складні мовні явища, такі як кількісне визначення, займенникові конструкції, складні часи тощо.

Пошук такого роду стає можливим шляхом написання формальних граматик і складається з налаштованих правил. Ці граматики дають змогу здійснювати пошук довільно складних контекстів у корпусі. Також завдяки настроюваним граматикам можна вказувати на словесні, прикметникові та субстантивні звороти. Під «граматикою» мається на увазі серія комбінованих пошуків. Можна продублювати заздалегідь визначену граматику, але не змінювати її. Заздалегідь визначені граматики є загальними для всіх користувачів. Граматики, які користувач створює для своїх пошукових запитів, є особистими та зберігаються у його браузері. Граматики корисні для розширеного пошуку та визначаються за допомогою мови, що відповідає певній платформі, і можуть містити списки та правила. Граматичні правила – це підмножини, які можна поєднувати між собою за допомогою логічних посилань.

Усі тексти корпусу класифікуються: вони позначені як частина мови або POS. Це означає, що кожному слову присвоюється граматичний тег: дієслово, прислівник, прикметник тощо. Оскільки будь-який граматичний тег має помилки, їх статистично більше буде у старих версіях текстів [3].

Морфосинтаксична анотація дає змогу розрізняти вживання слів. Саме позиція, узгодження і поєднання форм дають змогу визначити використання і вибрати найбільш відповідну граматичну категорію. У Frantext (https://www.frantext.fr), кожна форма пов'язана з категорією, тобто з унікальним граматичним тегом.

Для проведення діахронічних досліджень слід скористатися додатковим режимом пошуку, який полягає у використанні «флексії». У цьому разі пошук базується на лексиконах, у яких слово пов'язане з основною формою. Frantext (https://www.frantext.fr) дає можливість пошуку корпусу за допомогою трьох флексій: сучасної, середньовічної і флексії XVI–XVII століть. Останні дві присвоюють лемкові мітки, спираючись на лексикон LGeRM, адаптований для врахування графічних варіацій.

Вкладка «частотність» дає змогу розрахувати частоту слова чи регулярного виразу (за десятиліттями, століттями тощо) в особистому корпусі користувача. За допомогою встановленого таким чином обсягу роботи (твір, жанр, століття чи навіть весь корпус) можна з'ясувати частоту вживання слова, послідовності, співіснування слів або послідовностей слів. Вкладка «суміжність» дає змогу шукати збіг, одночасну появу двох або трьох послідовностей слова/слів, регулярних виразів або навіть граматики, а також переглядати їхній лівий та правий контексти. За допомогою цієї опції складається відсортований список слів в алфавітному порядку за зростанням або за спаданням частоти заданого слова. Вкладка «сусідство» дає змогу вивчити колокації слова, регулярного виразу, списку слів або граматики [3].

Можна створювати, редагувати та переглядати списки слів. Списки можуть містити окремі слова або послідовності слів, можуть створюватися за допомогою регулярних виразів або виразів CQL. Списки є особистими і зберігаються у браузері користувача. Заздалегідь визначений список можна продублювати, але змінювати його не можна. Попередньо визначені списки є загальними для всіх користувачів. Списки слів можуть бути використані повторно, зокрема для проведення пошуку чи вивчення сусідства. Після відображення результатів пошуку в конкордансі або контексті можна уточнити список, використовуючи інструменти, які дають змогу сортувати результати, фільтрувати метадані або змінювати пошук за допомогою контекстного меню. Оскільки простий, допоміжний та розширений пошуки вимагають великої кількості комп'ютерних процедур для кожного запиту, кількість результатів обмежена максимум 100 000. Тобто ступінь складності пошуку залежить виключно від бажання дослідника. Після проведення дослідження можна завантажити результати. Вони кодуються відповідно до стандарту ISO 8859-1, сумісного з використанням основних систем (MS-Windows, Unix, Mac-OS) [3].

Отже, Frantext є унікальною базою не лише за репрезентативністю текстового матеріалу, а й за глибиною представлення різних його аспектів. Її використання вже виходить далеко за межі, у яких була впроваджена система. Як і очікувалося, вона переважно використовується спеціалістами-мовниками як документальне джерело для вивчення, аналізу, спостереження, дослідження текстів. Проте користуються нею й дослідники суміжних дисциплін для проведення соціологічних, історичних, бібліографічних, юридичних розвідок.

Розгляд структури текстуальної бази Frantext, детермінативних характеристик, обсягу, хронологічних меж, сфери використання, текстового наповнення свідчить про те, що вона неухильно розвивається, нарощується її ефективність, зростає поле її застосування, збільшується коло користувачів. Існування в межах одного ресурсу французьких і франкомовних корпусів з інструментами для їх використання у наукових дослідженнях мови або для швидкої та ефективної перевірки особливостей вживання незнайомого слова чи граматичної форми, ознайомлення з будовою і функціонуванням мови становить інтерес для франкомовної спільноти – всіх, хто вивчає і викладає її як рідну або як іноземну.

Предметом наших подальших розвідок буде огляд та порівняльний аналіз сучасних французьких корпусів.

ЛІТЕРАТУРА

1. Демська-Кульчицька О. Дещо про класифікацію текстових корпусів. *Наукові записки. Серія «Мовознавство»*. 2004. Т. 1. №. 11. С. 153–157.
2. Жуковська В.В. Вступ до корпусної лінгвістики : навчальний посібник. 2013.
3. ATILF. Base textuelle Frantext (En ligne). ATILF-CNRS & Université de Lorraine. 1998–2020. URL: https://www.frantext.fr/ (дата звернення: 17.10.2020).
4. TLFi: Trésor de la langue Française informatisé. URL: http://www.atilf.fr/tlfi, ATILF – CNRS & Université de Lorraine.
5. Montémont V., Manea L.L. Lévolution de Frantext: quelles modifications et quels usages pour Frantext 2? *XXVIIe Congrès International de Linguistique et de Philologie Romanes*, Nancy, 2013.

6. Bernard P., Dendien J., Lecomte J., Pierrel J.M. Les ressources de lATILF pour lanalyse lexicale et textuelle: TLFi, Frantext et le logiciel Stella. *Actes des 8e Journées Internationales dAnalyse Statistique des Données Textuelles JADT*. 2002. P. 137–149.
7. Bernard P., Lecomte J., Dendien J., & Pierrel J.M. Computerized linguistic resources of the research laboratory ATILF for lexical and textual analysis: Frantext, TLFi, and the software Stella. In LREC. 2002, May.
8. Pierrel J.M. Un ensemble de ressources de référence pour létude du français: TLFi, Frantext et le logiciel Stella. *Revue québécoise de linguistique*. 2006. Vol. 32. №. 1. P. 155–176.
9. Martin, É. Frantext la base de données textuelles du français. *Revue Roumaine de linguistique*. 1992. Vol. 37. №. 5-6. P. 331–340.
10. Petitjean, É., Benzitoun, C., Husson, B., & Ollinger, S. Allegro: une plateforme «couteau suisse» pour lexploitation des ressources textuelles. 2019.

REFERENCES

1. Demska-Kulchytska O. (2004) Deshcho pro klasyfikatsiyu tekstovykh korpusiv [Something about the classification of text corpora]. *Naukovi zapysky. Seriya: Movoznavstvo*. Vol. 1. №. 11. P. 153–157.
2. Zhukovska, V.V. (2013). Vstup do korpusnoyi linhvistyky: navchalnyy posibnyk [Introduction to corpus linguistics: textbook].
3. ATILF. Base textuelle Frantext (En ligne). ATILF-CNRS & Université de Lorraine. 1998-2020. Retrieved from https://www.frantext.fr/ (17.10.2020).
4. TLFi : Trésor de la langue Française informatisé, http://www.atilf.fr/tlfi, ATILF – CNRS & Université de Lorraine.
5. Montémont, V., Manea, L. (2013). L. Lévolution de Frantext: quelles modifications et quels usages pour Frantext 2? *XXVIIe Congrès International de Linguistique et de Philologie Romanes*, Nancy.
6. Bernard, P., Dendien, J., Lecomte, J., & Pierrel, J.M. (2002). Les ressources de lATILF pour lanalyse lexicale et textuelle: TLFi, Frantext et le logiciel Stella. Actes des 8e Journées Internationales dAnalyse Statistique des Données Textuelles JADT, P. 137–149.
7. Bernard, P., Lecomte, J., Dendien, J., & Pierrel, J.M. (2002, May). Computerized linguistic resources of the research laboratory ATILF for lexical and textual analysis: Frantext, TLFi, and the software Stella. In LREC.
8. Pierrel, J. (2006). Un ensemble de ressources de référence pour létude du français : tlfi, frantext et le logiciel stella. *Revue québécoise de linguistique*. Vol. 32. № 1. P. 155–176.
9. Martin, É. (1992). Frantext la base de données textuelles du français. *Revue Roumaine de linguistique*. Vol. 37. №. 5-6. P. 331–340.
10. Petitjean, É., Benzitoun, C., Husson, B., & Ollinger, S. (2019). Allegro: une plateforme «couteau suisse» pour lexploitation des ressources textuelles.