

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
«ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»

Є. В. Купріянов

КОМП'ЮТЕРНА ЛЕКСИКОГРАФІЯ

Навчальний посібник
для бакалаврів напрямку підготовки
«035» філологія
денної та заочної форм навчання

Затверджено
редакційно-видавничою
радою університету,
протокол № 1 від 19.02.2026 р.

Харків
НТУ «ХПІ»
2026

УДК81'373:004
К 91

Рецензенти:

В. А. Висоцька, д-р техн. наук, проф. кафедри інформаційних систем та мереж, Національний університет «Львівська Політехніка»;

І. В. Остапова, канд. техн. наук, ст. наук. співробітник, Український мовно-інформаційний фонд НАН України

Купріянов Є. В.

К 91 Комп'ютерна лексикографія : навчальний посібник для бакалаврів
напрямку підготовки денної та заочної форм навчання / Є. В. Купріянов.
Харків : НТУ «ХП», 2026. 376 с.

ISBN 978-617-05-0614-6

У навчальному посібнику розглянуто теоретичні та прикладні засади сучасної комп'ютерної лексикографії, а також технології її застосування у словникарських проєктах різних типів. Особливу увагу приділено принципам укладання електронних словників, моделям лексикографічного опису, форматам репрезентації мовного матеріалу, онлайн-сервісам, корпусним технологіям і засобам штучного інтелекту.

Теоретичні положення ілюстровано численними прикладами з реальної лексикографічної практики; кожен тематичний розділ супроводжується контрольними запитаннями та завданнями для самостійної роботи, що сприяє формуванню практичних навичок аналізу й укладання словникових ресурсів.

Призначено для студентів спеціальності «Філологія» денної форми навчання, які опанували курси загального мовознавства, структурної лінгвістики та методів лінгвістичного аналізу, а також планують подальше навчання у сфері прикладної й комп'ютерної лінгвістики.

Іл. 83. Табл. 73. Бібліогр.: 64 назви.

УДК 81'373:004

ISBN 978-617-05-0614-6

© Купріянов Є. В., 2026
© НТУ «ХП», 2026

ЗМІСТ

ВСТУП.....	7
РОЗДІЛ 1. СЛОВНИКИ В ЦИФРОВОМУ СЕРЕДОВИЩІ.....	9
1. СЛОВНИК: ЙОГО СТРУКТУРА, ЗМІСТ І МЕТАМОВА	9
1.1. Про поняття словник та його будову.....	9
Питання для самоконтролю	18
Завдання для самоперевірки	18
1.2. Лексикографічна параметризація мовних одиниць.....	19
Питання для самоконтролю	24
Завдання для самоперевірки	25
1.3. Метамова словника	26
Питання для самоконтролю	35
Завдання для самоперевірки	36
Питання для поглибленого вивчення матеріалу.....	36
2. ІНФОРМАЦІЙНО-ЛІНГВІСТИЧНИЙ АСПЕКТ ЕЛЕКТРОННОГО СЛОВНИКАРСТВА	37
2.1. Картина мовного світу.....	38
Питання для самоконтролю	43
Завдання для самоперевірки	44
2.2. Поняття семантичного стану	44
Питання для самоконтролю	51
Завдання для самоперевірки	52
2.3. Лексикографічна система та лексикографічні середовища	53
Питання для самоконтролю	59
Завдання для самоперевірки	59
Питання для поглибленого вивчення матеріалу.....	60
3. СЛОВНИКИ В ГЛОБАЛЬНІЙ ЛЕКСИКОГРАФІЧНІЙ МЕРЕЖІ	61
3.1. Всесвітня лексикографічна мережа.....	61
Питання для самоконтролю	70
Завдання для самоперевірки	71
3.2. Локальні та мережеві лексикографічні ресурси.....	72
Питання для самоконтролю	81
Завдання для самоперевірки	82
3.3. Інтеграція онлайн-ресурсів у єдиний глобальний простір	83
Питання для самоконтролю	89
Завдання для самоперевірки	90
Питання для поглибленого вивчення матеріалу.....	91
РОЗДІЛ 2. КОМП'ЮТЕРНА ЛЕКСИКОГРАФІЯ: НАУКОВА І ПРАКТИЧНА ДІЯЛЬНІСТЬ.....	92
4. КОМП'ЮТЕРНА ЛЕКСИКОГРАФІЯ: ЇЇ ПРЕДМЕТ І ЗАВДАННЯ	92

4.1. Сучасна комп'ютерна лексикографія та її головні завдання	92
Питання для самоконтролю	107
Завдання для самоперевірки	108
4.2. Про поняття електронного словника	108
Питання для самоконтролю	119
Завдання для самоперевірки	120
4.3. Етапи створення комп'ютерних словників.....	121
Питання для самоконтролю	130
Завдання для самоперевірки	130
Питання для поглибленого вивчення матеріалу.....	131
5. МОВА ЯК ОБ'ЄКТ КОМП'ЮТЕРНОЇ ЛЕКСИКОГРАФІЇ	132
5.1. Моделі репрезентації мови в словнику.....	132
Питання для самоконтролю	140
Завдання для самоперевірки	141
5.2. Формати збереження лексикографічної інформації	142
Питання для самоконтролю	153
Завдання для самоперевірки	154
5.3. Поняття концептуальної моделі словника.....	155
Питання для самоконтролю	161
Завдання для самоперевірки	162
Питання для поглибленого вивчення матеріалу.....	163
6. ТИПОЛОГІЯ КОМП'ЮТЕРНИХ СЛОВНИКІВ	164
6.1. Принципи класифікації комп'ютерних словників.....	164
Питання для самоконтролю	169
Завдання для самоперевірки	170
6.2. Функціональна, структурна та технологічна типології	170
Питання для самоконтролю	183
Завдання для самоперевірки	183
6.3. Інші типи електронних словників	184
Питання для самоконтролю	188
Завдання для самоперевірки	189
Питання для поглибленого вивчення матеріалу.....	190
РОЗДІЛ 3. ТЕХНОЛОГІЇ У КОМП'ЮТЕРНІЙ ЛЕКСИКОГРАФІЇ	191
7. СЕРЕДОВИЩА ТА ІНСТРУМЕНТИ ДЛЯ УКЛАДАННЯ ЕЛЕКТРОННИХ СЛОВНИКІВ.....	191
7.1. Середовища розроблення словників (DWS)	191
Питання для самоконтролю	201
Завдання для самоперевірки	202
7.2. Віртуальні лексикографічні лабораторії (ВЛЛ)	203
Питання для самоконтролю	213
Завдання для самоперевірки	214

7.3. Інструменти й сервіси для лексикографів	214
Питання для самоконтролю	222
Завдання для самоперевірки	223
Питання для поглибленого вивчення матеріалу	223
8. КОРПУСНІ ТЕХНОЛОГІЇ У КОМП'ЮТЕРНІЙ ЛЕКСИКОГРАФІЇ.....	225
8.1. Текстові корпуси як основа для створення словників.....	225
Питання для самоконтролю	239
Завдання для самоперевірки	240
8.2. Корпусні інструменти для створення словників.....	242
Питання для самоконтролю	256
Завдання для самоперевірки	256
8.3. Електронні словники в корпусній лінгвістиці	257
Питання для самоконтролю	267
Завдання для самоперевірки	267
Питання для поглибленого вивчення матеріалу	268
9. ШТУЧНИЙ ІНТЕЛЕКТ В ЕЛЕКТРОННОМУ СЛОВНИКАРСТВІ	269
9.1. Можливості ШІ інструментів для створення словників	269
Питання для самоконтролю	280
Завдання для самоперевірки	281
9.2. Огляд ШІ інструментів для роботи з мовним матеріалом	282
Питання для самоконтролю	295
Завдання для самоперевірки	295
9.3. Створення словників і глосаріїв засобами ChatGPT	296
Питання для самоконтролю	311
Завдання для самоперевірки	312
Питання для поглибленого вивчення матеріалу	313
ПІСЛЯМОВА	314
ГЛОСАРІЙ.....	316
СПИСОК ДЖЕРЕЛ ІНФОРМАЦІЇ.....	345
КОРИСНІ ПОСИЛАННЯ	353
Онлайн-словники	353
Текстові корпуси.....	355
Додаткові Інтернет-ресурси	356
ДОДАТКИ	357
Д.1. Лексикографічні параметри словникової статті	357
Д.2. Метамова у різних типах словників	357
Д.3. Порівняльна таблиця лексикографічних середовищ	358
Д.4. Стандартизовані формати репрезентації даних.....	358
Д.5. Характеристика мережевих протоколів та API	359
Д.6. Внесок окремих дисциплін у комп'ютерну лексикографію	360
Д.7. Внесок лінгвістичних дисциплін у створенні словників	361

Д.8. Реалізація принципу розмічення і структурного принципу.....	362
Д.9. Зразок аналізу електронного словника.....	363
Д.10. Об'єднана система принципів створення електронного словника	363
Д.11. Порівняльна характеристика форматів PDF і HTML.....	364
Д.12. Порівняльна характеристика стандартів TEI та LMF.....	365
Д.13. Порівняльна характеристика форматів зберігання лексикографічної інформації	366
Д.14. Рівні лексикографічного моделювання	366
Д.15. Порівняльна таблиця типів комп'ютерних словників	367
Д.16. Порівняльна характеристика типологій комп'ютерних словників	368
Д.17. Класифікація середовищ укладання електронних словників	369
Д.18. Порівняльна характеристика різних систем розроблення словників (DWS / CPC).....	369
Д.19. Характеристика загальномовних корпусів	370
Д.20. Характеристика спеціалізованих корпусів.....	371
Д.21. Характеристика паралельних корпусів	372
Д.22. Приклади MSD тегів у MULTEXT-East	372
Д.23. Приклади дефініцій, створених різними ВЛМ.....	373
Д.24. Варіанти зредагованих дефініцій.....	374
Д.25. Типові помилки великих мовних моделей (ВЛМ) у лексикографічному тлумаченні	375

ВСТУП

Навчальний посібник укладено для студентів першого (бакалаврського) рівня вищої освіти, які здобувають спеціальність 035 «Філологія» за освітньою програмою «Прикладна та комп'ютерна лінгвістика». Він покликаний сформувати у студентів ґрунтовне уявлення про комп'ютерну лексикографію як важливий і динамічний напрям сучасної лінгвістики. У центрі уваги – принципи побудови комп'ютерних словників, методи організації та подання мовних даних, особливості моделювання лексикографічної інформації, а також новітні цифрові технології та інструменти, що забезпечують розвиток електронного словникарства. Посібник має на меті не лише ознайомити студентів із теоретичними засадами галузі, а й продемонструвати практичні підходи та сучасні рішення у створенні електронних лексикографічних ресурсів.

Виходячи з поставленої мети, матеріал посібника структуровано за розділами, кожен з яких висвітлює окремий аспект комп'ютерної лексикографії. Перший розділ – **«Словники в цифровому середовищі»** – присвячено загальним питанням про природу словника та принципи його побудови, відображення в ньому організації та функціонування мови, а також ролі історичних, культурних та когнітивних чинників у формуванні лексикографічних описів. Окрему увагу приділено феномену лексикографічної мережі – сукупності взаємопов'язаних словників і мовних ресурсів, що акумулюють лінгвістичний досвід людства у спільному інформаційно-цифровому просторі.

Другий розділ – **«Комп'ютерна лексикографія: наукова і практична діяльність»** – розкриває ключові завдання сучасного електронного словникарства, базові поняття галузі, етапи створення комп'ютерних словників, принципи моделювання та збереження лексикографічних даних, а також типологію електронних словників у контексті їх функціональних можливостей та технологічних рішень.

Третій розділ – **«Технології у комп'ютерній лексикографії»** – зосереджено на сучасних технічних засобах, що забезпечують створення, обробку та інтеграцію лексикографічних ресурсів. У ньому розглянуто інструменти

корпусного аналізу, віртуальні лексикографічні лабораторії, спеціалізовані сервіси на кшталт *ELEXIS* та *Sketch Engine*, а також системи штучного інтелекту, що дедалі активніше застосовуються для автоматичної екстракції даних, генерації словникових статей та оптимізації лексикографічних процесів. Розділ демонструє, як цифрові технології створюють нові можливості для лексикографа й трансформують традиційні методи роботи зі словниковою інформацією.

Структурно матеріал посібника поділено на три розділи, кожен із яких охоплює три тематичні розділи. Кожна тема, у свою чергу, складається з трьох підрозділів, присвячених окремим аспектам відповідної проблематики. Підрозділи містять теоретичний виклад, запитання для самоконтролю та завдання для самоперевірки, що дає змогу поєднати опрацювання теоретичного матеріалу з практичним застосуванням набутих знань. Кожна тема завершується переліком питань, що не увійшли до розгляду в посібнику, але є важливими для кращого її розуміння. Такі питання вивчаються студентами в межах самостійної роботи. Наприкінці посібника подано глосарій термінів, бібліографію та добірку корисних ресурсів для подальшого вивчення. Окремий розділ становлять додатки, у яких зібрано супровідні матеріали, зразки та приклади, що розширюють основний зміст.

У результаті опанування навчального матеріалу студенти здобудуть базові та поглиблені знання про принципи організації електронних словників, сучасні формати лексикографічних даних і технології їх опрацювання. Вони навчатимуться застосовувати інструменти корпусної лінгвістики, аналізувати мовні дані, працювати з віртуальними лексикографічними середовищами та використовувати методи штучного інтелекту для розв'язання практичних завдань словникарства.

РОЗДІЛ 1. СЛОВНИКИ В ЦИФРОВОМУ СЕРЕДОВИЩІ

1. СЛОВНИК: ЙОГО СТРУКТУРА, ЗМІСТ І МЕТАМОВА

План заняття:

- 1.1. Про поняття *словник* та його будову
- 1.2. Лексикографічна параметризація мовних одиниць
- 1.3. Метамова словника

Перш ніж переходити до вивчення проблематики сучасного електронного словникарства, необхідно чітко розуміти, що являє собою словник, з яких основних частин він складається та які функції покликаний виконувати. Словник – це не просто перелік слів чи термінів, а складна лексикографічна система, що має свою внутрішню структуру та організацію. Він виконує роль посередника між мовою та її користувачем, допомагаючи інтерпретувати значення, виявляти зв'язки між словами, з'ясовувати правила їх уживання. Одним із головних питань є також форма подання матеріалу. У традиційних друкованих словниках обсяг і спосіб подачі інформації значною мірою обмежувалися матеріальними рамками видання. Натомість електронні словники відкрили можливість для більшої гнучкості: інтерактивності, мультимедійного супроводу, гіпертекстових посилань та швидкого пошуку. Отже, розглядаючи словник як інструмент, ми повинні бачити його багатофункціональність: це водночас джерело знань, навчальний посібник, комунікативний ресурс і відображення мовної культури суспільства. Лише з розуміння цього фундаменту можна переходити до осмислення того, як сучасні інформаційні технології змінюють традиційні підходи у словникарстві.

1.1. Про поняття словник та його будову

Що таке словник? Здавалося б, відповідь проста: це книга зі словами. Але якщо зануритися глибше, то все не так однозначно. У різні часи мовознавці пропонували різні визначення: словник як *книга*, *довідковий ресурс* (паперовий чи електронний), *упорядковане зібрання* чи навіть просто *список*. Наведемо додаткові приклади визначень:

Словник – упорядкований в алфавітному чи тематичному порядку перелік заголовних слів, лексикографічно опрацьованих.

Словник є одним із засобів накопичення результатів пізнавальної діяльності людства, показником культури народу.

Словник – це не лише збірник алфавітно упорядкованих статей, а й також інструмент, що надає користувачеві всі необхідні ресурси для пошуку слова або фрази, які йому потрібні для здійснення комунікативного акту.

Чому ж так важко дати одне універсальне визначення? Тому що самі словники з часом змінювалися за формою, змістом, будовою та функціями. Розглянемо кілька основних чинників цієї різноманітності.

Перший пов'язаний з розвитком словникарської науки. У різні епохи уявлення про словник змінювалося: від простих списків слів до багаторівневих електронних баз. Важливо, що нині користувачами словників є не лише люди, а й комп'ютерні системи для роботи з текстами: програми автоматичного перекладу, засоби перевірки граматики, системи побудови онтологій, а також технології штучного інтелекту, наприклад, чат-боти. Це впливає і на побудову словників, і на спосіб подання інформації. Якщо раніше словники були лише писемними (паперовими), то зараз з'явилися нові формати: ілюстративні, аудіо- та відеословники, а також інтерактивні мультимедійні проєкти. У такий спосіб словник перетворюється не лише на текстове довідкове джерело, а й на повноцінний мультимедійний ресурс.

Другий чинник – це розширення об'єктів опису. Якщо раніше словники здебільшого містили лише окремі слова, то з часом спектр одиниць значно розширився: фразеологізми та сталі вирази (*фразеологічні словники*), морфеми (*словники морфем*), вимова та правила наголосу (*орфоепічні словники*), правопис (*орфографічні словники*). Крім того, з розвитком комп'ютерної лінгвістики (NLP – Natural Language Processing) з'явилися частотні словники, які показують, як часто те чи інше слово вживається у великих масивах текстів. Вони дають інформацію про частоту вживання як у писемній, так і в усній мові. Наприклад,

у корпусах сучасної української мови серед найуживаніших слів – *і, в, не, на, що*, тоді як рідкісні слова можуть зустрічатися лише кілька разів у всій базі даних.

Ще одна нова сфера – словники інших знакових систем. Зокрема, словники сурдоперекладу. Це візуальні посібники, які містять набір жестів для різних слів і понять. Вони використовують мову жестів (жести руками, міміка, рухи губ, положення тіла) для передачі інформації та є важливим інструментом комунікації між чуючими та людьми з порушеннями слуху. До цього переліку варто також додати електронні термінологічні глосарії, візуальні словники з ілюстраціями, емоційні та культурні словники (наприклад, словники символів, кольорів чи навіть мемів). Це підтверджує, що об'єктом лексикографічного опису сьогодні може бути не лише слово, а й будь-який елемент знакової системи, який відіграє роль у комунікації.

Третій пов'язаний з розширенням програмного інструментарію для пошуку та впорядкування лінгвістичного матеріалу. Якщо у друкованих словниках користувач шукав слово лише за абетковим реєстром, то сучасні електронні системи дають змогу здійснювати багатопараметровий пошук. Наприклад, у «Оксфордському словнику англійської мови» (<https://www.oed.com/>) чи у віртуальній лексикографічній лабораторії «Словник української мови» можна знайти слово за такими параметрами, як: «частина мови», «галузь використання», «наявність певного слова в тлумаченні» тощо. Завдяки цьому користувач може створити власну вибірку словникових статей, яка фактично постає як похідний словник. Наприклад, із бази даних «Словника української мови» можна сформувати список статей, у яких заголовкові слова мають іншомовне походження. Таким чином утворюється «Словник іншомовних слів української мови» – новий продукт, отриманий автоматизованим способом.

Четвертий чинник пов'язаний з інтеграцією словників різних типів у єдиний мовно-інформаційний об'єкт. Це означає, що користувачеві більше не потрібно звертатися до кількох окремих довідників – він отримує доступ до комплексної системи, яка поєднує одразу кілька словникових функцій. Так, інтегрована лексикографічна система «Словники України», яка розроблена в Українському

мовно-інформаційному фонді НАН України (<https://lcorp.ulif.org.ua/dictua/>) містить чотири базові словники, що описують: словозміну, синонімію, антонімію та фразеологію. Інший приклад – онлайн-ресурс «Горох» (<https://goroh.pp.ua/>), який поєднує ще ширший набір лексикографічних інструментів: **тлумачний, словозмінний, перекладний, орфоепічний, синонімічний, частотний, етимологічний, фразеологічний**, а також словники анаграм і слововживання. Отже, сучасний словник дедалі більше наближається до моделі універсального мовного порталу, що дає змогу комплексно працювати з лексикою й задовольняти різні потреби користувача – від перевірки правопису до пошуку походження чи статистики слова.

У світі того, що сучасне словникарство дедалі більше переходить в інтернет, словники перестають існувати як ізольовані ресурси: вони інтегруються між собою, утворюючи своєрідну «лексикографічну мережу». Показовим прикладом є інструмент SkELL корпорації *Sketch Engine* (<https://www.sketchengine.eu/skell/>), що надає студентам і викладачам можливість аналізувати реальне вживання конкретних слів і виразів у мовленні носіїв. SkELL підтримує англійську, німецьку, італійську, чеську та інші мови, пропонуючи три основні режими роботи: 1) конкорданс – добір автентичних прикладів уживання слова; 2) колокації – перелік сталих словосполучень, що формує слово; 3) синонімія. Важливо, що всі ці дані отримуються автоматично на основі багатомільйонних корпусів текстів, без ручного опрацювання, що підкреслює нову якість лексикографічних ресурсів.

Виходячи з вищесказаного, можна запропонувати таке визначення поняття **словник**: *упорядкований і систематизований мовно-інформаційний ресурс, здатний інтегруватися та взаємодіяти з іншими ресурсами з метою задоволення різноманітних потреб користувачів – від наукових до практичних. Він забезпечує доступ до лексичних одиниць і пов'язаної з ними інформації у зручній для пошуку та використання формі. Якщо говорити простіше, словник – це не просто перелік слів, а своєрідна база даних мови, де інформація впорядкована*

так, щоб користувачеві було легко знайти потрібне. У таблиці 1.1. подано типи лінгвістичної інформації, її характеристика і приклади словників, які її містять.

Таблиця 1.1 – Характеристика лінгвістичної інформації

Тип інформації	Що містить словник	Приклад словника
Лексична	Значення слова, тлумачення	СУМ у 20 т., OED, Duden
Граматична	Відмінювання, відмінки, рід, число, час тощо	СУМ у 20 т., OED, Duden, словозмінні словники (BESUM)
Орфографічна / орфоепічна	Правильне написання та вимова	Орфоепічні, орфографічні словники
Фразеологічна	Стійкі вирази, ідіоми	Тлумачні, фразеологічні словники
Частотна	Частота вживання у текстах	Частотні словники
Етимологічна	Походження слова	Етимологічний словник української мови в 7 т.
Перекладна	Відповідники іншими мовами	Українсько-іншомовні словники загальної і спеціальної лексики
Візуальна / аудіо	Ілюстрації, жести, вимова	Енциклопедичні словники, словник сурдоперекладу
Інтегрована	Поєднання різних типів інформації в одному ресурсі	«Словники України», «Горох», SkELL

Тобто словник – це універсальний інструмент для роботи зі словами, який потрібен і науковцю, і перекладачу, і студенту, і навіть комп’ютерним програмам. Тут варто уточнити поняття «упорядкований» та «систематизований». Перше означає, що мовні елементи розташовані у певному порядку: алфавітному, тематичному, частотному чи інверсійному. А друге підкреслює наявність внутрішньої структури словника: кожна словникова стаття має певні поля та взаємозв’язки, що полегшують пошук і аналіз інформації. Нижче (таблиця 1.2) подано характеристики різних типів реєстрів, які застосовуються у сучасних словниках.

Таблиця 1.2 – Характеристики типів реєстрів словників

Тип реєстру	Опис	Приклад використання
Абетковий	Слова розташовані за абеткою, що полегшує швидкий пошук конкретної лексеми	Більшість словників
Тематичний	Слова згруповані за темами або категоріями (професії, їжа, природа тощо)	Тезаурусні словники загальної та спеціальної лексики
Частотний	Слова упорядковані за частотою вживання у певному корпусі текстів	Частотні словники сучасної української мови
Інверсійний	Слова розташовані в абетковому порядку, але не за першою, а останньою літерою	Бази даних NLP-додатків для аналізу морфологічних та словотвірних моделей

Для виконання своїх функцій словник має бути організований відповідним чином. Його структура є складним і багаторівневим утворенням, яке складається з макроструктури та мікроструктури. **Макроструктура** – це організація словника в цілому, що визначає логіку його побудови й доступу до інформації. До неї належать словниковий реєстр, масив словникових статей та допоміжні розділи (передмова, правила користування, список умовних скорочень, покажчики тощо; див. рис. 1.1). Оскільки словник репрезентує мову як системно-структурне утворення, у його макроструктурі також можуть відображатися різні внутрішні зв'язки між мовними одиницями: синонімія, антонімія, омонімія, паронімія, а також зовнішні зв'язки – семантичні поля, лексико-семантичні групи та підгрупи. Наявність або відсутність окремих елементів макроструктури залежить від типу словника, його призначення та цільової аудиторії.

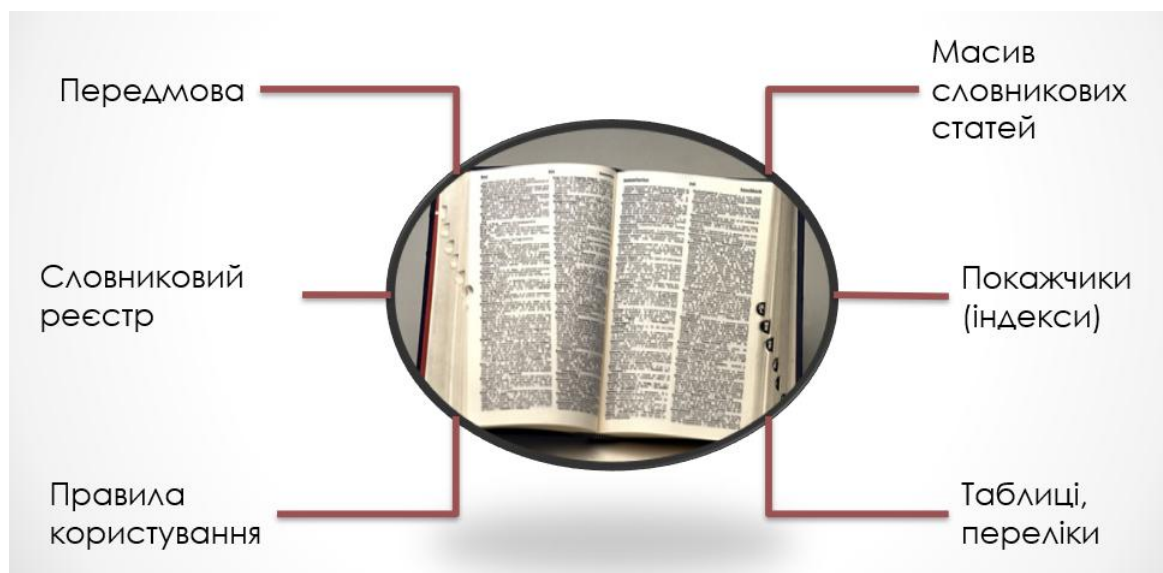


Рис. 1.1 – Елементи макроструктури словника

Мікроструктура словника – це внутрішня організація окремої *словникової статті*. Вона визначає, які саме елементи інформації про слово подаються та в якій послідовності. До складу мікроструктури можуть входити: заголовне слово, його граматичні характеристики (рід, число, відмінювання, дієвідміна тощо), наголос, фонетична транскрипція, стилістичні позначки, тлумачення значення, приклади вживання, переклади іншими мовами, синоніми, антоніми, фразеологізми, похідні форми та інші відомості. Конкретний набір і порядок цих компонентів залежить від типу словника та його функцій: у тлумачному словнику

домінує пояснення значення, у перекладному – відповідники іншою мовою, у частотному – кількісні дані щодо вживання слова. Отже, мікроструктура забезпечує повноту та зручність отримання інформації в межах окремої словникової статті. Відмінності між макро- та мікроструктурою наочно наведені у таблиці 1.3.

Таблиця 1.3 – Відмінності між макро- та мікроструктурою

Ознака	Макроструктура	Мікроструктура
Що описує	Словник у цілому	Окрему словникову статтю
Основні елементи	Реєстр (список заголовкових слів), система статей, допоміжні матеріали (передмова, правила користування, скорочення, покажчики)	Заголовне слово, граматична інформація, транскрипція, тлумачення, приклади, синоніми, антоніми, фразеологізми, переклади
Функція	Організує словник як цілісний ресурс	Надає повну інформацію про одне слово/одиначку
Зв'язки	Синонімія, антонімія, омонімія, паронімія, лексико-семантичні поля	Внутрішні зв'язки між елементами статті: значення ↔ приклади ↔ граматики
Залежність від типу словника	Може містити різні розділи чи відсутність деяких (наприклад, без передмови чи покажчика)	Вміст і порядок компонентів відрізняється у тлумачному, перекладному, частотному, орфографічному тощо

Незалежно від типу чи призначення словника, у структурі словникової статті зазвичай виокремлюють **ліву** та **праву** частини. **Ліва частина** – це зона ідентифікації заголовкової одиниці (**леми**). Тут подається саме слово чи вираз (лема), яке словник описує: це може бути окреме слово, словосполучення, морфема, термін чи навіть фразеологізм. Далі йде його характеристика:

- граматична (частина мови, рід, число, відмінювання тощо);
- фонетична (транскрипція, наголос, іноді інтонація);
- структурна (морфологічні варіанти, можливі форми).

У деяких словниках додається також етимологічна інформація. Наприклад, у *Словнику іспанської мови Королівської академії* (RAE) ліворуч вказується походження слова. В англійських словниках, як-от *Merriam-Webster Dictionary* чи *American Heritage Dictionary*, часто наводять розбиття слова на склади, а також уточнюють лексико-граматичні характеристики (збірне, злічуване чи незлічуване). Таким чином, ліва частина словникової статті виконує функцію «візитної картки»

слова: саме з неї користувач отримує перші ключові відомості для розпізнавання та правильного використання одиниці. На рисунку 1.2 подано приклад лівої частини словникової статті.

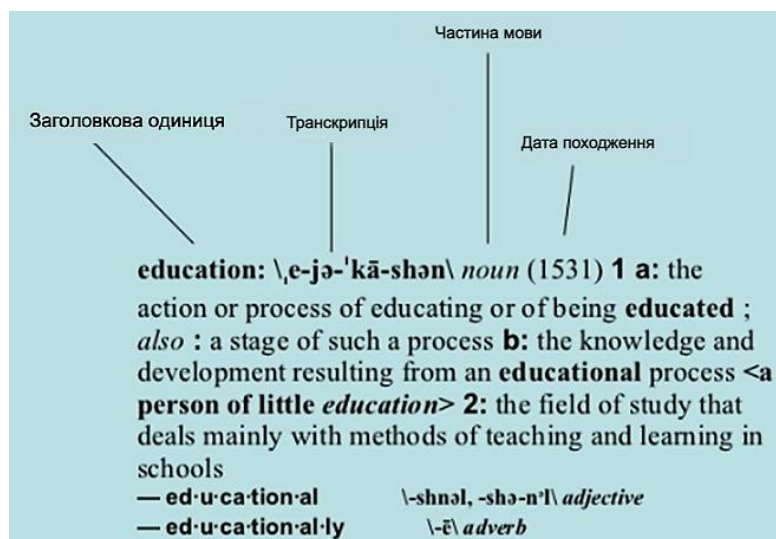


Рис. 1.2 – Елементи лівої частини словникової статті

Права частина словникової статті зазвичай містить опис семантичних, стилістичних, узуальних та інших характеристик леми. Її обсяг і структура можуть суттєво відрізнятися залежно від типу словника. У тлумачних словниках (наприклад, *Merriam-Webster Dictionary*) права частина складається з **дефініцій** (тлумачень значень) і **ілюстративних прикладів**, які демонструють уживання слова в мовленні, як це показано на рис. 1.3.

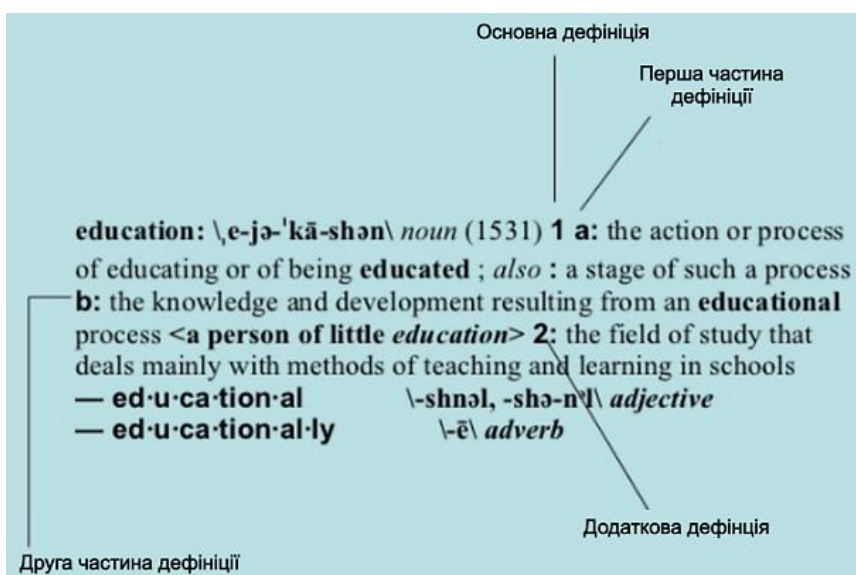


Рис. 1.3 – Елементи правої частини словникової статті

Зміст правої частини залежить від того, який рівень **мовної системи** описується:

– *фонетичний* – **орфоепічні** словники подають правильну вимову, а також позиційні й комбінаторні зміни звуків (асиміляція, акомодация, лабіалізація тощо);

– *морфологічний* – **словозмінні** словники наводять усі граматичні форми слова; **словотвірні** – моделі творення нових слів за допомогою префіксів, суфіксів тощо;

– *лексичний* – у **тлумачних** словниках розкривається значення слова (**сигніфікативне, конотативне, денотативне, узуальне**). Сюди ж належать словники синонімів, антонімів, омонімів, паронімів, **ідеографічні** словники;

– *синтаксичний* – словники сполучуваності, **колокацій** і стійких словосполучень показують, як слово функціонує у складі більших мовних одиниць.

Крім рівня опису, важливим є й часовий аспект, у якому здійснюється опис мови:

– в **етимологічних словниках** права частина розкриває походження слова, історію його форми у споріднених і неспоріднених мовах;

– в **історичних словниках** описуються слова певної епохи, засвідчені у писемних пам'ятках. Тут зазначаються значення, наводяться ілюстрації з текстів, а також дата (рік) першої появи слова.

Отже, права частина словникової статті виступає головним носієм спеціалізованої інформації, яка конкретизує та поглиблює відомості про лему. Вона може містити як мінімальний обсяг даних (наприклад, коротку дефініцію та приклад уживання), так і значно ширший опис, що включає кілька рівнів аналізу – від фонетичного до синтаксичного. Ключовим завданням цієї частини є забезпечення адресата максимально точними й корисними відомостями про слово залежно від типу словника та його комунікативної функції.

З погляду структурної організації права частина має блокову побудову: визначення, граматичні характеристики, семантичні відтінки, ілюстративний матеріал, стилістичні позначки, коментарі тощо. У двомовних словниках вона може містити перекладні відповідники, сферу вживання чи прагматичні

обмеження. У навчальних словниках, розрахованих на іноземців, праву частину часто доповнюють фонетичною транскрипцією, прикладами типових словосполучень та інформацією про контекстуальні варіанти.

Питання для самоконтролю

1. Що таке словник у вузькому та широкому розумінні?
2. Які основні функції виконують словники в мовній практиці?
3. Яка головна перевага інтегрованих словників у порівнянні з окремими довідниками?
4. Яку інформацію містить ліва частина словникової статті?
5. Яке значення має права частина словникової статті та які її структурні компоненти можна виділити?
6. Від чого залежить обсяг і зміст правої частини словникової статті?
7. Які рівні мовного опису можуть відображатися у словниках?
8. Які особливості подання матеріалу мають:
 - орфоепічні словники; – словники синонімів, антонімів, омонімів, паронімів;
 - словозмінні словники; – словники сполучуваності та стійких словосполучень;
 - словотвірні словники; – етимологічні та історичні словники.
 - тлумачні словники;
9. Чим відрізняється організація правої частини статті в одномовних, двомовних та навчальних словниках?
10. Які функції виконує права частина словникової статті, окрім пояснювальної?

Завдання для самоперевірки

1. Подумайте, які одиниці опису лежать в основі кожного з наведених типів словників:
- Тлумачний словник української – Частотний словник сучасної мови;
української мови;

- Фразеологічний словник;
- Орфоепічний словник;
- Словник сурдоперекладу;
- Візуальний словник для вивчення іноземної мови.

2. Визначте, які з цих словників найзручніше було б реалізувати у друкованій формі, а які – у електронній. Поясніть свою думку.

3. Уявіть, що ви створюєте «Словник сучасного студентського сленгу». Які одиниці опису ви включили б до нього: лише слова чи ще й сталі вирази, меми, емодзі? Чому?

4. Уявіть, що вам доручили створити новий інтегрований словниковий портал. Які словники ви об'єднали б у ньому, щоб зробити його максимально корисним для школярів і студентів?

5. Поясніть, що означає поняття похідний словник. Наведіть приклад, які ще вибірки можна було б створити (крім іншомовних слів).

1.2. Лексикографічна параметризація мовних одиниць

Для лексикографа ключовим завданням є визначення того, яка саме інформація має бути подана у словнику, аби задовольнити очікування й практичні потреби користувача. Тобто, необхідно визначити набір лексикографічних параметрів для відтворення цілісної картини про заголовкове слово. **Лексикографічний параметр** – це «квант» лінгвістичної інформації (частина даних про слово), що може мати самостійну цінність для користувача, але зазвичай функціонує в поєднанні з іншими такими «квантами» для формування повного уявлення про слово у словнику. Термін уперше запропонував Ю. М. Караулов: він застосував його для позначення окремих характеристик слова, що відображаються у словникових статтях. Параметри грають важливу роль у словниках, оскільки:

- вони допомагають відрізнити одне слово від іншого за граматичними, стилістичними, семантичними та іншими характеристиками;
- кожен параметр дає додаткову інформацію, необхідну для правильного розуміння та використання слова;

– у комплексі параметри забезпечують користувачеві цілісне й багатовимірне уявлення про слово.

Кількість і склад лексикографічних параметрів залежать від **композиції словника** та **методичної установки укладача** (рис. 1.4). Перша визначає склад його основних частин, обсяг поданої інформації, способи та засоби характеристики реєстрових слів, а друга – задає призначення словника: характер використання, принципи відбору та розташування лексики. Важливим елементом авторської установки є також орієнтація на певну цільову аудиторію користувачів.



Рис. 1.4 – Чинники, що впливають на вибір лексикографічних параметрів

Лексикографічні параметри можуть відображатися як у **лівій**, так і в **правій частині** словникової статті. Вони передають різні типи інформації, тому традиційно розрізняють такі групи: **формальні**, **етимологічні**, **атрибутивні**, **асоціативні**, **прагматичні**, **інтерпретаційні** та **ілюстративні**.

Формальні параметри репрезентують план вираження заголовкової одиниці, тобто її зовнішню оболонку в мові. Вони охоплюють різні характеристики: написання, вимову, наголос, граматичні відомості, словозмінні особливості, фонетичні, морфологічні та синтаксичні варіанти (див. таблицю 1.4).

Таблиця 1.4 – Відображення формальних параметрів у словнику

Тип формального параметра	Приклад	Коментар
Написання (орфографічний вигляд)	abad, desa	Подання заголовкових одиниць у базовій графічній формі.
Вимова (фонетична транскрипція)	acknowledge [ə'kɒlɪdʒ]	Використання міжнародної фонетичної абетки (IPA) для відтворення звучання.

Наголос граматичні відомості	+	СУМ'ЯТНИЙ , а, е, <i>розм.</i>	Указується місце наголосу, рід, закінчення, стилістична позначка.
Словозмінні особливості		jugar [...] Conjug. modelo. ♦ U. solo en pasado en acer. 13.	Інформація про дієвідміну або вживання лише в певних формах.
Фонетичні варіанти		water /'wɔ:tə(r)/ /'wɔ:tər/	Подання альтернативних варіантів вимови.
Морфологічні варіанти		ВХО́ДИТИ і ВВІХО́ДИТИ (УХО́ДИТИ, УВІХО́ДИТИ)	Відображення паралельних форм, що функціонують у мовленні.
Синтаксичні варіанти		ва́жити голово́ю (життя́м) чиєю (чийм)	Подання конструкцій, що демонструють особливості сполучуваності та варіативність у словосполученні.

Етимологічні параметри подають інформацію про походження слова, його джерельну мову, етимон (початкову форму) та первісне значення. Обсяг такої довідки залежить від типу словника, традицій лексикографічної школи та призначення видання (див. таблицю 1.5). У практиці виділяють:

(а) стислі етимологічні довідки, де вказано лише мову-першоджерело та етимон.

(б) детальні етимологічні довідки, де простежується розвиток слова в історії мови, його відповідники в інших мовах, первісне значення.

Таблиця 1.5 – Відображення етимологічних параметрів у словнику

Тип довідки	Приклад	Коментар
Стисла довідка	cómico , ca. Del lat. <i>comīcus</i> , y este del gr. <i>κωμικός</i> <i>kōmikos</i> .	Вказано походження через латину від грецької форми.
Стисла довідка	ДІЄТА 1 (від грец. <i>δίαιτα</i> – спосіб життя, режим харчування).	Український тлумачний словник: наведено першоджерело та його базове значення.
Детальна довідка	actino- A prefix that means 'radiating', derived from Greek <i>aktis</i> , <i>inos</i> , meaning 'ray'.	Англійський словник: пояснюється походження та значення етимона.
Детальна довідка	water. Origin: Old English <i>wæter</i> ... related to Dutch <i>water</i> , German <i>Wasser</i> , Russian <i>voda</i> ...	Історична реконструкція: вказано прагерманські, індоєвропейські зв'язки та паралелі в кількох мовах.

Атрибутивні параметри (таблиця 1.6) уточнюють належність заголовкового слова до певного шару лексики. Вони оформлюються у вигляді **ремарок** (скорочених позначок) і можуть указувати на:

– функціональний стиль (розмовна, книжна, поетична, офіційна, архаїчна лексика тощо);

- предметну галузь (терміни фізики, хімії, дипломатії, медицини, фінансів тощо);
- ареал поширення (регіональні варіанти, діалектизми, американізми, британізми тощо);
- соціальну сферу використання (жаргон, сленг, просторіччя).

Таблиця 1.6 – Відображення атрибутивних параметрів у словнику

Лема	Атрибутивна ремарка	Пояснення
ПЛЕНТАТИСЯ	<i>розм.</i>	Розмовна лексика, уживана у неформальному мовленні.
АДСОРБЕ́НТ	<i>фіз., хім.</i>	Термін, що належить до двох галузей науки: фізики та хімії.
ДЕПОЗИТА́РІЙ	<i>дипл., фін., юр.</i>	Багатогалузовий термін: дипломатія, фінанси, юриспруденція.
chamba	<i>coloq. Am. Cen., Ec. y Méx.</i>	Розмовна лексика; обмежена ареалом (Центр. Америка, Еквадор, Мексика).
thou²	<i>informal</i>	Розмовне, неофіційне вживання в сучасній англійській мові.

Атрибутивні параметри відображають належність лема до певного стилю, галузі, територіального чи індивідуального різновиду. Вони можуть фіксуватися як у ремарках, так і в лівій/правій частині словникової статті. Розрізняють кілька основних типів атрибуції: **функціонально-стильова**, **тематична**, **ареальна**, **ідіолектна** та **метаідіолектна** (див. таблицю 1.7).

Таблиця 1.7 – Відображення різних типів атрибуцій в словнику

Тип атрибуції	Позначення в словнику	Приклад
Функціонально-стильова	розм., офіц., роет.	ПЛЕНТАТИСЯ , недок., розм.
Тематична	фіз., хім., Geom.	АДСОРБЕ́НТ , хол., фіз., хім.
Ареальна	Am. Cen., Esp., Haw.	templar En Esp. ... En algunos lugares de Am.
Ідіолектна	іменне посилання	ЄВГЕНІКА ... термін Ф. Гальтона
Метадіалектна	назва школи/напряму	Accidental species ... Braun-Blanguet school

Асоціативні параметри показують зв'язки лема з іншими мовними одиницями, розкривають семантичне оточення слова. Основні типи параметрів наведено у таблиці 1.8. Вони включають **родові / видові номінації** (зв'язки «клас – підклас», «родове – видове»), **партитивні** або **частина ↔ ціле** (елементи, які

входять у склад іншого поняття), **синонімічні / дублетні зв'язки** (відсилання до рівнозначних лексем) й **асоціативні розгалуження значень** (від основного значення до різних контекстних реалізацій).

Таблиця 1.8 – Відображення різних типів зв'язків у словнику

Тип асоціативного зв'язку	Приклад
Родове – видові	Microfauna ↔ macrofauna, meiofauna, mesofauna
Частина – ціле	wheel → steering wheel, potter's wheel
Синонімія / дублети	ремарка ↔ примітка, помітка, нотатка, футнот
Розгалуження значень	wheel → torture device, bicycle, gambling disk

Прагматичні параметри (таблиця 1.9) показують, як саме слово функціонує в мовленні залежно від часу, частотності, стилістичної доречності та контексту вживання. До різновидів таких параметрів належать хронологічний статус (**архаїзми, історизми, неологізми**), ступінь уживаності (загальноновживане, рідковживане, застаріле) та узуальне значення (яке склалося в певній мовній спільноті: переносні, розмовні, вульгарні, іронічні відтінки). У статті можуть відображатися у вигляді ремарок або коротких коментарів.

Таблиця 1.9 – Відображення прагматичних параметрів у словнику

Тип прагматичного параметра	Приклад
Хронологічний статус	АЕРОПЛАН заст. Літак; СОСК archaic
Ступінь уживаності	КОБІЛА 5. рідко
Узуальне значення	КОБІЛА 2. перен., вульг.; HEN: WOMAN; acercar ... U. t. en sent. fig.

Інтерпретаційні параметри мають на меті пояснити значення реєстрового слова, тобто його семантичний зміст. Це центральний компонент правої частини словникової статті, який робить слово зрозумілим користувачеві. Основним засобом розкриття значення леми є тлумачення. Як це показано у таблиці 1.10, Складниками інтерпретаційних параметрів є **сигніфікативне значення** (узагальнене значення, що закріплено в мові), **денотативне значення** (вказівка на реальний об'єкт, явище, дію чи ознаку), **конотативне значення** (емоційні, оцінні чи стилістичні відтінки), **полісемія** та розмежування значень, тобто інтерпретаційні параметри структурують багатозначні слова, нумеруючи або розділяючи значення.

Таблиця 1.10 – Відображення інтерпретаційних параметрів у словнику

Тип параметра	Приклад
Сигніфікативне	ДИПЛОМАТ – офіційний представник держави...
Денотативне	WATER – a clear liquid...
Конотативне	РОБІТНИЧОК – зменш.-пестл.; HEN – WOMAN
Полісемія	BANK 1. Financial institution; BANK 2. River side

Ілюстративні параметри, як видно з таблиці 1.11, за своїм призначенням показують, як саме слово функціонує у реальному мовленні. Вони підтверджують значення прикладами, демонструють типові контексти та моделі вживання. За формою реалізації параметрів розрізняють цитати з художньої, публіцистичної чи наукової літератури (показують уживання слова в різних стилях), авторські приклади / штучно створені зразки (для ілюстрації граматичних форм, сполучуваності, переносних значень), приклади з усної мови чи фольклору (підкреслюють живе, розмовне використання) та приклади-словосполучення / колокації (дають відомості про синтаксичне оточення слова).

Таблиця 1.11 – Відображення ілюстративних параметрів у словнику

Тип прикладу	Ілюстрація
Літературна цитата	На столі лежав ясний да високий хліб. (П. Куліш)
Авторський приклад	He bumped his elbow on/against the table.
Усна мова / фольклор	Картопля – половина хліба. (прислів'я)
Словосполучення / колокації	steering wheel, potter's wheel

Отже, лексикографічні параметри, поєднуючись між собою, формують цілісну та багатовимірну картину «світу» кожного слова, представленого у словнику. Вони забезпечують комплексне уявлення про його форму, зміст, походження, вживання та контексти функціонування. Особливо показово, що в електронних словниках параметри можуть реалізовуватися у різних форматах – текстовому, графічному, звуковому чи мультимедійному, що значно розширює можливості користувача та робить словникову інформацію максимально доступною й наочною. Зведена характеристика для наведених типів лексикографічних параметрів наведена у додатку Д.1.

Питання для самоконтролю

1. Що таке лексикографічний параметр і яке його основне призначення?
2. Які два чинники визначають кількість і склад лексикографічних параметрів у словнику?

3. До яких частин словникової статті (лівої чи правої) можуть належати лексикографічні параметри?
4. Яку інформацію охоплюють **формальні параметри**? Наведіть приклади.
5. Чим відрізняється стислий та розгорнутий спосіб подання **етимологічних параметрів**?
6. Які різновиди атрибутивних параметрів ви знаєте (стильова, тематична, ареальна, ідіолектна, метадіалектна)? Проілюструйте прикладами.
7. Що входить до складу **асоціативних параметрів**? Як вони допомагають зрозуміти слово?
8. Яку інформацію відображають **прагматичні параметри**? Які ремарки можуть сигналізувати про їхню наявність?
9. Чим відрізняються **інтерпретаційні** параметри від **прагматичних**?
10. Яке значення мають **ілюстративні параметри** у словниковій статті?
11. Як поєднання різних лексикографічних параметрів формує цілісне уявлення про слово?
12. У яких формах (текстовій, графічній, звуковій, мультимедійній) можуть подаватися параметри в електронних словниках?

Завдання для самоперевірки

1. Проаналізуйте, які лексикографічні параметри містять наведені нижче словники для опису слова *вода, чорний, стояти*:

– Словник української мови (СУМ) – Словник синонімів української мови (ССУМ)	– Словник антонімів української мови (САУМ) – Фразеологічний словник української мови (ФСУМ)

Визначте, на яких саме характеристиках слова акцентує увагу кожен із наведених словників:

Словник	Характеристики слова
СУМ	
ССУМ	
САУМ	
ФСУМ	

Уточніть, яку саме інформацію подано в цих словниках та заповніть таблицю:

Слово / Словник	СУМ (тлумачний)	ССУМ (синонімів)	САМУМ (антонімів)	ФСУМ (фразеологізмів)
вода				
чорний				
стояти				

Для виконання цього завдання рекомендується використовувати такі ресурси:

- Словники України онлайн: <https://lcorp.ulif.org.ua/dictua/>
- Словник «Горох»: <https://goroh.pp.ua/>

2. Прочитайте наведені словникові статті та визначте, які лексикографічні параметри в них відображено:

КОШЕНЯ́, *-яти, сер.* Зменш.-пестл. до **кіт**.

thou *archaic* the second person singular subject of a verb.

ШАРІ́К, *-а, чол. розм.* Кулька.

3. Виберіть слово (*світло, рух, зелений* або інше на ваш розсуд) і порівняйте його опис у двох словниках різного типу (наприклад, тлумачному та фразеологічному). Визначте, які параметри в них збігаються, а які відрізняються.

4. Розподіліть наведені приклади за типами параметрів (формальні, семантичні, атрибутивні, асоціативні, прагматичні):

БА́БА 2. *перен., зневажл.* Про боягузливого чоловіка.

Microflora Compare **macroflora**.

КИМЧІ *кор.* Традиційна корейська страва з овочів.

gramercy *archaic*: thank you.

ЄВГЕНІКА Термін запропонував англійський вчений Френсіс Гальтон.

5. Укладіть міні-словникову статтю на 3–4 рядки для слова *смартфон*, додавши якомога більше параметрів (формальні, семантичні, атрибутивні, прагматичні).

1.3. Метамова словника

Лексикографічні параметри мають відповідні формати й засоби подання у статті, що забезпечують фіксацію та інтерпретацію лінгвістичних характеристик мовної одиниці. Сукупність таких форматів і засобів називається **метамовою словника**. У широкому сенсі термін **метамова** означає *знакову*

систему, призначену для опису іншої мови. У вузькому, лексикографічному сенсі він вказує на набір типових прийомів і формул, за допомогою яких укладачі словника описують значення слова й супровідні відомості. Компонентами метамови словника є: формули тлумачення, ремарки, скорочення, а також абеткові та неабеткові символи.

Формули тлумачення являють собою *типові синтаксичні конструкції* («те, що...», «той, хто...», «предмет, призначений для...»), *що використовують для опису лексичного значення або функціональних характеристик заголовкового слова*. Наприклад, для опису семантики таких слів, як *стіл*, *стілець* у СУМ 20 т. використовують такі формули тлумачення:

СТІЛ 1. Вид меблів у вигляді горизонтально укріпленої на ніжках широкої дошки.

СТІЛЕЦЬ 1. Вид меблів у вигляді короткої лави, перев. із спинкою, для сидіння однієї людини.

На вибір тієї чи іншої формули тлумачення впливають кілька чинників, зокрема, граматичний клас мовної одиниці, тип лексичного значення леми (сигніфікативне, конотативне, денотативне), а також тип одиниці за значенням (лексичні чи службові слова). Нижче подано приклади тлумачень для іменника, прикметника та дієслова:

СТІЛ 1. Вид меблів у вигляді горизонтально укріпленої на ніжках широкої дошки (іноді з ящиками, тумбочками), на яких розміщують різні предмети.

TABLE 1a: a piece of furniture consisting of a smooth flat slab fixed on legs.

ГЛУХИЙ 1. Який нічого не чує або недочуває.

DEAF 1: having total or partial hearing loss.

ЧЕРВОНІТИ 1. Ставати, робитися червоним.

REDDEN: to become red.

На відміну від лексичних одиниць, службові слова не позначають предметів чи ознак, а виражають відношення між словами в реченні або модальні відтінки. У словниках їх часто описують через спеціальні формули: «Указує на...», «Уживається в сполуч. з...», «Уживається при...», «Уживається для» тощо:

ДО *прийм.*, з *род. в.* 1. Указує на позначення місця, предмета або особи, в бік яких спрямована дія.

БИ *після голосного Б, част.* 1. Уживається для творення дієслівних форм умовного способу.

AND conjunction 1. used as a function word to indicate connection or addition especially of items.

Нижче у таблиці 1.12 наведено порівняльну характеристику формул тлумачення для різних частин мови.

Таблиця 1.12 – Формули тлумачення для різних частин мови

Частина мови	Типова формула тлумачення	Приклад (укр.)	Приклад (англ.)
Іменник	«Вид + ознака / призначення»	СТІЛ – вид меблів у вигляді дошки на ніжках.	TABLE – a piece of furniture consisting of a flat top on legs.
Прикметник	«Який (що) ...» (ознака предмета)	ГЛУХИЙ – який нічого не чує або недочуває.	DEAF – having total or partial hearing loss.
Дієслово	«Дія/стан: робити, ставати...»	ЧЕРВОНІТИ – ставати, робитися червоним.	REDDEN – to become red.
Службове слово	«Уживається для...»	БИ/Б – Уживається для творення дієслівних форм умовного способу.	AND – a word used to join words, phrases, or clauses.

Більшість формул тлумачення у словниках побудовані за класичною логіко-філософською моделлю: спершу вказується родове поняття (*вид меблів, процес, стан, службове слово*), а далі уточнюється видова відмінність, яка відокремлює дане слово від інших (*у вигляді дошки на ніжках; що вживається для вираження умовності*):

СТІЛ – вид меблів (pid), у вигляді дошки на ніжках (видова відмінність).

TABLE – a piece of furniture (pid) consisting of a smooth flat slab fixed on legs (видова відмінність).

Цей принцип забезпечує чіткість, системність і порівнюваність тлумачень у словниках. Проте крім родо-видової дефініції у словникарській практиці існують інші: **визначення через синонім, пояснювальна, контекстуальна, перелічувальна та відсилкова дефініція**. Вони є менш розповсюдженими і використовуються в обмежених випадках.

Визначення через синонім – прийом тлумачення, за яким лексичне значення описують за допомогою тотожного або близького за значенням слова. Здебільшого цей тип дефініції використовують здебільшого для тлумачення мовних одиниць, позначених як регіоналізми, слова певної сфери або застарілі слова:

nacela. ... f. *Arq. escocia.*

expremir. ... tr. desus. **expresar.**

water 2d: WATER SUPPLY

АЕРОПЛАН а, ч., *заст.* Літак

ВОДА́ *і, ж.* 1.4. *перен., розм.* Непотрібні, беззмістовні фрази і т. ін. у викладі матеріалу; **багатослів'я**.

Характерною рисою **пояснювальної дефініції** є відсутність першого номінативного елемента (гіпероніма). Такий тип тлумачення *має на меті відповісти на питання: «що виражає?», «для чого використовують?», «яким чином використовують?»* тощо:

ПРО *прийм., зі знах. в.* Сполучення з прийм. **про** 1. Уживається при вказуванні на конкретну особу, предмет або абстрактне поняття, що виступають об'єктом розмови, думки і т. ін.

ouch interjection used especially to express sudden pain.

clíc onomat. 1. U. para reproducir ciertos sonidos.

Таке визначення використовують переважно для службових частин мови (вигуки, займенники, прийменники сполучення).

Контекстуальне визначення не лише розкриває значення самого заголовкового слова, а й також указує на контекст його використання. Таке визначення потрібне особливо носіям інших мов, які не знають окремих нюансів щодо використання слова:

late 1b (1): of or relating to an advanced stage in point of time or development: occurring near the end of a period of time or series.

sweet 4c: free from excessive acidity — used especially of soil.

КІСЛИЙ а, е. 2.2. Який виділяють бродильні речовини; прокислий (про запах).

Крім наведених вище типів визначення, у словнику використовуються інші, зокрема **перелічувальні** (або описово-перелічувальні) дефініції, *призначені для семантизації слова шляхом називання предметів (понять) або характеристик, що йменуються цим словом:*

arte [...] m. o f. [...] О 9. pl. Lógica, física y metafísica.

МАШІНА и, ж. заст. Поїзд, паровоз.

craft 2a: an occupation, trade, or activity requiring manual dexterity or artistic skill.

Відсилкова дефініція – спосіб розкриття значення заголовкового слова шляхом указівки на іншу словникову статтю. Такий прийом застосовується для архаїзмів, регіональних та морфологічних варіантів слова, яке вже описано в іншій словниковій статті:

ozonósfera V. **ozonosfera**.

defence, defenceman chiefly British spellings of DEFENSE, DEFENSEMAN

ЛАНІТИ Див. ланіти.

Вибір того чи іншого типу дефініції зумовлюється як граматичними й семантичними особливостями заголовкових слів, так і завданнями конкретного словника (нормативного, перекладного, історичного, тлумачного тощо). Стисла характеристика наведених вище типів дефініцій подано у таблиці. 1.13.

Таблиця 1.13 – Типи дефініцій

Тип дефініції	Особливості	Приклад
Родо-видова	Тлумачення через вказівку на родові поняття (гіперонім) та видові ознаки. Найпоширеніший тип.	<i>СТИЛ</i> – предмет меблів для роботи, приймання їжі тощо.
Через синонім	Значення пояснюється близьким чи тотожним словом. Використовується для архаїзмів, регіоналізмів, варіантів.	<i>АЕРОПЛА́Н</i> – літак.
Пояснювальна	Без гіпероніма; відповідає на питання «що виражає?», «для чого вживається?». Характерна для службових частин мови.	<i>PRO</i> – уживається при вказуванні на об'єкт розмови.
Контекстуальна	Значення розкривається через приклади вживання або ситуації.	<i>late</i> – occurring near the end of a period of time.
Перелічувальна	Тлумачення через називання предметів чи ознак, які охоплює слово.	<i>craft</i> – occupation, trade, or activity requiring skill.

Відсилкова	Розкриття значення через посилання на іншу словникову статтю.	<i>ЛАНІТИ</i> – див. <i>ланіти</i> .

Отже, система словникових дефініцій не є одноманітною: у практиці лексикографії поєднуються різні підходи до семантизації слів. Родо-видова дефініція забезпечує чіткість і логічну впорядкованість, синонімічна й відсилкова – економію словникового простору, пояснювальна та контекстуальна – більшу наочність і доступність для користувача. Таким чином, різні типи дефініцій взаємодоповнюють одне одного, утворюючи гнучкий інструментарій опису значень. Кожна з наведених дефініцій має свої переваги та обмеження у використанні, як це показано у таблиці 1.14.

Таблиця 1.14 – Переваги та обмеження дефініцій

Тип дефініції	Переваги	Обмеження
Родо-видова	Чіткість, логічна структура, універсальність, можливість системного опису	Іноді звучить громіздко; не завжди придатна для службових частин мови
Синонімічна	Лаконічність; швидке розуміння через уже відоме слово	Потребує знання слова-синоніма; небажана для слів з обмеженою сферою вживання
Пояснювальна	Доступність; розкриває функцію чи спосіб використання слова	Менш системна; може бути багатослівною
Контекстуальна	Дає уявлення про ситуацію й умови вживання; особливо корисна для іноземців	Вимагає додаткового прикладу; не завжди придатна для стислого словника
Перелічувальна	Наглядність; показує різні прояви або приклади	Може втратити системність; не завжди зручно для великих лексичних груп
Відсилкова	Економія місця; уникання повторів; зручність при дублетах	Без іншої статті користувач не отримає пояснення

Словникові ремарки – це один із найважливіших інструментів метамови, що *розкриває специфіку вживання заголовкового слова*. Вони уточнюють його граматичний статус, сферу функціонування, стильову належність, ступінь уживаності тощо. Основними типами ремарок є:

(1) граматичні – указують на належність до частини мови чи відмінність у формах: **run**, *v.*; **добрий**, *прикм.*;

(2) статистичні (частотні) – позначають поширеність або рідкість вживання: *archaic*, *rare*, *common*, *заст.*, *рідко*;

(3) географічні – відображають територіальну обмеженість слова або значення: *Am. Eng., Can.*;

(4) стилістичні – уточнюють належність до функціонального стилю чи сфери спілкування: *розм., офіц., поет.*;

(5) прагматичні – характеризують емоційне забарвлення або оцінність: *вульг., жарт., ірон., taboo, offensive*;

(6) галузеві – указують на термінологічну належність до певної науки чи сфери діяльності: *фіз., хім., юр.*

У випадку з паперовими словниками перелік ремарок зазвичай подається після вступу або в розділі «Правила користування» перед масивом словникових статей. В електронних словниках вони можуть міститися у довідковому розділі або розшифровуватися безпосередньо під час роботи з текстом за допомогою контекстної підказки. Ті самі ремарки в електронних словниках можуть виконувати не лише функцію уточнення значення чи сфери вживання, а й слугувати пошуковим елементом. Завдяки цьому користувач отримує можливість робити вибірку слів, належних до певного граматичного класу або категорії, шару лексики, стилю чи предметної сфери. Такий підхід значно розширює дослідницькі можливості електронних словників, оскільки вони перетворюються не лише на засіб довідки, а й на інструмент для лінгвістичного аналізу, корпусних досліджень та створення навчальних вибірок. Як підсумок до вищесказаного, у таблиці 1.15 подано стислу характеристику для кожного типу словникових ремарок.

Таблиця 1.15 – Типи словникових ремарок

Тип ремарки	Призначення	Приклад
Граматичні	Вказують на частину мови, граматичну категорію чи форму	<i>ім., прикм., діал., pl.</i>
Статистичні (частотні)	Показують рівень уживаності слова у сучасній мові	<i>рідко, часто, застаріле</i>
Географічні	Обмежують ареал уживання	<i>діал. полісся, укр.-карп., канад.</i>
Стилістичні	Вказують на належність до певного стилю	<i>розм., офіц.-діл., поет., жарг.</i>
Прагматичні	Характеризують ставлення мовця або особливості комунікації	<i>ірон., зневажл., ласк.</i>
Галузеві	Уточнюють сферу науки, техніки чи діяльності	<i>фіз., мат., мед., муз.</i>

Окрім ремарок, важливим елементом метамови словника є **скорочення**, що виконують службову функцію в тексті словникової статті. Вони допомагають привернути увагу до тієї чи іншої інформації або ж використовуються з практичною метою – для економії місця у словнику. Цей прийом був особливо поширеним у паперових словниках, де обсяг видання обмежувався кількістю сторінок. Проте й у сучасних електронних словниках скорочення не втратили своєї актуальності: вони залишаються зручним засобом для компактного позначення повторюваних характеристик, а також для збереження усталених лексикографічних традицій. За формою виділяють кілька основних груп скорочень: **крапкові**, **дефісні** та **комбіновані**. У таблиці 1.16 наведено стислу характеристику скорочень разом із прикладами. Отже, скорочення поряд із ремарками становлять ще один інструмент словникової метамови, що підсилює інформативність статті та робить її більш компактною й зручною для користувача.

Таблиця 1.16 – Характеристика словникових скорочень

Тип скорочення	Приклади	Характеристика
Крапкові	<i>див.</i> (дивіться), <i>пор.</i> (порівняйте), <i>лат.</i> (латинська), <i>у знач.</i> (у значенні), <i>з наук.-попул. літ.</i> (з науково-популярної літератури)	Найпоширеніший тип. Використовується для позначення джерел, граматичних характеристик, стилістичних ремарок.
Дефісні	<i>літ-ра</i> (література), <i>р-н</i> (район), <i>ун-т</i> (університет)	Використовуються для скорочення довших слів, часто у довідкових словниках та енциклопедіях.
Комбіновані	<i>змениш.-пестл. до</i> , <i>півд.-зах.</i> (південно-західний)	Поєднують крапковий та дефісний способи, найчастіше вказують на граматичні, стилістичні або географічні особливості.

У словниковій практиці також застосовуються спеціальні символи, які допомагають структуровано подати інформацію та уникнути повторів. Вони виконують службову функцію: маркують основні та допоміжні значення, позначають приклади, відділяють елементи статті або замінюють повторювану частину слова. Система таких символів може відрізнятися від словника до словника і залежить від традицій конкретної лексикографічної школи, технічних можливостей видання (паперовий чи електронний формат), а також від принципів укладання словника. У сучасних електронних ресурсах символи

нерідко доповнюються кольоровим і шрифтовим виділенням, що робить їх ще більш інформативними та зручними для користувача. Найпоширеніші з них можна поділити на дві групи – **абеткові** та **неабеткові символи**. Короткий огляд найтиповіших символів наведена у таблиці 1.17.

Таблиця 1.17 – Абеткові та неабеткові символи

Тип символу	Приклад	Функція
Абеткові	а), б), в)...	позначають підпункти значення чи прикладів
	I, II, III...	нумерація значень заголовкового слова
	<i>adj.</i> , <i>subst.</i>	граматичні скорочення (частини мови)
Неабеткові	– (тире)	виділення прикладів або відокремлення тлумачення
	• (крапка/кружечок)	маркування синонімів, варіантів
	=	вказує на рівнозначність або синонімічність
	→	позначає відсилання до іншої статті
	~	замінює повторювану частину слова (наприклад, у словотворчих гніздах)

Отже, функція метамови як знакової системи словника полягає в тому, щоб забезпечити:

1. *Компактність подання інформації.* Використання скорочень, символів і формульних конструкцій дозволяє значно зменшити обсяг словникової статті без втрати її інформативності. Це особливо важливо для паперових словників, де обмеженням є обсяг друкованих сторінок.

2. *Одноманітність опису в межах словника.* Метамова створює стандартизовану систему, яка дає змогу однаково описувати різні лексичні одиниці. Завдяки цьому користувач швидко звикає до умовних позначень і з легкістю «зчитує» інформацію.

3. *Швидке орієнтування користувача.* Завдяки ремаркам, скороченням, символам і формульним моделям користувач може швидко знайти потрібні дані – граматичну характеристику, сферу вживання, ареал поширення тощо — навіть не заглиблюючись у все тлумачення.

4. *Можливість інтернаціонального використання.* Значна частина скорочень і символів у словниках має міжнародний характер (лат., zool., fig., Am. Engl. та ін.), що робить словник зрозумілим для користувачів різних мов і полегшує міжмовні зіставлення.

5. *Гнучкість у різних форматах.* У паперових словниках метамова переважно виконує функцію економії місця, тоді як в електронних — забезпечує інтерактивність (контекстні підказки, гіперпосилання, кольорове маркування, мультимедійні засоби).

Для порівняння у додатку Д.3 наведено метамовні особливості словників різного типу: тлумачного, перекладного, синонімічного та фразеологічного.

Питання для самоконтролю

1. Що таке **метамова словника** і яку функцію вона виконує?
2. Які основні елементи належать до метамови словника?
3. Що таке **формули тлумачення**? Наведіть приклади різних формул для опису іменника, прикметника та дієслова.
4. Які існують типи **дефініцій** у словниках, крім родо-видової? У яких випадках їх застосовують?
5. Для чого використовуються **словникові ремарки**? Які їхні основні різновиди можна виокремити?
6. Чим відрізняються **стилістичні** та **прагматичні ремарки**?
7. Які функції виконують **скорочення** у словниковій статті? Чому вони були більш поширеними у паперових словниках?
8. Які існують типи скорочень за формою (наприклад, крапкові, дефісні, комбіновані)?
9. У чому полягає роль символів у словниковій практиці? Наведіть приклади алфавітних і неабеткових символів.
10. Як відрізняється використання ремарок, скорочень і символів у **паперових** та **електронних** словниках?
11. Чому можна стверджувати, що метамова забезпечує інтернаціональний характер словника?
12. Сформулюйте приклади, які ілюструють компактність, одноманітність і зручність метамови на практиці.

Завдання для самоперевірки

1. Візьміть 2–3 словникові статті з різних словників (наприклад, тлумачного, перекладного та галузевого):

- Визначте тип поданої дефініції (родо-видова, через функцію, через опис).
- Оцініть, наскільки дефініція відповідає вимогам до наукової точності й компактності.

2. Проаналізуйте словникову статтю (з тлумачного або діалектного словника):

- Випишіть усі наявні ремарки.
- Визначте їх тип (граматичні, стилістичні, географічні, прагматичні, частотні тощо).
- Поясніть, як ці ремарки допомагають зрозуміти вживання слова.

3. Порівняйте одну й ту саму лексему у двох словниках різного типу (наприклад, «Словник української мови» та перекладний словник).

- Які елементи метамови подано в кожному словнику?
- Чи є різниця у використанні ремарок та скорочень?
- Як відрізняється спосіб подання дефініцій?

4. Виберіть 3–4 словникові статті й выпишіть усі скорочення та символи:

- Розподіліть їх за видами (крапкові, дефісні, комбіновані; символи відсилки, економії місця тощо).
- Спробуйте відновити «повний текст» статті без скорочень.

5. Уявіть, що ви укладаєте невеликий словник молодіжного сленгу. Складіть статтю на 2–3 слова, обов’язково використавши: дефініцію; ремарку (наприклад, стилістичну чи прагматичну); та хоча б одне скорочення чи символ.

Питання для поглибленого вивчення матеріалу

1. Лексикографія як розділ лінгвістики.
2. Теоретична і практична лексикографія.
3. Критерії класифікації словників. Функції словників.
4. Основні поняття словникарства.
5. Ключові принципи укладання словників.
6. Провідні українські та зарубіжні лексикографи в історії та сьогоденні.

2. ІНФОРМАЦІЙНО-ЛІНГВІСТИЧНИЙ АСПЕКТ ЕЛЕКТРОННОГО СЛОВНИКАРСТВА

План заняття:

1. Картина мовного світу
2. Поняття семантичного стану
3. Лексикографічна система та лексикографічні середовища

Словник – це не просто довідковий ресурс, а складна лексикографічна система, яка відображає організацію й функціонування мови як продукт не лише історичних та культурних впливів, а й насамперед когнітивних процесів. Він постає як своєрідна модель картини світу, у якій закарбовано спосіб мислення й пізнання певної спільноти. Кожний словник відображає погляд носіїв мови на оточуючу дійсність крізь призму власних мовних засобів. Через словникові дефініції, ремарки, приклади та семантичні зв'язки мова виявляє те, що для певної культури є суттєвим, виокремлює значуще і репрезентує його у впорядкованій системі.

Сучасне електронне словникарство виходить за межі традиційної книжкової форми й перетворює словник на інформаційно-лінгвістичний ресурс, здатний динамічно відображати семантичні стани мовних одиниць, їхні варіації, контекстуальні реалізації та асоціативні зв'язки. Воно інтегрує можливості інформаційних технологій із лінгвістичними моделями, створюючи нові середовища взаємодії користувача з мовним матеріалом. Таким чином, словник у сучасному розумінні постає як відкрита система, де поєднуються:

- **картина мовного світу**, яка формує концептуальне тло;
- поняття **семантичного стану**, що дозволяє врахувати багатовимірність значення;
- **лексикографічна система та середовища**, які забезпечують організацію та інтерактивність репрезентації мовних даних.

У наступних розділах ми детально ознайомимось з наведеними вище поняттями, на яких ґрунтується теоретико-методичний апарат сучасного електронного словникарства, запропонований та реалізований провідним

українським ученим у галузі комп'ютерної лінгвістики, академіком В. А. Широковим. Його підхід органічно поєднує класичні традиції лексикографії з новітніми інформаційними технологіями, що дало змогу сформувати нову модель представлення мовної інформації в електронних ресурсах. Таким чином, електронний словник постає не лише як засіб пошуку й упорядкування лексичних одиниць, а як інтелектуальна система, що моделює картину світу, відображену у мові.

2.1. Картина мовного світу

Розглянемо малюнок 2.1, що графічно репрезентує поняття картини мовного світу. Самий овал позначає навколишнє середовище, у якому знаходяться суб'єкти або учасники (сині кружечки), що отримують враження від навколишнього середовища та здійснюють між собою комунікацію, ділячись цими враженнями. Їх називають учасниками когнітивно-комунікативного процесу. Жовті стрілки демонструють інформаційну взаємодію учасників з середовищем через перцептивно-сенсорний апарат: зір, слух, чуття. Сукупність червоних стрілок позначає комунікативне середовище – інфраструктуру мовно-розумового (когнітивно-комунікативного) процесу. У цьому контексті людина як суб'єкт мовно-розумового процесу отримує зовнішні враження предметного світу, що опрацьовуються у його мовно-розумовому апараті і, як результат, набувають певної мовної форми. За допомогою комунікативної системи, яку дослідник розглядає як інфраструктуру мови, суб'єкти мовно-розумового процесу здійснюють обмін вербально вираженою інформацією.

У мовно-розумовому і комунікативному процесах виконуються, по суті, різні завдання (див. таблицю 2.1), що вимагають різних засобів для свого виконання. Якщо функцією комунікативної системи є передання мовних даних без або з мінімальними спотвореннями, то в мовно-розумовому апараті виконуються набагато різноманітніші операції: кодування і декодування, аналіз і синтез, перетворення, інтерпретація, концептуалізація, порівняння, запам'ятовування, коротко- і довгострокове зберігання інформації тощо. Ці

процеси здійснюються в різних середовищах і реалізуються у відмінних «апаратно-програмних» формах.

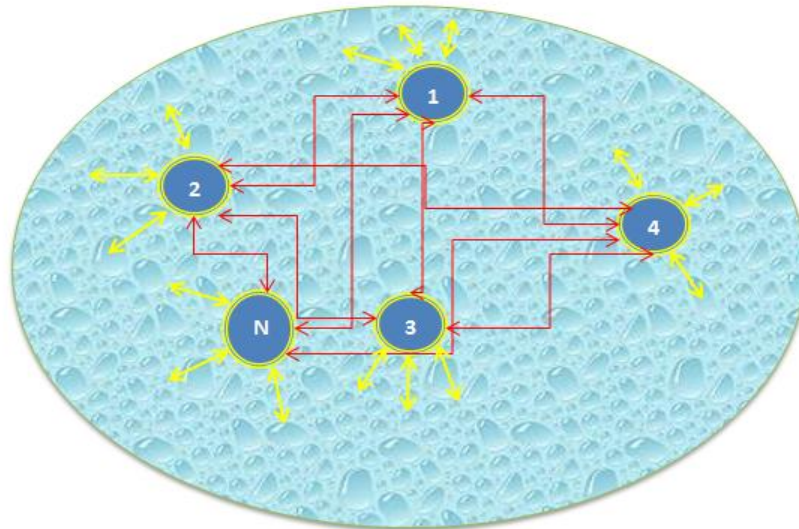


Рис. 2.1 – Картина мовного світу

У цьому контексті стає очевидним: людина засвоює світ через мову з єдиною метою – укоренитися в світі. Це прагнення є настільки сильним, що створення мовного словника – це практика, спрямована на процес облаштування. Людина укорінюється в світі через мову. Отже, картина мовного світу є не лише когнітивною моделлю відображення дійсності, але й способом забезпечення гармонійного співіснування людини з навколишнім середовищем.

Таблиця 2.1 – Мовно-розумовий апарат vs комунікативна система

Ознака / Функція	Мовно-розумовий апарат	Комунікативна система
Призначення	Опрацювання та організація знань, формування мовних структур	Передача мовних даних між суб'єктами
Основні процеси	Кодування/декодування, аналіз, синтез, інтерпретація, концептуалізація, запам'ятовування, зберігання	Відтворення і передача інформації з мінімальними спотвореннями
Рівень складності	Високий: багатоступеневі когнітивні операції	Порівняно простий: технічний перенос сигналів
Середовище функціонування	Внутрішній «мовно-розумовий апарат» людини	Зовнішня інфраструктура мови та каналів комунікації
Результат	Формування мовної картини світу, осмислення досвіду	Обмін вербально вираженою інформацією
Метафора	«Програмне забезпечення» мислення	«Канал зв'язку» між учасниками

Таким чином, мовно-розумовий апарат і комунікативна система виконують різні, але взаємодоповнювальні функції. Перший забезпечує глибинну

когнітивну обробку інформації, перетворюючи чуттєві враження у концепти й мовні структури, тоді як друга реалізує технічний рівень передання цих структур від однієї особи до іншої. Разом вони формують цілісний когнітивно-комунікативний процес, у якому знання не лише народжується та структурується, а й стає доступним для колективного використання.

Висвітлені вище положення щодо картини мовного світу вимагають інше бачення мови та її системного характеру, оскільки воно суттєво відрізняється від прийнятого в традиційній лінгвістиці. Наявні визначення системи мови зводяться до інтерпретації її як структури мовного об'єкта, що визначається набором його елементів та зв'язків між ними. Проте така інтерпретація виключає властивості мови як засобу передання інформації. До них належать, передусім, субстантивні властивості мови, оскільки мова має реальні «матеріальні» втілення в інформаційно-комунікаційних системах будь-якого характеру та рівня. Тому, субстанція, так само як і структура, є іманентними ознаками поняття системи. Проте, В. А. Широков зазначає, що в означенні поняття системи активну роль відіграє ще один елемент, а саме – суб'єкт, здатний ідентифікувати, інтерпретувати структурно-субстанційні властивості й елементи системи і здійснювати з ними операції за певними правилами та алгоритмами. Отже, будь-яка система репрезентується символічною тріадою «структура – субстанція – суб'єкт», де:

- **структура** забезпечує упорядкованість і взаємозв'язки елементів, які складають мову як організовану цілісність;

- **субстанція** надає мові матеріальне втілення – у звуковій, графічній, цифровій або мультимедійній формах, що дозволяє їй функціонувати в різних інформаційно-комунікаційних середовищах;

- **суб'єкт** (носій мови, користувач, інтерпретатор) є активним елементом, без якого система не може існувати, адже саме він наділяє мовні структури значенням, інтерпретує їх і застосовує у практичних комунікативних актах.

У межах такого підходу мова розглядається не лише як об'єкт із внутрішньою організацією, а як динамічна **лексикографічна система**, що

перебуває в постійній взаємодії з носіями та середовищем. Це дає змогу осмислити її не лише в категоріях традиційної лінгвістики, але й у контексті сучасних інформаційно-лінгвістичних технологій, де важливими стають процеси обробки, кодування, зберігання і трансляції мовних даних. Якщо подати лексикографічний процес у вигляді символічної тріади, то це буде виглядати так, як показано в таблиці 2.2.

Таблиця 2.2 – Системна тріада в лексикографічному процесі

Елемент системи	Зміст	Приклади реалізації
Суб'єкт	Активний учасник, що ідентифікує, інтерпретує й організовує мовний матеріал	– окрема людина (лексикограф, мовознавець) – мовна спільнота (носії мови) – авторський колектив укладачів словника
Структура	Організація та взаємозв'язки елементів мови	– системні відношення: синонімія, антонімія, полісемія – ієрархії: гіперо-гіпонімія – структурні рівні: фонетика, морфологія, синтаксис – внутрішня організація словника: статті, рубрики, гасла
Субстанція	Матеріальне втілення мовних одиниць	– мовний матеріал: лексеми, словосполучення, тексти – словотвірні моделі (префіксація, суфіксація тощо) – граматичні форми словозміни (відмінювання, дієвідмінювання) – синтаксичні конструкції як база для прикладів

Системна тріада також застосовна для опису граматичного рівня, що охоплює словотвір, словозміну та синтаксис (див. табл. 2.3). **Структура** на цьому рівні репрезентується формальною моделлю, яка демонструє розклад мовної одиниці на незмінну (основу) та змінну (флексію) частини слова. Наприклад, у слові *стіл-а* виділяється основа *стіл-* та флексія *-а*. **Субстанція** визначається набором мовних елементів, із яких формуються змінні частини відмінюваних одиниць. Це можуть бути закінчення (*-а, -у, -ом*), суфікси словотвору (*-ник, -ість*), службові слова тощо. **Суб'єкт** застосовує правила і алгоритми, за допомогою яких здійснюється побудова словозмінної парадигми або словотвірної моделі за заданою структурою з використанням субстанційних елементів. Наприклад, мовець за певними правилами формує форми *стіл – стола – столові*

– *столами* тощо, або творить нове слово (*стільчик*) шляхом додавання зменшувального суфікса.

Таблиця 2.3 – Застосування системної тріади на граматичному рівні

Рівень	Структура	Субстанція	Суб'єкт
Словотвір	Моделі творення нових слів: додавання суфіксів, префіксів, складання основ (<i>читати</i> → <i>читач</i> , <i>підписати</i> → <i>підпис</i>)	Афікси (суфікси: -ник, -ість; префікси: під-, над-, без-), словотворчі основи	Мовець застосовує правила словотвору, формує нові слова для позначення понять (<i>літак</i> → <i>літакобудування</i>)
Словозміна	Парадигми відмінювання і дієвідмінювання (<i>стіл</i> – <i>стола</i> – <i>столові</i> ; <i>йти</i> – <i>йду</i> – <i>йшов</i>)	Закінчення (-а, -ові, -ами), службові морфеми, граматичні форманти	Мовець формує граматично правильні форми слова залежно від контексту (<i>працюю</i> – <i>працював</i> – <i>працюватиму</i>)
Синтаксис	Схеми побудови словосполучень і речень (підмет – присудок – додаток)	Лексеми та словоформи, сполучники, прийменники, службові елементи	Мовець конструює висловлення за синтаксичними правилами (<i>Діти читають книжки</i> ; <i>Якби не дощ, ми пішли б у парк</i>)

Отже, **системна тріада** відображає систему мови як абстрактний механізм, що здатний породжувати лінгвістичні об'єкти будь-якого ступеня складності – від окремих звуків і морфем до складних синтаксичних конструкцій і текстів. У її основі закладено три ключові компоненти: **структуру, субстанцію та суб'єкта**:

– **структура** виступає головним організуючим елементом системи, оскільки саме вона визначає схему й правила побудови мовних одиниць. Це абстрактна модель, яка задає відношення між елементами і впорядковує їх у цілісний механізм;

– **субстанція** постачає матеріал для реалізації структури. Нею є мовний матеріал у всіх його формах: фонемі, морфемі, лексемі, граматичні форми, синтаксичні моделі. Субстанція є «будівельним матеріалом» мовної системи.

– **суб'єкт** забезпечує активну сторону функціонування системи. Це носій мови (індивід чи колектив), який засвоює правила, оперує структурними схемами та реалізує їх у комунікативній практиці. Саме суб'єкт надає мовній системі динамічного, творчого виміру.

У такій інтерпретації система мови не зводиться лише до статичної структури чи матеріальної субстанції, а постає як живий, динамічний процес,

у якому взаємодіють абстрактна організація, матеріальні ресурси та діяльність суб'єкта. Ця концепція має особливе значення для сучасних комп'ютерних систем опрацювання природної мови (NLP). Вони відтворюють системну тріаду у штучному середовищі:

- структура – це алгоритми та моделі аналізу й синтезу;
- субстанція – лінгвістичні корпуси, бази даних, словники;
- суб'єкт – користувач або програмний агент, що ініціює й контролює процес обробки.

Виходячи з вищесказаного, системна тріада стає методологічною основою для створення електронних словників, систем машинного перекладу, автоматичного аналізу текстів та інших інформаційно-лінгвістичних технологій. У таблиці 2.4 подано порівняльну характеристику реалізації системної тріади у традиційній мовній системі та комп'ютерному середовищі.

Таблиця 2.4 – Порівняльна характеристика системної тріади

Компонент тріади	У традиційній мовній системі	У комп'ютерному мовному середовищі
Структура	Правила граматики, моделі словотвору, словозміни та синтаксису; лексико-семантичні відношення (синонімія, антонімія, гіперо-гіпонімія тощо).	Алгоритми та формальні моделі аналізу й синтезу: морфологічні парсери, синтаксичні дерева, семантичні мережі.
Субстанція	Матеріал мови: звуки, морфеми, лексеми, словникові статті, тексти.	Лінгвістичні корпуси, електронні словники, бази даних, семантичні онтології.
Суб'єкт	Людина-носій мови або мовна спільнота, що застосовує мовні ресурси у комунікації.	Користувач системи, лінгвіст-розробник, а також програмний агент, що виконує обробку даних.

Питання для самоконтролю

1. Що таке «картина мовного світу» і які її основні складники?
2. Чим картина мовного світу відрізняється від «картини світу» загалом?
3. Як мова відображає культурно-історичний досвід народу?
4. Назвіть три компоненти системної тріади.
5. Яку функцію виконує **структура** у системі мови?
6. Що становить собою **субстанція** у мовній системі? Наведіть приклади.

7. Хто або що може виступати як **суб'єкт** у процесі породження мовних об'єктів?
8. Чому структура вважається головним елементом системної тріади?
9. Як реалізується системна тріада на рівні словотвору?
10. Як вона проявляється у словозміні та синтаксисі?
11. Наведіть приклад зіставлення тріади в природній мові та у комп'ютерних мовних системах.
12. Як тріада допомагає моделювати мовні процеси у штучному інтелекті?

Завдання для самоперевірки

1. Візьміть одне слово української мови та розпишіть його за тріадою:
 - структура (що є незмінним, що змінюваним),
 - субстанція (які конкретні мовні елементи залучаються),
 - суб'єкт (хто використовує слово і як).

2. Заповніть таблицю. Вставте приклади до таблиці:

Рівень мови	Структура	Субстанція	Суб'єкт
Словотвір	?	?	?
Словозміна	?	?	?
Синтаксис	?	?	?

3. Співвіднесіть наведені приклади з елементами тріади: а) корпус текстів, словники, фонemi; б) алгоритм відмінювання іменників; в) мовна спільнота, програміст, штучний інтелект.

4. Уявіть, що ви створюєте «мовний конструктор» (аналог LEGO).

- Що в ньому буде виконувати функцію **структури**?
- Що становитиме **субстанцію**?
- Хто виступить **суб'єктом**?

5. Побудуйте приклад синтаксичної конструкції (речення) та розпишіть її за тріадою.

2.2. Поняття семантичного стану

Відповідно до концепції системної тріади, викладеної у підрозділі 2.1, предметом моделювання у сучасному лінгвістичному описі мають бути не стільки самі мовні одиниці конкретного рівня, скільки інтелектуальні процеси,

що пов'язані з побудовою лінгвістичних об'єктів. У результаті цих процесів одиниці набувають комплексу лінгвістичних характеристик. Цей комплекс прийнято позначати терміном **семантичний стан**. Під семантичним станом розуміється *сукупність граматичних і лексичних значень, які реалізує мовна одиниця у різних мовних контекстах*. Інакше кажучи, кожне слово чи форма має не одне-єдине значення, а спектр реалізацій, що актуалізується залежно від ситуації комунікації, контексту і комбінацій з іншими одиницями. Як приклад, розглянемо семантичні стани іменника *ключ* та дієслова *йти*.

Приклад 1. Іменник *ключ*

– У контексті *металевий ключ* воно реалізує семантичний стан «предмет для замикання».

– У контексті *ключ до розв'язання задачі* – «засіб, спосіб».

– У контексті *лісовий ключ* – «джерело води».

Отже, семантичний стан слова *ключ* формується як сукупність його можливих значень, які активуються в різних мовних ситуаціях (див. таблицю 2.5).

Приклад 2. Дієслово *йти*

– *йти пішки* – фізичне пересування

– *йти до школи* – напрямок руху.

– *йти час* – метафоричне вживання

– *йти дощ* – природне явище.

У кожному випадку активується різний сегмент семантичного стану

дієслова, хоча форма залишається незмінною (див. таблицю 2.5).

формується як сукупність його можливих значень, які активуються в різних мовних ситуаціях (див. таблицю 2.5).

Таблиця 2.5 – Приклади реалізації семантичних станів

Лема	Контекст	Реалізоване значення	Тип семантичної варіації
ключ	<i>металевий ключ</i>	Прилад для замикання / відмикання	Пряме предметне значення
ключ	<i>ключ до задачі</i>	Засіб, спосіб для розв'язання	Метафоричне переносне значення
ключ	<i>лісовий ключ</i>	Джерело води	Омонімічне/лексичне значення
йти	<i>йти пішки</i>	Фізичне пересування у просторі	Базове значення
йти	<i>йти до школи</i>	Напрямок руху до мети	Конкретизація базового значення
йти	<i>йде час</i>	Плин, перебіг процесу	Метафоричне узагальнене
йти	<i>йде дощ</i>	Відбувається природне явище	Лексична полісемія / метафора

Наведена вище таблиця демонструє, що **семантичний стан** – це не одне зафіксоване значення, а комплекс можливостей, які активуються у мовленні залежно від контексту. У наведених вище контекстах одиниці реалізують не лише лексичне, а й граматичне значення. Особливо це спостерігається у дієслова *йти*. Наприклад, у значенні «пересування» воно має повну парадигму, тобто усі можливі граматичні форми. У значенні «рухатися в конкретному напрямі» воно може мати або повну парадигму, якщо перед ним стоїть іменник на позначення істоти, або неповну у випадку неістоти (наприклад, *тролейбус*). Що стосується значень «перебігу часу» та «відбуватися про природне явище», аналізоване дієслово матиме лише форми третьої особи однини та множини, як це показано у таблиці 2.6.

Таблиця 2.6 – Параметри семантичного стану дієслова *йти*

Реалізоване значення	Приклад контексту	Лексичний аспект	Граматичний аспект (парадигма)
Пересування у просторі	<i>Він іде вулицею</i>	Фізичний рух суб'єкта-істоти	Повна парадигма (усі особи, числа, часи, способи)
Рух у конкретному напрямі	<i>Тролейбус іде до центру</i>	Переміщення транспорту/неістоти	Неповна парадигма (відсутні деякі форми; напр., наказовий спосіб)
Перебіг часу	<i>Час іде швидко</i>	Метафоричне значення «плин часу»	Лише форми 3-ї особи (одн./мн.): <i>йде/йдуть</i>
Природне явище	<i>Йде дощ</i>	Подія, що відбувається у природі	Лише форми 3-ї особи (одн./мн.): <i>йде/йдуть</i>

Тобто, у межах одного слова семантичний стан визначає обсяг граматичних можливостей: від повної парадигми до обмеженої. Отже, семантичний стан – це динамічний багатовимірний простір значень, що задає потенціал мовної одиниці у різних контекстах. Ключовими характеристиками семантичного стану є:

- (1) *динамічність* – значення не є раз і назавжди закріпленими, а варіюються у мовленні.
- (2) *контекстуальність* – актуалізація відбувається лише в певному оточенні.
- (3) *комплексність* – поєднання лексичних і граматичних характеристик.
- (4) *когнітивна обумовленість* – реалізація семантичного стану залежить від когнітивних процесів суб'єкта, який інтерпретує ситуацію.

Найповнішу колекцію семантичних станів певної мови традиційно репрезентують національні **тлумачні словники**, які фіксують і систематизують лексичний фонд у всій його повноті – від загальноновживаної лексики до спеціалізованих та периферійних одиниць. У таких словниках відображаються всі відомі значення слова, їхні граматичні характеристики, стилістичні відтінки та приклади функціонування в різних комунікативних ситуаціях. Для української мови найбільш репрезентативним є «Словник української мови» у 20 томах (СУМ-20), який упродовж кількох десятиліть залишався головним джерелом семантичних станів української лексики. В англійській лексикографічній традиції цю роль виконує «Oxford English Dictionary», у німецькій – багатотомний DUDEN, в іспанській – авторитетний Diccionario de la lengua Española, підготовлений Королівською академією (RAE). Таким чином, національні тлумачні словники можна розглядати як лінгвістичні «репозиторії» семантичних станів, що відтворюють мовну картину світу певної спільноти й забезпечують можливість для подальшого моделювання у сфері електронного словникарства. Порівняльна характеристика наведених словників подана в таблиці 2.7.

Таблиця 2.7 – Порівняльна характеристика словників

Мова	Назва словника	Обсяг / томи	Установа / видавець	Головна функція
Українська	Словник української мови (СУМ у 20 томах)	20 томів (бл. 134 тис. слів)	Інститут мовознавства ім. О. Потебні НАН України	Повна репрезентація лексичного складу української мови XX ст.
Англійська	Oxford English Dictionary (OED)	понад 600 тис. слів, багатотомне видання	Oxford University Press	Історико-етимологічний опис англійської лексики, розвиток значень
Німецька	DUDEN. Das große Wörterbuch der deutschen Sprache	10–12 томів (різні видання)	Bibliographisches Institut	Кодифікація сучасної німецької літературної мови, нормативність
Іспанська	Diccionario de la lengua Española (DLE, RAE)	1 том (бл. 100 тис. слів)	Real Academia Española (RAE)	Нормативна кодифікація іспанської мови, загальнонаціональна стандартизація

Розглянемо словникову статтю лексеми *хліб* у тлумачному словнику української мови разом із її семантичними станами:

ХЛІБ

1. Харчовий продукт, що випікається з борошна. *Хліб усьому голова.*
2. *тільки одн.* Зерно, з якого виготовляють борошно. Швидше й швидше летять ешелони, хліб, і вугіль, і нафту везуть...
3. *перев. мн.* хліба, хліби, *ів.* Зернові культури. *Хліби зійшли, Радіють люде, бога просять...*
// перев. одн. Необмолочені зернові. *Коли б я хліба не возив, То що б стеріг тоді кудлатий.*
// перев. одн. Урожай зернових. *Моя мати між двома хлібами вмерла: одного не дочекала, а другий уже минувся.*
4. *перев. одн., перен., розм.* Засоби до існування; заробіток. *Коли людина має цікаву роботу, ...коли вона не має побоювань про завтрашній хліб, фізично здорова...*
5. *перен., розм.* Харчі, їжа. *Чий хліб їси, під того й дудочку скачи!*

Як бачимо, одна й та сама форма *хліб* реалізує низку різних семантичних станів – від конкретного харчового продукту до метафоричного позначення заробітку чи засобів існування. Це свідчить, що в словниковій практиці ми маємо справу не з «багатозначністю одного слова», а з низкою співіснуючих лексем, кожна з яких має власне, чітке значення. Таке трактування перегукується з відомим застереженням Л. В. Щерби: «Неправильно думати, що слова мають кілька значень: це, власно кажучи, формальна і навіть просто типографська точка зору. Насправді ми маємо завжди стільки слів, скільки фонетичне слово має значень (між іншим, так друкували в старих словниках: заголовкове слово повторювали стільки, скільки у нього було значень)»¹.

¹ Цит. за Бурковська О. Б. Українська термінологія лісівництва: становлення, розвиток та системно-структурна організація: дис. канд. філол. н. 10.02.01, 035. Львів. 2018. 392 с. С. 150. URL: https://science.dspu.edu.ua/wp-content/uploads/2018/08/dis_burkovska2018-ost.pdf.

Отже, теорія семантичних станів створює підґрунтя для побудови концептуальної моделі інтегрального опису мови. У цій моделі кожна мовна одиниця розглядається не ізольовано, а в нерозривному поєднанні своїх лексичних та граматичних характеристик. Граматичний компонент семантичного стану відображає належність мовної одиниці до певного граматичного класу. Наприклад, дієслово *йти* входить до класу дієслів руху, що мають парадигму часу, способу, особи тощо. Лексичний компонент реалізує конкретну лексичну семантику, яка завжди співвідноситься з граматичним класом. Так, значення *йти* «пересуватися» і *йти* «процес триває (йде час)» належать до різних семантичних станів, кожен з яких має свій набір граматичних обмежень. У такий спосіб забезпечується принцип інтегральності, сформульований Ю. Д. Апресяном²:

– з одного боку, у словнику повинна міститися інформація, що передбачає звернення до правил граматики;

– з іншого боку, формулювання граматичних правил має відбуватися з урахуванням словникових даних про конкретні мовні одиниці.

Цей підхід дає можливість створювати єдину інтегральну систему опису мови, у якій словникова і граматична інформація взаємно узгоджуються й доповнюють одна одну, як це показано у таблиці 2.8.

Таблиця 2.8 – Лексичний і граматичний компоненти семантичного стану

Компонент	Що репрезентує	Приклад	Взаємозв'язок
Граматичний	Клас мовних одиниць, що мають спільні граматичні характеристики (парадигму, категорії, правила вживання)	<i>йти</i> як дієслово руху: має час, спосіб, особу, число	Визначає формальну можливість реалізації семантичного стану
Лексичний	Конкретну семантику, яка реалізується в контексті й співвідноситься з граматичним класом	<i>йти</i> «пересуватися», <i>йти</i> «минати (про час)», <i>йти</i> «відбуватися (про подію)»	Задає змістову реалізацію граматичного потенціалу
Інтегральність	Поєднання обох аспектів: у словнику подається значення з	У словниковій статті «йти» зазначено можливість/неможливість	Словник і граматика

² Цит. за Спецкурс з основ укладання галузевих глосаріїв / О.В. Іванова. – К.: ЦП «Компринт», 2015. – 142 с. С. 44–45. URL: https://nubip.edu.ua/sites/default/files/u137/speckurs_z_glosariiv_praktichni_zanyattya.pdf.

	урахуванням граматичних обмежень, а граматику спирається на словникові дані	певних граматичних форм для різних значень	утворюють єдину систему

Варто зазначити, що словник описує не лише **повнозначні слова** (іменники, дієслова, прикметники, прислівники), а й **службові частини мови**. У такому випадку **лексичний компонент** семантичного стану має специфіку: він відображає не предметно-поняттєву семантику, а насамперед функціональне призначення цих одиниць у мовленні. Наприклад:

– **Прийменники** (*в, на, до, із* тощо) не мають власної номінативної семантики, але задають відношення між словами (*йти до школи, сидіти в кімнаті*).

– **Сполучники** (*і, але, тому що*) виражають логіко-граматичні відношення між частинами висловлення (*Я читаю, а він пише*).

– **Частки** (*би, ж, лише*) модифікують висловлення, підкреслюючи його модальні або стилістичні відтінки (*Хотів би відпочити*).

– **Займенники** (*я, ти, хтось, цей*) виконують функцію заміщення інших лексем і реалізують відсилання до мовного чи позамовного контексту.

Тобто, під час опису службових слів у словнику **лексичний компонент семантичного стану** репрезентує не лексичне значення у традиційному сенсі, а функціональне навантаження, яке вони несуть у мовній структурі (пор лексичні компоненти для повнозначних і службових слів у таблиці 2.9).

Таблиця 2.9 – Лексичний компонент для повнозначних і службових слів

Тип слова	Приклади	Лексичний компонент семантичного стану	Особливості реалізації
Повнозначні	<i>дім, вода, бігти, зелений, швидко</i>	Виражає предметне або ознакове значення (що називає слово: об'єкт, дію, властивість, ознаку, обставину).	Має лексичну семантику, яка може поєднуватися з граматичними характеристиками (рід, число, відмінок, час, спосіб тощо).
Службові: прийменники	<i>в, на, до, із</i>	Виражають відношення між повнозначними словами.	Не мають власної предметної семантики, їх значення – синтаксичне (місце, напрям, час, спосіб).

Службові: сполучники	<i>і, але, тому що</i>	Виражають логіко-граматичні відношення (зіставлення, протиставлення, причиновість тощо).	Лексичний компонент визначається через функцію зв'язку частин висловлення.
Службові: частки	<i>би, ж, лише</i>	Виражають модальні та стилістичні відтінки висловлення.	Лексичний компонент полягає у модифікації вислову (емоційність, підсилення, умовність).
Службові: займенники	<i>я, ти, він, цей, хтось</i>	Виконують функцію заміщення іменників чи прикметників, відсилання до учасників комунікації або контексту.	Лексичний компонент відображає деїктичну функцію (вказування, відсилання).

Виходячи з вищесказаного, семантичний стан є ключовим поняттям у сучасній лексикографічній та комп'ютерно-лінгвістичній парадигмі, адже він відображає не лише полісемію чи граматичні особливості слова, а й цілісний результат інтелектуальної діяльності мовця. Це не статична одиниця, а динамічна модель, що поєднує у собі:

- лексичний компонент (предметно-поняттєве значення або функціональне призначення слова);
- граматичний компонент (парадигматичні й синтаксичні характеристики слова);
- контекстуальну реалізацію (конкретне вживання у мовленні, що визначає актуальний зміст).

Отже, семантичний стан можна уявити як модель інтеграції граматики та лексики, де кожна мовна одиниця набуває конкретного змісту і форми у процесі комунікації. Це дозволяє описати мову як динамічну систему, у якій взаємодія лексичного та граматичного рівнів забезпечує породження нескінченного різноманіття мовленнєвих форм. У перспективі ця концепція є фундаментом для створення електронних словників нового покоління, де опис слова повинен включати не лише його значення, а й повний спектр його семантичних станів.

Питання для самоконтролю

1. Що таке семантичний стан і чим він відрізняється від полісемії?
2. Які основні компоненти входять у структуру семантичного стану?
3. Як змінюється семантичний стан дієслова залежно від його контексту?

4. Яку роль відіграє граматичний компонент у формуванні семантичного стану?
5. Чому, за концепцією Ю. Д. Апресяна, правила граматики й словникова інформація мають інтегруватися?
6. Як описуються семантичні стани у службових частин мови (прийменники, сполучники, частки, займенники)?
7. Чому семантичний стан вважають моделлю інтеграції граматики та лексики?
8. Зіставте позицію Л. В. Щерби з традиційною лексикографічною практикою. У чому полягає відмінність?

Завдання для самоперевірки

1. Визначте, скільки семантичних станів має слово *хліб* у наведеній статті.
2. Підберіть приклад із сучасної мови, де *хліб* вживається у переносному значенні. Вкажіть, до якого з поданих семантичних станів він належить.
3. Виділіть семантичні стани дієслова *йти* у наведених реченнях і вкажіть граматичні обмеження:
 - а) *Він щодня йде до школи.*
 - б) *Годинник іде правильно.*
 - в) *Літо йде швидко.*
 - г) *Дощ іде без утину.*
4. Розгляньте словникову статтю до слова **ключ** (українською або в іншій мові за вашим вибором). Визначте, які семантичні стани тут представлені й чи всі вони мають повні парадигми.
5. Випишіть із **СУМу** або іншого тлумачного словника три багатозначні слова. Для кожного з них:
 - сформулюйте семантичні стани,
 - зазначте, як співвідносяться лексичні та граматичні компоненти.
6. Порівняйте опис семантичних станів у двох словниках (наприклад, **СУМ** та **Oxford English Dictionary**). Визначте:
 - чи подано там усі можливі стани,

- чи обмежено лише типовими контекстами,
- чи співпадає представлення граматичних характеристик.

7. Для службових слів (**і, але, щоб, як**) опишіть семантичні стани у термінах функціонального призначення (наприклад: «і» — єднальний зв'язок; «але» — протиставлення).

2.3. Лексикографічна система та лексикографічні середовища

Словник, як найповніший репозиторій семантичних станів, являє собою складну **лексикографічну систему**. Водночас словник є лише окремим випадком таких систем. Лексикографічною системою може виступати і сама природна мова, функціонування якої відбувається у вигляді тріади «структура + суб'єкт + субстанція». Ба більше, аналогічні властивості можуть мати й інші інформаційні системи, організовані за принципом, подібним до природної мови. Будь-яка лексикографічна система включає:

(1) *Множину елементарних інформаційних одиниць* – мінімальних компонентів, що описують дійсність або її фрагмент та відповідають певному **лексикографічному ефекту** (див. нижче про лексикографічний ефект).

(2) *Множину описів цих одиниць* – їхніх формальних (граматичних, структурних) та змістових (лексико-семантичних) характеристик.

(3) *Три типи відношень*:

- за першим одиниця набуває форми,
- за другим – змісту,
- за третім встановлюється відповідність між формою та змістом.

Детальну характеристику елементів лексикографічної системи подано у таблиці 2.10.

Таблиця 2.10 – Основні компоненти лексикографічної системи

Компонент	Зміст	Функція
Елементарні інформаційні одиниці	мінімальні одиниці опису (слова, морфеми, терміни, умовні знаки)	репрезентують фрагменти дійсності
Описи одиниць	формальні (морфологія, синтаксис) та змістові (лексичне значення, семантичні відтінки) характеристики	забезпечують системність та стандартизацію

Відношення	форма ↔ зміст ↔ відповідність	формують цілісну модель мовної/інформаційної системи

Як можна побачити, така система завжди передбачає наявність певного аналога **граматики** та **тезауруса (словника)**, які взаємодіють між собою у процесі функціонування. Результатом цієї взаємодії стають різнорівневі явища, що, за концепцією В. А. Широкова, визначаються як **лексикографічні ефекти**. Прикладами таких ефектів у мові виступають *об'єкти мовної системи* (звуки, морфеми, слова, колокації, мовленнєві акти та ін.) або *лінгвістичні відношення* (синонімія, антонімія, гіперо-гіпонімія, модальність, узгодження, керування, метонімія тощо). **Лексикографічний ефект**, за своєю сутністю, *виступає репрезентантом певного мовного феномену, який має власні форму, зміст і механізм утворення* (див. таблицю 2.11). Залежно від природи феномену, механізм його утворення буде відмінним. Так, наприклад, якщо розглядати мовленнєві ефекти (речення) як лексикографічні, то їхній механізм породження відповідатиме моделям, описаним у теорії формальних граматик Н. Хомського. Якщо ж йдеться про словозміну, то утворення словоформ являтиме собою механізм послідовного приєднання до основи слова флексій з відповідного парадигматичного набору.

Таблиця 2.11 – Лексикографічні ефекти та механізми їх утворення

Рівень мови	Приклади лексикографічних ефектів	Механізм утворення
Фонетичний	звуки, фонеми, алофони	артикуляційні процеси, фонологічні правила
Морфологічний	морфеми, словоформи, моделі словозміни	приєднання афіксів (флексій, префіксів, суфіксів) до основи відповідно до парадигми
Лексичний	слова, фразеологізми, колокації	комбінація морфем; усталене сполучення слів; фіксація значення у словнику
Синтаксичний	речення, мовленнєві акти, синтаксичні конструкції	правила поєднання слів за моделями формальних граматик (у т. ч. трансформаційних)
Семантичний	синонімія, антонімія, гіперо-гіпонімія, метонімія	встановлення семантичних зв'язків між одиницями
Прагматичний	модальність, мовленнєві жанри, комунікативні функції	співвіднесення мовних засобів із ситуацією та інтенцією мовця

Таким чином, лексикографічна система може бути охарактеризована через такі ключові компоненти:

- *концептуальний* – визначає загальну модель опису мови чи інформаційної системи;
- *методологічний* – окреслює підходи та принципи упорядкування одиниць;
- *структурний* – відображає внутрішню організацію даних (граматика, словник, формат представлення).
- *функціональний* – забезпечує можливість взаємодії суб'єкта з інформаційною субстанцією.

З огляду на це постає поняття **лексикографічного середовища** – *простору, у межах якого відбувається практична реалізація функцій системи*. У традиційному розумінні воно охоплює друковані словники та довідкові видання, а в сучасних умовах поширюється на електронні бази даних, онлайнві платформи й інтерактивні корпуси. Саме середовище визначає, якими є способи доступу, форми репрезентації та можливості користувача щодо навігації в семантичному полі. Розрізняють традиційне (друковане), електронне, онлайнове інтегроване та динамічне середовище. Наведемо їхню характеристику разом із перевагами та недоліками.

Традиційне (друковане) середовище словникарства охоплює тлумачні, орфографічні, етимологічні, перекладні словники, енциклопедії, атласи. Його ключова характеристика – статичність: структура словника фіксована, обсяг обмежений, а оновлення можливе лише з виходом нових видань. Перевагами традиційного середовища є надійність, культурний і науковий авторитет, а також зручність для цитування, що забезпечує сталість і точність посилань у науковій практиці. Водночас воно має суттєві недоліки: відсутність динамічності, складність оперативного оновлення та обмежений доступ для широкого кола користувачів.

На відміну від традиційного, **електронне (офлайнове) середовище** представлене CD/DVD-словниками, комп'ютерними програмами для інсталяції та локального використання. Основними рисами цього середовища є цифровий формат, зручний пошук за ключовими словами, можливість мультимедійного супроводу (звукових чи візуальних ілюстрацій). Це робить словник більш

компактним, динамічним у використанні, а також значно пришвидшує доступ до інформації. Проте обмеженням залишається потреба у технічному носії, а також складність в оперативному оновленні, яке повноцінно реалізується лише в онлайн-середовищі. Порівняльна характеристика традиційного і електронного середовища подана в таблиці 2.12.

Таблиця 2.12 – Традиційне vs. електронне середовище

Ознака	Традиційне середовище (друковане)	Електронне середовище (офлайнове)
Форми	Тлумачні, орфографічні, етимологічні, перекладні словники; енциклопедії, атласи	CD/DVD-словники, встановлювані програми
Характер	Статичність, фіксована структура	Цифровий формат, інтерактивність
Обсяг	Обмежений, залежить від фізичного носія (книги)	Компактний, може містити значно більші масиви даних
Пошук інформації	Лінійний (за алфавітом, індексами)	Автоматизований пошук за ключовими словами
Мультимедійність	Відсутня	Є можливість додавання звуку, зображень, відео
Оновлення	Лише через нові видання	Обмежене, можливе лише з новими версіями програм/дисків
Доступність	Потрібна фізична наявність книги	Потребує технічного носія (диск, комп'ютер)
Переваги	Надійність, культурний авторитет, зручність для цитування	Швидкість пошуку, компактність, мультимедійність
Недоліки	Відсутність динаміки, обмежений доступ, складність оновлення	Залежність від технічних засобів, складність регулярного оновлення

З виходом лексикографії в Інтернет почало активно розвиватися **онлайн-середовище**, яке репрезентують сайти-словники, гіпертекстові системи та електронні версії друкованих словників. Його основними характеристиками є: доступність через Інтернет, інтерактивність, постійне оновлення, що уможлиблює широке охоплення аудиторії. Воно також забезпечує багатомовність, швидке редагування і створює умови для комунікації користувача з базою даних. Недоліками онлайн-середовища є залежність від Інтернету та інколи нерівномірна якість контенту. Для вирішення цієї проблеми необхідне створення якісної ресурсної бази. У традиційній лексикографії такою базою були картотеки, а в сучасних умовах – **корпуси текстів**.

Корпусне середовище охоплює загальномовні та спеціалізовані корпуси текстів, як-от: *Український національний лінгвістичний корпус*, *Генеральний регіональний анований корпус (ГРАК)*, *Броунівський корпус*, *Британський національний корпус*. Воно надає великий емпіричний матеріал, у якому можна простежити слововживання, контексти, граматичні зв'язки. Головними перевагами корпусного середовища є достовірність даних, їхня прив'язаність до реального мовного вжитку та можливість статистичного аналізу. Водночас воно вимагає спеціалізованого інтерфейсу, а для непідготовленого користувача робота з корпусом може виявитися складною (див. порівняльну характеристику у таблиці 2.13).

Таблиця 2.13 – Онлайнове vs. корпусне середовище

Ознака	Онлайнове середовище	Корпусне середовище
Форми	Сайти-словники, гіпертекстові системи, онлайн-версії друкованих словників	Загальномовні та спеціалізовані корпуси текстів
Доступність	Через Інтернет, у будь-який час і з будь-якої точки світу	Через спеціалізовані платформи, потребує технічної підготовки
Характер	Інтерактивність, постійне оновлення	Емпірична база мовних даних, відображення реального вжитку
Оновлення	Швидке редагування, динамічне оновлення контенту	Оновлення корпусів через додавання нових текстових масивів
Можливості	Багатомовність, гіпертекстові зв'язки, інтеграція мультимедіа	Статистичний аналіз, пошук контекстів, граматичних зв'язків
Переваги	Широке охоплення користувачів, доступність, оперативність	Достовірність, реальний мовний ужиток, аналітичний потенціал
Недоліки	Залежність від Інтернету, нерівномірна якість контенту	Складність для непідготовлених користувачів, потреба у спеціалізованому інтерфейсі

Інтегроване середовище являє собою поєднання електронного словника, корпусу, тезауруса, граматики в єдиній платформі – багатофункціональній системі, орієнтованій на комплексний аналіз мови. До переваг такого середовища можна віднести повноту опису, можливість поєднання різних джерел знань, використання гнучких інструментів для дослідників. Водночас для реалізації подібних рішень потрібні значні ресурси.

Частково цю проблему вирішують **динамічні (соціально-мережеві) середовища** – краудсорсингові платформи (Urban Dictionary, Вікісловник), інтерактивні лексикографічні сервіси, наповнення яких здійснюється

спільнотою. Дані тут змінюються у режимі реального часу, що робить такі ресурси особливо чутливими до мовних новацій. Однак вони не завжди гарантують академічну стандартизацію матеріалу, можуть містити помилки або суб'єктивні тлумачення (див. порівняльну характеристику у таблиці 2.14).

Таблиця 2.14 – Інтегроване vs. динамічне середовище

Середовище	Приклади	Переваги	Недоліки
Інтегроване	Єдині платформи, що об'єднують словник, корпус, тезаурус, граматику	Повнота опису, комплексність аналізу, доступ до різних джерел знань в одному інтерфейсі	Висока вартість і складність реалізації, потреба у значних ресурсах
Динамічне (соціально-мережеве)	Urban Dictionary, Вікісловник, інтерактивні сервіси	Постійне оновлення, участь спільноти, чутливість до нових слів і мовних тенденцій	Відсутність академічної стандартизації, суб'єктивність, можливі помилки

Розвиток лексикографічних середовищ від традиційних друкованих словників до інтегрованих і динамічних платформ відображає загальну еволюцію інформаційних систем та способів взаємодії з мовними даними. Кожне середовище має власні переваги та обмеження, пов'язані з рівнем технологічного розвитку, доступністю ресурсів і потребами користувачів.

Якщо традиційні словники вирізняються авторитетністю та стабільністю, то електронні й онлайнові системи забезпечують динаміку, інтерактивність і широке охоплення. Корпусні ресурси відкривають можливості для статистичного й емпіричного аналізу мовних фактів, тоді як інтегровані середовища прагнуть синтезувати різні підходи в єдиній багатофункціональній платформі. Нарешті, динамічні (соціально-мережеві) середовища демонструють потенціал колективної взаємодії користувачів, проте залишаються чутливими до питань академічної точності. Отже, сучасна лексикографія перебуває у фазі співіснування різних середовищ, кожне з яких виконує власні завдання і може взаємодоповнювати інші. Оптимальним підходом видається інтеграція їхніх можливостей із забезпеченням балансу між науковою достовірністю, технологічною зручністю та відкритістю до мовних інновацій (див порівняльну характеристику лексикографічних середовищ у додатку Д.3).

Питання для самоконтролю

1. У чому полягає відмінність між словником і лексикографічною системою?
2. Чому природну мову можна розглядати як лексикографічну систему?
3. Які три обов'язкові компоненти входять до складу будь-якої лексикографічної системи?
4. Що розуміється під елементарними інформаційними одиницями? Наведіть приклади.
5. Які характеристики описують елементарні інформаційні одиниці?
6. Які три типи відношень формуються у лексикографічній системі?
7. Як співвідносяться між собою форма і зміст у межах лексикографічної системи?
8. Чим лексикографічні системи, організовані за принципом мови, відрізняються від інших інформаційних систем?
9. Як, на вашу думку, можна застосувати принципи побудови лексикографічних систем у сучасних комп'ютерних технологіях (наприклад, у базах даних або штучному інтелекті)?

Завдання для самоперевірки

1. Зіставте наведені нижче компоненти лексикографічної системи з їхнім описом:
 - (1) Елементарні інформаційні одиниці
 - (2) Описи одиниць
 - (3) Типи відношень
 - а) відповідність між формою і змістом
 - б) характеристики форми та значення
 - в) одиниці, що відображають дійсність
2. Побудуйте власний приклад тріади «структура – субстанція – суб'єкт» у межах лексикографічної системи (наприклад, для словотвірної моделі чи для конкретного словникового фрагмента).
3. Складіть таблицю порівняння двох електронних лексикографічних середовищ (наприклад: СУМ20, Горох, Merriam-Websters Dictionary, Oxford

English Dictionary) за параметрами: інтерфейс, пошук, обсяг даних, навчальні можливості.

4. Уявіть, що ви створюєте нову електронну платформу для укладання словників. Опишіть, які інструменти та функції ви б у ній передбачили.

5. Які переваги та виклики, на вашу думку, має поєднання штучного інтелекту й лексикографії?

Питання для поглибленого вивчення матеріалу

1. Системність у мові та лінгвістиці.

2. Теорія семантичних станів як інструментальний засіб експлікації лексичного значення.

3. Лексикографічна замкненість і лексикографічна повнота словників.

4. Архітектура лексикографічних систем. Інтегровані лексикографічні системи.

5. Теорія лексикографічних систем як концептуальна основа для створення електронних словників.

3. СЛОВНИКИ В ГЛОБАЛЬНІЙ ЛЕКСИКОГРАФІЧНІЙ МЕРЕЖІ

План заняття:

3.1. Глобальна лексикографічна мережа.

3.2. Локальні та мережеві лексикографічні ресурси.

3.3. Інтеграція онлайн-ресурсів у єдиний глобальний простір.

Упродовж останніх десятиліть лексикографія зазнала фундаментальних змін: від статичних друкованих словників вона поступово перемістилася у цифровий простір, де словник перестає бути самодостатнім об'єктом і перетворюється на елемент складної мережі знань. Інтернет виявився не просто платформою для публікації словників, а ідеальним середовищем для взаємодії, інтеграції та мережевого обміну мовними ресурсами, створюючи можливість формування **Всесвітньої лексикографічної мережі (Global Lexicographic Network)**.

3.1. Всесвітня лексикографічна мережа

Сучасна комп'ютерна лексикографія дедалі більше нагадує мережу знань, у якій словники, корпуси, онтології, морфологічні розділи та інші мовні ресурси взаємодіють між собою через спільні формати, протоколи й стандарти. Ця модель різко відрізняється від традиційної парадигми, де словник існував як замкнений об'єкт, ізольована книга або автономний електронний файл. Інтернет надав словникам нову «екологічну нішу»: можливість бути взаємопов'язаними, динамічними, оновлюваними і відкритими до зовнішніх даних. Отже, термін **лексикографічна мережа** використовується на позначення множини лексичних ресурсів, які:

- взаємопов'язані семантично, технічно або функціонально,
- обмінюються даними автоматично,
- опираються на спільні технічні стандарти,
- можуть інтегрувати зміни одне одного в реальному часі.

Інакше кажучи, йдеться про екосистему, де жоден словник не існує ізольовано; навпаки, він посилається на інші ресурси, отримує від них визначення, приклади, колокації, морфологічну інформацію, частотні дані,

семантичні зв'язки – і водночас може служити джерелом даних для інших. Це породжує явище дистрибутивної лексикографії, коли знання розподілені між різними вузлами і з'єднані між собою мережевими зв'язками. Структурними елементами лексикографічної мережі є:

1. Словники (тлумачні, перекладні, спеціалізовані)

Вони є основними «вузлами знань».

- BECUM
- Oxford Dictionaries
- Wiktionary
- Duden
- Lexicala API

Служать джерелом прикладів, частот та колокацій:

- ГРАК
- BNC
- COCA
- Sketch Engine corpora.

3. Семантичні мережі

Вони забезпечують концептуальні зв'язки:

- WordNet
- BabelNet
- ConceptNet
- FrameNet

4. Лінгвістичні розділі

Морфологічні, синтаксичні, семантичні аналізатори:

- Apertium
- spaCy models
- Stanza
- UDPipe

5. Платформи інтеграції

Інструменти, що збирають та з'єднують словники та корпуси:

- Sketch Engine
- Lexonomy
- LinguaLibre
- Lexicala
- Wikidata Lexemes.

Становлення лексикографічної мережі стало можливим завдяки трьом групам технологій, до яких належать: стандартизовані формати репрезентації

лексичних даних (додаток Д.4), мережеві протоколи й API (додаток Д.5), а також корпусні інструменти (див. тему 8). Формати репрезентації (рис. 3.1) дають можливість різним словникам «розуміти» один одного, забезпечуючи спільні правила подання лексичних даних та їхню взаємну сумісність. Завдяки стандартизованим моделям структурування інформації лексикографічні ресурси можуть інтегруватися, обмінюватися даними та використовувати єдині інструменти для обробки. Основні формати та їхні властивості докладно розглянуто у підрозділі 5.2. У цьому ж розділі зосередимося на мережевих протоколах і API, які забезпечують технічну взаємодію між словниками, відкриваючи можливості для розподілених, динамічно оновлюваних лексикографічних систем.



Рис. 3.1 – Формати репрезентації лексичних даних

У сучасній лексикографії API відіграють ключову роль у забезпеченні взаємодії між словниками, корпусами та іншими мовними ресурсами. Різні типи API служать різним завданням – від швидкого доступу до словникових статей до побудови складних семантичних мереж. Нижче подано короткий огляд найбільш поширених API та сфер їхнього застосування в лексикографічній інфраструктурі.

Нижче наведено короткий огляд мережевих протоколів, які використовуються в комп'ютерній лексикографії.

REST API залишається найпоширенішим інструментом публікації словникових даних. Його використовують провідні онлайн-словники (Oxford Dictionaries API, Merriam-Webster API), корпусні платформи (Sketch Engine) та сервіси машинного перекладу. Він підходить передусім для задач, де потрібні прості, швидкі запити типу «отримати статтю», «знайти лему», «повернути морфологічний аналіз». REST забезпечує ефективну інтеграцію з мобільними застосунками, веб-сервісами та чат-ботами. Так, наприклад, словник, що розробляється на платформі Lexonomy (рис. 3.2), можна під'єднати через API до сервісів Google Search, Pixabay та Voice RSS.

🏠 > DICTIONARIES > УКРАЇНСЬКО-АНГЛІЙСЬКИЙ СЛОВНИК ЛЕКСИЧНИХ ПАРАЛЕЛІ > CONFIGURE > MULTIMEDIA API

MULTIMEDIA API

IMAGE LICENCE
any licence ▼
Select licence type when searching for images.

GOOGLE CUSTOM SEARCH API KEY

Insert your Google Custom Search API key to allow multimedia search.

GOOGLE CUSTOM SEARCH ID

Insert ID of your Custom Search - see [documentation](#)

PIXABAY API KEY

Insert your [Pixabay API key](#).

VOICERSS API KEY

Insert your [VoiceRSS](#) API key to enable text-to-speech.

Рис. 3.2 – Інтеграція словника з різними сервісами через API

GraphQL API застосовується там, де лексикографічні дані мають багаторівневу структуру та численні взаємозв'язки. На відміну від REST, який повертає цілі структури, GraphQL надає клієнту можливість запитувати рівно ті фрагменти даних, які йому потрібні. Це робить його корисним у великих багатокомпонентних ресурсах на кшталт сучасних багатомовних платформ (на зразок Glosbe) або інтегрованих термінологічних баз. GraphQL добре підходить для побудови інтерфейсів, у яких необхідно збирати пов'язані елементи: лему, значення, приклади, джерела, семантичні мітки тощо (рис. 3.3).

```

class Meaning(models.Model):
    id = models.UUIDField(primary_key=True, default=uuid_lib.uuid4, editable=False)
    meaning = models.TextField(null=True, blank=True, db_index=True)
    lemma_related = models.BooleanField(default=True)
    language = models.CharField(max_length=10, blank=True, null=True, db_index=True)
    related_meanings = models.ManyToManyField('self', blank=True, related_name='meaning_related_meanings')
    created_at = models.DateTimeField(auto_now_add=True)

    def related_lemmas(self):
        return self.lemma_related_meanings.all()

history = HistoricalRecords()
class Meta:
    constraints = [
        models.UniqueConstraint(
            fields=['meaning', 'language'], name='meaning_language_meaning'
        )
    ]
    ordering = ['meaning']
    indexes = [
        models.Index(fields=['meaning', 'language']),

```

Рис. 3.3 – Використання GraphQL для збору лексикографічної інформації

SPARQL endpoint – це основний механізм доступу до ресурсів, що організовані як Linked Data. Він використовується в таких глобальних мовних і енциклопедичних проєктах, як Wikidata чи DBpedia, а також у термінологічних банках на основі онтологічних моделей (IATE, EuroTermBank, OntoLex-Lemon). SPARQL дозволяє виконувати складні запити, наприклад: знайти всі лексеми певного семантичного класу, порівняти відповідники між мовами чи перевірити походження терміна (рис. 3.4). Це центральний компонент формування Всесвітньої лексикографічної мережі.

```

PREFIX wikibase: <http://wikiba.se/ontology#>
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX p: <http://www.wikidata.org/prop/>
PREFIX q: <http://www.wikidata.org/prop/qualifier/>
PREFIX v: <http://www.wikidata.org/prop/statement/>

SELECT DISTINCT ?gene ?start ?stop WHERE {
  ?gene wdt:P1057 wd:Q840604 .
  ?gene p:P644 ?statement .
  ?statement v:P644 ?start .
  ?statement q:P659 wd:Q20966585 .

  ?gene p:P645 ?statement2 .
  ?statement2 v:P645 ?stop .
  ?statement2 q:P659 wd:Q20966585 .

  FILTER (xsd:integer(?start) > 21000000 && xsd:integer(?start) < 30000000)
}

```

← RDF prefixes for all resources

← select WD items with subclass protein-coding gene

← and has a genomic start position (P644) for genome build GRCh38 on chromosome (P659) 9 (Q20966585)

← and has a genomic stop position (P645) for genome build GRCh38 on chromosome (P659) 9 (Q20966585)

← with coding region start between 21 Mb and 30 Mb

Рис. 3.4 – Приклад використання запити SPARQL в Wikidata

Webhooks працюють за принципом «подія → повідомлення». Їх використовують у динамічних словникових системах, де дані постійно оновлюються. Наприклад, редакторська система (Lexonomy, TLex Cloud) може надсилати сигнал зовнішнім сервісам, коли з'являється нова стаття або редагується наявна. Це дає змогу синхронізувати локальні та віддалені ресурси без ручного втручання.

Протокол **OAI-PMH** застосовується для масового збору метаданих і використовується переважно у наукових, архівних і корпусних проєктах. Він не надає повних словникових даних, але дозволяє отримувати опис ресурсів: авторство, структуру, формат, дату створення, мовні параметри. Його часто використовують цифрові бібліотеки та репозитарії, що хочуть інтегрувати словники або корпуси у спільні каталоги.

SOAP API зберігся у ряді старих академічних систем, де історично вже існували інтерфейси для обміну структурованими даними. Хоча сьогодні він майже не використовується в нових лексикографічних проєктах, деякі національні корпуси або термінологічні бази продовжують покладатися на нього через складність міграції.

Кожен тип API обслуговує свою ділянку лексикографічної екосистеми. REST відповідає за оперативний доступ до статей, GraphQL – за точне витягування структурованих даних, SPARQL – за міжресурсні зв'язки, Webhooks – за синхронізацію оновлень, OAI-PMH – за метадані, а SOAP – за легасі-системи. У сукупності вони й формують інфраструктуру, у якій словник перестає бути ізольованим об'єктом і перетворюється на частину глобальної пов'язаної мережі знань. Як приклад, розглянемо схему (рис. 3.5), що демонструє, як різні лексикографічні ресурси можуть взаємодіяти між собою, утворюючи цілісну мережу довідкових джерел, через які проходить користувач у процесі пошуку значення, перекладу чи контекстної інформації про слово.

1. Нормативний та спеціалізований словник → корпус. Пошук починається у словнику ВЕСУМ, який пропонує морфологічний аналіз, перелік словоформ і нормативні словникові дані. Від цього ресурсу логічно перейти до корпусу ГРАК, що дозволяє побачити, як ці словоформи функціонують у реальних текстах різних жанрів. У такий спосіб створюється зв'язок *лексикографічне джерело → корпусний контекст*, що допомагає перевірити частотність, синтаксичні моделі та живе вживання слова.

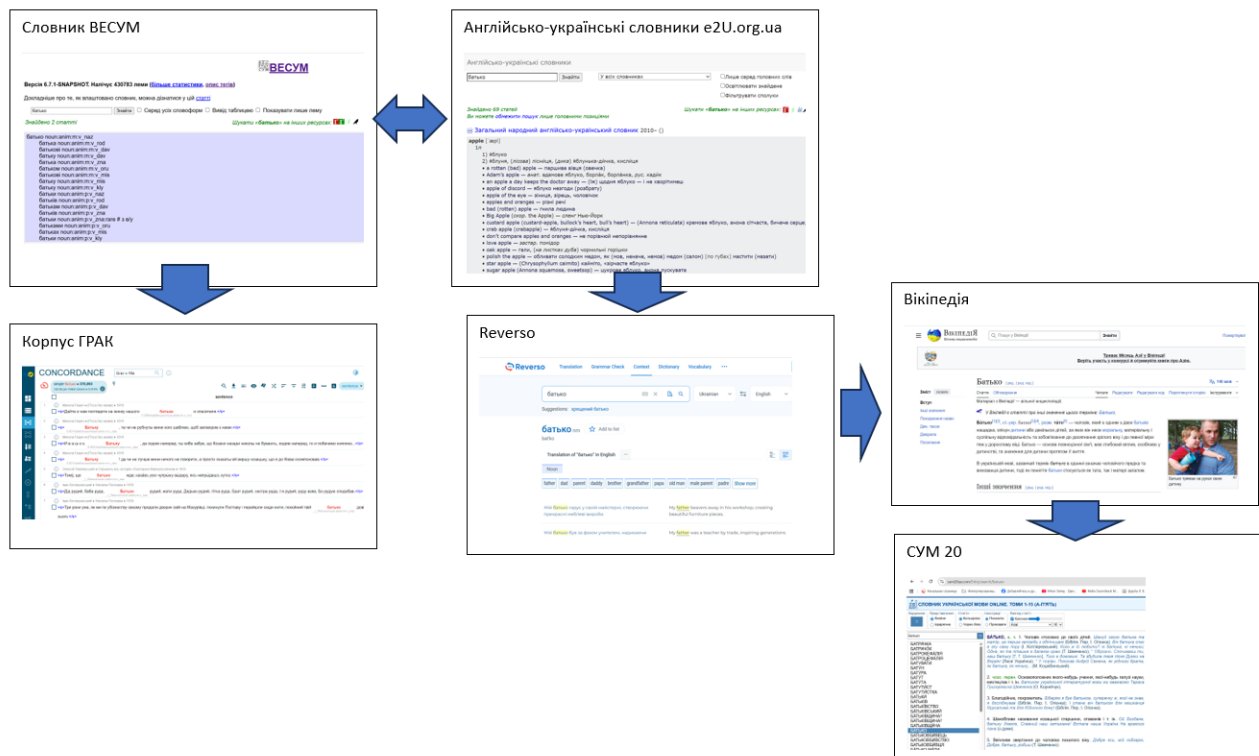


Рис. 3.5 – Приклад лексикографічної мережі

2. *Перекладні словники → автоматизовані перекладачі.* Другий блок об'єднує перекладні словники (e2u.org.ua) та автоматизовані онлайн-сервіси на кшталт Reverso. Тут реалізується двосторонній зв'язок:

- користувач може уточнити переклад у традиційному словнику, спираючись на багаторівневу структуру статті;
 - або перейти до перекладача, який надає варіанти перекладу, приклади вживання та додаткові значення, що ґрунтуються на машинній обробці корпусів.
- Цей рух показує, як класичні лексикографічні джерела працюють у парі з автоматизованими сучасними сервісами.

3. *Енциклопедичні джерела → тлумачні словники нового покоління.* Третій сегмент демонструє перехід від Вікіпедії до СУМ-20, що представляє сучасний стандарт української тлумачної лексикографії. Енциклопедичний опис дає користувачеві:

- загальносвітову інформацію про поняття,
- культурний, соціальний та історичний контекст,
- можливість швидкого ознайомлення з концептом.

Після цього користувач переходить до тлумачного словника, щоб отримати інформацію про:

- нормативне визначення,
- граматичні характеристики,
- стилістичні помітки,
- фразеологічні звороти та приклади.

4. *Взаємні зв'язки між ресурсами.* На схемі видно, що окремі блоки не ізольовані: наприклад, перекладні словники та корпуси можуть взаємодіяти між собою опосередковано через пошук прикладів; тлумачні й перекладні – через уточнення значення; енциклопедичні та корпусні – через перевірку вжитку назв або термінів. У підсумку формується лексикографічний маршрут користувача, який може виглядати так: 1) нормативні дані → 2) корпус → 3) переклад → 4) машинний аналіз → 5) енциклопедична інформація → 6) тлумачне уточнення або в будь-якому іншому порядку залежно від потреби. Хоч схема демонструє лише інтерфейси окремих сервісів, у реальності їхня взаємодія стає можливою завдяки використанню стандартизованих протоколів, форматів даних і API.

HTTP(S) та REST: основа зв'язків між ресурсами. Усі показані джерела – ВЕСУМ, e2u, Reverso, ГРАК, Вікіпедія, СУМ-20 – функціонують через HTTP(S). Це дає змогу: виконувати запити з браузера, посилатися на конкретні лема або сторінки, інтегрувати їх у сторонні сервіси. Схематично: клік по слову в одному ресурсі → перехід за URL → отримання даних із сервера іншого ресурсу.

API як спосіб автоматичного зв'язку. Деякі з цих ресурсів мають відкриті або напіввідкриті API, що уможливорює отримання словникових або корпусних даних не вручну, а програмно. Крім цього, API дають змогу включати ці дані у власні словники, мобільні застосунки, навчальні платформи, а також узгоджувати словникові статті між різними системами. Наприклад, Вікіпедія має повноцінний REST API, завдяки чому є можливість витягувати значення, категорії та метадані для словникових статей. Корпус ГРАК надає API для пошуку вживань слова в реальних текстах → словники або онлайн-курс можуть автоматично показувати приклади. Reverso використовує власний API для

перекладів і контекстів (обмежений) → цим користуються інші перекладні інструменти. У результаті лексикографічна мережа стає не просто набором сторінок, а системою взаємодіючих сервісів.

Формати даних як «спільна мова» між словниками. Наведені вище стандартизовані формати (TEI/XML, JSON-LD, RDF тощо), виконують роль єдиної структури для опису статей, граматичних ознак, значень, а також механізму інтероперабельності між різними словниковими базами. Завдяки цьому словник ВЕСУМ може експортувати дані у форматі TEI, інший словник (наприклад, перекладний або навчальний) імпортує ці дані без втрат структури, а корпус або семантична база (напр. DBpedia) автоматично зіставляє лєми, значення та концепти.

«Невидима мережа» під видимою схемою. На поданій схемі 3.5 користувач переходить між сервісами вручну, але технічно ці переходи стали можливими тому, що:

- ресурси публікуються у відкритій мережі (HTTP/S);
- використовують уніфіковані формати запитів (REST);
- частина сервісів надає API, завдяки чому їх можна під'єднувати один до одного;
- формати репрезентації даних уніфікуються, що дає змогу автоматично зіставляти дані.

Усе це формує фундамент всесвітньої лексикографічної мережі, де словник перестає бути ізольованим ресурсом і стає вузлом у мережі даних. Нижче у таблиці 3.1 наведено типи даних, що передає кожен API в межах наведеної вище лексикографічної мережі.

Таблиця 3.1 – API та типи даних для лексикографічної мережі

Ресурс	Тип API / Формат	Які дані передає	Рівень доступу
ГРАК	REST / JSON	Контексти (KWIC), грамтеги, лєми, метадані	Відкритий (ключ)
Вікіпедія (MediaWiki)	REST / JSON, XML	Текст статей, структура, категорії, міжмовні посилання	Повністю відкритий
Reverso	REST / JSON	Переклади, приклади, граматична інформація	Закритий (внутрішній)

e2u.org.ua	Неофіц. JSON-endpoints	Переклади, значення, словникові списки	Частково відкритий
СУМ-20	XML (TEI), частк. REST	Словникові статті, розмітка, приклади	Частково відкритий
БЕСУМ	Без API / HTML, TXT	Леми, словоформи, морфологія	Немає API
UD / UDPipe	REST / JSON, CoNLL-U	Леми, теги, синтаксичні залежності	Відкритий
Apertium	REST / JSON, XML	Морфаналіз, генерація, переклад	Відкритий

Узагальнюючи, сучасні протоколи та API перетворюють словники з ізольованих інформаційних «островів» на взаємопов'язану екосистему, де кожен ресурс може доповнювати інший у реальному часі. Корпуси постачають контексти, машинні словники – морфологію та переклад, енциклопедії – довідкові знання, а універсальні формати на кшталт JSON, XML чи TEI забезпечують технічну взаємодію між ними. Саме завдяки цим механізмам лексикографічна мережа стає не статичною колекцією словників, а динамічним середовищем, здатним оперативно комбінувати й перерозподіляти дані, створюючи якісно новий рівень доступності та гнучкості мовних ресурсів у цифрову епоху.

Питання для самоконтролю

1. Що таке лексикографічна мережа і чому саме Інтернет став оптимальним середовищем для її формування?
2. У чому полягає різниця між локальними та зв'язаними словниковими ресурсами? Наведіть приклади.
3. Які властивості Інтернету забезпечують можливість інтеграції словників між собою?
4. Чому стандартизовані формати репрезентації даних є ключовими для побудови лексикографічної мережі?
5. Які формати обміну даними найчастіше використовуються у сучасних словниках і яку проблему кожен із них вирішує?
6. Чим відрізняється XML TEI від JSON-LD у контексті лексикографії?
7. Яку роль відіграють такі технології, як REST, GraphQL, SPARQL у структурі словникових API?

8. Які типи даних зазвичай передають словникові API (на прикладі схеми з DWDS → ELEXIS → KDE → Wikidata)?

9. Чому навіть великі проєкти (DWDS, Wikidata, Apertium) переходять до відкритих API?

10. Як інтеграція словників через API впливає на дослідницькі завдання корпусної та комп'ютерної лінгвістики?

Завдання для самоперевірки

1. Порівняйте три формати: **XML TEI, JSON-LD та RDF Turtle**. Вкажіть:

- у яких випадках кожен формат оптимальний,
- яку проблему він вирішує,
- яке обмеження має.

2. Поясніть, як працювала б ваша власна міні-лексикографічна система, якби ви інтегрували:

- локальний словник у форматі XML TEI,
- API Wikidata,
- SPARQL-запити,
- JSON-словник перекладів.

Опишіть:

- які дані звідки надходитимуть,
- який формат використовуватиметься для обміну,
- що буде «центральною вузлом» системи.

3. Є чотири твердження. Визначте, до якої категорії вони належать:

- (А) формат репрезентації даних
- (Б) мережевий протокол
- (В) тип API
- (Г) функція лексикографічної мережі

Твердження:

- (1) «Дозволяє формувати запити до графових баз знань».
- (2) «Використовується для представлення онтологічних зв'язків між об'єктами».

(3) «Повертає лише ті поля, які запитує клієнт».

(4) «Забезпечує структуроване подання словникових статей і глибоку текстову анотацію».

4. Уявіть, що ви маєте словник у форматі **XML TEI** і хочете інтегрувати його з open-source перекладним словником, що доступний через **REST API** у форматі **JSON**. Потрібно описати кроками (5–7 пунктів):

- що треба конвертувати,
- які протоколи та формати знадобляться,
- як буде виглядати кінцева структура об'єднаних даних.

5. Подивіться на таку фразу з підручника: «Сучасні словники дедалі частіше стають не окремими продуктами, а частинами комплексної мережі мовних ресурсів». Переоформіть її так, щоб стало зрозуміло, *чому* це відбувається — використавши терміни:

- інтероперабельність,
- відкриті формати,
- API,
- зв'язні дані.

3.2. Локальні та мережеві лексикографічні ресурси

У сучасній лінгвістиці особливого значення набуває питання способу організації лексикографічних ресурсів. Співіснують дві моделі – **локальна** та **мережева**, і різниця між ними визначає можливості подальшої інтеграції, автоматичної обробки та взаємодії з іншими мовними даними. У цьому підпункті розглянемо ключові характеристики обох підходів та окреслимо їхню роль у сучасній інфраструктурі цифрової лексикографії. Нижче наведено детальний огляд характеристик цих моделей.

Локальні ресурси – це словники, які існують як окремі, самодостатні об'єкти. Вони найчастіше розробляються у форматах, орієнтованих на людину-читача, а не на машиночитність. Такі словники можуть мати високу якість та значну лексикографічну цінність, проте їхні можливості в інтеграції з іншими мовними ресурсами є суттєво обмеженими. Такі ресурси відрізняються від

мережевих форматом реалізації, відсутністю механізмів обміну даними, обмеженою або відсутньою інтеграцією з корпусними платформами та невикористанням стандартів відкритих лінгвістичних даних.

Щодо формату реалізації, лексикографічні дані здебільшого подано у вигляді файлів PDF, HTML, DOCX або інших статичних форматів. У ряді випадків використовуються власні формати словникових редакторів чи програм (наприклад, StarDict, Lingvo DSL, Turbolex), що ще більше ускладнює доступність для зовнішніх систем. Локальні словники не передбачають API, потоків оновлень або структурованих інтерфейсів доступу. Дані можна отримати лише шляхом ручного копіювання або експорту з конкретного програмного продукту (якщо він взагалі існує). У більшості випадків локальний словник не має функції автоматичного зв'язку з корпусними збереженнями, не забезпечує зіставлення словникових форм із корпусними контекстами. Локальні ресурси, як правило, не підтримують RDF, TEI-XML, JSON-LD або інших форматів, що дозволяють включати словникові дані в ширші лінгвістичні мережі.

Як правило, до локальних належать традиційні офлайнові словники (Turbolex, Інтегрована лексикографічна система «Словники України» на CD), старі електронні видання СУМ чи OED, внутрішні тематичні глосарії окремих установ, а також словники, розміщені на сайтах без відкритих структур (рис. 3.6).

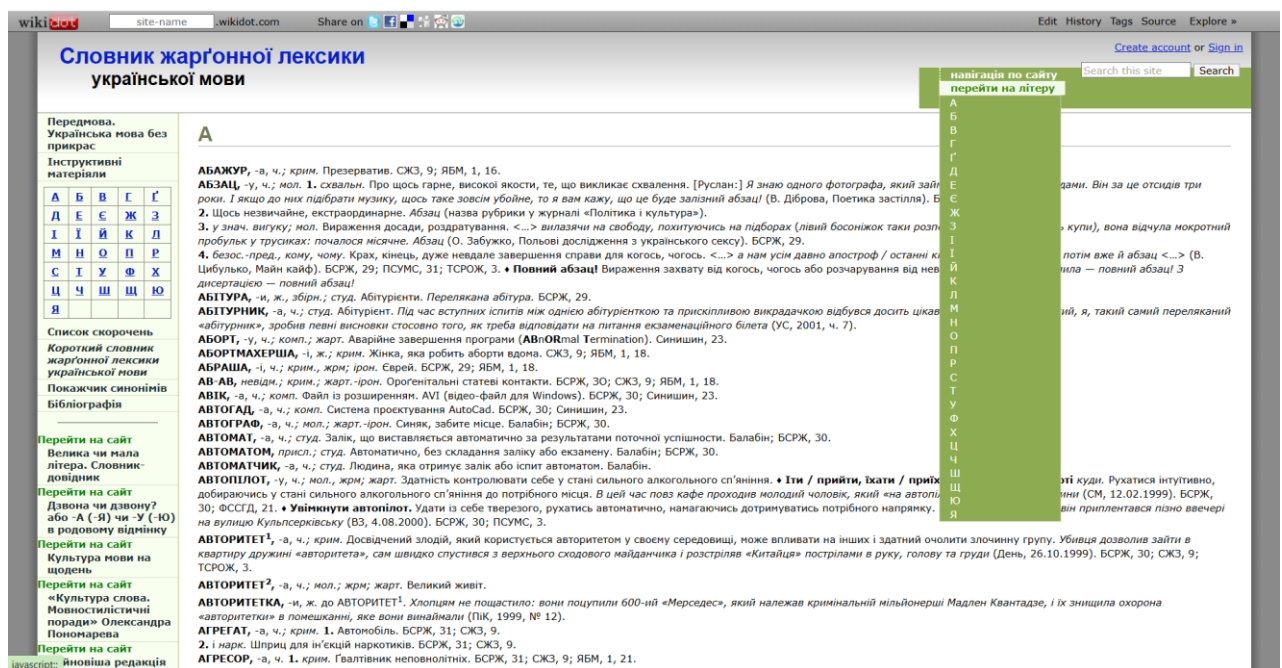


Рис. 3.6 – Словник жаргонної лексики на ресурсі wikidot

Мережеві (зв'язані) лексикографічні ресурси – ресурси, що орієнтовані на інтегрованість та відкритість. Вони створюються з урахуванням можливості повторного використання, автоматичного доступу та включення до спільної інфраструктури мовних даних. Вони передбачають активну взаємодію зі словниковими базами, корпусами, морфологічними та синтаксичними інструментами.

Використання відкритих структурованих форматів. У мережевих лексикографічних ресурсах дані зазвичай подаються у формальних моделях, таких як RDF, XML TEI, JSON-LD, LMF та інші стандартизовані нотації. Їхня головна перевага полягає в тому, що всі структурні елементи словникових статей є формально описаними та машиночитаними. Завдяки цьому забезпечуються уніфікація, можливість автоматичної обробки, сумісність із корпусами та інструментами NLP, а також просте включення даних у ширші лінгвістичні екосистеми. Це – одна з ключових ознак, яка чітко відмежовує мережеві лексикографічні моделі від локальних, де дані залишаються закритими, слабо структурованими або непристосованими до інтеграції.

Підтримка API та веб-сервісів. До словникових даних можна отримувати доступ через REST API, SPARQL-endpoints, JSON-інтерфейси тощо. Це дозволяє програмно формувати запити до словника, автоматично оновлювати дані та включати їх у зовнішні застосунки чи корпуси. Огляд мережевих протоколів було наведено у підрозділі 3.1. Завдяки структурованим і відкритим інтерфейсам API істотно розширюють можливості розробників, дослідників і користувачів щодо інтеграції й повторного використання лексичних ресурсів. У результаті мережеві словники перестають бути самодостатніми «островами» інформації й стають активними вузлами глобальної лексикографічної інфраструктури.

Можливість взаємодії з корпусами та інструментами аналізу є однією з ключових переваг мережевих лексикографічних ресурсів. На відміну від локальних словників, що існують окремо від мовних даних, мережеві системи інтегруються з корпусами (наприклад, Sketch Engine), морфологічними аналізаторами, тегерами та іншими інструментами. Це дає змогу зіставляти

словникові статті з реальними вживаннями, будувати частотні моделі, перевіряти гіпотези щодо сполучуваності та виявляти закономірності мовної поведінки.

Одним із найпоказовіших інструментів такої інтеграції є SkELL – сервіс, створений на базі корпусної платформи Sketch Engine спеціально для студентів і викладачів (рис. 3.7). Він надає інформацію про реальне вживання слів і словосполучень у сучасних мовах, працюючи з англійською, німецькою, італійською, чеською естонською та іншими мовами. SkELL доступний онлайн і пропонує три режими:

- (1) конкорданс (concordance) – приклади вживання слова в реальних текстах;
- (2) колокації (collocations) – типові словосполучення, характерні для цього слова;
- (3) синоніми (similar words), обчислені автоматично на основі статистичних моделей.

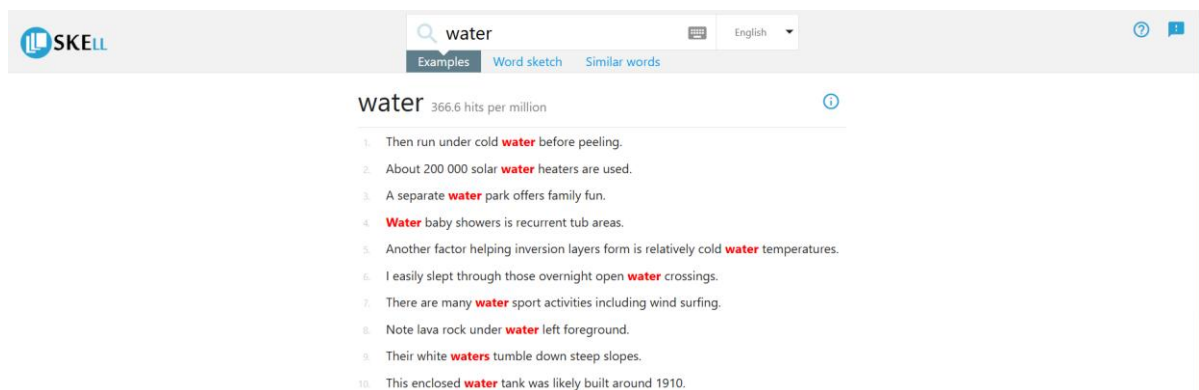


Рис. 3.7 – Інтерфейс Skell на базі Sketch Engine

Важливо, що всі приклади, колокації та синоніми генеруються автоматично з багатомільйонних корпусів без ручної розмітки. Це показує, наскільки потужним може бути поєднання мережевого словника й корпусної інфраструктури, коли лексикографія переходить від статичного опису до динамічного моделювання мовних даних.

Ще одним прикладом масштабної інтеграції лексичних ресурсів у форматі зв'язаних даних є проєкт VerbaAlpina (Університет Людвіга-Максиміліана, Мюнхен), показаний на рисунку 3.8. Цей ресурс об'єднує матеріали з різних словників, атласів, діалектологічних описів і польових досліджень, присвячених мовам та говіркам альпійського регіону. Дані подаються у стандартизованому,

семантично зв'язаному вигляді (Linked Open Data), що дозволяє порівнювати лексеми між різними мовами та діалектними ареалами, відстежувати історичні відповідники, будувати ареальні карти та виконувати комплексний регіональний аналіз.

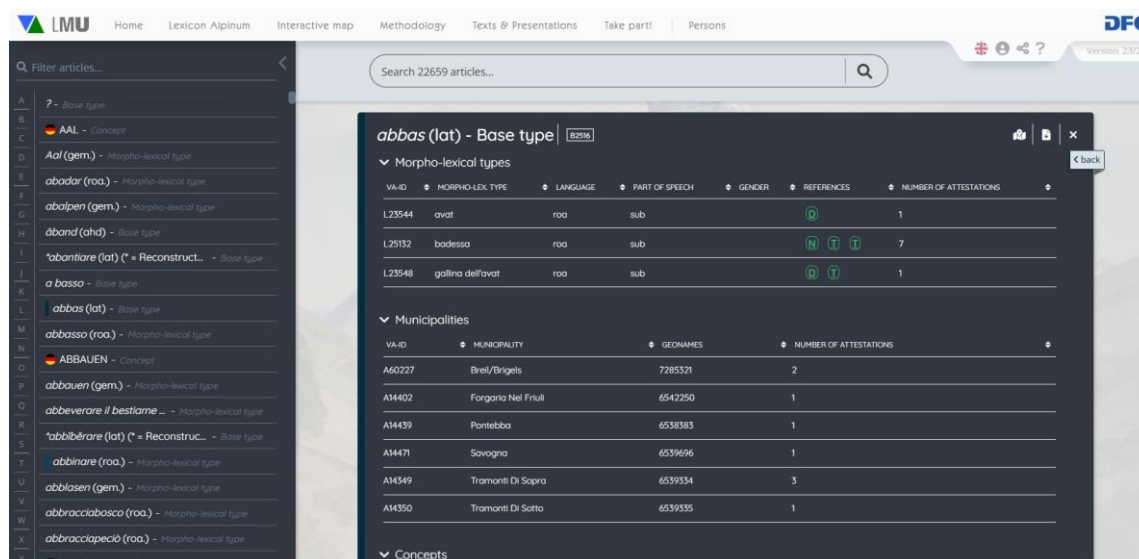


Рис. 3.8 – Інтерфейс ресурсу VerbaAlpina

Розподілений характер ресурсів, що інтегруються в межах VerbaAlpina, демонструє потенціал мережевих словників у наукових дослідженнях: цей проєкт не лише агрегує дані з різних країн і наукових традицій, а й забезпечує доступ до них через відкриті інтерфейси, підтримує реплікацію, збагачення й подальше використання в інших лінгвістичних системах. Таким чином, VerbaAlpina є показовою моделлю того, як сучасні технології дозволяють переосмислити регіональні та діалектні словники як частину великої екосистеми лексичних зв'язаних даних. Проєкт VerbaAlpina демонструє кілька принципово важливих технологічних і методологічних рішень, які роблять його зразком для розвитку мережевих словників та лексичних платформ:

1. Використання онтологій та RDF-графів. Лексеми, топоніми, фонетичні варіанти, семантичні категорії й ареальні характеристики описуються у формі зв'язних RDF-триплетів. Це дозволяє не просто зберігати дані, а формувати семантично насичену мережу знань, у якій кожен елемент може бути знайдений, зіставлений та інтерпретований автоматично.

2. Геокодування і ареальний аналіз. Кожна одиниця прив'язана до конкретної географічної точки або ареалу. У результаті лексикографія стає просторовою

дисципліною: дослідник може одразу бачити, як слово поширене регіоном, які є його географічні варіанти, і як ці варіанти співвідносяться з політичними кордонами, історичними областями чи міграціями населення.

3. *Мультимодальність даних.* VerbaAlpina інтегрує історичні словники, діалектні атласи, польові записи, фонетичні транскрипції, варіанти написання й ареальні мапи. Усе це поєднується в одному інтерфейсі, що робить ресурс одночасно словником, корпусом і геолінгвістичним інструментом.

4. *Відкриті стандарти та повторне використання.* Дані доступні через SPARQL-endpoint та інші інтерфейси, що дає змогу:

- вбудовувати їх у сторонні системи,
- автоматично зіставляти з іншими LOD-ресурсами,
- розширювати матеріал новими джерелами.

Це важливо, бо перетворює регіональний словник на частину глобальної мережі лексичних знань. Проєкт поєднує лексикографію, діалектологію, етнолінгвістику, географію, інформатику, а подекуди й історію та антропологію. Це робить VerbaAlpina моделлю того, як майбутні лексикографічні ресурси виходитимуть за межі вузької словникової парадигми та перетворюватимуться на дослідницькі екосистеми.

Частина ширших лінгвістичних екосистем. У сучасному лексикографічному середовищі дані дедалі рідше існують у межах одного ізольованого ресурсу. Натомість вони стають частиною ширших мереж знань, де різні словники, корпуси та онтології взаємодіють між собою. Дані можуть одночасно бути представлені в кількох пов'язаних проєктах і автоматично «збагачуватися» інформацією з інших мовних баз. Одним із найвідоміших прикладів такої взаємодії є багаторівнева інтеграція Wiktionary з DBpedia, Wikidata та WordNet. Ці ресурси поєднуються завдяки застосуванню Linked Data та спільних форматів обміну, зокрема моделі **lemon**, що дозволяє описувати лексичні дані у формі семантичних графів. Унаслідок цього окремі проєкти можуть:

- автоматично витягувати дані з Вікісловника;
- перетворювати їх у спільні структури;

- зіставляти поняття й сутності з WordNet та Вікіданими;
- формувати розширені багатомовні лексичні бази.

Хоча багато етапів можуть бути автоматизовані, для досягнення високої точності зазвичай потрібна ручна валідація. Така багаторівнева інтеграція передбачає реалізацію таких методів і проєктів, мова про які піде нижче.

1. *Формат lemon*. Формальна модель для репрезентації лексичних даних у форматі Linked Data. Забезпечує сумісність між різними лексикографічними проєктами, зокрема WordNet та Wiktionary.

2. *NULEX*. Інтеграційний проєкт, що поєднує WordNet, VerbNet та Wiktionary. З WordNet беруться базові семантичні структури, з VerbNet – рамки підкатегоризації дієслів, а з Wiktionary – граматична інформація, зокрема часові форми. Це створює багатший узагальнений ресурс.

3. *OntoWiktionary*. Проєкт зі створення онтології на основі даних Вікісловника. Спочатку збирається структурована інформація, після чого вона онтологізується та інтегрується в хмару пов'язаних даних.

4. *RDF-екстракція через DBpedia*. DBpedia надає механізми для вилучення частин мови, визначень, синонімів та перекладів із Вікісловника й представлення їх у форматі RDF. Це робить ресурс придатним для подальшого автоматичного аналізу.

5. *Поєднання WordNet із Вікіданими*. Зіставлення синсетів WordNet з елементами Wikidata здійснюється за допомогою NLP-методів та ручної перевірки. Це забезпечує коректне узгодження понять між двома великими семантичними системами.

6. *Wikidata як центральний хаб*. Wikidata виступає центральною платформою, що акумулює лексикографічні дані з Вікісловника та забезпечує автоматичні багатомовні посилання, перетворюючи словник на частину глобальної бази фактів.

При цьому, слід зауважити, що технічна реалізація зустрічає певні виклики. Насамперед, має місце несумісність і варіативність даних: оскільки Wiktionary створюється спільнотою, записи можуть сильно відрізнятися в структурі,

деталізації та розмітці. Слід також зважати на так звані «острови даних» – до впровадження форматів, таких як **lemon**, багато ресурсів існували у закритих або нестандартних форматах, що ускладнювало інтеграцію та повторне використання.

Однак така інтеграція має низку переваг. По-перше, поєднання Вікісловника з WordNet дозволяє отримати набагато ширший, багатомовний і семантично структурований лексикон. По-друге, Інтеграція відкриває доступ до інформації, якої немає в окремих ресурсах: синтаксичні рамки VerbNet, часові форми з Wiktionary, семантичні групи WordNet тощо. І, по-третє, зв'язування синонімів ImageNet/WordNet з Wikidata покращує моделі глибокого навчання, зокрема у задачах семантичного пошуку, комп'ютерного зору та багатомовної обробки тексту. Порівняльна характеристика різних типів інтеграції подана у таблиці 3.2.

Таблиця 3.2 – Типи інтеграції між ресурсами

Ресурс / зв'язка	Тип інтеграції	Що дає (результат / користь)
Wiktionary → DBpedia	Автоматичний RDF-витяг структурованих даних	Частини мови, визначення, переклади, синоніми у форматі, придатному для семантичного аналізу
Wiktionary → Wikidata	Мапування лексичних статей на елементи Wikidata	Центральне сховище лексем, багатомовні зв'язки, уніфікація термінів
Wiktionary → WordNet	Зіставлення значень (sense alignment)	Узгодження лексичних одиниць зі семантичними синсетами; розширення покриття WordNet
WordNet → Wikidata	Зв'язування синсетів з об'єктами бази знань	Збагачення WordNet фактичними даними; можливість зв'язувати лексеми з реальними сутностями
WordNet → VerbNet (через NULEX)	Інтеграція семантики WordNet із синтаксичними рамками VerbNet	Повніші лексико-синтаксичні моделі дієслів; краща підтримка NLP-додатків
Wiktionary → VerbNet (через NULEX)	Автоматичне доповнення граматичними характеристиками	Часові форми, додаткові морфологічні дані, яких немає у VerbNet
Wiktionary → OntoWiktionary	Побудова онтології на основі словникових статей	Можливість інтеграції у Linked Data Cloud, чіткіша категоризація лексем
DBpedia + Wikidata + WordNet (комплексно)	Семантична федерація ресурсів	Потужні багатомовні графи знань для лексикографії, NLP, пошукових систем
Wikidata як хаб	Централізоване зберігання лексичних даних	Уніфікація, стабільні ідентифікатори, зв'язність між усіма проектами

Ще одним із практичних і дуже показових прикладів мережевої взаємодії є інтеграція Lexonomy зі службами Sketch Engine, що спирається на API-орєнтовану взаємодію, як про це було зазначено у підрозділі 3.1. У середовищі Lexonomy працює інструмент OneClick Dictionary, який дозволяє автоматично витягувати з корпусу найважливіші дані для укладання словникової статті – приклади вживання, колокації, частотні профілі, синонімічні ряди, типові граматичні конструкції тощо. Взаємодія відбувається через Sketch Engine API, тому лексикограф може працювати в одному інтерфейсі, але при цьому тягнути матеріал з будь-якого доступного корпусу. Це демонструє, як мережеві моделі дозволяють не просто поєднувати ресурси, а створювати динамічні словникові статті, які напряду опираються на корпусні дані і можуть автоматично оновлюватися.

У подібний спосіб Open Multilingual WordNet, ресурси Universal Dependencies, Lexicala та інші онлайн-платформи забезпечують стандартизований доступ до лексичних і граматичних даних, дозволяють зіставляти інформацію між мовами та створювати багатомовні зв'язки. Усе це розширює лексикографічні можливості, роблячи словники частиною великої мережі пов'язаних мовних ресурсів.

Отже, локальні словники забезпечують автономність, стабільність та простоту поширення, проте вони обмежені у функціональності й не здатні ефективно взаємодіяти з іншими даними. Натомість мережеві ресурси характеризуються високим ступенем відкритості, підтримують стандарти обміну, надають інтерфейси програмного доступу та функціонують як частина складнішої цифрової мовної інфраструктури. Саме завдяки мережевим моделям словники перестають бути статичними зібраннями значень і перетворюються на активні елементи великої системи мовних даних, у якій лексичні, синтаксичні, енциклопедичні та корпусні відомості взаємопов'язуються між собою. Як підсумок, у таблиці 3.3. наведено порівняльну характеристику локальних та мережевих ресурсів.

Таблиця 3.3 – Порівняльна характеристика ресурсів

Критерій	Локальні ресурси	Мережеві (зв'язані) ресурси
Формат даних	Статичні файли (PDF, DOCX, HTML), власні формати	Стандартизовані структури (XML TEI, RDF, JSON-LD, LMF)
Можливість обміну даними	Відсутня або мінімальна; ручний експорт/копіювання	API, SPARQL, REST-запити, автоматичні оновлення
Інтеграція з корпусами	Обмежена або відсутня; немає автоматичного зв'язку	Повна інтеграція з корпусами, тегерами, аналізаторами
Можливість повторного використання	Низька; залежить від формату	Висока; дані машиночитні та придатні для об'єднання
Зв'язність з іншими ресурсами	Немає механізмів зв'язування	Підтримка linked data, міжсловникові та міжмовні зв'язки
Оновлення	Ручні, періодичні, несинхронізовані	Автоматизовані, синхронізовані з іншими сервісами
Сфера застосування	Переважно офлайнове використання, індивідуальний доступ	Веб-платформи, мовні API, дослідницькі інфраструктури
Типові приклади	Офлайн СУМ, локальні глосарії, PDF-словники	Wiktionary+Wikidata, Lexonomy, Sketch Engine, OMW, UD-ресурси

Порівняння локальних і мережевих лексикографічних ресурсів показує глибоку трансформацію сучасної лексикографії: від ізольованих статичних файлів до динамічних, взаємопов'язаних систем даних. Локальні словники працюють як автономні довідники – зручні для читання, але малоприсадибні для автоматичного аналізу чи інтеграції з іншими інструментами. Натомість мережеві ресурси створюють розширювану екосистему, у якій словники, корпуси, морфологічні аналізатори та онтології обмінюються даними в режимі реального часу. Завдяки стандартизованим форматам (XML TEI, RDF) та API такі ресурси стають частиною глобальної лексикографічної інфраструктури, що забезпечує масштабованість, інтероперабельність і багатомовну сумісність.

Питання для самоконтролю

1. Що відрізняє мережеві словники від традиційних локальних словників?
2. Які формати структурованих даних найчастіше використовуються у мережевих лексичних ресурсах?
3. Чому RDF-структури важливі для зв'язування словників між собою?
4. У чому полягає перевага інтеграції словника з корпусами?
5. Які можливості дає інструмент SkELL? Назвіть три режими роботи.

6. Як Sketch Engine взаємодіє зі словниковими платформами (наприклад, Lexonomy)?

7. Що таке OneClick Dictionary і яку роль він відіграє у створенні словникових статей?

8. Які види лексичних зв'язків найчастіше використовуються в мережевих ресурсах (семантичні, словотвірні, колокаційні тощо)?

9. Чому мережеві словники особливо важливі для менших або регіональних мовних спільнот?

10. Що таке Linked Open Data і як ця концепція впливає на сучасну лексикографію?

Завдання для самоперевірки

1. Складіть коротку таблицю (3–4 рядки), де порівняєте:

- локальний словник,
- онлайн-словник без зовнішніх інтеграцій,
- мережевий словник із підтримкою LOD.

Орієнтуйтеся на критерії: структура, оновлення, інтеграції, юзабіліті.

2. Візьміть будь-яке слово з вашої мовної спеціалізації та:

- знайдіть його у корпусі (або в SkELL, якщо доступний),
- визначте 3 типові колокації,
- зробіть короткий висновок: як ці дані могли би змінити словникову статтю.

3. Уявіть, що вам потрібно створити словниковий запис у Lexonomy.

Опишіть кроки, як за допомогою OneClick Dictionary ви отримаєте:

- 5 прикладів уживання,
- 5 найчастотніших колокацій,
- найближчий синонімічний ряд.

4. Поясніть, чому один і той самий інструмент (наприклад, SkELL) інколи пропонує приклади, які не зовсім відповідають очікуваному значенню слова. Що це показує про роботу корпусних алгоритмів?

5. Оберіть один мережевий ресурс з лекції (Lexiconista, Verba Alpina, Open Multilingual WordNet тощо) і коротко опишіть:

- який тип даних він надає,
- які формати використовує (RDF, TEI, JSON-LD...),
- як ці дані можуть бути інтегровані в інший лексичний ресурс.

6. Створіть схему «мережі зв'язків» для одного слова:

- корпусні дані → колокації,
- семантична мережа → синоніми/антоніми,
- морфологічний аналіз → парадигма,
- енциклопедичні ресурси → культурна/регістрова інформація.

Зробіть висновок, як це слово «живе» у мовній системі.

3.3. Інтеграція онлайн-ресурсів у єдиний глобальний простір

Розвиток Інтернету докорінно змінив лексикографію, перетворивши її з дисципліни, що раніше спиралася на статичні друковані видання, на динамічну мережеву систему. Сучасний словник уже не існує ізольовано. Він функціонує у тісній взаємодії з іншими мовними ресурсами, семантичними мережами, корпусами текстів, інструментами автоматичного аналізу та освітніми платформами. Ця інтеграція сприяє появі глобального мовного простору, у межах якого лексичні дані постійно оновлюються й збагачуються іншими джерелами.

Одним із найважливіших напрямів розвитку є взаємозв'язок між окремими словниковими ресурсами. Сьогодні багато онлайн-словників обмінюються даними завдяки використанню єдиних технічних стандартів, таких як RDF, JSON-LD або LMF, а також завдяки застосуванню відкритих API. Окремі словникові бази можуть бути пов'язані через централізовані сховища знань на кшталт Wikidata, де кожна лексична одиниця має стабільний ідентифікатор, що дозволяє легко зіставляти дані різних проєктів. Семантичні онтології, зокрема WordNet, BabelNet та ConceptNet, створюють спільний концептуальний простір, у якому значення, переклади та довідкова інформація різних мов комбінуються в єдину мережу.

Не менш важливим чинником є поєднання словників із корпусами текстів. Завдяки такій інтеграції словник перестає бути лише набором дефініцій і перетворюється на інструмент доступу до реального мовного вживання.

Корпусні системи забезпечують автоматичний пошук колокацій, частотних профілів, прикладів уживання та контекстних інтерпретацій слів. Фактично словникова стаття починає функціонувати як інтерфейс між змістовим описом та мовною поведінкою слова у природних текстах. Це дає змогу формувати опис значень на основі реальних емпіричних даних, а не лише інтуїтивних суджень укладача.

Інтеграція з інструментами обробки природної мови (NLP) є наступним кроком у розвитку сучасних словникових ресурсів. Морфологічні аналізатори, синтаксичні парсери, моделі векторного подання слів та системи машинного перекладу використовують словникові дані як основу власної роботи, одночасно збагачуючи їх результатами обробки текстів. Таким чином, словник стає елементом складної лінгвістичної інфраструктури, де традиційні лексикографічні описи поєднуються з алгоритмами автоматичного аналізу.

Доволі важливу роль відіграє також взаємодія словників з освітніми та тренувальними платформами. Цифрові словники інтегруються зі системами онлайн-перекладу, мовними курсами, гейміфікованими застосунками та сервісами для тренування словникового запасу. У такому форматі словник не лише надає інформацію, а й стає частиною навчального процесу, автоматично підлаштовуючись під потреби користувача та створюючи індивідуальні маршрути роботи зі словом.

Ще один важливий напрям – це інтеграція словникових ресурсів з інструментами редагування текстів. Системи на кшталт Grammarly або LanguageTool поєднують словникові дані з корпусними показниками та машинними моделями, забезпечуючи комплексний аналіз правопису, граматики й стилістики. Завдяки цьому словник, який раніше був окремим довідником, тепер працює «за лаштунками» більшості редакторів, тим самим допомагаючи користувачеві у реальному часі.

Усі ці процеси сприяють формуванню динамічних словникових середовищ. На відміну від традиційних видань, що оновлювалися раз на кілька років, сучасні онлайн-словники здатні безперервно змінюватися: отримувати нові дані

з корпусів, збагачуватися інформацією з інших словників, узгоджувати семантичні моделі з NLP-інструментами та підтримувати взаємодію з освітніми ресурсами. Це означає, що лексикографія перестає бути закритою системою і перетворюється на відкриту, міждисциплінарну галузь, у якій кожен ресурс є частиною великої мережі.

У цифрову епоху словник уже не може існувати «сам по собі». Він є вузлом глобальної семантичної інфраструктури, інтерфейсом між мовою і технологіями, інструментом наукового аналізу та освітнім компонентом. Саме Інтернет створив умови, за яких лексикографія стала гнучкою, мережевою і взаємодоповнюваною дисципліною, де кордони між окремими ресурсами майже зникають, а дані вільно циркулюють між різними системами.

Сучасна лексикографія дедалі частіше спирається не на окремі ізольовані інструменти, а на зв'язки між платформами, які дозволяють автоматизувати рутинні операції, синхронізувати дані та робити словники «живими». Один із прикладів – інтеграція редакторів словникових статей із корпусами або базами знань. У низці університетських проєктів та відкритих словників застосовується пов'язка Lexonomy (як редактор словникових статей) та SketchEngine (як корпусний інструмент). У розділі налаштувань «Sketch Engine Connection» редактор:

(1) обирає корпус в полі «Corpus Name», наявний у Sketch Engine, а також підключає до корпусу відповідні елементи словникової статті: ілюстративні приклади, колокації, синоніми та дефініції (рис. 3.9);

HOME > DICTIONARIES > CLONING OF ENGLISH DICTIONARY OF CRIMINAL TERMS > CONFIGURE > SKETCH ENGINE CONNECTION

SKETCH ENGINE CONNECTION

CORPUS NAME
British National Corpus (BNC)

Select a Sketch Engine corpus from the list of corpora available to you.
Currently selected corpus: preloaded/bnc2_t131, show [detailed corpus info](#)

CONCORDANCE QUERY

The CQL query that will be used to obtain concordance from Sketch Engine. You can use placeholders for elements in the form of "% (element)", e.g. "[lemma~"% (headword)"]". If left empty the 'simple' query type will be used as configured for the respective corpus. Please note that you cannot use CQL syntax with default attribute because it is not specified.

SAMPLE SIZE
0

Whether to apply automatic sampling of the concordance. Any non-zero value means to automatically create a random sample of that size.

ADDITIONAL SEARCH ELEMENTS

You can select any textual elements here whose content you would like to search for in Sketch Engine. A menu will be displayed next to all these elements like for the root entry element.

EXAMPLES

EXAMPLE CONTAINER
example

Select the element which should wrap each individual example. When you pull example sentences automatically from a corpus, LEXONOMY will insert one of these elements for each example sentence.

COLLOCATIONS

COLLOCATION CONTAINER
(not set)

Select the element which should wrap each collocation (the collocate plus any other data). When you pull collocations automatically from a corpus, LEXONOMY will insert one of these elements for each collocation.

THESAURUS ITEMS

THESAURUS ITEM CONTAINER
label

Рис. 3.9 – Підключення словника до корпусу в LEXONOMY

(2) переносить потрібні дані у статтю, не перемикаючись між вікнами, при цьому шукати корпусні дані може безпосередньо у Lexonomu через меню Sketch Engine (рис. 3.10):

- Find examples: знайти ілюстративні приклади для заголовкового слова в обраному корпусі;
- Find thesaurus items: знайти синоніми або споріднені терміни в текстовому корпусі;
- Find collocations: знайти колокації (словосполучення, фраземи) для заголовкового слова в обраному корпусі.

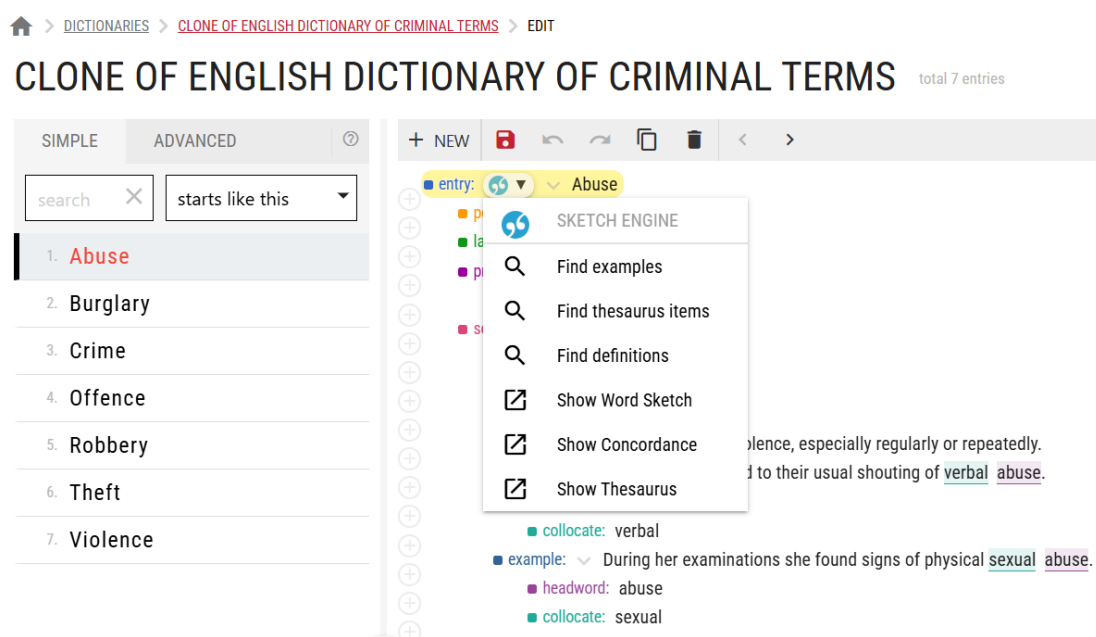


Рис. 3.10 – Пошук інформації з корпусу в Lexonomu

(3) може задавати контекстне оточення для заголовкового слова під час роботи з ілюстративними прикладами, використовуючи мову CQL.

Завдяки такій інтеграції темп роботи над словником зростає у 2–3 рази, виключається ризик втратити відповідність прикладів і даних, і, крім цього, лексикограф бачить живу мовну поведінку слова прямо під час укладання. Інший приклад – інтеграція двох великих відкритих проєктів: *Wiktionary* та *Wikidata*. Взаємодія між ними здійснюється за такою схемою:

- кожна лексема у Wikidata має власний уніфікований ідентифікатор (наприклад, L12345);

– один запис у Wikidata може забезпечувати одразу кілька мовних випусків Вікісловника;

– зміни в описі лексеми чи її граматичних властивостей у Wikidata автоматично відображаються в мовних розділах Вікісловника, де налаштовані шаблони імпорту.

У результаті інтеграції граматичні парадигми, списки значень, етимології перестають дублюватися вручну. Також знижуються помилки та розбіжності між мовними розділами, і, найголовніше, спільноти можуть працювати з єдиним, централізованим набором лексичних даних. Таким чином, інтеграція двох ресурсів дає змогу підтримувати багатомовні словники без гігантських команд редакторів, швидко масштабувати контент на нові мови та імпортувати дані в інші проєкти: навчальні застосунки, корпуси, чат-боти.

Ще одним показовим прикладом інтеграції лінгвістичних та інформаційних сервісів є платформа ULISS (Ukrainian Lingua-Information Semantic Service), на основі якої створено Український національний лінгвістичний корпус (УНЛК). ULISS належить до класу Л-систем, тобто систем, що поєднують багат шарові мовно-інформаційні ресурси й забезпечують їхню взаємодію в єдиному середовищі. ULISS розроблено за принципами сервіс-орієнтованої архітектури (SOA), рекомендованої консорціумом W3C. Це означає, що:

- окремі словникові та корпусні сервіси працюють як веб-сервіси,
- система функціонує як віртуальна лінгвістична лабораторія (ВЛЛ),
- взаємодія між сервісами йде через Інтернет-інфраструктуру,
- кожен розділ є автономним, але доступним через єдиний інтерфейс.

ULISS об'єднує функціональні можливості низки попередніх інформаційно-лексикографічних систем:

- електронної бібліотеки (тексти та джерела),
- граматичного словника (морфологічні дані),
- інтегрованої лексикографічної системи «Словники України»,
- розділя повнотекстового пошуку.

Для кожної зі згаданих систем були створені клієнтські програми з власними інтерфейсами, але ULISS надає уніфікований доступ до всіх сервісів одразу. У поєднанні цих сервісів у межах платформного комплексу ULISS формується єдине середовище для обробки та збереження мовних ресурсів, що забезпечує:

- цілісність даних між різними розділами;
- автоматичну взаємодію корпусних і словникових ресурсів;
- можливість розвивати корпус і словники синхронно;
- масштабування нових сервісів без переробки ядра системи.

Таким чином, УНЛК на базі ULISS є прикладом того, як архітектурна єдність і сервісна розділність дозволяють створити потужну національну мовно-інформаційну платформу.

Щоб зрозуміти, як саме відбувається технічна інтеграція сучасних лінгвістичних ресурсів (таблиця 3.4), важливо бачити не лише напрямки взаємодії, а й механізми, які роблять її можливою. Після таблиці наведено короткий огляд ключових технологій – від API до онтологічних моделей, – що забезпечують обмін даними, семантичну узгодженість та роботу ресурсів у спільному цифровому середовищі. Саме ці інструменти формують інфраструктуру, на якій ґрунтується мережевість сучасної лексикографії.

Таблиця 3.4 – Кейси інтеграції лінгвістичних ресурсів

Кейс інтеграції	Тип інтеграції	Що поєднується	Основний механізм	Що це дає користувачеві / лінгвісту
Lexonomy → SketchEngine	Словник ↔ корпус	Лексикографічний редактор + корпусні дані	API, OneClick Dictionary, автоматичне витягнення даних	Приклади, колокації, частоти, синоніми; автоматизоване наповнення словникових статей; скорочення часу укладання
Wiktionary → Wikidata	Глобальна база знань ↔ краудсорсинговий словник	Семантичні дані + багатомовні словникові статті	RDF, структуровані властивості Wikidata, посилання між елементами	Узгоджені значення, переклади, гіпероніми/гіпоніми; можливість автоматичного імпорту та валідації даних
ULISS → УНЛК	Інтегрована лінгвістична	Граматика, словники,	Сервіс-орієнтована	Єдиний інтерфейс до всіх мовних

	платформа національний корпус ↔	бібліотека, корпусний пошук	архітектура (SOA), веб- сервіси W3C	ресурсів; синхронізація словникових і корпусних даних; масштабованість та розділність

Як пояснення до таблиці 3.4, нижче подано роль основних форматів та технологій у процесах інтеграції лінгвістичних ресурсів:

– API забезпечують інтеграцію словників із корпусами та іншими ресурсами, дозволяючи отримувати дані в реальному часі (напр., *Lexonomy* ↔ *SketchEngine*).

– RDF та SPARQL поєднують словникові бази в єдину семантичну мережу, що підтримує зв'язність даних і міжмовну взаємодію (напр., *Wiktionary* ↔ *Wikidata*, *WordNet* ↔ *BabelNet*).

– Сервіс-орієнтована архітектура (SOA) інтегрує різні розділи мовних систем у спільне інформаційне середовище, забезпечуючи доступ до них через єдиний інтерфейс (напр., *ULISS* ↔ *УНЛК*).

– Онтологічні моделі створюють основу для масштабованості, формального опису значень та автоматичної інтероперабельності між ресурсами (напр., *BabelNet*, *ConceptNet*).

Питання для самоконтролю

1. Що означає інтеграція лексикографічних ресурсів у глобальний інформаційний простір?
2. Які основні типи інтеграції між онлайн-словниками сьогодні існують?
3. У чому полягає роль Wikidata як централізованої бази знань у лексикографії?
4. Як корпуси підтримують автоматичне збагачення словникових даних?
5. Назвіть приклади NLP-розділів, які можуть бути безпосередньо інтегровані зі словниковими ресурсами.
6. У чому полягає перевага використання RDF і SPARQL для об'єднання лексичних даних?

7. Чим відрізняються можливості API від можливостей онтологічних моделей (WordNet, BabelNet)?

8. Які технічні принципи лежать в основі сервіс-орієнтованої архітектури ULISS?

9. Як онлайн-ресурси типу Duolingo або Grammarly використовують словникові та корпусні дані?

10. Чому інтеграцію мовних ресурсів розглядають як ключовий чинник формування “динамічного словникового середовища”?

Завдання для самоперевірки

1. Оберіть один ресурс (Lexonomy, Wiktionary, SketchEngine, ULISS, WordNet, Wikidata) і коротко опишіть:

- які дані він надає;
- через які механізми інтегрується з іншими;
- які нові можливості створює ця інтеграція для користувачів.

2. На основі матеріалу розділу поясніть різницю між: а) API та RDF; б) SPARQL та REST-запитом; в) онтологією та корпусом. Подати короткі приклади використання кожного.

3. На основі опису Lexonomy ↔ SketchEngine або Wiktionary ↔ Wikidata:

- визначте, які саме типи даних передаються між системами;
- покажіть, як це впливає на кінцевий словниковий запис;
- у чому полягає користь автоматичного збагачення.

4. Запропонуйте умовну модель взаємодії: **словник** → **корпус** → **NLP-розділ** → **освітня платформа**. Пояснити, які саме дані передаються між компонентами і які формати/технології потрібні.

5. Використовуючи таблицю з попереднього підрозділу, побудуйте власну (текстову або графічну) «мережу взаємодії» ресурсів. Описати щонайменше 3 вузли та 3 типи зв'язків між ними.

6. Поясніть, як сервіс-орієнтована архітектура ULISS дозволяє об'єднати бібліотеку, граматичний словник, повнотекстовий пошук і лексикографічні

розділі. Підкресліть, що саме інтегрується: формат даних, інтерфейси, функції чи семантика.

7. Чи можливо створити універсальний глобальний словник, який охоплював би всі мови і всі типи даних? Пояснити, які технічні та концептуальні бар'єри цьому заважають (і що з них можна подолати).

8. Вибрати будь-який онлайн-словник і простежити:

- які дані він отримує з Інтернету;
- чи використовує корпуси або зовнішні API;
- чи генерує щось назад у глобальну мережу (open data, RDF, JSON API).

Питання для поглибленого вивчення матеріалу

1. Досвід УМІФ НАН України в інтеграції лінгвістичних ресурсів.
2. Порівняння моделей взаємодії словників у середовищах open data та закритих комерційних екосистем.
3. Роль RDF-тристок та SPARQL-запитів у побудові семантично зв'язаних лексикографічних платформ.
4. Еволюція лексичних онтологій: від WordNet до BabelNet і Wikidata Lexemes.
5. Механізми автоматичної синхронізації словникових баз у різних мовах і проєктах (multilingual linking).
6. Використання API як інструмента для створення динамічних словників у реальному часі.
7. Концепція «єдиної лінгвістичної інфраструктури» та перспективи створення глобальних лексичних хабів.
8. Соціально-інфраструктурний вимір: як інтегровані словники впливають на освіту, переклад і мовну політику.

РОЗДІЛ 2. КОМП'ЮТЕРНА ЛЕКСИКОГРАФІЯ: НАУКОВА І ПРАКТИЧНА ДІЯЛЬНІСТЬ

4. КОМП'ЮТЕРНА ЛЕКСИКОГРАФІЯ: ЇЇ ПРЕДМЕТ І ЗАВДАННЯ

План заняття:

4.1. Сучасна комп'ютерна лексикографія та її головні завдання

4.2. Про поняття *електронного словника*

4.3. Етапи створення комп'ютерних словників

Застосування комп'ютерних технологій у словникарстві суттєво вплинуло на розвиток лексикографії, автоматизувавши багато трудомістких процесів – від створення електронних картотек до формування оригінал-макетів словників. Спочатку комп'ютерна лексикографія виконувала переважно допоміжні функції, проте поступово її роль зросла: з інструменту для технічної підтримки вона перетворилася на самостійну наукову галузь зі своїм теоретико-методичним апаратом.

Із розвитком комп'ютерної лексикографії змінилися й підходи до розуміння словника. Він перестав сприйматися виключно як друкований об'єкт і отримав нове середовище існування – цифрове. Це зумовило появу в науковій літературі різних номінацій (комп'ютерний словник, електронний словник, автоматичний словник), які вказують на словник у цифровому просторі, але не завжди точно відображають його сутність.

Варто наголосити: комп'ютерна лексикографія не заперечує традиції словникарства, а, навпаки, розвиває їх, надаючи нові можливості для реалізації завдяки сучасним технологіям. Таким чином, створення комп'ютерних словників – це не лише технологічний процес, а й продовження культурної та наукової традиції укладання словників у новому форматі.

4.1. Сучасна комп'ютерна лексикографія та її головні завдання

Комп'ютерна лексикографія сьогодні постає як *міждисциплінарна галузь, що поєднує традиційні принципи словникарства з можливостями цифрових технологій*. Вона охоплює як практичні завдання (створення електронних

словників, автоматизація пошуку, інтеграція мультимедійних даних), так і науково-методичні (розробка моделей репрезентації знань, формалізація мовних процесів, вивчення впливу цифрового середовища на словник).

Щоб краще зрозуміти еволюцію цієї галузі, можна виокремити кілька етапів її розвитку: початковий, формувальний, онлайновий, сучасний (див. таблицю 4.1).

Таблиця 4.1 – Етапи розвитку комп'ютерної лексикографії

Етап	Характеристика	Особливості
Початковий (1970–1980-ті рр.)	Використання комп'ютерів для створення електронних картотек та частотних словників	Допоміжна функція, обмежені ресурси, орієнтація на корпуси текстів
Формувальний (1990-ті рр.)	Створення перших CD/DVD-словників і комп'ютерних програм	Зручний пошук, мультимедійність, але відсутність динамічного оновлення
Онлайновий (2000-ті рр.)	Активний вихід словників в Інтернет	Доступність, інтерактивність, оновлюваність, але різна якість контенту
Сучасний (2010-ті рр. – теперішній час)	Інтегровані та динамічні лексикографічні середовища	Поєднання словників, корпусів, тезаурусів, участь спільнот, краудсорсинг, штучний інтелект

Отже, сучасна комп'ютерна лексикографія виконує не лише функцію збереження та подання мовних даних, а й функцію інтеграції, аналізу та моделювання мовних явищ у цифровому середовищі. Залежно від характеру завдань, комп'ютерна лексикографія може одночасно розглядатися і як *наукова*, і як *виробнича* галузь.

1. *Наукова сфера* орієнтована передусім на пошук нових моделей репрезентації знань і розробку інноваційних способів роботи з мовними даними. З огляду на це, наукові дослідження зосереджені на:

- застосуванні методів і алгоритмів комп'ютерних наук у лексикографічних практиках;
- моделюванні словникових структур у цифровому форматі;
- створенні комп'ютерних систем для укладання, підтримки та оновлення лексикографічних ресурсів;
- вивченні специфіки функціонування словників у цифровому середовищі;
- формуванні теоретико-методичного апарату комп'ютерної лексикографії.

2. *Виробнича сфера* комп'ютерної лексикографії пов'язана з практичною реалізацією розробок:

- створення програмного забезпечення для словників;
- розроблення комп'ютерних систем, що використовують лексикографічний опис як базу;
- забезпечення користувачів ефективними інструментами для отримання, обробки й передавання мовної інформації;
- інтеграція словників з іншими інформаційними системами (корпуси, пошукові системи, машинний переклад).

Таким чином, наукова сфера визначає теоретичні основи та методи, а виробнича сфера забезпечує їх прикладне втілення у вигляді конкретних цифрових продуктів. Методичний апарат сучасної комп'ютерної лексикографії є органічним поєднанням методів споріднених із нею дисциплін (див. рисунок 4.1).



Рис. 4.1 – Споріднені дисципліни

Теорія мови формує базові уявлення про мовні одиниці та їхні властивості (фонетичні, морфологічні, синтаксичні, семантичні). Забезпечує термінологічну та концептуальну основу опису. У цьому контексті для опису мовних одиниць використовуються такі методи: описовий (узагальнення, класифікація та упорядкування), зіставний (виявлення спільних та відмінних рис двох або кількох мов у разі створення перекладних словників), порівняльно-історичний (для етимологічних словників або формулювання довідки щодо походження слова), метод компонентного аналізу (для **філіації** тобто розкладу слова на значення),

лексико-семантичного поля (створення ідеографічних словників, для опису семантики слів, що входять до того самого поля), функціональний метод (подання функціональних характеристик мовними одиницями у комунікативній ситуації). Взаємозв'язок між теорією мови і комп'ютерною лексикографією полягає в тому, що перша надає методичний апарат для опису мовного матеріалу, а друга забезпечує зворотний зв'язок, надаючи інформацію щодо ефективності застосування цих методів у практиці електронного словникарства.

Традиційна лексикографія сформувала фундаментальні принципи добору, опису та систематизації лексики, які й сьогодні залишаються методологічною основою для укладання словників у будь-якому форматі. Вона визначає як загальні засади побудови словника (відбору словникового складу, розроблення системи реєстрових одиниць), так і конкретні елементи його структури – макроструктуру (порядок розташування реєстрових одиниць) та мікроструктуру (внутрішній устрій словникової статті, набір обов'язкових та факультативних параметрів опису). Саме традиційна лексикографія виробила типологію словників (тлумачні, перекладні, термінологічні, ідеографічні, енциклопедичні, корпусні тощо) та окреслила функції кожного з них.

Електронне словникарство наслідувало ці напрацювання, але надало їм нового втілення. Так, усталені принципи упорядкування словникових статей (алфавітний, гніздовий, частотний) отримали реалізацію у вигляді гнучких алгоритмів сортування та можливостей багатовимірного пошуку. Типологічні різновиди словників у цифровому середовищі доповнилися новими форматами, наприклад інтерактивними корпусними чи мультимедійними словниками, які розширюють межі традиційних жанрів. Структура словникової статті, вироблена традиційною лексикографією, також зазнала модифікацій: тепер вона може включати гіпертекстові посилання, багаторівневі ієрархії значень, мультимедійні ілюстрації.

Окрім цього, саме термінологічний апарат традиційної лексикографії став основою для електронного словникарства. Такі поняття, як **реєстрове слово**, **словникова стаття**, **дефініція**, **ремарка**, **ілюстрація**, зберігають свою

актуальність і сьогодні, хоча в цифровому середовищі вони доповнюються новими термінами (**гіпертекст, тег, база даних, інтерфейс, парсинг, пошуковий індекс, вибірка**). Таким чином, традиційна лексикографія забезпечує тяглість і наступність, а комп'ютерна лексикографія розширює її потенціал, адаптуючи перевірені принципи до нових технологічних умов.

Взаємозв'язок між традиційною та комп'ютерною лексикографією полягає в тому, що перша забезпечує методичну й термінологічну основу, а друга розширює її межі, використовуючи цифрові технології для гнучкішого, багатовимірного та інтерактивного представлення словникової інформації.

Структурна лінгвістика надає інструменти для опису відношень між мовними одиницями – парадигматичних, синтагматичних та ієрархічних. Це сприяє побудові мережових ієрархій у словниках, тезаурусах і семантичних мережах. Вплив структурної лінгвістики на лексикографію виявився визначальним: словники перестали бути простими зібраннями слів та перетворилися на наукові інструменти системного аналізу мови. Вона заклала концептуальні підвалини системності в описі мови (ідеї Х. Касареса, І. Мельчука, Ю. Апресяна, В. Широкова), що передбачають аналіз лексичних одиниць не ізольовано, а в системі зв'язків. Парадигматичні та синтагматичні відношення, а також опис фонематичного й морфологічного рівнів стали ключовими у створенні словників нового типу. Саме завдяки цим засадам відбувся перехід від описової лексикографії до системної, у якій кожен елемент подається у взаємозв'язках з іншими. Це забезпечило появу тезаурусів, ідеографічних та електронних словників, де відображено багатовимірність мовної структури.

Взаємозв'язок між структурною лінгвістикою і комп'ютерною лексикографією полягає в тому, що перша надала інструменти для формалізованого опису мовних відношень, а друга забезпечила технічні засоби для їхньої практичної реалізації у вигляді складних словникових систем.

Корпусна лінгвістика забезпечує комп'ютерну лексикографію надійним емпіричним підґрунтям, оскільки корпуси текстів виступають джерелом реального мовного матеріалу. Вони надають можливість досліджувати слова

й вирази у контексті, відстежувати їхню частотність, уживаність у різних жанрах і стилях, а також фіксувати зміни в мові. Саме корпуси дозволяють переходити від інтуїтивного або фрагментарного опису до статистично підтверджених висновків. Для укладання словників корпусна лінгвістика надає:

(1) *репрезентативний матеріал* – збалансовану вибірку текстів, що відображає реальне функціонування мови;

(2) *частотні дані* – основа для створення частотних словників та словників мінімальної лексики;

(3) *контекстуальні приклади* – ілюстрації для словникових статей, які відображають природне вживання;

(4) *динамічний моніторинг* – можливість простежити появу неологізмів і зміну значень слів у режимі реального часу.

Завдяки корпусній лінгвістиці укладачі отримали змогу створювати словники нового покоління – корпусні, частотні, контекстуальні, а також інтерактивні ресурси, що дозволяють користувачеві самостійно досліджувати мовний матеріал. Взаємозв'язок між корпусною лінгвістикою і комп'ютерною лексикографією полягає в тому, що перша постачає автентичні дані про функціонування мови, а друга надає інструменти для їхнього упорядкування, обробки та представлення у словниковій формі.

Математична лінгвістика забезпечує комп'ютерну лексикографію формальними методами аналізу мови та інструментами побудови моделей. Її внесок полягає у створенні строгих описів мовних одиниць і відношень між ними, що дозволяє відобразити мову у вигляді системи правил, алгоритмів та структур. Основні напрями застосування:

– *формальні граматики* (Н. Хомський та його послідовники) дають змогу описати породження речень та структуру словоформ;

– *теорія множин і графів* застосовується для побудови семантичних мереж, тезаурусів і онтологій;

– *теорія семантичних станів* для побудови формальних моделей взаємодії граматичного та лексичного значення;

– *теорія лексикографічних систем* допомагає побудувати формальну модель системи мови як механізм породження того чи іншого мовного феномена (лексикографічного ефекту);

– *теорія ймовірностей і статистика* дозволяє моделювати частотність лексем і визначати закономірності їхнього вживання;

– *методи обчислювальної семантики* допомагають відтворювати зв'язки між формою та значенням слова;

– *алгоритмічні підходи* забезпечують створення програм для автоматичного аналізу та генерації словникових даних.

Для укладання та підтримки комп'ютерних словників математична лінгвістика надає:

– моделі опису словникової статті, що дозволяють автоматизувати структурування;

– засоби для розробки інтерфейсів, які працюють із формалізованими даними;

– методи статистичної обробки, що гарантують об'єктивність відбору лексики.

Взаємозв'язок між математичною лінгвістикою і сучасною комп'ютерною лексикографією полягає в тому, що перша формує строгий теоретико-методологічний апарат, а друга використовує його для створення цифрових ресурсів, здатних не лише зберігати, а й моделювати знання про мову.

Комп'ютерна лінгвістика є найближчою до комп'ютерної лексикографії дисципліною, оскільки безпосередньо поєднує мовознавчі знання з інструментами програмування. Вона забезпечує технічний та алгоритмічний інструментарій для реалізації лексикографічних проєктів у цифровому середовищі. Основні напрями застосування:

– розроблення програмного забезпечення для функціонування словників, пошуку, фільтрації та візуалізації даних;

– автоматичний морфологічний та синтаксичний аналіз, що дозволяє інтегрувати у словники розділі для розпізнавання форм і структур;

- технології машинного перекладу, які спираються на словникові бази й одночасно стимулюють їхнє вдосконалення;
- системи обробки природної мови (NLP), що охоплюють пошук ключових слів, аналіз семантики та генерацію текстів;
- інтерфейси взаємодії (веб-платформи, мобільні додатки, API), які дають змогу користувачам ефективно працювати зі словниковими ресурсами.

Комп'ютерна лінгвістика, надаючи свій широкий інструментарій робить можливим:

- інтеграцію словників у ширші інформаційні системи;
- автоматичне оновлення та поповнення лексикографічних баз;
- застосування штучного інтелекту для аналізу та збагачення словникових даних;
- створення інтелектуальних пошукових систем і мовних помічників.

Взаємозв'язок між комп'ютерною лінгвістикою і комп'ютерною лексикографією полягає в тому, що перша надає програмно-алгоритмічний інструментарій, а друга визначає мовний матеріал і принципи його організації. Разом вони формують сучасне цифрове словникарство як прикладну і наукову дисципліну.

У таблиці 4.2 подано стислий підсумок щодо внеску кожної спорідненої дисципліни в електронне словникарство (див. також додатки Д.6 і Д.7).

Таблиця 4.2 – Внесок споріднених дисциплін

Дисципліна	Методи / інструменти	Що саме надає дисципліна
Теорія мови	Описовий метод, зіставний метод, порівняльно-історичний метод, компонентний аналіз, метод лексико-семантичних полів, функціональний метод	Базові уявлення про мовні одиниці та їхні властивості; концептуально-термінологічний апарат для опису; принципи класифікації та упорядкування лексики
Традиційна лексикографія	Принципи добору лексики; укладання макро- та мікроструктури; типологія словників (тлумачні, перекладні, енциклопедичні тощо)	Систематизація словникових даних; термінологічна база; структура словникової статті; спадкоємність традицій
Структурна лінгвістика	Аналіз парадигматичних, синтагматичних і ієрархічних відношень; побудова мережових ієрархій	Моделювання системних зв'язків між одиницями; системний підхід у лексикографії; поява тезаурусів та системних словників

Корпусна лінгвістика	Створення і використання корпусів текстів; статистичний аналіз; автоматизований пошук контекстів	Емпірична база даних для словників; реальні приклади слововживання; уточнення характеристик мовних одиниць
Математична лінгвістика	Формалізація мовних структур; побудова моделей; теорія множин і відношень	Формальні моделі словника; основа для баз даних і цифрових інтерфейсів; алгоритмічні схеми опису
Комп'ютерна лінгвістика	Морфологічний, синтаксичний і семантичний аналіз; алгоритми NLP; програмні інструменти	Автоматичне опрацювання лексики; створення ПЗ для електронних словників; інтеграція словників у цифрові системи

Сучасна комп'ютерна лексикографія, спираючись на поєднання напрацювань мовознавчих дисциплін і комп'ютерних наук, має можливість вирішувати широкий спектр завдань, про які детально викладено нижче.

1. Створення та підтримка фундаментальних лексиконів у цифровому середовищі. Фундаментальні електронні лексикони – це масштабні словникові бази, які охоплюють значні сегменти лексики та призначені для довготривалого використання як у науковій, так і в прикладній сфері. Їхніми характерними рисами є:

– *масштаб і всеохопність*: охоплюють як активну лексику сучасної мови, так і історичні, діалектні, професійні пласти.

– *динамічність*: завдяки автоматизованим механізмам оновлення відбувається постійна адаптація словникового матеріалу до змін у мові.

– *цифровий формат*: забезпечує інтерактивність, зручність пошуку, можливість інтеграції з іншими інформаційними системами.

У цифровому середовищі вони можуть набагато ефективніше виконувати свої функції:

(1) *аккумуляція знань про мову*: збирання та систематизація лексичних одиниць.

(2) *забезпечення актуальності*: регулярне поповнення неологізмами, змінами у вживанні.

(3) *збереження культурної пам'яті*: великі словники є пам'ятками мовознавства, які відображають мовну свідомість і цінності суспільства певної доби.

(4) *методологічна роль*: виступають «школою лексикографії» для наступних поколінь, адже в їхній структурі закладені нетлінні зразки роботи з мовним матеріалом.

Прикладами таких лексиконів, що функціонують в цифровому середовищі, є:

– у світовій практиці: [*Oxford English Dictionary*](#), [*Trésor de la langue française*](#), [*Diccionario de la lengua Española*](#), [*DUDEN*](#);

– в українській практиці: *Словник української мови* в 11 томах (СУМ-11), нині – [*СУМ-20 онлайн*](#).

У сучасному інформаційному світі значення фундаментальних лексиконів важко переоцінити. По-перше, вони є базою для машинного перекладу, семантичних пошукових систем, мовних моделей ШІ. По-друге, служать орієнтиром для мовної політики, нормування правопису, термінології. І, по-третє, формують підґрунтя для освітніх і культурних програм, адже вони закарбовують національну мовну ідентичність. У таблиці 4.3. перелічено методи / інструменти, а також результати їхнього застосування.

Таблиця 4.3 – Створення та підтримка фундаментальних лексиконів

Завдання	Методи / Інструменти	Результати
Укладання та підтримка великих електронних словників (глумачних, перекладних, термінологічних)	– Лінгвістичне опрацювання корпусів – Автоматизоване оновлення баз даних – Використання XML/TEI-форматів – Семантична розмітка	– Стабільні базові ресурси для мовних технологій – Актуалізація словників у реальному часі – Відображення мовних змін – Взірцеві лексикографічні об'єкти для науки

2. Переведення у цифрову форму паперових словників – це одне з ключових завдань комп'ютерної лексикографії, яке поєднує культурологічний, освітній і технологічний виміри. Оцифрування паперових словників здійснюється з метою:

(1) *збереження культурної спадщини* – паперові словники старіють, псуються, тиражі зникають, проте цифровий формат гарантує довготривале збереження;

(2) *формування національної словникової бази* – створення єдиної бази даних, що відображає історичний розвиток мови;

(3) *забезпечення доступності* – відкритий онлайн-доступ для науковців, студентів і широкої аудиторії;

(3) *інтеграція в сучасні інформаційні системи* – можливість використання в мовних технологіях (машинний переклад, пошукові системи, семантичні бази знань).

Для переведення друкованих словників у цифровий формат Український мовно-інформаційний фонд НАН України під керівництвом В. Широкова розробив спеціальну технологію **парсингу** – *автоматизованого процесу перетворення тексту словника в комп'ютерну лексикографічну базу даних із чітко визначеною структурою*. Як показано на рисунку 4.2, технологія передбачає такі етапи:

1. *Сканування та OCR (оптичне розпізнавання тексту)* – базовий етап переведення.

2. *Редакторська коректура розпізнаного тексту* для виявлення та усунення помилок.

3. *Розмічення словникового тексту мовою розмічення HTML* за допомогою спеціального програмного забезпечення.

4. *Конвертація тексту HTML у формат бази даних, у якій він зберігається з XML-розміченням*.

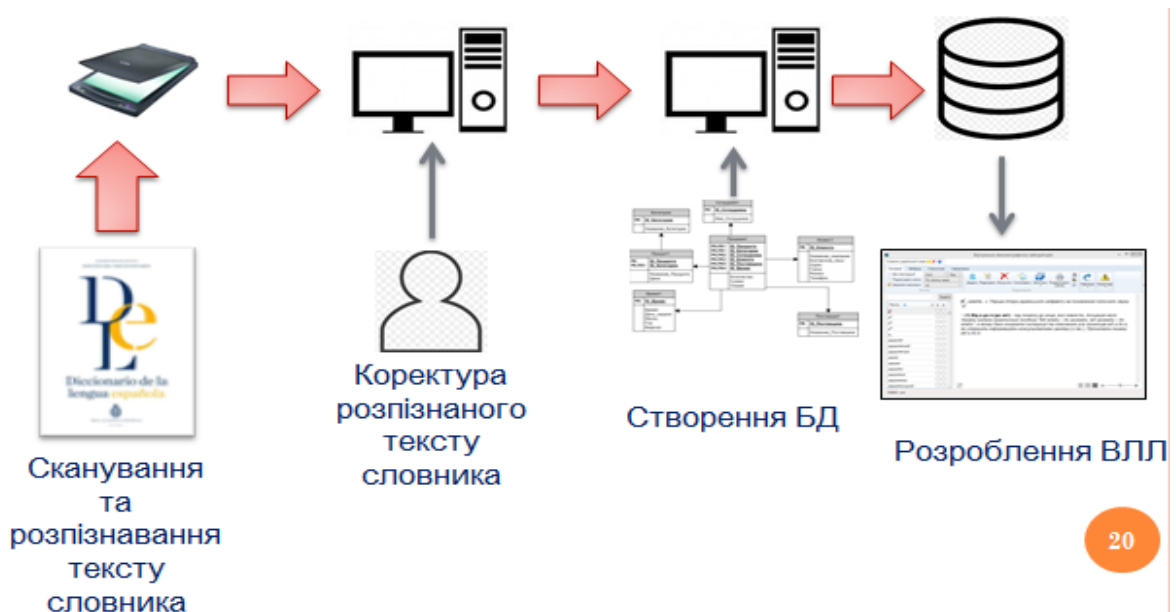


Рис. 4.2 – Схема парсингу словника

Доступ до словника в БД забезпечується через інтерфейс віртуальної лексикографічної лабораторії, про детально висвітлено нижче. У такий спосіб було переведено низку фундаментальних праць, зокрема:

- *Словник української мови* в 11 томах (нині триває робота над СУМ-20),
- *Етимологічний словник української мови* у 7 томах.

Цей досвід став підґрунтям для формування національної словникової бази України та забезпечив можливість інтеграції українських словників у світові інформаційні ресурси.

Отже, технології конвертації національно-культурного словникового надбання:

- створюють міст між традицією і сучасними мовними технологіями;
- дає змогу повторно використовувати класичні словники як базу для лінгвістичних досліджень та цифрових інструментів;
- сприяє популяризації національної культури та науки у світовому інформаційному просторі.

Проте, слід мати на увазі, що процес конвертації паперових словників у цифровий формат може викликати певні складнощі:

- (1) розмаїття форматів і структур словників різних епох – від рукописних картотек до друкованих багатотомників;
- (2) якість розпізнавання – особливо для старих шрифтів, дореформеної орфографії, рукописних приміток.
- (3) авторські права – потреба врегулювання доступу до сучасних словників.
- (4) фінансування та масштабність проєктів – оцифрування великих корпусів вимагає значних ресурсів.

У таблиці 4.4. перелічено методи / інструменти, а також результати їхнього застосування.

Таблиця 4.4 – Переведення у цифрову форму паперових словників

Завдання	Методи / Інструменти	Результати
Переведення друканих словників у цифровий формат	– Сканування та OCR – Коректура – HTML-розмічення словникового тексту	– Доступність і збереження словників – Формування національної словникової бази

	– Конвертація в XML-бази даних- Технологія парсингу (УМІФ, В. Широков)	– Можливість пошуку та повторного використання- Інтеграція в автоматичні системи обробки інформації

3. Інтеграція словників у онтологічні системи та системи екстракції знань (knowledge mining). Словники в електронному форматі перестають бути лише пасивними сховищами лексики. Завдяки сучасним інформаційним технологіям вони виконують функцію **семантичних баз даних**, на основі яких будуються **онтологічні системи** та системи **екстракції знань**.

Онтологія в інформатиці – *це формалізована модель певної предметної галузі, у якій задаються поняття (класи), відношення між ними, властивості та правила інтерпретації.* У цьому контексті електронні словники надають лексичні та семантичні ресурси, які допомагають створювати такі моделі. Так, ідеографічні словники та тезауруси дають готові структури парадигматичних і синтагматичних відношень, які можна формалізувати як мережу понять. У такий спосіб створюються бази знань для систем семантичного вебу, автоматичного перекладу, інтелектуального пошуку інформації.

Системи екстракції знань (knowledge mining) – *системи призначені для автоматичного вилучення релевантної інформації з великих масивів текстів.* Застосування таких систем охоплює галузі від пошукових механізмів та систем бізнес-аналітики до юридичної експертизи та біомедичних досліджень. У таких системах лексикографічні ресурси (тлумачні словники, термінологічні бази, словники синонімів і антонімів, словники сполучуваності) забезпечують **семантичний апарат**, завдяки якому машини можуть інтерпретувати значення слів у конкретному контексті. Вони допомагають уникати омонімії та багатозначності, формують основу для розпізнавання сутностей (імен, назв, термінів) та встановлення смислових відношень між ними. Для роботи в онтологічних системах словники мають бути:

- структурованими (XML, RDF, OWL),
- семантично багатими (з описом синонімії, антонімії, ієрархії понять),
- машиночитаними (пристосованими до інтеграції у програмні розділи).

Таким чином, комп'ютерна лексикографія виходить за межі «традиційного словникарства» і стає ключовим інструментом у побудові інтелектуальних систем нового покоління. У таблиці 4.5. перелічено методи / інструменти, а також результати їхнього застосування.

Таблиця 4.5 – Системи екстракції знань

Завдання	Методи / Інструменти	Результати
Інтеграція словників у онтології та системи knowledge mining	– Формалізація лексики у вигляді RDF/OWL-онтологій – Використання тезаурусів та ідеографічних словників – Автоматична семантична розмітка текстів – Алгоритми вилучення знань	– Побудова інтелектуальних систем – Семантичний пошук та розпізнавання сутностей – Усунення омонімії та полісемії – Використання у штучному інтелекті, семантичному вебі, експертних системах

4. Створення віртуальних лексикографічних лабораторій. Віртуальні лексикографічні лабораторії (ВЛЛ) є одним із найновіших напрямів комп'ютерної лексикографії, що поєднують функції електронних словників, систем автоматизованого аналізу текстів та наукових дослідницьких середовищ. Термін **«віртуальна лексикографічна лабораторія»** запровадив В. А. Широков для позначення інтегрованої лексикографічної системи, яка є розподіленою у фізичному просторі й забезпечує віддалений доступ до лексикографічної бази даних словника. Основна ідея ВЛЛ полягає у створенні середовища, де дослідник або користувач може:

- працювати безпосередньо з великими електронними лексиконами;
- проводити лінгвістичні експертизи (перевірка правильності вживання мовних одиниць, аналіз нормативності тощо);
- тестувати нові методи опису мовних одиниць та структурування лексикографічного матеріалу;
- створювати і перевіряти моделі автоматизованого аналізу текстів.

У практиці Українського мовно-інформаційного фонду (УМІФ) усі ВЛЛ, розроблені для новостворених або конвертованих словників, функціонують передусім як **dictionary-writing systems (DWS)**, тобто як спеціальні редакторські платформи для роботи над словником. Водночас їхній інструментарій набагато ширший, адже система забезпечує повний доступ до структурованої бази даних

словника. На рисунку 4.3. зображено зовнішній вигляд інтерфейсу ВЛЛ для роботи з СУМ 20.

Ключовою особливістю ВЛЛ є **індексування** елементів словникової статті: кожна одиниця – слово, форма, дефініція, приклад чи ілюстрація – включається до пошукової системи. Це відкриває можливості не лише для швидкої навігації та пошуку, але й для проведення лінгвістичних і статистичних досліджень на основі словникових даних.

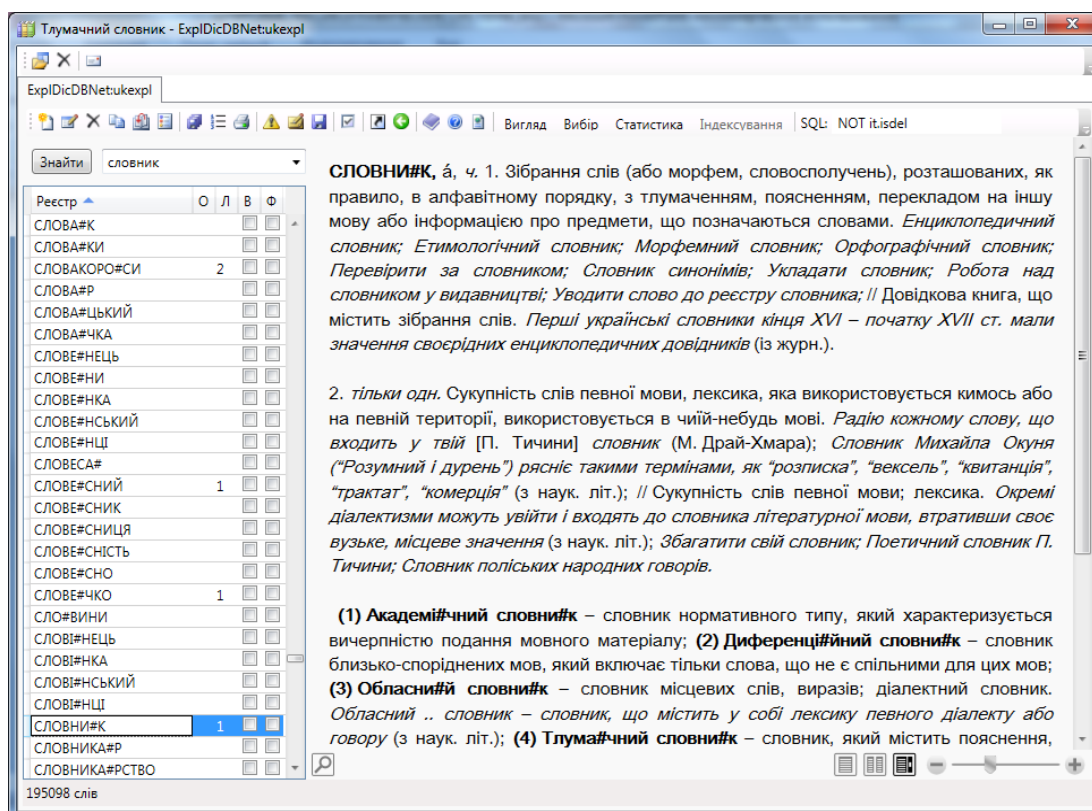


Рисунок 4.3 – Інтерфейс ВЛЛ СУМ 20

Таким чином, ВЛЛ перетворюється на універсальний науково-дослідний інструмент, що поєднує лексикографічні традиції з сучасними інформаційними технологіями. У таблиці 4.6. перелічено методи / інструменти, а також результати їхнього застосування.

Таблиця 4.6 – Створення віртуальних лексикографічних лабораторій

Завдання	Методи / Інструменти	Результати
Організація інтегрованих лексикографічних систем із віддаленим доступом до баз даних словників	– Концепт ВЛЛ (В. Широков) – Використання DWS (dictionary-writing systems) – Розподілена архітектура з веб-доступом	– Дистанційна робота з великими електронними лексиконами – Проведення лінгвістичних і статистичних досліджень

	– Індексуювання елементів словникової статті – Інтерактивний інтерфейс користувача	– Лінгвістичні експертизи на базі словника – Тестування нових методів опису мовних одиниць – Створення моделей автоматизованого аналізу текстів

Як можна побачити, завдання сучасної комп'ютерної лексикографії значно ширші, ніж у традиційної. Вона не обмежується укладанням словників, а виходить у площину цифрових гуманітарних наук та інтелектуальних систем. Комп'ютерна лексикографія сьогодні:

- поєднує фундаментальні й прикладні дослідження;
- забезпечує збереження лексикографічної спадщини;
- інтегрується з інформаційними технологіями, штучним інтелектом та knowledge mining;
- створює інтерактивні середовища для дослідження та викладання мови;
- відкриває нові перспективи для міждисциплінарної взаємодії.

Таким чином, вона є не лише продовженням словникової традиції, а й важливим інструментом розвитку сучасної лінгвістики та цифрових технологій.

Питання для самоконтролю

1. Які основні завдання вирішує сучасна комп'ютерна лексикографія?
2. Чим відрізняється створення фундаментальних лексиконів від завдань традиційної лексикографії?
3. У чому полягає значення переведення лексикографічної спадщини у цифрову форму?
4. Як застосовуються словники в онтологічних системах?
5. Які завдання виконує комп'ютерна лексикографія у сфері knowledge mining (видобування знань)?
6. Які функції можуть мати віртуальні лексикографічні лабораторії?
7. Чому для комп'ютерної лексикографії важлива співпраця з іншими дисциплінами (наприклад, корпусною чи математичною лінгвістикою)?
8. Наведіть приклад, як комп'ютерна лексикографія може бути корисною у сфері освіти.

9. Опишіть можливості використання словників у штучному інтелекті.
10. Які труднощі можуть виникнути під час парсингу паперових словників або словників у форматі pdf?

Завдання для самоперевірки

1. Поясніть, чим відрізняється використання словників у традиційній лексикографії та в онтологічних системах.
2. Наведіть приклад, як комп'ютерна лексикографія може сприяти розвитку штучного інтелекту.
3. Розкрийте роль knowledge mining у застосуванні лексикографічних ресурсів.
4. Установіть відповідність між завданням комп'ютерної лексикографії та його прикладом:
 - (1) Створення фундаментальних лексиконів
 - (2) Переведення словників у цифрову форму
 - (3) Використання словників у системах екстракції знань
 - (4) Віртуальні лексикографічні лабораторії
 - а) інтеграція словника у пошукову систему для автоматичного вилучення фактів;
 - б) створення електронної версії багатотомного тлумачного словника;
 - в) побудова базових словників сучасної української мови;
 - г) онлайн-платформа для проведення експертиз та аналізу мовних даних.
5. Складіть короткий план проєкту «Віртуальна лексикографічна лабораторія Словник українського сленгу» (3–4 пункти), зазначивши її основні функції, інструменти та користувачів.

4.2. Про поняття електронного словника

У підрозділі 1.1 було розглянуто поняття «словник» та подано власну дефініцію, що відображає його основні ознаки. Наступним кроком постає питання: **що таке електронний словник?** Чим він відрізняється від своїх паперових аналогів і чи можна застосувати до нього ту саму дефініцію?

У науковій літературі існують різні трактування цього поняття (див. рис. 4.4). Найчастіше вживаються такі терміни, як *комп'ютерний словник*, *комп'ютерна копія словника* та *комп'ютерна версія словника*. Хоча ці варіанти близькі за змістом, вони не є абсолютно тотожними, оскільки кожен з них відображає лише окремі аспекти аналізованого феномену. Так, *комп'ютерна копія словника* зазвичай позначає цифрове відтворення друкованого видання без суттєвих змін його структури. *Комп'ютерна версія словника* натомість може передбачати певні технічні модифікації – наприклад, інтерактивний пошук чи можливість гіпертекстових переходів. Водночас термін *електронний словник* має ширше значення: він окреслює не лише спосіб зберігання інформації, а й принципово новий формат її організації, який дозволяє реалізувати функції, недоступні традиційним друкованим словникам. Крім цього, електронні словники здатні об'єднуватися між собою та з іншими інформаційно-лінгвістичними ресурсами, утворюючи лексикографічну мережу, як це було зазначено у підрозділі 4.1.

Отже, визначення електронного словника потребує уточнення з урахуванням його специфічних характеристик: інтерактивності, динамічності структури, можливостей оновлення, інтеграції з іншими інформаційними системами та розширених засобів пошуку.



Рис. 4.4 – Лексико-семантичне поле *електронний словник*

Головна відмінність між паперовим і електронним словником, як це показано у таблиці 4.7. полягає не лише у формі носія, а й у функціональних можливостях. Паперовий словник є завершеним, статичним продуктом, що відображає стан мови на певний історичний момент. Електронний словник, навпаки, характеризується динамічністю та інтерактивністю: він може оперативно оновлюватися, інтегруватися з іншими інформаційними системами та надавати користувачу інструменти для швидкого пошуку й аналізу.

Таблиця 4.7 – Паперовий vs. електронний словник

Ознака	Паперовий словник	Електронний словник
Матеріальна форма	Друковане видання на папері	Цифровий файл / база даних
Структура	Статична, жорстко закріплена	Динамічна, може змінюватися та оновлюватися
Пошук інформації	Лінійний (за алфавітом, через перегортання сторінок)	Швидкий автоматизований пошук за словом, частиною слова, граматичними ознаками
Функціональність	Обмежена текстом та ілюстраціями	Може містити гіперпосилання, мультимедійні матеріали (звук, відео, зображення)
Можливість оновлення	Вимагає нового видання (перевидання)	Оновлюється в режимі реального часу або шляхом завантаження нових версій
Доступність	Потрібен фізичний доступ до книги	Доступ із будь-якого пристрою (ПК, планшет, смартфон) через інтернет або локально
Інтерактивність	Відсутня	Підтримка інтерактивних інструментів: пошукові фільтри, статистичний аналіз, інтеграція з іншими системами
Обсяг	Обмежений фізичним форматом (кількість сторінок, розмір видання)	Практично необмежений, можливе включення дуже великих масивів даних
Збереження та поширення	Потребує друку і фізичної логістики	Легко копіюється, поширюється й архівується в цифровому вигляді
Вартість виробництва	Висока (друк, папір, розповсюдження)	Нижча, після створення цифрової бази даних витрати на поширення мінімальні

Таким чином, електронний словник не є простим дублюванням друкованого аналога у цифровому форматі, а являє собою якісно новий тип лексикографічного ресурсу. На підставі розглянутих ознак можемо запропонувати таке визначення: **електронний словник** – це динамічна та інтерактивна лексикографічна система, яка у цифровому середовищі забезпечує оперативний доступ до упорядкованого лінгвістичного матеріалу фактично необмеженого обсягу та

надає можливість інтеграції з іншими інформаційними ресурсами. У таблиці 4.8 детально розписано характеристики електронного словника та приклади їхньої реалізації.

Таблиця 4.8 – Ознаки електронного словника

Ознака	Характеристика	Приклад реалізації
Динамічність	Можливість постійного оновлення та редагування словникового матеріалу	Автоматичне додавання нових слів, виправлення помилок
Інтерактивність	Активна взаємодія користувача з ресурсом через пошук, фільтри, гіперпосилання	Пошук за частиною слова, клік на лему відкриває статтю
Необмеженість обсягу	Відсутність фізичних обмежень на кількість словникових статей чи ілюстративного матеріалу	Розширені корпусні словники з мільйонами прикладів
Інтегрованість	Можливість підключення до інших інформаційних ресурсів і мовних технологій	Використання словника в системах машинного перекладу, онтологіях
Мультимедійність	Подання мовного матеріалу у різних форматах (текст, звук, відео, графіка)	Озвучування слів, відео з прикладами вживання

В електронній лексикографічній практиці на сьогодні закріпилися три базові принципи побудови словників: **принцип розмічення, структурний та комбінований**. Вони формують основу сучасних підходів до укладання й подання словникової інформації в цифровому середовищі. Кожен із принципів відображає певний рівень розвитку комп'ютерних технологій та еволюцію уявлень про електронний словник: від простого перенесення друкованого матеріалу у формат HTML-сторінки – до створення складних моделей даних, які дозволяють інтерактивну взаємодію з будь-яким елементом словникової статті. Умовно можна сказати, що принцип розмічення зорієнтований на зовнішню форму подання словника, структурний принцип – на внутрішню організацію даних, тоді як комбінований прагне поєднати їхні переваги. Особливості їхньої реалізації залежать як від завдань укладачів словника, так і від потреб користувачів: одні моделі роблять акцент на простоті представлення матеріалу, інші – на пошукових можливостях і інтеграції з іншими інформаційними системами.

Принцип розмічення. Його сутність полягає у використанні спеціальних мов розмітки (передусім XML та HTML), які дають змогу:

- чітко визначати структуру словникової статті;
- відокремлювати окремі елементи (лему, граматичні характеристики, тлумачення, приклади, етимологію тощо);
- забезпечувати коректне відображення матеріалу у веб- та програмних інтерфейсах;
- гарантувати сумісність з іншими цифровими системами.

Завдяки цьому електронний словник не обмежується «плоским текстом», а перетворюється на багаторівневий структурований ресурс. Одним із прикладів застосування принципу розмічення є Словник гідротурбінних термінів *Turbolex* (укладач Є. В. Купріянов), зовнішній вигляд якого подано на рисунку 4.5. Його основу становить масив словникових статей, що реалізовані у вигляді окремих HTML-сторінок із блоковою версткою та системою стилів CSS.

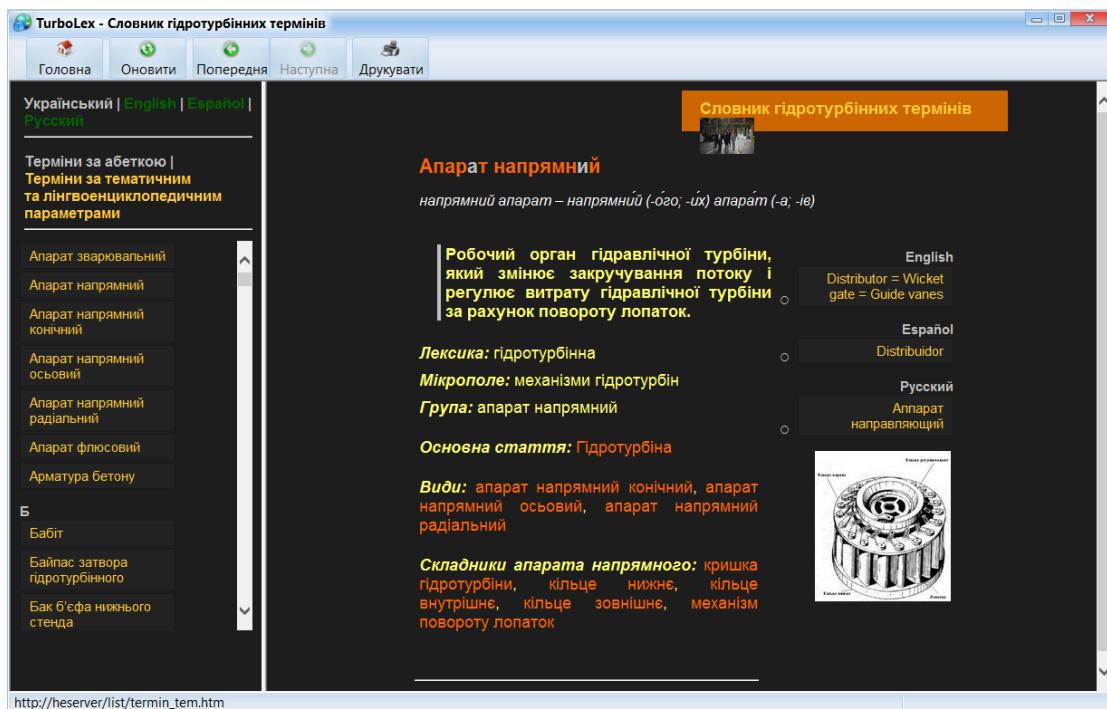


Рис. 4.5 – Словник *Turbolex*, укладений за принципом розмічення

Крім цього, термінологічні реєстри (як алфавітний, так і тематичний) також функціонують як HTML-сторінки з відповідним форматуванням і навігаційними елементами. Використання гіпертекстових технологій дає змогу не лише структурувати словникову статтю (лема, тлумачення, приклади, переклади тощо), а й організувати матеріал у вигляді **тезауруса**. Це дозволяє чітко простежити місце кожного терміна в термінологічній системі та візуалізувати

його зв'язки з іншими одиницями. Таким чином, *Turbolex* є прикладом словника, де принцип розміщення поєднано з гіпертекстовою організацією, що надає користувачеві додаткові інструменти для навігації й аналітики. Проте словник має обмежені пошукові можливості, інакше кажучи, пошук здійснюється лише за реєстром. Для забезпечення пошуку за іншими елементами словникової статті використовується **структурний принцип**. Якщо словники на зразок *Turbolex* реалізують інформацію через розмічені HTML-сторінки, то структурний підхід передбачає збереження даних у вигляді бази з чітко визначеними відношеннями.

Структурний принцип. Тут словник розглядається не як набір текстів, а як модель даних. Кожна одиниця словникової інформації (тлумачення, приклад вживання, граматична форма, перекладний еквівалент) є окремим елементом бази даних. Це дозволяє:

- здійснювати пошук за будь-яким параметром (наприклад, за граматичним класом, семантичним полем, походженням слова);
- автоматично формувати нові вибірки даних;
- інтегрувати словникову базу у зовнішні лінгвістичні та інформаційні системи (корпуси, онтології, системи машинного перекладу).

Структурний принцип активно використовують в Українському мовно-інформаційному фонді. Зокрема, у ВЛЛ СУМ-20 та інших словниках, створених у цифровому середовищі, він не лише візуалізує структуру словникової статті у вигляді ієрархії елементів, а й дозволяє редагувати кожен елемент окремо. Більше того, це уможливлює створення нового елемента в межах заданої структури, що суттєво розширює функціонал електронного словника порівняно з «плоским текстом» (див. рис. 4.6). Подібний підхід демонструє і WordNet, де значення слова зберігаються не як статті, а як вузли у семантичній мережі. Користувач має можливість виконувати пошук не лише за реєстром, а й за семантичними відношеннями (синонімія, антонімія, гіпонімія тощо). Завдяки цьому WordNet став основою для створення багатомовних проєктів, зокрема EuroWordNet, де окремі національні лексикони інтегровані в єдину мережу через

спільні міжмовні зв'язки. Таким чином, словник набуває ознак динамічної інформаційної системи, а не лише лексикографічного довідника.

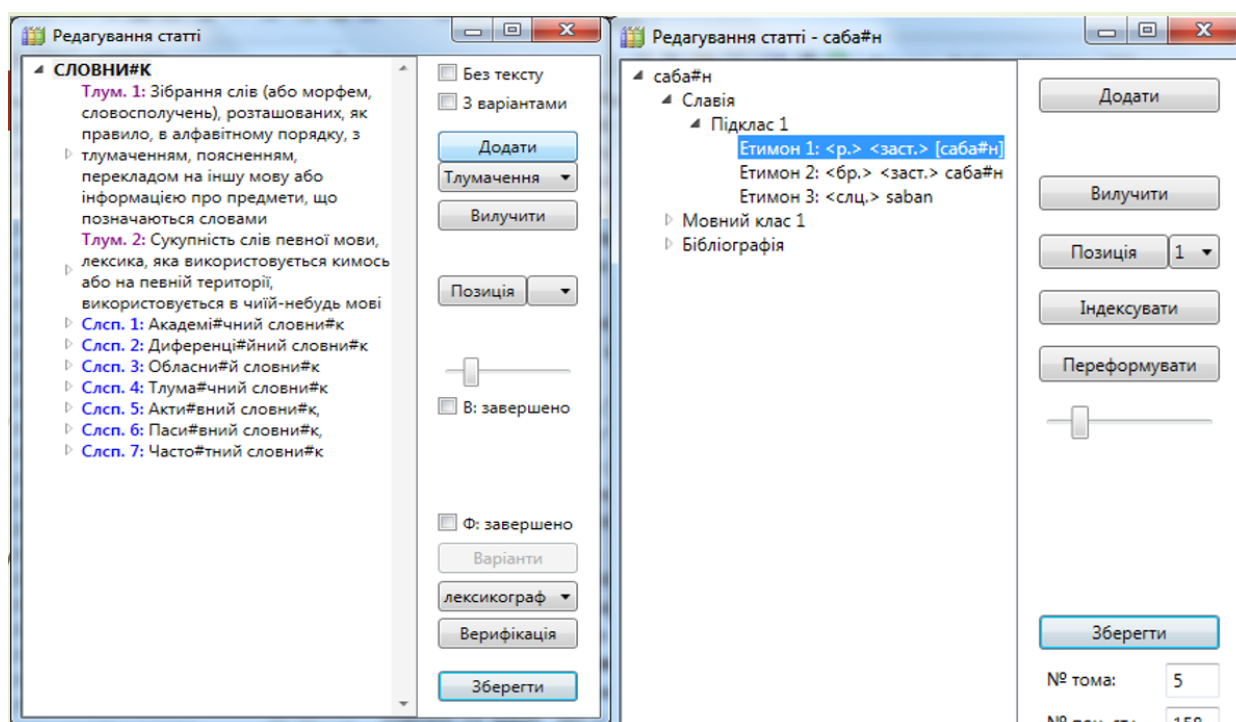


Рис. 4.6 – Вікна редагування статей СУМ 20

У таблиці 4.9 подана порівняльна характеристика обох принципів створення словників за сутністю, основною одиницею, рівнем організації матеріалу, пошуковими можливостями, інтеграцією та результатом застосування. Ключові особливості реалізації двох принципів на наведених вище прикладах (Turbolex, СУМ-20, WordNet, EuroWordNet) подано в додатку Д.8.

Таблиця 4.9 – Порівняння принципів створення електронних словників

Ознака	Принцип розмічення	Структурний принцип
Сутність	Використання мов розмітки (XML, HTML) для структурування словникового тексту	Подання словника як моделі даних у базі даних
Основна одиниця	Словникова стаття з внутрішньою розміткою	Інформаційний елемент (лема, граматика, приклад, переклад тощо)
Рівень організації	Текст з розміткою, що відображає структуру статті	Реляційна/об'єктна структура даних із множиною зв'язків
Можливості пошуку	Пошук переважно за лемами та ключовими полями	Пошук за будь-яким параметром (граматичним, семантичним, стилістичним)
Гнучкість	Обмежена – текст структурований, але статичний	Висока – можна створювати нові запити та автоматичні вибірки
Інтеграція	Сумісність із веб-форматами та іншими XML-базованими системами	Легка інтеграція у корпуси, онтології, системи машинного перекладу

Результат	Словник із чітко оформленими статтями	Динамічна інформаційна система з необмеженим доступом до даних

У сучасній електронній лексикографії дедалі частіше використовується комбінований підхід, який поєднує можливості структурного принципу та принципу розміщення. За структурним принципом формується **модель даних словника**, що містить усі інформаційні елементи (реєстрове слово, граматичні характеристики, тлумачення, приклади, ілюстрації тощо). Кожен елемент виступає окремим **індексом**, що забезпечує точний і гнучкий пошук за будь-яким параметром. Принцип розміщення, у свою чергу, дозволяє репрезентувати лінгвістичний матеріал у зручному для користувача вигляді, забезпечуючи коректне відображення структури словникової статті та її форматування. Яскравим прикладом такого підходу є Віртуальна лексикографічна лабораторія DLE 23 (рис. 4.7) – проєкт для роботи з *Diccionario de la lengua española* (тлумачним словником іспанської мови), розроблений Є. Купріяновим, І. Остаповою та М. Яблочковим. У цьому ресурсі структурна модель словника поєднана з HTML- та XML-розміткою, що забезпечує як потужний пошук, так і зручний інтерфейс для користувача.

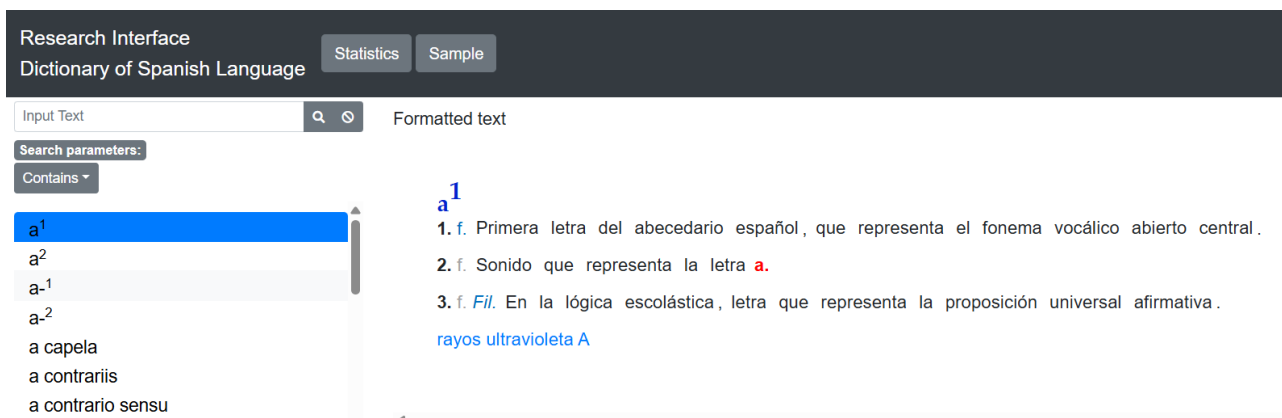


Рис. 4.7 – Візуалізація словникової статті у ВЛЛ DLE 23

Робота зі словником у цій ВЛЛ, створеній за комбінованим принципом, відбувається переважно через інструмент «Вибірка» (див. рис. 4.8). Це діалогове вікно містить перелік доступних пошукових індексів і забезпечує формування сукупності словникових статей, які відповідають конкретному запиту користувача. Наприклад, дослідник може налаштувати пошук словникових статей, у яких:

- заголовкове слово належить до іншомовної лексики;
- фіксується омонім;
- реєстрова одиниця є словом або морфемою;
- наявний акронім чи інша специфічна ознака.

Рис. 4.8 – Параметри для формування вибірки у ВЛЛ DLE 23

У результаті опрацювання запиту у реєстрі ВЛЛ автоматично формується вибірка статей, яка дозволяє здійснювати аналіз певної підмножини лексикону. Такий підхід відкриває нові можливості для лінгвістичних досліджень, порівняння та статистичної обробки даних. Як підсумок, у таблиці 4.10 наведено порівняльну характеристику всіх трьох принципів, що були розглянуті вище.

Таблиця 4.10 – Принципи побудови електронних словників

Принцип	Характеристика	Приклади
Принцип розмічення	Словниковий текст подається у вигляді HTML/XML-розмітки; забезпечується відтворення структури та форматування статей, але пошук обмежений реєстром.	«Turbolex» (гідротурбінна термінологія), перші електронні версії друкованих словників
Структурний принцип	Словник реалізується як модель даних; кожен елемент статті (реєстрове слово, значення, приклади, граматичні ознаки) проіндексовано для пошуку й редагування.	ВЛЛ СУМ-20 (Український мовно-інформаційний фонд), WordNet, EuroWordNet
Комбінований принцип	Поєднання двох підходів: структурний принцип відповідає за пошук та індексацію, а розмічення – за відображення матеріалу у зручному форматі.	ВЛЛ DLE-23 (тлумачний словник іспанської мови, Купріянов, Остапова, Яблочков)

Принципи побудови електронних словників визначають спосіб репрезентації лексикографічного матеріалу та характер взаємодії користувача з ним. **Принцип розмічення** забезпечує візуальне відтворення структури словникової статті й зручність представлення інформації, проте має обмежені пошукові можливості. **Структурний принцип** створює модель даних, у якій кожен елемент статті може бути індексований та доступний для пошуку чи редагування, що робить словник інтерактивним і багатофункціональним ресурсом. **Комбінований принцип** поєднує переваги обох підходів, дозволяючи одночасно реалізувати потужні пошукові механізми й гнучке відображення матеріалу у форматі, зручному для користувача. Отже, сучасна електронна лексикографія дедалі більше тяжіє до комбінованої моделі, яка забезпечує баланс між формальною структурованістю даних та зрозумілим представленням мовного матеріалу.

Показовою рисою електронних словників є надання користувачеві різноманітних інструментів доступу користувача до лінгвістичного матеріалу. У лексикографічній традиції такий інструмент прийнято називати **входом до словника**. На відміну від друкованих аналогів, де домінує лише один тип входу (зазвичай алфавітний порядок реєстрових слів), у цифровому середовищі може існувати ціла система альтернативних входів, які відображають багатомірність мови й дозволяють користувачу обирати найбільш зручний шлях навігації. Серед поширених типів входів у електронних словниках можна виокремити:

- *алфавітне розташування заголовкових одиниць*, що забезпечує класичний, найбільш інтуїтивний спосіб пошуку;
- *план понятійної класифікації лексики* (ідеографічний підхід), який дозволяє переходити від значення чи тематичної сфери до відповідних лексем;
- *словникові статті* як структуровані одиниці доступу, де користувач може переходити не лише до всього матеріалу статті, а й до її окремих елементів (визначення, приклади вживання, граматичні характеристики тощо);
- *історія перегляду словникових статей*, що виконує роль навігаційного сліду та полегшує повторне звернення до вже опрацьованих одиниць;

– *індекси* (морфологічні, частотні, тематичні, мовні та інші), які створюють додаткові шляхи пошуку і вибірки.

Завдяки цьому електронний словник функціонує як динамічна система входів і виходів, що значно розширює можливості його використання в науковій, навчальній та перекладацькій практиці. Характеристика різних типів входів наведена у таблиці 4.11.

Таблиця 4.11 – Типи входів до електронного словника

Тип входу	Характеристика	Приклади словників / проєктів
Алфавітне розташування заголовкових одиниць	Класичний спосіб доступу, що відтворює друковану традицію. Забезпечує швидкий пошук за початковою літерою.	<i>CYM-20 (ВЛЛ), Oxford English Dictionary Online</i>
План понятійної класифікації лексики	Ідеографічний підхід: доступ до лексем через тематичні групи чи семантичні поля. Дозволяє бачити слова у системних зв'язках.	<i>WordNet, EuroWordNet, Roget's Thesaurus Online</i>
Словникові статті	Структуровані блоки, що дають змогу працювати як з усією статтею, так і з окремими елементами (тлумачення, приклади, граматика).	<i>CYM-20 (ВЛЛ), Merriam-Webster Online</i>
Історія перегляду статей	Навігаційний інструмент, що дозволяє користувачу повернутися до раніше відкритих статей.	<i>DLE 23 (ВЛЛ), Cambridge Dictionary Online</i>
Індекси (морфологічні, частотні, тематичні тощо)	Дають можливість формувати спеціалізовані вибірки: за частотою, формою, темою, походженням, стилістикою тощо.	<i>CYM-20 (інструмент «Вибірка»), Sketch Engine</i>

Як видно з таблиці, електронні словники надають можливість користуватися різними типами входів – від пошуку за словоформами до введення транскрипції чи перекладу. Така множинність входів є однією з головних переваг електронних словників над паперовими, адже вона значно прискорює доступ до потрібної інформації та робить роботу зі словником більш гнучкою й зручною.

Ще однією характерною рисою електронних словників є використання індексів, які забезпечують пошук у їхній структурі. Пошуковим індексом може виступати будь-який елемент словникової статті, включно з метамовними компонентами. Наприклад, у OED online (рис. 4.9.) пошуковими індексами є галузь використання, мова походження, частина мови та лінгвопрагматичні характеристики. У ВЛЛ ЕСУМ реалізовано індекси мов, етимонів та реєстрових слів. У ВЛЛ DLE 23 до індексів віднесено не лише елементи словникової статті,

а й метамовні показники, зокрема тип мовної одиниці (морфема, слово, акронім, запозичення тощо), кількість тлумачень, наявність та види фразеологізмів.

Рис. 4.9 – Пошукові індекси OED online

Таким чином, система індексів у електронних словниках значно розширює можливості користувача, оскільки дозволяє здійснювати пошук не лише за реєстровим словом, а й за будь-яким пов'язаним параметром. Це забезпечує багатовимірний доступ до лексикографічного матеріалу, що принципово відрізняє електронні словники від паперових і робить їх ефективним інструментом для наукових досліджень і навчальної практики.

Питання для самоконтролю

1. Які три базові принципи побудови електронних словників закріпилися у сучасній лексикографічній практиці?
2. У чому полягає специфіка принципу розмічення, структурного та комбінованого підходів?
3. Наведіть приклади словників або проєктів, побудованих за кожним із принципів.
4. Що таке «вхід до словника» і які його основні різновиди в електронних лексикографічних системах?
5. Чому множинність входів вважається однією з головних переваг електронних словників над паперовими?

6. Яку роль відіграють пошукові індекси у функціонуванні електронних словників?

7. Які типи індексів реалізовані у таких словниках, як OED online, ВЛЛ ЕСУМ, ВЛЛ DLE 23?

8. У чому полягає відмінність між пошуком за реєстровим словом та пошуком за метамовними елементами?

9. Як наявність індексів і множинних входів впливає на зручність роботи користувача зі словником?

Завдання для самоперевірки

1. Проведіть аналіз та заповніть таблицю для таких словників: *Gemet thesaurus online*, *Collins dictionary Online*, *Словники України Online*. Для кожного словника визначте:

- кількість і типи входів (алфавітний, тематичний, історія перегляду тощо);
- які пошукові індекси доступні користувачу;
- який принцип реалізації (розмічення, структурний, комбінований) переважає.

2. Порівняйте результати. Визначте, які функціональні можливості наявні у всіх словниках, а які є унікальними.

3. Проаналізуйте власний досвід користування. Який зі словників видався найбільш зручним для роботи і чому?

4. Додаткове завдання (на вибір):

- Оберіть ще один електронний словник (наприклад, *Cambridge Dictionary Online*, *Wiktionary* чи *Linguee*). Проведіть такий самий аналіз і порівняйте його з уже розглянутими.

- Складіть коротку таблицю «Переваги – Недоліки» для кожного словника.

- Спробуйте змоделювати власний варіант ідеального входу та пошукової системи для електронного словника.

Назва словника	Входи (алфавітний, тематичний, історія	Пошукові індекси (частина мови, галузь,	Принцип побудови (розмічення, структурний, комбінований)	Переваги	Недоліки

	перегляду тощо)	етимологія, інші)			
Gemet thesaurus online					
Collins dictionary Online					
Словники України Online					
Інший словник (на вибір)					

Приклад заповнення таблиці аналізу електронного словника наведено у додатку Д.9.

5. Уявіть, що вам потрібно укласти тлумачно-перекладний електронний словник юридичної лексики.

(1) Визначте, який принцип побудови словника (розмічення, структурний чи комбінований) буде найбільш доречним у цьому випадку та чому.

(2) Запропонуйте можливі **входи до словника** (наприклад: алфавітний реєстр, тематична класифікація – кримінальне право, цивільне право, адміністративне тощо).

(3) Сформулюйте набір **пошукових індексів**, які були б корисними користувачу (наприклад: мова оригіналу, частина мови, галузь права, наявність прикладів у законодавчих текстах, переклад на іншу мову).

(4) Складіть невелику схему (опис) того, як користувач працюватиме зі словником (алгоритм дій при пошуку потрібного терміна).

4.3. Етапи створення комп'ютерних словників

Створення комп'ютерних словників може відбуватися двома шляхами, а саме:

(а) *на основі вже існуючого словника* – коли паперове видання або його електронна копія перетворюється на структуровану базу даних;

(б) *з нуля* – якщо словник укладається без попереднього друкованого аналога, одразу в електронному середовищі.

Незалежно від обраного шляху, у процесі створення електронного словника можна виокремити три головні етапи: попередній, лінгвістичний та технічний. Залежно від того, який варіант було обрано, завдання на цих етапах можуть

включати додаткові процедури, як лінгвістичного, так і технічного характеру. Перелік завдань, виконуваних на цих етапах, незалежно від обраного варіанту, подано в таблиці 4.12.

Таблиця 4.12 – Етапи створення електронних словників

Етап	Основні дії	Результат
Попередній	– визначення мети й завдань словника; – встановлення цільової аудиторії; – вибір типу словника і лексикографічних параметрів; – розробка концепції словникової статті; – добір корпусу текстів;	Концепція словника, загальна структура та план укладання
Лінгвістичний	– укладання словникових статей; – розробка системи ремарок і класифікацій; – формування реєстру слів і термінів; – забезпечення внутрішньої узгодженості матеріалу; – побудова концептуальної моделі словника як лексикографічної системи; – побудова моделі семантичного стану для узгодженості граматичної й лексичної інформації;	Лексикографічна база даних (зміст словника), узгоджена й готова для технічної реалізації
Технічний	– оцифрування друкованих матеріалів (за потреби); – структуроване кодування словникових статей (XML, TEI, HTML тощо); – створення та налаштування бази даних; – розробка інтерфейсу доступу й пошукових інструментів; – створення індексів і механізмів фільтрації; – інтеграція словника з іншими ресурсами; – тестування, підтримка та оновлення	Повнофункціональний електронний словник із пошуковими та інтеграційними можливостями

На попередньому етапі формується **авторська установка словника**, що визначає концепцію, структуру й принципи укладання. Усі пункти, наведені в таблиці, є обов’язковими для словника, який створюється вперше. У випадку підготовки електронного аналога, частина параметрів може зазнати корекцій чи уточнень. Це зумовлюється низкою чинників, серед яких основними є: хронологічний, ідеологічний та технологічний.

1. *Хронологічний чинник*. Він тісно пов’язаний зі ступенем актуальності описуваної лексики. Лексикограф має враховувати:

– наявність застарілих одиниць, що можуть залишитися лише як марковані (*застар.*, *діал.*, *істор.*);

- появу нових слів, не представлених у попередніх реєстрах;
- розширення цільової аудиторії словника, що вимагає модифікації принципів подання лінгвістичної інформації;
- необхідність введення додаткових лексикографічних параметрів (семантичних, стилістичних, прагматичних).

2. *Ідеологічний чинник.* Лексикографічна практика завжди відображає культурні та соціально-політичні установки суспільства. Це стосується як відбору лексем, так і формування дефініцій та добору ілюстративного матеріалу. Для прикладу, СУМ-11 (1970–80-х рр.) і його осучаснений аналог СУМ-20 демонструють різні рівні відкритості й актуалізації словникового корпусу. Таким чином, створюваний словник має враховувати зміни в суспільному дискурсі, аби відображати мовну картину світу адекватно сучасності.

3. *Технологічний чинник.* Сучасна лексикографія тяжіє до інтеграційних рішень:

- забезпечення сумісності електронних словників на рівні форматів;
- можливість створення «лексикографічної мережі» через взаємопов'язані бази даних;
- використання інтероперабельних моделей (XML, TEI, LMF тощо), що дозволяють конвертувати інформацію та розширювати її функціонал у цифровому середовищі.

Усі ці параметри в комплексі визначають як функціональність майбутнього словника, так і його відповідність науковим та суспільним запитам. Узгодженість цих принципів дає можливість створити словник як динамічну лексикографічну систему, що водночас зберігає наукову традицію, відображає актуальний стан мови та відкриває перспективи цифрової інтеграції.

Лінгвістичний етап має свої особливості залежно від того, чи створюється словник з нуля, чи на основі вже існуючого видання. У випадку розроблення нового електронного словника всі завдання, перелічені у попередній таблиці, виконуються від початку: від побудови концептуальної моделі до формування бази даних і структурування словникових статей. Натомість для паперових

аналогів цей етап передбачає їх підготовку до процедури **парсингу** (див. підрозділ 3.1), тобто конвертації готової концепції у цифровий формат. На цьому етапі колектив лексикографів проводить попередній аналіз тексту з кількома завданнями:

- виявлення та опис будови словникових статей (див. підрозділ 1.1);
- аналіз і систематизація метамови словника (див. підрозділ 1.3);
- чітка ідентифікація структурних елементів і лексикографічних параметрів для подальшої роботи парсера (див. підрозділ 1.2).

Отже, лінгвістичний етап у разі перенесення паперового словника в електронний формат виконує функцію перехідної ланки між друкованим текстом і цифровою базою даних. На рисунку 4.10 подано приклад результату лексикографічного аналізу паперового тексту *Словника української мови* перед процедурою парсингу. Важливо зазначити, що результати цього аналізу використовуються не лише для технічної підготовки до оцифрування, а й для побудови концептуальної моделі словника (лексикографічної системи) – формальної моделі, яка на абстрактному рівні відображає його структуру разом із тими елементами, що мають бути доступними для користувача у базі даних через інтерфейс.

Структурний елемент	Формальна ознака структурного елемента
Рєєстрове слово N	Початок: абзац, літера (українська, півжирний шрифт, верхній регістр). Кінець: кома; крапка; пробіл, індекс, цифра, кома; пробіл «див.»; пробіл, індекс, цифра, пробіл «див.»; двокрапка.
Права частина статті, початок	Крапка, пробіл, цифра «1», крапка (прямий шрифт); крапка, пробіл, літера (українська, прямий шрифт, верхній регістр); двокрапка; відсилкове слово «див.» (курсив).
Тлумачення, початок	Крапка, пробіл, літера (українська, прямий шрифт, верхній регістр).

Рис. 4.10 – Фрагмент таблиці лексикографічного аналізу СУМ

Побудова такої моделі спирається на математичний апарат, зокрема алгебру кінцевих предикатів, теорію графів, а також спеціалізовані підходи на кшталт теорії лексикографічних систем. Саме остання (розроблена В. Широковим)

продемонструвала високу ефективність у реалізації масштабних проєктів Українського мовно-інформаційного фонду – *СУМ-20*, *ЕСУМ*, інтегрованої лексикографічної системи «Словники України» та низки граматичних словників. У таблиці 4.13 показано етапи переходу від паперового словника до електронної версії.

Таблиця 4.13 – Перехід від паперового словника до електронної версії

Рівень	Зміст	Результат
Аналітичний	Лексикографічний аналіз тексту: виявлення структури словникових статей, метамови, параметрів опису.	Система чітко ідентифікованих елементів (заголовкове слово, граматичні характеристики, дефініція, приклади тощо).
Концептуальний	Побудова формальної моделі словника за допомогою математичного апарату (теорія графів, алгебра предикатів, теорія лексикографічних систем).	Абстрактна модель словника як системи взаємопов'язаних елементів.
Технічний (реалізаційний)	Перенесення концептуальної моделі в базу даних і розробка інтерфейсу доступу.	Електронний словник із пошуковими індексами та множинними входами.

Отже, лінгвістичний етап у створенні електронного словника суттєво залежить від способу його укладання. Якщо словник створюється з нуля, то основна увага зосереджена на розробленні нової концепції та структури. Якщо ж електронний ресурс ґрунтується на паперовому аналозі, то ключовим завданням стає адаптація наявного матеріалу до цифрового формату через парсинг і попередній аналіз тексту. Обидва підходи є рівнозначними з точки зору кінцевого результату, адже забезпечують формування цілісної електронної моделі словника, придатної для подальшої технічної реалізації. Як підсумок, у таблиці 4.14 показано різницю в реалізації лінгвістичного етапу для існуючих та вперше створюваних електронних словників.

Таблиця 4.14 – Лінгвістичний етап у створенні електронних словників

Варіант створення	Основні завдання	Результат
З нуля	– побудова концептуальної моделі словника; – визначення структури словникової статті; – формування моделі семантичного стану; – відбір та опис лексичного матеріалу;	Повністю сформована концепція електронного словника та його структура.

	– розробка бази даних і лінгвістичних параметрів.	
На основі паперового аналога	– підготовка матеріалу до парсингу; – аналіз будови словникових статей (див. підрозділ 1.1); – опис метамови словника (див. підрозділ 1.3); – ідентифікація структурних елементів і параметрів (див. підрозділ 1.2).	Конвертована в цифровий формат концепція друкованого словника, готова до автоматичної обробки.

На основі побудованої концептуальної моделі розробляються **внутрішня модель** (база даних) і **зовнішня модель** (інтерфейси) лексикографічної системи відповідно до стандарту **ANSI/SPARC**. Основні завдання цього етапу:

- створення бази даних для збереження лексикографічної інформації згідно з концептуальною моделлю;
- розробка інтерфейсу користувача, що забезпечує доступ до словникових статей і пошукових інструментів;
- налаштування пошукової системи, яка реалізує роботу індексів і множинних входів до словника;
- інтеграція мультимедійних компонентів (аудіо, відео, графіка, ілюстрації), якщо цього вимагає концепція;
- забезпечення сумісності з іншими електронними лексикографічними ресурсами та платформами з метою формування «лексикографічної мережі»;
- тестування й відлагодження, спрямоване на перевірку коректності пошуку, відображення словникових статей та роботи інтерфейсу.

Для створення бази даних можуть застосовуватися різні технології (Access, SQLite, MySQL, NoSQL, JSON тощо), вибір яких залежить від масштабу словникового проєкту та технічних можливостей, оскільки кожна з них має свої переваги та обмеження. Так, MS Access зручний для невеликих проєктів і швидкого прототипування, проте має обмежену масштабованість. SQLite відзначається простотою і портативністю, що робить його ефективним для автономних застосунків. MySQL є оптимальним варіантом для середніх і великих систем із багатокористувацьким доступом, забезпечуючи високу продуктивність та стабільність. NoSQL-рішення (наприклад, MongoDB) дозволяють працювати з великими обсягами неструктурованих або напівструктурованих даних, що

особливо актуально для мультимедійних словників. Натомість JSON-структури добре підходять для зберігання і обміну словниковими статтями у форматі, придатному для інтеграції з веб-застосунками та іншими лінгвістичними сервісами.

Особливе значення має зовнішня модель, що реалізується у вигляді інтерфейсу словника. Вона надає користувачеві інструменти для роботи з лінгвістичним матеріалом і може відрізнятися залежно від категорії користувачів. Для лінгвістів-професіоналів, які проводять дослідження на базі тексту словника, особливо якщо це фундаментальний лексикон (академічний тлумачний словник), інтерфейси мають в ідеалі забезпечувати доступ до всіх структурних елементів словникової статті, у тому числі метамовних, для формування пошукових вибірок; різні точки доступу (не лише через реєстр, а й через словникові статті та індекси); повнотекстовий пошук по словнику; а також надавати інструменти для здійснення статистичних підрахунків по текстовому масиву (див. таблицю 4.15).

Таблиця 4.15 – Відмінності інтерфейсів словника

Категорія користувачів	Основні функції інтерфейсу	Характерні особливості
Пересічні користувачі	– пошук за реєстром; – перегляд словникових статей у зручному для читання форматі; – доступ до мультимедійних ілюстрацій (за наявності);	Простий і зрозумілий дизайн, інтуїтивна навігація, мінімум додаткових опцій.
Лінгвісти-професіонали	– пошук за кількома точками доступу (реєстр, індекси, словникові статті); – доступ до всіх структурних елементів словникової статті, включно з метамовними; – повнотекстовий пошук; – статистичний аналіз словникового корпусу; – можливість формування вибірок та експорту даних;	Розширений функціонал, багатоступенева система пошуку, наявність інструментів для наукових досліджень.

Для пересічних користувачів, навпаки, інтерфейс має бути спрощеним і зручним у використанні, з акцентом на швидкому пошуку та зрозумілому відображенні статей. Так, на рисунку 4.11 зображено різні інтерфейси СУМ-20 для:

– пересічних користувачів зі спрощеним функціоналом (праворуч): пошук за реєстром, перегляд словникових статей, зміна режиму вигляду статей;

– для експертів та лексикографів з розширеним інструментарієм для створення, редагування та наукового аналізу (ліворуч).

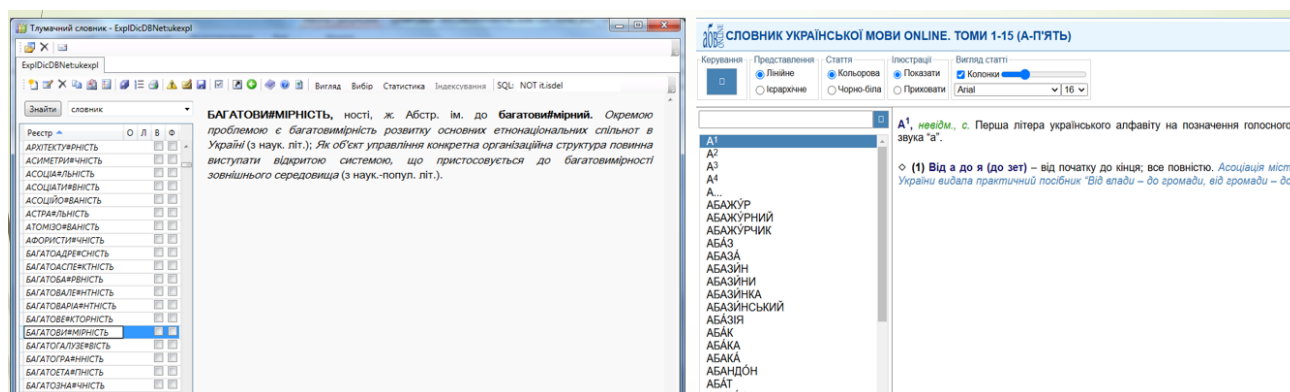


Рис. 4.11 – Види інтерфейсів для СУМ-20

Для лексикографів інтерфейс забезпечує не лише операції з укладання словника, а й організацію процесу роботи над словником (розподілення роботи по кожному лексикографу, облік виконаної роботи, історія внесення правок). Так, ВЛЛ СУМ-20 за допомогою мережових технологій забезпечує роботу над укладанням словника української мови, навіть якщо словникарі знаходяться в різних точках країни або світу. Коло функцій, що надає інтерфейс, залежить від рівні доступу: адміністратор, лексикограф, читач. У таблиці 4.16 подано характеристику функціональних можливостей інтерфейсу залежно від категорії користувачів.

Таблиця 4.16 – Функціональні можливості інтерфейсу

Рівень доступу	Основні функції	Особливості використання
Адміністратор	– керування правами доступу; – налаштування інтерфейсу та структури БД; – моніторинг роботи користувачів; – збереження та відновлення даних;	Повний контроль над системою, відповідальність за безпеку й стабільність роботи словника.
Лексикограф	– створення та редагування словникових статей; – додавання прикладів, ілюстрацій, мультимедійних матеріалів; – ведення історії правок; – участь у колективній роботі над статтями;	Інструменти для укладання словника, розподіл завдань між колективом.
Читач (кінцевий користувач)	– пошук і перегляд словникових статей; – використання індексів та точок входу; – доступ до мультимедійних компонентів (за наявності);	Спрощений інтерфейс, орієнтований на зручність користування.

Як видно, функціональні можливості інтерфейсу електронного словника чітко розподіляються залежно від рівня доступу користувача. Це забезпечує як

організацію колективної роботи лексикографів, так і зручність використання словника для кінцевих читачів, водночас гарантує цілісність і безпеку лексикографічного ресурсу. Багаторівнева організація інтерфейсу дозволяє одночасно підтримувати ефективну роботу над словником і забезпечувати зручність для кінцевих користувачів, зберігаючи при цьому цілісність і безпеку ресурсу.

Технічний етап є завершальною фазою створення електронного словника, у межах якої концептуальна модель набуває практичної реалізації. Систематизація його основних завдань і процедур у таблиці 4.17 дає можливість побачити взаємозв'язок між теоретичними принципами та їх технічним втіленням, що зрештою забезпечує повноцінне функціонування словника як сучасного цифрового ресурсу.

Таблиця 4.17 – Технічний етап у створенні електронних словників

Завдання	Інструменти / технології	Результати
Створення бази даних	Access, SQLite, MySQL, NoSQL, JSON	Збереження лексикографічної інформації у формалізованій структурі
Розробка інтерфейсу користувача	Web-технології (HTML, CSS, JavaScript), UI-frameworks, API	Інтерфейс для доступу до статей і пошукових функцій
Налаштування пошукової системи	Алгоритми індексації, SQL-запити, full-text search	Пошук за індексами, множинні входи до словника
Інтеграція мультимедійних компонентів	Формати аудіо/відео (MP3, MP4), графіка (SVG, PNG), інтерактивні розділи	Багатоканальна репрезентація словникових даних
Забезпечення сумісності з іншими системами	Стандарти даних (XML, RDF, JSON-LD), API, протоколи обміну	Можливість інтеграції у «лексикографічну мережу»
Тестування і відлагодження	Unit-тести, юзабіліті-тестування, beta-версії	Перевірка коректності пошуку, інтерфейсу, стабільності роботи

Система принципів, притаманних кожному етапу, подана в додатку Д.10. Ця система окреслює багаторівневу структуру побудови електронного словника, де лінгвістичний аналіз поєднується з інформаційною організацією даних і технологічними рішеннями. Такий підхід дає змогу розглядати словник не лише як зібрання лексем, а як цілісну лексикографічну систему, здатну забезпечити узгоджене й динамічне подання мовного матеріалу в різних форматах і для різних користувацьких потреб.

Питання для самоконтролю

1. Які етапи створення електронного словника можна виокремити?
2. У чому полягає відмінність попереднього етапу при створенні словника «з нуля» та при перенесенні паперового аналога?
3. Які завдання виконує лінгвістичний етап?
4. Яку роль відіграє побудова концептуальної моделі словника?
5. Для чого здійснюється парсинг паперових словників?
6. Які основні завдання виконує технічний етап створення електронних словників?
7. Які переваги та обмеження мають різні технології (наприклад, Access, SQLite, MySQL, NoSQL, JSON)?
8. Чим відрізняються внутрішня та зовнішня моделі лексикографічної системи?
9. Які технології можуть застосовуватися для створення баз даних словників?
10. Які інтерфейсні можливості мають бути передбачені для лінгвістів-професіоналів?
11. Які функції забезпечує інтерфейс для лексикографів під час роботи над словником?
12. Які рівні доступу користувачів можуть бути передбачені у ВЛЛ?

Завдання для самоперевірки

1. Обрати будь-який сучасний електронний словник (наприклад, *Oxford English Dictionary online*, *Словники України online*, *WordNet*) і:
 - визначити, які рішення реалізовані на попередньому етапі (авторська установка, концепція словника);
 - описати лінгвістичний етап (наявність концептуальної моделі, побудова словникових статей, метамова);
 - проаналізувати технічний етап (база даних, інтерфейс, пошукова система, індекси, мультимедійність).
2. Порівняти електронний словник, створений «з нуля», і електронну версію паперового словника.

- Визначити, які завдання на лінгвістичному етапі виконуються по-різному.
 - Пояснити, які переваги та недоліки має кожен із підходів.
3. Уявіть, що вам потрібно укласти електронний словник сучасного українського молодіжного сленгу.
- Визначте завдання попереднього, лінгвістичного та технічного етапів.
 - Які індекси та входи ви передбачите?
 - Які технології бази даних ви обрали б і чому?
4. Проаналізуйте будь-який відомий електронний словник і визначте:
- які можливості інтерфейсу доступні **звичайному користувачеві**,
 - які – **лексикографу**,
 - які – **адміністратору**.

Питання для поглибленого вивчення матеріалу

1. Українська комп'ютерна лексикографія: сучасний стан і перспективи.
2. Історія розвитку світової та української комп'ютерної лексикографії.
3. Основні поняття комп'ютерної лексикографії.
4. Теорія лексикографічних систем як основа створення електронних словників.
5. Технологічні інструменти для реалізації електронного лексикографічного проєкту СУМ 20 та інших цифрових словників.

5. МОВА ЯК ОБ'ЄКТ КОМП'ЮТЕРНОЇ ЛЕКСИКОГРАФІЇ

План заняття:

5.1. Моделі репрезентації мови в словнику

5.2. Формати збереження лексикографічної інформації

5.3. Поняття концептуальної моделі словника

Мови як об'єкт словникового опису відзначаються унікальними, специфічними рисами, що зумовлені їхньою структурою, історією розвитку та культурним контекстом, а також спільними або відносно спільними властивостями, притаманними різним мовам. Це безпосередньо впливає на вибір моделі опису, адже кожна з них має свої переваги й обмеження. При цьому важливим чинником є не лише особливості самої мови та картини світу, яку вона відображає, а й традиції певної лексикографічної школи (української, англійської, німецької, французької тощо). У результаті в комп'ютерній лексикографії поєднуються як універсальні підходи до опису мовних систем, так і специфічні, притаманні окремій науковій традиції. Вибір моделі репрезентації мови стає визначальним чинником у розробленні форматів збереження даних, побудові концептуальної моделі словника та формуванні інтерфейсу, який забезпечує доступ до інформації.

5.1. Моделі репрезентації мови в словнику

У словникарській практиці застосовують дві основні моделі репрезентації мови – **універсальну** та **специфічну**. Кожна з них відповідає певним завданням, що постають перед словником, і визначає характер подання лексичного матеріалу. Вибір тієї чи іншої моделі зумовлюється передусім авторською установкою словника, у якій конкретизуються такі параметри:

- тип словника та його призначення (одномовний, двомовний, багатомовний; перекладний, тлумачний, навчальний тощо);
- лексичний різновид описуваної мови (мова загального вжитку чи мова спеціальної сфери – наукової, технічної, юридичної тощо);
- набір лексикографічних параметрів, що визначають структуру і наповнення словникової статті;

– особливості картин світу та понятійних систем, які впливають на ступінь еквівалентності між мовними одиницями різних мов.

Отже, авторська установка задає рамки для майбутньої концепції словника і водночас виступає критерієм вибору моделі репрезентації: універсальної – для уніфікованого опису кількох мов, або специфічної – для відображення індивідуальних властивостей окремої мовної системи.

Універсальна модель передбачає єдину систему опису для двох або більше мов, що особливо важливо у перекладних словниках та багатомовних термінологічних глосаріях. Її переваги полягають у тому, що:

- забезпечується охоплення кількох мов в межах одного словника;
- застосовується уніфікований набір параметрів у словниковій статті;
- підтримується типова структура опису для всіх репрезентованих мов.

Така модель репрезентації мови застосовується переважно у багатомовних та перекладних словниках, де головним завданням є забезпечення максимальної відповідності між одиницями різних мов. Така модель характерна для багатомовних енциклопедій (*Wiki-ресурси*), тематичних словників (наприклад, [*Мультимедійний словник з інфомедійної грамотності*](#), [*Лексикон польської та української активної фразеології*](#)) та термінологічних глосаріїв (*Багатомовний словник-довідник зі зварювання на CD*, [*Interactive Terminology for Europe*](#), [*Terminology of European education and training policy*](#)). Схема реалізації універсальної моделі може мати кілька варіантів, як це показано у таблиці 5.1. Це далеко не всі можливі реалізації, і в деяких випадках вони можуть поєднуватися.

Таблиця 5.1 – Схеми реалізації універсальної моделі

Варіант	Схема	Характеристика
(а)	Лема_мова ₁ → Лема_мова ₂	Найпростіша форма перекладного словника без дефініцій; зв'язок установлюється лише між реєстровими одиницями.
(б)	Лема_мова ₁ : Дефініція_мова ₁ → Лема_мова ₂ : Дефініція_мова ₂	Двосторонній переклад з урахуванням дефініцій, що забезпечує більшу точність відповідників.
(в)	**Лема_мова ₁ : Дефініція_мова ₁ → Лема_мова ₂	Лема_мова ₁ **
(г)	Лема_мова ₁ + Варіації : Дефініція_мова ₁ → Лема_мова ₂ + Варіації : Дефініція_мова ₂	Найповніша модель, що охоплює всі формальні та семантичні варіанти лексеми.

На рисунку 5.1 показано приклад реалізації універсальної моделі у *Лексиконі польської та української активної фразеології*, укладений в Українському мовно-інформаційному фонді. У цьому ресурсі словникова стаття складається з двох частин – польської та української, які уніфіковані за структурою та набором лексикографічних параметрів. У ньому реалізовано схему (г): у кожній частині подано лему, дефініцію та ілюстративний матеріал. Дефініції в обох частинах збігаються, але приклади вживання відрізняються, що дає змогу показати національну специфіку вживання фразеологізмів у різних мовних контекстах.

Polsko-ukraiński słownik jednostek frazeologicznych

Польсько-український словник фразеологічних одиниць

A B C Ć D E Ę F G H I J K L Ł M N Ń O Ó P (Q) R S Ś T U (V) W (X) Y Z Ż Ź

Українсько-польський словник фразеологічних одиниць

Ukraińsko-polski słownik jednostek frazeologicznych

A B B V G Ą D E Ę Ж З И І Ї Й К Л М Н О П Р С Т У Ф Х Ц Ч Ш Щ Ь Ю Я

a nuż, widelec kol. 'wtedy, gdy ktoś ma nadzieję, że coś może się udać' <i>Dziś można wygrać piętnaście milionów w totka. Kup trzy losy.</i> <i>A nuż, widelec się uda.</i>	чого́ дóброго 'тоді, коли хтось вважає щось імовірним, можливим' <i>Нам не завадить заздалегідь зібрати валізу. Чого доброго щось забудемо</i>
albo rybki, albo akwarium / albo rybka, albo pipka kol. 'wtedy, gdy trzeba się zdecydować na jedną z dwóch rzeczy' <i>Nie stać nas na dwa wypadki w te ferie. Zdecyduj się: albo góry, albo morze – albo rybka, albo akwarium.</i>	одне́ з двох 'тоді, коли потрібно визначитися з однією з двох речей' <i>Доведеться вибирати одне з двох: або кар'єра, або стосунки.</i>
ani ziębi, ani grzeje (kogoś) kol. 'wtedy, gdy ktoś ma obojętny stosunek do czegoś' <i>Ten film ani mnie ziębi, ani grzeje.</i>	ні хо́лодно ні жа́рко кол. 'тоді, коли комусь байдуже до чогось' <i>Мені від усіх цих розмов ні холодно ні жарко.</i>

Рис. 5.1 – Універсальна модель у двомовному фразеологічному словнику

Ще одним прикладом реалізації універсальної моделі є *Багатомовний словник зі зварювання* (рис. 5.2), у якому відображено принцип опису мови спеціального призначення за схемами (а) – ліворуч і (г) – праворуч. Словникова стаття у випадку схеми (г) поділяється на кілька зон, де кожна з них присвячена опису заголовкового терміна відповідною мовою. Кожна зона включає заголовковий термін і його дублетні варіанти, дефініцію та позначення тематичної категорії. Створення такого спеціального словника з універсальною моделлю стало результатом співпраці лексикографів і фахівців зі зварювання, що сприяло стандартизації термінології та уніфікації словникового опису.

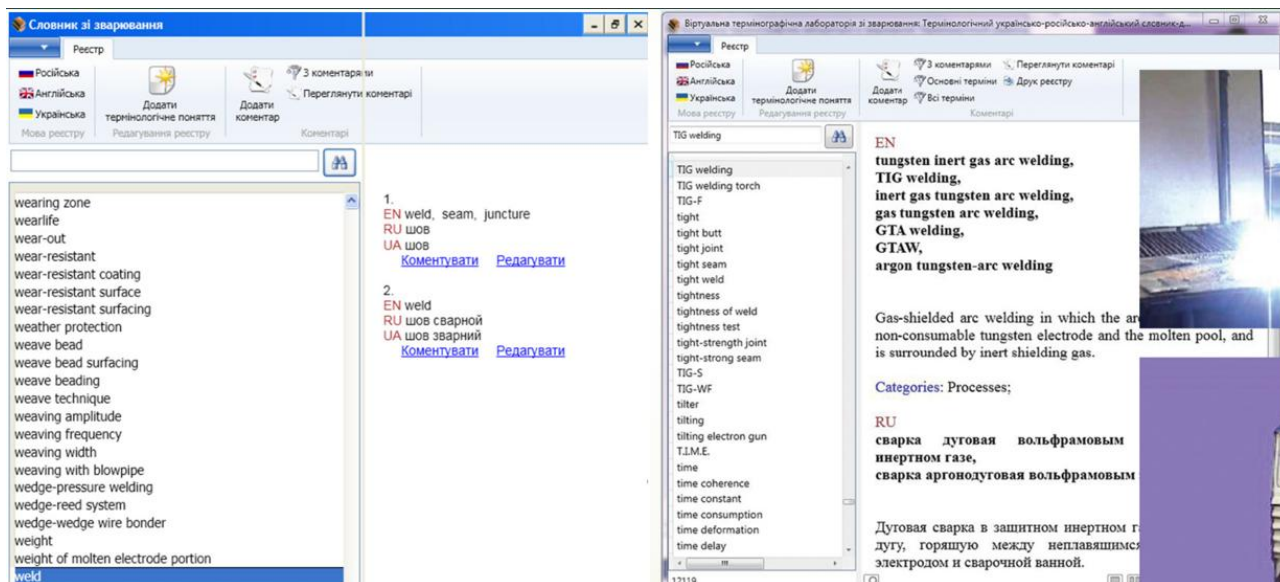


Рис. 5.2 – Універсальна модель у багатомовному словнику зі зварювання

Перехідний варіант між універсальною та специфічною моделями представлено в *Багатомовному термінологічному глосарії з архівної справи (Multilingual Archival Terminology)*, показаному на рисунку 5.3. У таких глосаріях вхідне слово або словосполучення має розширену описову частину, тоді як у перекладній частині подаються еквіваленти різними мовами. Цей тип репрезентації передбачає, що основна увага приділяється вхідній мові, тоді як інші розглядаються допоміжно. Водночас обов'язковою умовою є повна еквівалентність між термінами в усіх мовах, що забезпечує узгодженість і точність перекладу.

archival bond (Edit Term)

Add Definition(s)

Definitions:

noun

1. The network of relationships that each record has with the records belonging in the same records aggregation.

Citation:

[Duranti, Luciana, Terry Eastwood, and Heather MacNeil. Preservation of the Integrity of Electronic Records. Dordrecht, Netherlands: Kluwer Academic Publishers, 2002.]

Region: Canada

Related:

Arabic: رابطَة أرشيفية(ar),

Belarusian: ўзаемасувязь(be), ўнутрыфндавая ўзаемасувязь(be),

Croatian: arhivska veza(hr), arhivska veza(hr),

Dutch: archival bond(nl),

English: archival bond(en),

Farsi: همبستگی آرشیوی(fa),

French: archival bond(fr),

German: [Archivische Verbindung](de),

Greek: Αρχείο δεσμόс(el),

Italian: vincolo archivistico(it),

Japanese: アーカイバル・ボンド(ja),

[Link Definition](#) | [Edit Definition](#) | [Report an Error](#)

Рис. 5.3 – Універсальна модель у *Multilingual Archival Terminology*

Як можна бачити, перевагою універсальної моделі є її стандартизованість і зручність для багатомовної інтеграції: вона забезпечує уніфікований опис мовних одиниць незалежно від їхньої національної специфіки. Завдяки цьому така модель ефективно використовується у великих електронних базах даних, термінологічних ресурсах та інтегрованих лексикографічних системах, де важливими є узгодженість і технічна сумісність даних.

Водночас обмеження універсальної моделі полягає в тому, що вона не завжди здатна адекватно відобразити національно-культурні або семантичні відтінки слів, особливо у випадках безеквівалентної лексики – як у мові загального, так і спеціального призначення. Так, на рисунку 5.4 показано словникову статтю *адвокат*, у якій подано численні англійські еквіваленти до українського терміна в [Українсько-англійському юридичному словнику](#) (укладачі І. Борисенко, В. В. Саєнко, Н. М. Конончук, Т. І. Конончук). Наявність різних відповідників (*lawyer, attorney, barrister, solicitor* тощо) свідчить про розбіжності в поняттєвих системах юридичної сфери української та англійської культур. Це один із чинників, через які універсальну модель неможливо застосувати без суттєвої адаптації до специфіки кожної мови.

АДВОКАТ

ім counsel; lawyer; legal adviser; [англ тж] (баристер; адвокат вищого рангу, що виступає в суді) barrister; (повірений; соліситор) solicitor; [рідк] advocate; [амер] attorney (at law); counsellor (at law) • бути представленим адвокатом to be represented by a counsel (by a lawyer) • викликати адвоката to summon a counsel • виступати як головний ~ сторони to lead in the case • відмовлятися від адвоката to waive a counsel (legal aid / assistance / defence); (захищати себе в суді) to act as attorney pro se (as one's own attorney / counsel / lawyer); conduct one's own defence; serve one's own counsel • захищати себе (захищатися) через адвоката to defend oneself by a counsel • консультуватися в на to consult a counsel; take a counsel's opinion • наймати на to hire (retain) a counsel (a lawyer) • позбавляти звання на to disbar • призначати адвоката to appoint (assign) a counsel (a lawyer) • стати ~ом to be admitted (called) to the Bar; become a lawyer; go to (join) the Bar • колегія ~ів the Bar; [амер] the Bar Association • право на представництво ~ом right to a counsel • призначення на appointment (assignment) of a counsel (of a lawyer) •

Рис. 5.4 – Специфічна модель у перекладному юридичному словнику

Недоліки універсальної моделі зумовлюють потребу у використанні більш гнучких і диференційованих підходів до опису мовних одиниць, здатних враховувати структурну, семантичну й культурну своєрідність кожної мови. Саме таку можливість забезпечує специфічна модель репрезентації мови, яка

орієнтована на детальний, внутрішньомовний опис і відображає індивідуальні риси мовної системи.

На відміну від універсальної, **специфічна модель** орієнтована на докладний опис конкретної мови з урахуванням її внутрішніх закономірностей і культурного контексту. Вона більш адекватно відображає унікальність мовної системи, проте обмежує можливості прямого зіставлення з іншими мовами. **Специфічна модель** ґрунтується на принципі максимально точного відтворення особливостей окремої мови. Вона найчастіше застосовується у тлумачних словниках, енциклопедіях та спеціалізованих лексикографічних виданнях. До її переваг зараховують:

- можливість врахування внутрішньої структури мови та її граматичних і семантичних особливостей;
- адекватне відображення національно-культурної специфіки мовної картини світу;
- гнучкість у формуванні словникових статей, що дозволяє розкривати додаткові відтінки значень і контексти вживання.

Для прикладу розглянемо реалізацію специфічних моделей опису дієслова *бути* у двох словникових джерелах – *СУМ-20* та *Merriam-Webster's Dictionary* (рис. 5.5). Порівняння їхніх словникових статей засвідчує суттєву різницю у підходах до репрезентації мовного матеріалу. Перший момент, у *СУМ-20* аналізоване слово трактується як лексема, тобто як носій лексичного значення з визначеним набором словоформ, що належить до системи української мови. Натомість у *Merriam-Webster's Dictionary* воно розглядається передусім як грамема – допоміжне дієслово, що виконує граматичні функції у структурі висловлення.

Ці відмінності у тлумаченні демонструють різні лексикографічні традиції та моделі опису мови. Українська лексикографічна школа, представлена зокрема *СУМ-20*, дотримується структурно-семантичного підходу, у межах якого слово розглядається як єдність форми й змісту, а головною метою є розкриття його значення в системі мови. Для англomовної традиції, яку репрезентує *Merriam-*

Webster's Dictionary, характерний функціонально-граматичний підхід, що орієнтований на роль слова у синтаксичній структурі та його граматичну поведінку. Таким чином, відмінність між двома моделями полягає не лише у способі лексикографічного представлення, а й у самій концепції того, що становить «мовну одиницю» — лексему чи граему.

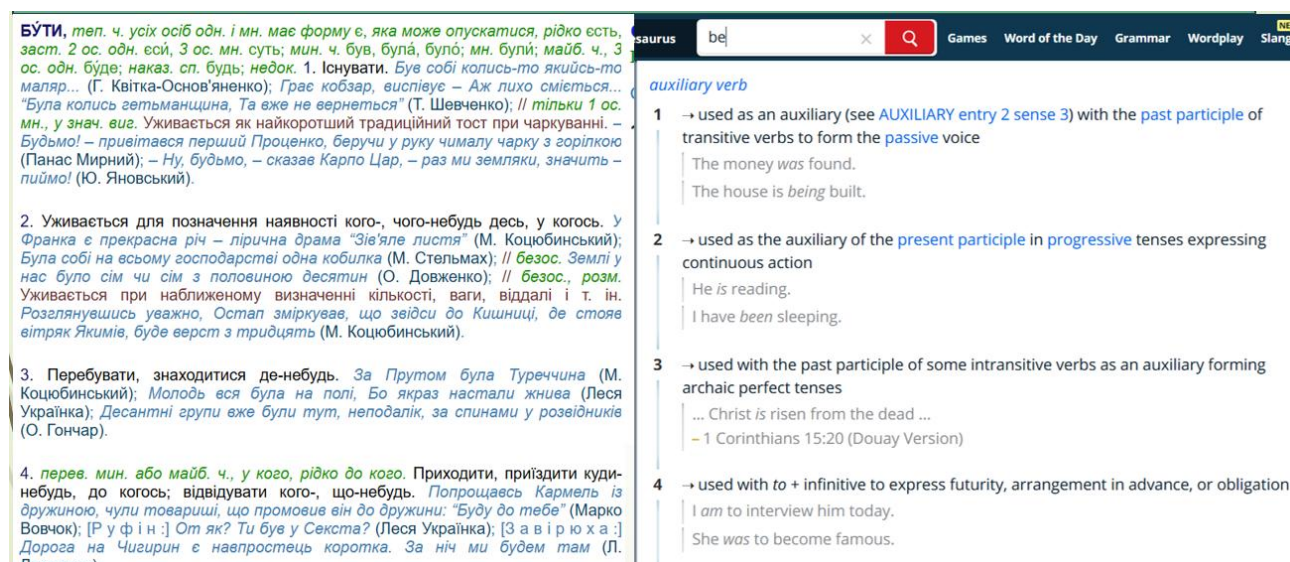


Рис. 5.5 – Специфічна модель у *SUM-20* та *Merriam-Webster's Dictionary*

У контексті електронної лексикографії вибір моделі опису безпосередньо впливає на структуру бази даних і архітектуру пошукової системи. Якщо у випадку специфічної моделі (як у *SUM-20*) головним є збереження лексико-семантичних зв'язків між словоформами, значеннями, фразеологізмами та прикладами, то у моделі типу *Merriam-Webster's* основну увагу приділено морфологічній і синтаксичній маркованості одиниць, що дозволяє швидко реалізовувати граматичний пошук і формувати гнучкі мовні корпуси. Відповідно, різні лексикографічні моделі визначають не лише формат представлення мовних даних, а й способи взаємодії користувача з ними в електронному середовищі.

Другий момент полягає у відмінності функцій ілюстративного матеріалу. У *SUM-20* приклади виконують доказову та контекстуально-репрезентативну функції: вони демонструють уживання слова в аутентичних джерелах, що засвідчують реальне функціонування лексеми в живому мовленні або художньому тексті. Завдяки цьому ілюстрації виступають не лише як елемент тлумачення, а й як засіб фіксації мовної норми та мовної картини світу. Натомість у *Merriam-Webster's Dictionary* приклади мають переважно навчально-прикладне

призначення: вони демонструють, як правильно побудувати речення з використанням заголовкового слова у певному значенні, тобто допомагають користувачеві оволодіти моделями уживання. У такому підході відбивається орієнтація англomовної лексикографічної традиції на практичну сторону слововживання й комунікативну ефективність, тоді як українська лексикографічна школа тяжіє до дескриптивності й нормативності в подачі матеріалу.

Разом із тим, специфічна модель репрезентації мови має і свої обмеження. По-перше, вона ускладнює пряме порівняння лексичних одиниць різних мов, оскільки опис ґрунтується на унікальних граматичних і семантичних характеристиках кожної окремої мовної системи. Це робить складним автоматичне співвіднесення лексем у багатомовних ресурсах. По-друге, така модель потребує значно більших лінгвістичних і технічних ресурсів для укладання та опрацювання, адже передбачає детальний опис не лише формальної, а й когнітивної, прагматичної та культурної складових мовної одиниці. По-третє, вона може бути надмірно деталізованою для швидкого практичного використання, зокрема у навчальних або перекладних словниках, де користувач очікує стислого і функційно орієнтованого тлумачення. Отже, специфічна модель забезпечує глибину й точність опису, але ціною складнішої реалізації та обмеженої універсальності.

Виходячи з вищесказаного, універсальна і специфічна моделі репрезентації мови у словнику відображають різні підходи до організації словникового матеріалу. Якщо універсальна модель зорієнтована на уніфікованість опису кількох мов, то специфічна, навпаки, акцентує на індивідуальних властивостях кожної мовної системи. Подальший аналіз дає змогу детальніше розглянути переваги й обмеження кожної з моделей. У таблиці 5.2. стисло викладено основні характеристики універсальної та специфічної моделей.

Таблиця 5.2 – Характеристики моделей репрезентації мови

Критерій	Універсальна модель	Специфічна модель
Сфера застосування	Перекладні словники, багатомовні глосарії	Тлумачні словники, енциклопедії, спеціалізовані словники
Основний принцип	Єдиний набір параметрів для кількох мов	Максимальне відображення особливостей конкретної мови

Переваги	– Охоплює дві чи більше мов; – Уніфікована структура статей; – Зручність міжмовного зіставлення	– Враховує граматичні й семантичні особливості мови; – Відображає національно-культурну специфіку; – Гнучкість у поданні значень і контекстів
Обмеження	– Можлива лише для абсолютно еквівалентних слів; – Не працює для безеквівалентної лексики; – Ігнорує національно-культурні розбіжності	– Ускладнює зіставлення між мовами; – Потребує більших ресурсів для укладання; – Іноді надто деталізована для швидкого використання
Мета застосування	Забезпечення універсального зіставлення мов	Повне й точне відображення мовної системи

Вибір між універсальною та специфічною моделями репрезентації мови у словнику не є випадковим – він визначається метою словникового проєкту, типом мовного матеріалу та цільовою аудиторією. Якщо універсальна модель забезпечує зручність багатомовної інтеграції та стандартизації, то специфічна дає змогу глибше відобразити структурно-семантичні особливості мови, культурні маркери і функціональні відтінки значень. У контексті електронної лексикографії цей вибір набуває ще більшого значення, адже від нього залежить архітектура бази даних, формати подання інформації та можливості пошукового інтерфейсу.

Питання для самоконтролю

1. У чому полягає суть універсальної моделі лексикографічного опису?
2. Які типи словників найчастіше реалізують універсальну модель?
3. Які переваги універсальної моделі у процесі порівняння лексики різних мов?
4. Чому універсальну модель вважають зручнішою для перекладачів та студентів?
5. Які обмеження притаманні універсальній моделі під час роботи з багатозначними словами або фразеологізмами?
6. У чому полягає головна відмінність специфічної моделі від універсальної?
7. Які переваги має специфічна модель у плані глибини опису слова?
8. Які ресурси потребує укладання словника за специфічною моделлю?
9. Які чинники ускладнюють використання специфічної моделі в навчальних цілях?

10. Як специфічна модель впливає на можливість порівняння лексем різних мов?

11. Яка модель – універсальна чи специфічна – є ефективнішою для академічних досліджень, а яка – для практичного користування?

12. Як поєднання обох моделей може підвищити якість сучасного тлумачного словника?

13. Які словники української та англомовної традицій ілюструють універсальну, а які – специфічну модель?

14. У яких випадках ілюстрації до словникової статті виконують навчальну, а в яких – демонстраційну функцію?

15. Як цифрові технології можуть допомогти подолати обмеження обох моделей у майбутніх електронних словниках?

Завдання для самоперевірки

1. Проаналізуйте наведені електронні словники та визначте, яку модель лексикографічного опису (універсальну чи специфічну) реалізовано у кожному випадку.

№	Назва електронного словника	Тип моделі (універсальна / специфічна)	Аргументація (2–3 речення)
1	Collins English-German Dictionary		
2	Англо-український словник		
3	Словник «Горих»		
4	Merriam-Webster's Dictionary		

Обґрунтуйте свій вибір, зазначивши основні ознаки моделі: тип подання даних, рівень деталізації, кількість мов, наявність перекладних еквівалентів, ілюстрацій, граматичних або енциклопедичних компонентів тощо.

2. Проаналізуйте наведені типи словників і визначте, яка модель – універсальна чи специфічна – є для них більш придатною. Обґрунтуйте свій вибір, зважаючи на мету словника, тип опису та очікувану аудиторію.

№	Тип словника	Оберіть модель (У / С)	Обґрунтування
1	Тлумачний словник сучасної української мови		
2	Перекладний англо-український технічний словник		
3	Багатомовний термінологічний глосарій ЄС		
4	Етимологічний словник індоєвропейських мов		

5	Навчальний словник для вивчення іноземної мови		

Рекомендація. Після виконання спробуйте сформулювати коротке правило: які параметри (мета, обсяг, тип лексики, кількість мов, користувач) визначають вибір між універсальною та специфічною моделлю.

3. Ознайомтеся з наведеними нижче прикладами опису слова в різних словниках і виконайте завдання після таблиці.

№	Джерело	Фрагмент словникової статті	Модель (У / С)	Пояснення
1	Collins English-German Dictionary	<i>advocate</i> – <i>der Anwalt, die Anwältin</i>		
2	Словник української мови в 20 томах (СУМ-20)	Бути, дієсл. 1. Існувати. <i>Був собі колись-то якийсь-то маляр...</i>		
3	Multilingual Archival Terminology	<i>record</i> – англ. → <i>documento</i> (ісп.), <i>Akte</i> (нім.), <i>document</i> (фр.),		
4	Merriam-Webster's Dictionary	to be (verb) – 1 a: to equal in meaning: have the same connotation as: symbolize		

(1) Визначте, до якої моделі (універсальної чи специфічної) належить кожний приклад.

(2) Обґрунтуйте свій вибір (звертайте увагу на наявність дефініцій, перекладів, прикладів уживання, граматичних характеристик).

(3) Сформулюйте узагальнення:

- які риси характерні для універсальної моделі,
- які – для специфічної,
- у яких випадках можлива їхня комбінація.

5.2. Формати збереження лексикографічної інформації

Словник у цифровому середовищі може підтримувати різні формати збереження інформації, кожен з яких має власне призначення, переваги та обмеження. До найпоширеніших належать:

– **PDF:** формат, що походить із настільних видавничих систем і призначений насамперед для збереження статичного вигляду словника (макета, шрифтового оформлення, графіки);

- **HTML**: мова розмітки для структурування контенту у веб-середовищі, яка забезпечує інтерактивність і гіперпосилання між словниковими статтями;
- **XML**: універсальний формат представлення, зберігання й обміну даними, придатний як для машинної обробки, так і для читання людиною;
- **Формати баз даних**: структури організації та взаємодії даних, що забезпечують функціонування словникових систем на рівні програмних та серверних рішень (детальніше розглянуті у підрозділі 4.3).

У цьому підрозділі буде детально проаналізовано особливості застосування цих форматів в електронному словникарстві, їхню роль у формуванні структури словникових ресурсів, а також наведено приклади використання у сучасних лексикографічних проєктах.

Серед форматів зберігання лексикографічної інформації важливу роль відіграють **PDF** і **HTML**, які використовуються переважно для статичного представлення словникових матеріалів. Їх основне завдання – забезпечити *зручність читання* та *візуальне відтворення* структури словникової статті, а не складну інтерактивну взаємодію з даними.

Формат PDF (Portable Document Format) спочатку створювався як засіб збереження макету друкованого видання у цифровій формі. Тому його перевагою є висока точність відтворення поліграфічного дизайну – шрифтів, колонок, нумерації, таблиць, ілюстрацій тощо. У лексикографії цей формат часто застосовують для репрезентації друкованих словників у цифровому вигляді, наприклад:

- *Oxford English Dictionary* (деякі архівні версії), *German-English Bilingual Visual Dictionary*, *Large English Dictionary* та інші;
- *Словник української мови в 11 томах* (у вигляді PDF-реплік на офіційних платформах), *Словник української біологічної термінології*, *Словник фізичних термінів-синонімів* тощо.

Перевагою PDF є збереження автентичного вигляду сторінки, що важливо для наукового цитування та для роботи з рідкісними чи історичними словниками. Водночас його недолік полягає у відсутності машиночитаності: дані у PDF важко

обробляти автоматично, здійснювати пошук за структурними елементами (напр., за лемою чи граматичним тегом).

Формат HTML, на відміну від PDF, формат HTML (HyperText Markup Language) створено для структурування інформації у веб-середовищі. HTML забезпечує розміщення тексту за тегамі, що описують логічну структуру сторінки – заголовки, абзаци, списки, таблиці, посилання, мультимедіа тощо. У контексті словникарства HTML дає можливість:

- публікувати онлайн-словники з гіперпосиланнями між статтями;
- організовувати внутрішню навігацію (напр., між граматичними формами);
- додавати мультимедійні елементи, такі як аудіо вимови, ілюстрації, приклади вживання тощо.

Для задання відповідних стилів відображення кожного структурного елемента словникової статті, а також її макету, може використовуватися **таблиця каскадних стилів CSS (Cascading Style Sheets)**. Вона визначає кольори, шрифти, відступи, вирівнювання, тобто зовнішній вигляд лексикографічного контенту. Таким чином, з одного боку забезпечується чітке виокремлення структури словникової статті, а з іншого – зручне сприйняття матеріалу користувачем. Одночасне використання **HTML** і **CSS** активно застосовується в сучасних електронних словниках.

Наприклад, у словнику «Turbolex» кожна словникова стаття представлена у вигляді окремого HTML-документа, до якого прив'язана таблиця стилів. Це дає змогу автоматично контролювати форматування та забезпечувати візуальну узгодженість усіх елементів словника – від лем до перекладних еквівалентів. Іншими прикладами лексикографічних проєктів, що підтримують формат мови розмітки, є *Cambridge Dictionary Online*, Словник української мови онлайн (sum.in.ua), *Glosbe*, *Wiktionary* та інші ресурси. Усі вони демонструють можливості гіпертекстової організації словника, коли статті взаємопов'язані посиланнями, а структура даних легко адаптується під різні пристрої та формати відображення.

Основна перевага **HTML** – гнучкість і інтерактивність, що дозволяє створювати словники з динамічним пошуком, фільтрацією, мультимедіа-елементами тощо. Водночас дані у цьому форматі не мають чіткої лексикографічної розмітки, якщо не застосовано додаткових схем або стандартів, таких як XML-розширення чи TEI (Text Encoding Initiative). У таблиці 5.3 подано порівняльну характеристику форматів **PDF** і **HTML**, що відображає їхні відмінності за основними параметрами структурування та використання у словникарстві. Більш розгорнуту порівняльну характеристику подано в додатку Д.11.

Таблиця 5.3 – Порівняння PDF і HTML у лексикографічному контексті

Параметр	PDF	HTML
Призначення	Репрезентація друкованого словника у цифровій формі	Інтерактивне онлайн-представлення
Структура	Візуальна, сторінкова	Логічна, гіпертекстова
Пошук	Обмежений, за простим текстом	Розширений, за тегамі та полями
Оновлення даних	Ускладнене	Легке, динамічне
Машиночитаність	Низька	Висока (за наявності розмітки)
Застосування	Архівні, наукові, друковані репліки	Сучасні онлайн-словники

Отже, формати **PDF** і **HTML** виконують різні ролі у цифровій лексикографії: **PDF** – це носій фіксованого, поліграфічно точного представлення словникового тексту (архіви, репліки, друк), **HTML** – робочий формат для інтерактивних онлайн-ресурсів, що забезпечує гнучкість, пошук, мультимедіа й інтеграцію з іншими сервісами. Для повноцінної електронної лексикографії найбільш ефективним є комбінований підхід: зберігати (архівувати) макети у **PDF**, а для повсякденної роботи й доступу використовувати **HTML** (переважно – у поєднанні з семантичною розміткою **XML/TEI** для машинного опрацювання).

XML (Extensible Markup Language) як універсальний формат для зберігання та обміну словниковими даними – це універсальна мова розмітки, розроблена для структурування, зберігання та обміну даними у машиночитаному форматі. На відміну від **HTML**, який призначений переважно для візуального відображення контенту, **XML** зосереджується на структурі та змісті. У лексикографії **XML** став основним форматом представлення словникових даних, оскільки забезпечує

їхню гнучкість, формалізованість і сумісність між різними системами. Нижче викладено принципи використання **XML** у комп'ютерній лексикографії.

XML-файли складаються з **тегів**, які описують логічні частини словникової статті. Кожен тег має відкривальний і закривальний елемент (наприклад, `<entry>` ... `</entry>`), що обмежує певну зону тексту. Усередині тегів можуть бути вкладені інші теги – це дає можливість створювати ієрархічну структуру опису. На малюнку 5.6. зображено фрагмент XML-опису для заголовкового слова *гора*.

```
<entry id="001">
  <form>
    <orth>гора</orth>
    <pos>іменник</pos>
  </form>
  <sense n="1">
    <def>Підвищення земної поверхні з чітко вираженою вершиною.</def>
    <example>Карпати – мальовничі гори України.</example>
  </sense>
</entry>
```

Рис. 5.6 – Словникова стаття у форматі XML

Така структура дає змогу точно відокремлювати формальну, семантичну й ілюстративну частини словникової статті, а також легко переносити ці дані в інші системи (онлайн-словники, корпуси, бази даних). Для уніфікації опису словникових даних у межах XML-формату розроблено кілька стандартів, які визначають обов'язкові та рекомендовані елементи опису.

TEI (Text Encoding Initiative) – міжнародний стандарт розмітки текстових та лексикографічних даних. TEI має спеціальний розділ *Dictionaries*, який описує структуру словникової статті через елементи `<entry>`, `<form>`, `<sense>`, `<gramGrp>`, `<cit>` тощо. Наприклад:

```

<entry xml:id="бігти">
  <form type="lemma">
    <orth>бігти</orth>
    <gramGrp>
      <pos>дієслово</pos>
    </gramGrp>
  </form>
  <sense n="1">
    <def>Пересуватися швидко, роблячи короткі кроки.</def>
    <cit type="example">
      <quote>Діти бігли до школи.</quote>
    </cit>
  </sense>
</entry>

```



Перевага TEI полягає у її сумісності з бібліотечними, архівними і лінгвістичними форматами, що робить її універсальною платформою для академічних словників.

LMF (Lexical Markup Framework) – стандарт ISO 24613, розроблений спеціально для представлення лексичних ресурсів. **LMF** описує модель лексикону через основні об'єкти: *LexicalEntry*, *Form*, *Sense*, *Example* тощо. Структура подібна до TEI, але більш формалізована й орієнтована на автоматизовану обробку (наприклад, у корпусних системах чи нейронних моделях перекладу).

В українському електронному словникарстві мова **XML** використовується в кількох масштабних проєктах:

1. *СУМ-20 (Словник української мови, 20 томів, електронна версія)* – внутрішня структура словникових статей реалізована у форматі **XML**, що забезпечує можливість пошуку за граматичними ознаками, значеннями, прикладами та джерелами.

2. *Словник української мови онлайн (sum.in.ua)* — дані імпортовано з **XML**-бази, створеної на основі **TEI**-моделі.

3. *Національний корпус української мови* використовує **XML** як базовий формат для зберігання та обробки корпусних і лексикографічних даних.

4. *Словник українського молодіжного сленгу* також застосовує спрощену **TEI**-модель з метою структурування реєстрових одиниць і прикладів уживання.

Використання мови **XML** у комп'ютерній лексикографії забезпечує низку переваг, зокрема:

- чітку структуруваність та можливість обробки даних різними програмами;
- масштабованість – легке розширення опису без втрати сумісності;
- інтероперабельність – можливість обміну даними між словниками, корпусами, перекладацькими системами;
- підтримку семантичного пошуку (за грамемами, частинами мови, тематичними групами тощо);
- інтеграцію з сучасними веб-стандартами (RDF, JSON-LD, TEI Publisher, Lexonomy).

XML став універсальною платформою для опису лексикографічних даних, що поєднує гнучкість людського читання з точністю машинної обробки. Його стандарти – **TEI** та **LMF** – заклали основу для побудови сучасних електронних словників, корпусів і термінологічних систем. У поєднанні з **HTML** і базами даних **XML** забезпечує єдину інформаційну екосистему сучасної цифрової лексикографії. Якщо **TEI** можна розглядати як «мову розмітки для людей», яка фіксує лінгвістичну структуру у текстовій формі, то **LMF** — це «модель для машин», яка забезпечує логічну і семантичну інтерпретацію лексичних даних у цифрових системах. Порівняльна характеристика стандартів **TEI** та **LMF** подано в додатку Д.12.

Показовим є багатий досвід Українського мовно-інформаційного фонду НАН України у комплексному використанні різних форматів для конвертування паперових словників (або електронних словників у форматі PDF) у базу даних за допомогою власної технології парсингу. На малюнку 5.7 наведено технологічну схему конвертації PDF-версії *Словника української біологічної термінології*. Ця схема демонструє послідовні кроки, на яких використовується той чи інший формат. Розглянемо детально роль форматів на кожному з етапів.

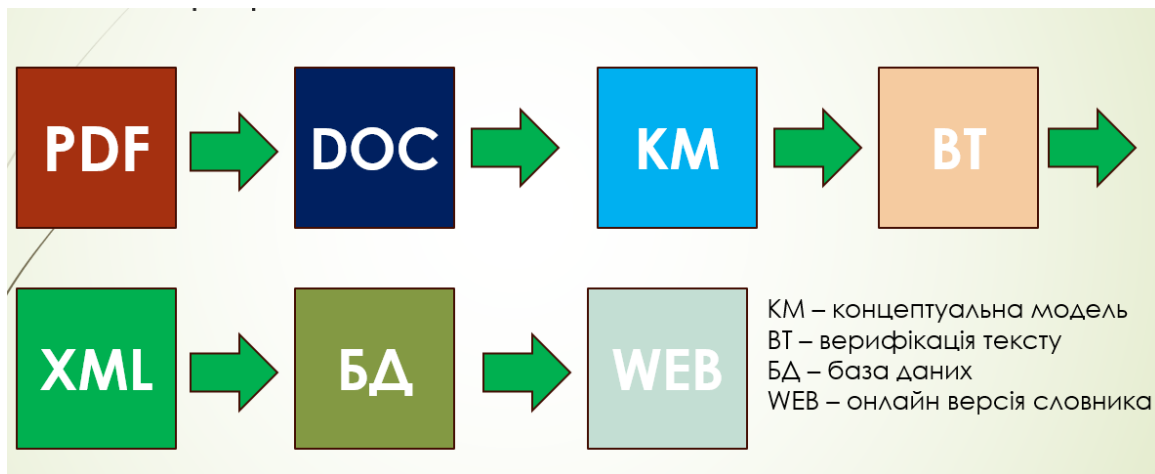


Рис. 5.7 – Технологічна схема створення словника на основі PDF формату

Етап 1. Оцифрування та розпізнавання тексту. Початковим джерелом є друкована або PDF-версія словника. За допомогою системи OCR (наприклад, *ABBYY FineReader*) виконується розпізнавання тексту, у результаті чого створюється текстовий файл у форматі Word із базовим форматуванням (шрифти, абзаци, нумерація). На цьому етапі структура словникової статті ще не позначена тегами, тому її потрібно формалізувати.

Етап 2. Робота з текстовим файлом Word. На цьому етапі у сканований текст вносяться правки, здійснюється нормалізація розпізнаного матеріалу (виправлення помилок OCR, приведення до єдиних форматних норм). Після цього, за допомогою спеціальних програмних засобів, текст розмічається HTML-тегами для збереження оригінальної структури словникових статей і позначення елементів (заголовне слово, дефініція, граматична характеристика, приклади тощо). Отриманий HTML-документ (рис. 5.8) стає основою для побудови концептуальної моделі словникової статті на наступному етапі.

```

<B>а#апа</B>· (<I>рос.</I>· аа#па, <I>англ.</I>· аара)· 1· Безлісий,· дуже· обводнений,·
  грядково-мочарний вид болота на півночі Євразії.¶
<B>аба#ка</B>· (<I>рос.</I>· аба#ка, <I>англ.</I>· abacus)· 1· Багаторічна трав'яниста
  рослина родини бананових, яку культивують у тропіках; використовують у
  народному господарстві для виготовлення канатів, мотузок, корабельних снастей
  тощо.¶
<B>абаксія#льний</B>· (<I>рос.</I>· абаксия#льный, <I>англ.</I>· abaxial)· 1· Який
  стосується нижньої поверхні листового органа.¶
<B>абера#нтний</B>· (<I>рос.</I>· абера#нтный, <I>англ.</I>· aberrational)· 1·
  Незвичайний, що відхилився від норми· Аберантні приклади паразитних відносин
  представлені гіперпаразитизмом.¶
<B>абдо#мен</B>,· -а· (<I>рос.</I>· абдо#мен, <I>англ.</I>· abdomen)· 1· Протилежна від
  спини частина тіла людини і тварини, в якій містяться живіт, черевце, задній відділ
  тіла у членистоногих тварин.¶
<B>абдоміна#льний</B>· (<I>рос.</I>· абдомина#льный, <I>англ.</I>· abdominal)· 1· Який

```

Рис. 5.8 – Фрагмент словника, розміченого мовою HTML

Етап 3. Побудова концептуальної моделі словника. На цьому етапі створюється концептуальна модель словника (лексикографічної системи) – формальна модель, що на абстрактному рівні відображає його структуру та взаємозв'язки між основними елементами. Вона визначає, які саме об'єкти й атрибути (заголовкове слово, граматичні характеристики, дефініції, приклади, ілюстрації тощо) мають бути доступними користувачеві через інтерфейс бази даних. У результаті лінійний текст перетворюється на ієрархічну структуру, де кожен компонент має свою позицію у системі відношень. Саме на цьому етапі готується основа для подальшої розмітки словникових одиниць мовою XML, що забезпечує структурну єдність між текстом і базою даних.

Етап 4. Формальна розмітка та створення XML-моделі. Після розроблення концептуальної моделі здійснюється формалізація у форматі XML, у межах якої кожен елемент словникової статті отримує чітко визначений тег відповідно до прийнятої схеми розмітки. Наприклад, тег <entry> визначає межі словникової статті, <form> – форму слова, <def> – дефініцію, а <example> – ілюстративний матеріал. Така XML-структура дає змогу уніфіковано відображати словникові дані, забезпечує їх машиночитаність, взаємну сумісність і придатність до інтеграції у бази даних або конвертування в інші формати (наприклад, TEI чи LMF). На рисунку 5.9 показано фрагмент XML-структури *Українського словника біологічної термінології*.

```
<?xml version="1.0" encoding="windows-1251"?>
<словникові_статті>
  <СС>
    <текст_СС><![CDATA[<B>а#апа</B> (<I>рос.</I> аа#па, <I>англ.</I> аара) 1. Безлісний, дуже обводнений, грядково-мочарний вид болота на півночі Євразії.]]></текст_СС>
    <ЗТ номер_омоніма="0">а#апа</ЗТ>
    <ТВ номер="1">
      <ТК_У номер="1">
        <Т_У>а#апа</Т_У>
        <ММ>укр.</ММ>
      </ТК_У>
      <ТК_Р номер="1">
        <Т_Р>аа#па</Т_Р>
        <ММ>рос.</ММ>
      </ТК_Р>
      <ТК_А номер="1">
        <Т_А>аара</Т_А>
        <ММ>англ.</ММ>
      </ТК_А>
    </ТВ>
    <СМБ номер="1">
      <БТ номер="1">
        <НТ>1</НТ>
        <ТЛ>Безлісний, дуже обводнений, грядково-мочарний вид болота на півночі Євразії</ТЛ>
      </БТ>
    </СМБ>
  </СС>
</словникові_статті>
```

Рис. 5.9 – Фрагмент XML-структури словника

Етап 5. Інтеграція у базу даних. Після формальної розмітки XML-документ імпортується у **реляційну або документно-орієнтовану базу даних** (наприклад, *PostgreSQL, MongoDB, BaseX*), де забезпечується довготривале збереження, індексація, пошук і оновлення словникових записів. Завдяки цьому словникові дані набувають статусу динамічного ресурсу, з яким можна працювати не лише як із текстом, а як із системою взаємопов'язаних об'єктів. Такий підхід створює передумови для формування інтегрованих лексикографічних платформ, які об'єднують кілька словників або глосаріїв у спільному середовищі. Прикладами таких систем є «Словники України on-line» та електронна версія СУМ-20, де XML-структури зберігаються у базах даних і динамічно відтворюються на вебінтерфейсі.

Для *Українського словника біологічної термінології* (рис. 5.10) було обрано документно-орієнтовану NoSQL-базу MongoDB, оскільки вона забезпечує гнучкішу структуру даних порівняно з реляційними системами. Це особливо важливо для словників спеціального призначення, у яких кількість атрибутів і типи інформації (визначення, синоніми, латинські назви, ілюстрації тощо) можуть відрізнятися від запису до запису. У MongoDB дані зберігаються у вигляді документів, подібних до **JSON-об'єктів**, що містять вкладені поля, відповідні XML-тегам. Така організація спрощує взаємодію між словниковою структурою та користувацьким інтерфейсом, а також забезпечує можливість швидкої модифікації й масштабування системи без потреби у зміні загальної архітектури бази.



Рис. 5.10 – Фрагмент бази даних, зробленої на основі XML розмітки

Етап 6. Створення інтерфейсу для Web-версії словника. На заключному етапі розробляється користувацький інтерфейс, який забезпечує взаємодію користувача з базою даних і дозволяє зручно працювати з електронним словником у режимі онлайн. Основним завданням цього етапу є візуалізація структурованої лексикографічної інформації, представленої у форматі XML або у вигляді документів бази даних, у зрозумілій і доступній для користувача формі. Інтерфейс вебверсії словника зазвичай реалізують за допомогою **HTML, CSS і JavaScript**, а обробку запитів до бази даних забезпечують серверні технології – **PHP, Python (Django/Flask), Node.js** тощо. Залежно від типу словника передбачаються функції пошуку, фільтрації, навігації між реєстровими словами, а також візуальні позначення граматичних, стилістичних або семантичних характеристик.

У випадку спеціалізованих чи академічних словників можуть бути реалізовані розширені функції:

- контекстний пошук і морфологічний аналіз;
- паралельне відображення статей різними мовами;
- гіпертекстові посилання між термінами;
- вбудовані мультимедійні ілюстрації (звукові приклади, зображення, відео);
- динамічне оновлення контенту без перезавантаження сторінки (технологія AJAX).

Для проєктів Українського мовно-інформаційного фонду (зокрема, *СУМ-20 online* та *Словники України on-line*) створення інтерфейсу базується на принципах динамічного представлення даних: під час пошуку система звертається безпосередньо до бази даних, витягає **XML**-структуру статті, перетворює її за допомогою **XSLT-шаблонів** у **HTML** і подає користувачеві вже у зформатованому вигляді. Отже, на цьому етапі завершуються всі попередні процеси – від оцифрування паперового словника до його інтерактивного представлення у веб просторі, що робить словник не лише джерелом знань, а й елементом цифрової мовної екосистеми, придатним для подальшої інтеграції у національні та міжнародні лексикографічні ресурси.

Цей етап завершує цикл створення електронного словника, переводячи його з технічного рівня зберігання даних у площину активної взаємодії з користувачем. Завдяки вебінтерфейсу словник стає не лише джерелом інформації, а й динамічним простором для мовного спостереження, поповнення і обговорення новітньої лексики.

Процес створення електронного словника новітньої української лексики є комплексним і поєднує лінгвістичний аналіз із цифровими технологіями. Він охоплює послідовні етапи – від оцифрування джерел і розпізнавання тексту до побудови концептуальної моделі, **XML**-розмітки, інтеграції у базу даних та розроблення вебінтерфейсу. Застосування форматів **XML**, **TEI** і **LMF** забезпечує формальну структурування й машиночитаність словникових статей, що відкриває можливості для автоматизованої обробки, пошуку та обміну даними між лексикографічними системами.

У підсумку створюється не лише електронний словник у традиційному розумінні, а цілісна цифрова платформа, здатна відображати динаміку сучасної української лексики, підтримувати її оновлення та інтегруватися у глобальні мовні ресурси. Порівняльна характеристика форматів подана у додатку Д.13.

Питання для самоконтролю

1. У чому полягає головна мета використання XML у сучасній електронній лексикографії?
2. Які основні етапи технологічного процесу створення електронного словника можна виокремити?
3. Яке призначення має концептуальна модель словника і чому вона передуює XML-розмітці?
4. Які переваги надає XML-розмітка порівняно з HTML при формалізації структури словникових статей?
5. Які міжнародні стандарти XML-розмітки використовують у лексикографії (наприклад, TEI, LMF)?
6. Які типи баз даних можуть використовуватись для зберігання словникових даних?

7. Чому для Українського словника біологічної термінології було обрано документоорієнтовану базу даних MongoDB?
8. Які функції має інтерфейс вебверсії електронного словника?
9. Як забезпечується зв'язок між елементами XML-документа і структурою бази даних?
10. Які переваги комплексного підходу до використання PDF, HTML і XML у лексикографічних проєктах?

Завдання для самоперевірки

1. Припустимо, що потрібно створити Українсько-англійсько-французький словник із комп'ютерної лінгвістики та англійсько-український словник із майнового права. Визначте, яку модель репрезентації мови доцільно застосувати в кожному випадку (семантичну, онтологічну, концептуальну, формальну тощо) і як вона реалізуватиметься у форматах зберігання даних (наприклад, XML, база даних, HTML-інтерфейс).
2. Поясніть, на якому етапі технологічного процесу створення електронного словника здійснюється XML-розмітка та чому саме там. Наведіть приклад мінімального XML-фрагмента словникової статті (3-5 тегів).
3. Опишіть, які зміни доведеться внести у структуру XML-документа, якщо словник з однієї мови перетворюється на багатомовний. Укажіть, які нові елементи або атрибути можуть бути додані (наприклад, <lang>, <translation>, <equivalent>).
4. Порівняйте роботу реляційної та документоорієнтованої баз даних під час інтеграції словникових матеріалів. Яка з них буде ефективнішою для словника, що швидко оновлюється, і чому?
5. Запропонуйте структуру веб інтерфейсу електронного словника, побудованого на основі XML і MongoDB. Опишіть, які основні функції має виконувати користувацький інтерфейс і як він взаємодіє з базою даних.
6. Розпишіть переваги та недоліки кожного з форматів (PDF, HTML, XML, база даних) у контексті створення електронних словників. Подайте результати у вигляді короткої таблиці (4 стовпці × 2 рядки: переваги/недоліки).

7. Створіть мінімальний XML-фрагмент для словникової статті слова “**data**”, який містить елементи: *форма слова, частина мови, визначення, переклад українською*. Використайте власні теги (наприклад, <entry>, <form>, <pos>, <definition>, <translation>).

8. Складіть невеликий HTML-фрагмент словникової статті для слова “**network**” і додайте короткий CSS-стиль, який зробить заголовне слово синім, а визначення – курсивом.

9. Проаналізуйте, на яких етапах технологічного процесу (від OCR до веб інтерфейсу) використовуються формати PDF, HTML, XML і база даних. Побудуйте коротку таблицю або схему, де для кожного етапу зазначено, який формат і для чого застосовується.

5.3. Поняття концептуальної моделі словника

Вибір моделі репрезентації мови та відповідних форматів збереження лексикографічної інформації безпосередньо залежить від того, як побудована **концептуальна модель словника**, що *визначає принципи його структури, способи організації мовного матеріалу та характер взаємозв'язків між його компонентами*.

Концептуальна модель словника може бути реалізована у різних формальних структурах залежно від мети, обсягу та типу лексикографічного ресурсу. Вибір моделі визначає, яким чином організовуються відношення між лексичними одиницями, як представлено семантичні, граматичні та прагматичні зв'язки, а також якою мірою забезпечується гнучкість пошуку й інтеграція з іншими ресурсами. У сучасній електронній лексикографії застосовуються такі основні типи концептуальних моделей:

– **ієрархічна модель**: відображає структуру словника як дерево, де вищі рівні (наприклад, макрорівень) підпорядковують нижчі (мікро- та медіорівень). Вона зручна для опису систематичних словників або тезаурусів, де важливе родовидове групування понять;

– **сіткова (мережева) модель**: описує взаємозв'язки між лексичними одиницями у вигляді графа, де вузли відповідають словам або поняттям, а ребра

– типам відношень (синонімія, антонімія, гіпонімія тощо). Цей підхід використано у проєктах типу *WordNet*;

– **фреймова модель**: орієнтована на представлення знань у вигляді сценаріїв або фреймів, що описують типові ситуації, пов’язані з певним концептом. Цей підхід дає змогу моделювати не лише лексичні, а й когнітивні зв’язки між одиницями;

– **онтологічна модель**: найбільш формалізований тип, де лексичні одиниці та їхні відношення описуються як об’єкти і класи у спеціальній онтологічній схемі (OWL, RDF). Вона забезпечує сумісність з іншими базами знань та системами штучного інтелекту.

Отже, тип концептуальної моделі визначає не лише технічну реалізацію словника, а й його **епістемологічну основу** – тобто спосіб, у який знання про мову організуються, зберігаються та подаються користувачу. Порівняльна характеристика концептуальних моделей подана в таблиці 5.4.

Таблиця 5.4 – Порівняльна характеристика концептуальних моделей

Тип моделі	Структурна основа	Переваги	Недоліки	Приклади застосування
Ієрархічна	Дерево (parent–child) з підпорядкуванням елементів	Чітка структурованість; зручна для тематичних, енциклопедичних, термінологічних словників	Обмежена гнучкість при описі складних міжпоняттєвих зв’язків	Тезауруси, класифікаційні словники, Словник української біологічної термінології, Словники України on-line
Сіткова (мережева)	Орієнтований або неорієнтований граф	Відображає багатовимірні зв’язки; зручна для семантичного пошуку	Вимагає складної структури даних; важко підтримувати цілісність	<i>WordNet</i> , <i>BabelNet</i> , Український WordNet (UWN)
Фреймова	Структура фреймів (сценарії, ролі, ситуації)	Враховує контекст, типові ситуації, когнітивні зв’язки; підходить для навчальних і когнітивних словників	Не завжди придатна для великих формалізованих баз даних	Когнітивні та тлумачні словники, електронні лінгвоплатформи
Онтологічна	Формальна онтологія (класи,	Сумісність із системами III,	Складність створення,	<i>OntoLex-Lemon</i> , <i>DBpedia</i>

	об'єкти, властивості, відношення)	формальна точність, можливість інтеграції семантичну павутину	потреба онтологічному інжинірингу	в

Ієрархічна модель традиційно вважається початковою формою організації словникового матеріалу. Вона передбачає жорстку підпорядкованість елементів: кожна одиниця належить до певного класу, який, у свою чергу, підпорядкований ширшій категорії. Такий підхід забезпечує чіткість структури, однак не завжди може відобразити багатовимірні зв'язки між словами, наприклад, синонімічні, антонімічні чи асоціативні. З часом ієрархічна модель почала еволюціонувати в сіткову (мережеву), де лексичні одиниці з'єднуються за різними типами відношень – семантичними, граматичними, прагматичними тощо.

Сіткова модель дозволяє представити складну систему зв'язків між словами у вигляді графа, що особливо ефективно в електронному середовищі, де можлива гіперпосилальна навігація. Саме ця модель лежить в основі таких проєктів, як *WordNet*, *BabelNet* або Український WordNet (UWN).

Фреймова модель, у свою чергу, постає як розвиток сіткової, але з орієнтацією на когнітивний і ситуаційний аспект. У фреймовій структурі слова описуються не лише через системні відношення, а й через типові контексти вживання – сценарії, дії, ролі учасників. Така модель особливо ефективна для навчальних або перекладних словників нового покоління, які демонструють значення слова у динамічному контексті.

Найбільш формалізованою та перспективною є онтологічна модель. Вона ґрунтується на принципах семантичного вебу та машинно-читаних онтологій, де словникові одиниці описуються як об'єкти з чітко визначеними властивостями та відношеннями. Онтологічні словники забезпечують інтероперабельність між лінгвістичними базами даних різних мов, що особливо важливо для систем машинного перекладу, штучного інтелекту та автоматизованої обробки текстів.

Таким чином, можна спостерігати трансформаційний ланцюжок розвитку моделей:

Ієрархічна → Сіткова → Фреймова → Онтологічна

Кожен наступний етап не заперечує попередній, а розширює його, додаючи новий рівень опису, гнучкості та інтегрованості в цифровому просторі. У структурному аспекті концептуальну систему складають кілька рівнів з численними внутрішніми відношеннями, у межах якої виділяють три основні рівні:

- **мікрорівень:** рівень словникової статті, де представлено опис окремих одиниць (лем, дефініцій, прикладів, граматичних і стилістичних характеристик);
- **медіорівень:** рівень групування словникових статей за тематичними, частиномовними або функціональними ознаками;
- **макрорівень:** рівень загальної організації словника, що охоплює принципи упорядкування, тип користувацького доступу, пошукову систему та формат подання даних.

На кожному з цих рівнів існують свої об'єкти й відношення між ними, які формують цілісну систему. Починаючи роботу над словником, укладач насамперед має визначити його **концепцію**, тобто призначення, коло потенційних користувачів, способи використання, орієнтовний обсяг, кількість мов і характер подання мовних входів. Вибір об'єктів і відношень на всіх рівнях формує абстрактну модель словника, яка є концептуальною основою для його подальшої структурної та технічної реалізації.

На **мікрорівні** розв'язуються питання, пов'язані зі словниковою статтею: визначення її структури, способів подання та розкриття семантики реєстрових одиниць, побудови ієрархії значень, взаємодії між окремими компонентами. У межах моделі виділяють два типи відношень:

- *основні відношення:* між реєстровою одиницею і лексикографічними параметрами, що її описують (дефініцією, граматичними, стилістичними, семантичними характеристиками тощо);
- *внутрішні відношення:* між самими лексикографічними параметрами в межах однієї словникової статті, які визначають її внутрішню організацію.

Система внутрішніх зв'язків формує порядок розміщення елементів у словниковій статті. Цей порядок, як і сам набір параметрів, залежить від типу

словника (тлумачний, перекладний, термінологічний, навчальний тощо) й авторської концепції, проте в межах одного словника він є стабільним та уніфікованим, що забезпечує структурну цілісність та однорідність подання матеріалу. Так, наприклад³, структуру словникової статті СУМ-20 можна подати у вигляді формальної моделі типу: $ENTRY(x) = \{\Lambda(x), P(x)\}$, де $V(x)$ – словникова стаття, $\Lambda(x)$ – ліва частина, $P(x)$ – права частина. А символ x позначає реєстрову одиницю, що є об'єктом лексикографічного опису. Загальна форма $\Lambda(x)$ має такий вигляд: $x, Пар(x), Грам(x), Рем(x), Синт(x)$, де x – реєстрова одиниця; $Пар(x)$ – парадигматичні (словозмінні) характеристики; $Грам(x)$ – граматичні параметри, такі як частина мови, вказівка на рід для іменників, вид для дієслова тощо; $Рем(x)$ – словникові ремарки різного типу; $Синт(x)$ – показники синтаксичних зв'язків та функцій.

Більш складну структуру має права частина, що зображена на рисунку 5.11 у вигляді графа. Символами позначено такі елементи: $P(x)$ – права частина, що розкриває багатозначність реєстрової одиниці; $S(x)$ – комплекс рубрик лексичного значення; $K(x)$ – комплекс колокацій; $J(x)$ – комплекс ілюстрацій всіх категорій; $V(x)$ – комплекс відтінків значень; $R(x)$ – комплекс стійких словосполучень; $F(x)$ – комплекс фразеологізмів; $T(x)$ – комплекс термінологічних словосполучень; $E(x)$ – комплекс еквівалентів слова. Стрілками позначено відношення належності.

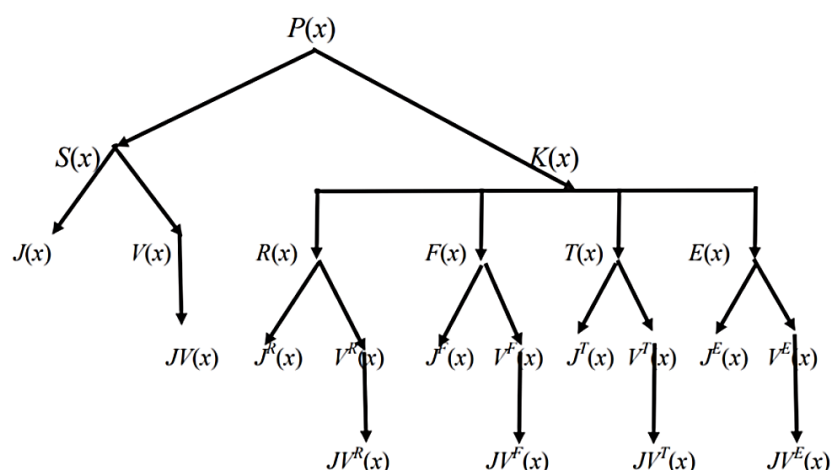


Рис. 5.11 – Концептуальна схема словникової статті СУМ-20

³ Приклад взято з: Лінгвістично-інформаційні студії: Праці Українського мовно-інформаційного фонду НАН України : у 5 т. Т. 3 : Тлумачна лексикографія. Кн. 1 : Словник української мови у двадцяти томах / В. А. Широков та ін. Київ : Український мовно-інформаційний фонд НАН України, 2018. 276 с. С. 42–43. URL: https://movoznavstvo.org.ua/files/tom_3_B5_print.pdf

У тлумачних словниках (як-от СУМ-20) домінує семантична та ілюстративна компонента, тоді як у перекладних чи галузевих словниках більшу вагу мають параметри валентності та відповідників у цільовій мові. У електронних словниках нового покоління структура $\Lambda(x)$ і $P(x)$ може бути реалізована у вигляді **XML- або JSON-моделі**, що забезпечує машинну обробку лексикографічних даних і гнучке відображення інформації у веб інтерфейсі. Отже, формальна модель $ENTRY(x) = \{\Lambda(x), P(x)\}$ дозволяє поєднати традиційну лінгвістичну системність із сучасними технічними принципами репрезентації даних, що відкриває можливості для створення інтегрованих лексикографічних платформ та автоматизованих систем обробки мови.

На **медіорівні** відбувається не просто опис окремих лексем, а систематизація словникового матеріалу. Тут важливо показати, як слова взаємодіють між собою, утворюючи парадигматичні, синтагматичні та дериваційні зв'язки. Саме на цьому рівні здійснюється групування статей за різними принципами:

- *тематичним*: об'єднання лексем за спільною семантичною сферою, наприклад, «інформаційні технології», «військова справа», «екологія»;
- *частиномовним*: класифікація за граматичними характеристиками, іменники, дієслова, прикметники тощо;
- *семантичними відношеннями*: омонімія, полісемія, синонімія, антонімія, гіпонімія, словотвірна спорідненість.

Отже, медіорівень виконує роль посереднього етапу між мікроаналізом (описом окремої одиниці) та макроструктурою словника (узагальненням у межах усього корпусу). Він дозволяє побачити мережу лексичних зв'язків, що формують мовну систему як живий організм.

На **макрорівні** здійснюється узагальнення всієї лексикографічної системи – формування концептуальної моделі словника як цілісного організму. Тут розглядаються взаємозв'язки між групами слів, їхня функціональна роль у межах тематичних, стилістичних або ідеологічних блоків. Саме на цьому рівні визначаються основні напрями репрезентації мовної картини світу, що постає через сукупність словникових статей. Макрорівень охоплює:

- структуру словника в цілому (розділи, підрозділи, ієрархія тематичних груп);
- принципи подання матеріалу (алфавітний, тематичний, комбінований);
- метамовні зв'язки між словниковими одиницями (цитування, перехресні посилання, блоки синонімів тощо);
- ідейно-культурний вимір, тобто яким чином словник відображає сучасні тенденції, національну ментальність, світоглядну систему суспільства.

Як можна побачити, макрорівень є підсумковим етапом лексикографічного моделювання, який інтегрує результати попередніх рівнів (мікро- та медіоаналізу) у єдину лексико-семантичну картину мови. Порівняльну характеристику рівнів подано у додатку Д.14.

Питання для самоконтролю

1. У чому полягає відмінність між універсальною та специфічною моделями репрезентації лексики?
2. Які основні відмінності між форматами PDF, HTML, XML та форматами баз даних?
3. Для чого використовують стандарти ISO 24613 (LMF) у комп'ютерній лексикографії?
4. Як XML-теги забезпечують ієрархічність і гнучкість структури словникової статті?
5. У чому полягає перевага використання баз даних (SQL, NoSQL) для збереження великих лексикографічних ресурсів?
6. Як здійснюється візуалізація лексикографічних даних у цифрових словниках – які формати це підтримують найкраще?
7. Що розуміють під *концептуальною моделлю словника* і яку роль вона відіграє в процесі проєктування електронних лексикографічних ресурсів?
8. Які основні рівні структури словника (мікро-, медіо-, макрорівень) і які об'єкти описуються на кожному з них?
9. У чому полягає відмінність між основними та внутрішніми відношеннями лексикографічних параметрів?

10. Чому уніфікація структури словникових статей є важливою для цілісності та автоматизованої обробки лексикографічної інформації?

11. Які переваги має XML-репрезентація словникової структури порівняно з традиційною текстовою моделлю?

Завдання для самоперевірки

1. Побудуйте **формальну модель** для *українсько-англійського перекладного словника* технічних термінів. Укажіть, які параметри повинні входити до:

– лівої частини $\Lambda(x)$: реєстрова одиниця, граматичні ознаки, ремарки, частина мови тощо;

– правої частини $P(x)$: перекладні відповідники, тематична група, приклади вживання, контекст, транскрипція.

2. Запропонуйте схему XML-репрезентації цієї моделі (з використанням власних тегів або елементів TEI).

3. Проаналізуйте відмінності між моделями перекладного словника та тлумачного словника за структурою $\Lambda(x)$ і $P(x)$. Які параметри є спільними, а які – специфічними для кожного типу?

4. Побудуйте формальну модель для *перекладного словника технічних термінів* (наприклад, *українсько-англійський словник інженерної термінології*), дотримуючись рівневої структури:

– Опишіть **макроструктуру** (тип словника, цільова аудиторія, принцип подання).

– Змодельуйте **медіорівень**: групування статей за темами (електрика, механіка тощо).

– Розробіть **зразок мікрорівня** — одну статтю (з позначенням частини мови, перекладом, прикладом уживання).

– Додайте **метарівень** — систему позначок для автоматизованої обробки (наприклад, XML-теги).

Питання для поглибленого вивчення матеріалу

1. Технологія парсингу в Українському мовно-інформаційному фонді НАН України.
2. Електронні фреймові словники та принципи їх побудови.
3. Розроблення WordNet-подібного словника української мови.
4. Інтеграція української мови у світовий мовний простір через WordNet.
5. Модель репрезентації мови OntoLex-Lemon Model та її застосування в електронних словниках.
6. Використання JSON, TEI та LMF у комп'ютерній лексикографії.

6. ТИПОЛОГІЯ КОМП'ЮТЕРНИХ СЛОВНИКІВ

План заняття:

- 5.1. Принципи класифікації комп'ютерних словників
- 5.2. Функціональна, структурна та технологічна типології
- 5.3. Інші типи електронних словників

Вибір типу словника є одним із ключових етапів його проектування, адже саме від цього залежить формування авторської концепції – визначення обсягу словникового матеріалу, кола передбачуваних користувачів, принципів побудови макро- та мікроструктури, а також добір системи лексикографічних параметрів. Тип словника безпосередньо впливає на розроблення його концептуальної моделі, яка, у свою чергу, зумовлює вибір відповідних технологій створення бази даних, програмних інтерфейсів, моделей опису мови та форматів збереження лексикографічної інформації. Отже, для ефективного проектування сучасних електронних лексикографічних систем необхідно мати чітку типологію комп'ютерних словників, що враховує як традиційні лексикографічні принципи, так і специфіку цифрового середовища.

6.1. Принципи класифікації комп'ютерних словників

Вироблення єдиної класифікації, що поєднувала б усі типи електронних словників, є доволі складним завданням. На відміну від друкованих, комп'ютерні словники можуть одночасно належати до кількох категорій, виконувати різні функції та поєднувати різні технології представлення даних. Вони здатні бути перекладними й навчальними водночас, або ж поєднувати тлумачний компонент із термінологічною базою. Це зумовлює потребу у гнучкому підході до типології таких ресурсів. Класифікація комп'ютерних словників зазвичай спирається на кілька взаємопов'язаних критеріїв, що відображають різні аспекти лексикографічної системи. Серед основних критеріїв виокремлюють:

– *функційний критерій*: визначає призначення словника та його роль у комунікаційній діяльності користувача. За цим принципом розрізняють **тлумачні, перекладні, термінологічні, навчальні, корпусні, комбіновані** тощо;

– *структурний критерій*: стосується способів організації матеріалу, тобто побудови мікро-, медіо- і макроструктури, а також способів подання словникових статей. За структурою словники можуть бути **алфавітними, гніздовими, тематичними, онтологічними** тощо;

– *технологічний критерій*: пов'язаний із форматом подання, способом збереження та доступом до інформації. За цим критерієм розрізняють локальні, мережеві, мобільні, гібридні словники, а також системи, побудовані на базах даних;

– *критерій користувацького призначення*: враховує, для кого створено словник: для фахівців, студентів, перекладачів, дослідників, мовознавців, або для широкої публіки. Від цього залежить ступінь деталізації, складність термінології та інтерфейс;

– *критерій мовного охоплення*: визначає кількість мов та їхню роль у структурі словника: **одномовні** (монолінгвальні), **двомовні** (білінгвальні), **багатомовні** (мультилінгвальні) словники;

– *критерій способу створення*: поділяє словники на **автоматизовані** (згенеровані засобами корпусної лінгвістики) та **експертно укладені** (створені вручну лексикографами);

– *критерій ступеня інтерактивності*: визначає рівень взаємодії користувача зі словником: від статичних баз даних до інтерактивних ресурсів із функціями пошуку, коментування, редагування чи машинного перекладу.

Отже, класифікація комп'ютерних словників є багатовимірною, де кожен із критеріїв описує окремий зріз системи. У практичному словникарстві найчастіше використовують комбінований підхід, який дозволяє описати конкретний ресурс у кількох площинах – функційній, технологічній та структурній (таблиця 6.1).

Таблиця 6.1 – Критерії класифікації комп'ютерних словників

Критерій	Сутність	Основні типи / класи	Приклади електронних словників
1. Функційний	Визначає головне призначення словника у комунікації	Тлумачні, перекладні, термінологічні, навчальні, корпусні, комбіновані	<i>СУМ-20 online, Cambridge Dictionary Online, Lingvo Live</i>

2. Структурний	Відображає організацію матеріалу у словнику	Алфавітні, гніздові, тематичні, онтологічні, мережеві	<i>Wiktionary, EuroTermBank, Словники України on-line</i>
3. Технологічний	Визначає спосіб зберігання і подання даних	Локальні, мережеві, мобільні, гібридні, базові (на БД)	<i>ABBYY Lingvo, Glosbe, Oxford Learner's Dictionary</i>
4. Користувацького призначення	Орієнтується на потреби конкретної групи користувачів	Для перекладачів, дослідників, студентів, широкої публіки	<i>Reverso Context, Multumpan, WordReference</i>
5. Мовного охоплення	Кількість і роль мов у структурі словника	Одномовні, двомовні, багатомовні	<i>Oxford English Dictionary, Collins English–German, IATE</i>
6. Способу створення	Метод формування словникової бази	Експертно укладені, автоматизовані (з корпусів, машинного аналізу)	<i>Sketch Engine, Google Translate Dictionary, СУМ-11 (електронна версія)</i>
7. Ступеня інтерактивності	Рівень взаємодії користувача з контентом	Статичні, динамічні, інтерактивні, краудсорсингові	<i>Wiktionary, Urban Dictionary, Glosbe</i>

Для конкретизації наведених вище положень розглянемо приклад відомого лексикографічного ресурсу «Словники України online» УМІФ НАН України (<https://lcorp.ulif.org.ua/dictua/>), зображеного на рисунку 6.1. Цей ресурс наочно демонструє, як різні типологічні параметри поєднуються в межах однієї системи, як це показано у таблиці 6.2.

Таблиця 6.2 – Приклад класифікації ресурсу Словники України online

Критерій класифікації	Тип / Характеристика	Пояснення
1. Функціональний тип	Комбінований	Об'єднує кілька словників різного призначення: граматичний (словозмінний), фразеологічний, синонімів та антонімів.
2. Структурний тип	Абетковий принцип упорядкування	Реєстрові слова розташовані за алфавітом, що відповідає традиційним принципам лексикографії.
3. Технологічний тип	Онлайновий вебресурс	Доступ до словника здійснюється через Інтернет; інсталяція на жорсткий диск не потрібна (раніше існувала локальна версія на CD).
4. Користувацьке призначення	Загальнодоступний (масовий)	Орієнтований на широке коло користувачів: студентів, викладачів, науковців, фахівців.
5. Мовне охоплення	Одномовний український	Зосереджується на українській мові, відображаючи явища словозміни, синонімії, антонімії та фразеології.

Водночас така модель має певні обмеження: рівень наукової верифікації контенту залежить від активності та фаховості спільноти, а структура й формат даних не завжди є уніфікованими. Таким чином, *Wiktionary* і подібні ресурси займають окрему нішу в сучасній електронній лексикографії, поєднуючи властивості динамічних баз знань і відкритих краудсорсингових систем, що розширюють традиційне розуміння словника як замкненого, статичного продукту. Порівняльну характеристику експертно укладених та краудсорсингових систем подано у таблиці 6.3.

Таблиця 6.3 – Експертно укладені vs краудсорсингові системи

Критерій	Експертно укладені системи (наприклад, <i>Словники України online</i>)	Краудсорсингові системи (наприклад, <i>Wiktionary</i>)
1. Принцип укладання	Професійна робота лексикографів і наукових установ	Відкрите редагування спільнотою користувачів
2. Рівень наукової верифікації	Високий, проходить редакційне рецензування та контроль якості	Варіюється, залежить від активності та компетентності користувачів
3. Тип структури	Стала, уніфікована, з фіксованою мікро- та макроструктурою	Динамічна, може змінюватися залежно від мовної версії
4. Оновлення даних	Регулярне, але відбувається централізовано, з певною періодичністю	Постійне, завдяки відкритому редагуванню в режимі реального часу
5. Формати зберігання	XML, SQL, HTML із чіткими схемами розмітки	Вікі-формат, із власною розміткою MediaWiki
6. Тип доступу	Онлайн-доступ або інсталяційна програма	Онлайн-доступ (через браузер)
7. Цільова аудиторія	Фахівці, студенти, науковці, перекладачі	Широке коло користувачів, у т.ч. нефахівці
8. Приклади	<i>Словники України online</i> , <i>СУМ-20</i> , <i>Cambridge Dictionary Online</i>	<i>Wiktionary</i> , <i>Urban Dictionary</i> , <i>OmegaWiki</i>
9. Переваги	Точність, системність, наукова достовірність	Динамічність, багатомовність, відкритість, швидке оновлення
10. Недоліки	Повільне оновлення, потреба у великих ресурсах	Варіативність якості, відсутність стабільної редакційної політики

Якщо порівнювати «Словники України online» та *Wiktionary* із комерційними лексикографічними проєктами, такими як *ABBYY Lingvo*, можна зробити такі узагальнення:

– «Словники України online»: класичний приклад академічного, експертно створеного ресурсу, що забезпечує високий рівень достовірності та стандартизованості даних;

– *Wiktionary*: динамічний, відкритий і гнучкий у мовному охопленні ресурс, який репрезентує сучасну тенденцію до «соціальної лексикографії», тобто колективного наповнення та редагування словникового контенту користувачами;

– *ABBYY Lingvo*: приклад професійного комерційного продукту, що поєднує традиційні лексикографічні підходи з цифровими та мультимедійними можливостями.

Узагальнюючи наведене, можна зазначити, що академічні лексикографічні ресурси (як-от «*Словники України online*») характеризуються високою науковою достовірністю, стабільністю структури й термінологічною точністю, однак їхня оновлюваність і інтерактивність є обмеженими. Натомість краудсорсингові платформи типу *Wiktionary* вирізняються оперативністю поповнення та широким мовним охопленням, але водночас – варіативністю якості контенту, що залежить від активності та компетентності користувачів. Комерційні словники (наприклад, *ABBYY Lingvo*) займають проміжну позицію між академічною та соціальною лексикографією: вони поєднують надійність експертного наповнення з мультимедійними та інтерактивними функціями, проте доступ до них зазвичай є платним або ліцензійно обмеженим. Більш детальну порівняльну характеристику перелічених вище ресурсів подано в додатку Д.15.

Отже, сучасна електронна лексикографія демонструє плюралізм моделей укладання, де кожен тип ресурсу виконує власну функцію в інформаційному середовищі – від академічного еталону до відкритої колективної практики мовного опису.

Питання для самоконтролю

1. Чому класифікація електронних словників є складнішою, ніж паперових?
2. Які основні критерії застосовують для типології комп'ютерних словників?
3. У чому полягає різниця між функціональним і структурним критерієм?
4. Які ознаки дозволяють визначити технологічний тип словника?
5. Як класифікуються словники за мовним охопленням? Наведіть приклади.

6. Що означає ступінь інтерактивності словника і як він впливає на користувацький досвід?
7. Які переваги й недоліки має краудсорсний спосіб укладання порівняно з експертним?
8. Чим відрізняються академічні ресурси (як-от «Словники України online») від комерційних систем (як *ABBYY Lingvo*)?
9. Яке місце займають відкриті лексикографічні платформи типу *Wiktionary* у сучасній електронній лексикографії?
10. Як розвиток вебтехнологій змінив підходи до створення та класифікації словників?

Завдання для самоперевірки

1. Оберіть два електронні словники (українські або міжнародні) та зробіть порівняльний аналіз:

– класифікуйте їх за наведеними критеріями (функціональний, структурний, технологічний, користувацьке призначення, мовне охоплення, спосіб створення, ступінь інтерактивності);

– зробіть короткий висновок: які типологічні риси є спільними, а які – унікальними для кожного ресурсу.

Форма виконання: таблиця або структурований опис (до 1 сторінки).

2. Уявіть, що ви створюєте новий електронний словник для молоді – «Словник сучасного українського сленгу».

- Визначте, до яких типів він належатиме за кожним критерієм класифікації.
- Обґрунтуйте вибір формату даних, способу створення (експертний чи краудсорсний) і рівень інтерактивності.
- Опишіть коротко, як виглядатиме інтерфейс користувача.

6.2. Функціональна, структурна та технологічна типології

Як зазначено у підрозділі 6.1, типологія комп'ютерних словників ґрунтується на сукупності критеріїв, що відображають функційне призначення, структурну організацію, ступінь автоматизації обробки даних та способи взаємодії з користувачем. Відповідно, можна виокремити **функціональну, структурну** та

технологічну типології електронних словників. Нижче наведено їхній детальний огляд.

Функціональна типологія – це класифікація словників за їх призначенням і виконуваними функціями. Вона в основному збігається з класифікацією традиційних друкованих словників, однак у цифровому середовищі кожен тип може бути реалізований у більш складних або комбінованих формах. Основними функціональними типами словників є:

1. **Тлумачні словники** – подають дефініції, граматичні характеристики, приклади вживання, іноді також етимологічні відомості. *Приклади:* «Словник української мови» (СУМ-20 online), система «Горих», *Oxford English Dictionary*, *Diccionario de la lengua Española* тощо. Цей тип словників базується на монолінгвальній моделі опису, орієнтованій на внутрішню системність мови.

2. **Перекладні словники** – забезпечують міжмовну відповідність лексичних одиниць. Вони можуть спиратися як на **універсальні моделі** репрезентації мови (на основі семантичних зв'язків), так і на **специфічні**, орієнтовані на двомовні пари. *Приклад:* колекція англійсько-українських словників загальномовної та спеціальної лексики на порталі e2u.org.ua.

3. **Термінологічні словники** – орієнтовані на фахову лексику певної галузі знань, містять дефініції, переклади та структурні зв'язки між термінами. *Різновиди:* дво- та багатомовні **термінологічні глосарії**, що подають прямі еквіваленти до вхідної мови та ґрунтуються на **універсальній моделі опису** (понятійно-семантичній). *Приклад:* перекладні словники спеціальної лексики на e2u.org.ua.

4. **Енциклопедичні словники** – фокусуються не на слові як мовній одиниці, а на реаліях, поняттях, подіях і персоналіях. Такі ресурси часто поєднують текстову, ілюстративну та мультимедійну інформацію. *Приклади:* *COSMOS*, *Encyclopedia Britannica*, *Internet Encyclopedia of Ukraine*, *Wikipedia* тощо.

Великі академічні тлумачні словники містять багатий лінгвістичний матеріал і найбільш повно описують усю систему національної мови, її закономірності в системно-структурній організації та функціонуванні. У зв'язку

з цим постає завдання створення програмних інструментів для екстракції лінгвістичної інформації з текстів словників. Такі інструменти можуть виконувати роль своєрідного «лінгвістичного мікроскопа», що дозволяє дослідникам проникати у найтонші структурні зв'язки словникових статей, аналізувати їх з погляду лексикографічної системності, функціональних зв'язків і типологічних особливостей (див. підрозділ 2.3).

Одним із напрямів розвитку таких програмних інструментів є **віртуальні лексикографічні лабораторії (ВЛЛ)** – *інтерактивні цифрові середовища, які надають користувачеві (зокрема професійному лінгвісту) можливість автоматизованого аналізу словникових даних*. На рисунку 6.2 представлено діалогове вікно «Вибірка» ВЛЛ DLE 23, створеної на базі словника іспанської мови. У цьому інтерфейсі відображено елементи словникової статті, що можуть бути активовані для автоматичного формування пошукового запиту в базу даних. У результаті виконання запиту користувач отримує вибірку словникових статей, які містять зазначені лінгвістичні характеристики.

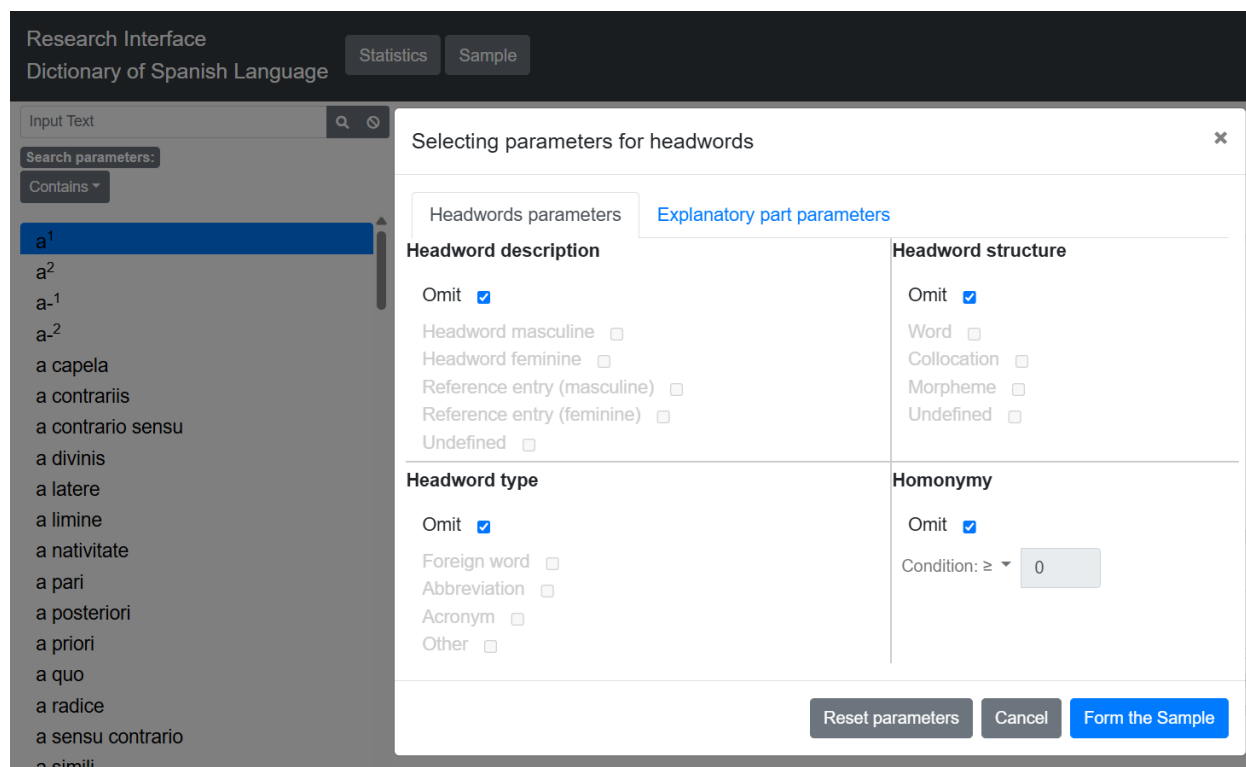


Рис. 6.2 – Діалогове вікно «Вибірка» ВЛЛ DLE 23

Застосування віртуальних лексикографічних лабораторій до перекладних словників відкриває нові можливості для міжмовного аналізу та порівняння

термінологічних систем (див. рис. 6.3⁴). Завдяки чітко структурованій рубрикації та інтерактивним пошуковим механізмам користувач може оперативно зіставляти лексеми різних мов, виявляти відповідники, відмінності у поняттєвому наповненні та варіативність перекладу.

Особливістю ВЛЛ Українського мовно-інформаційного фонду є те, що вона не лише відтворює функції традиційного перекладного словника, а й моделює взаємозв'язки між термінами у межах галузевої терміносистеми, що наближає її до інструментів онтологічного типу. Такі системи сприяють побудові єдиних лінгвістичних просторів знань, у межах яких здійснюється інтеграція лексикографічних, термінологічних і корпусних ресурсів. Отже, віртуальні лексикографічні лабораторії виступають не лише засобом представлення словникової інформації, а й ефективним інструментом наукового аналізу, автоматизованої обробки даних і побудови складних мовних моделей, що відображають динаміку розвитку терміносистем у різних галузях знань.

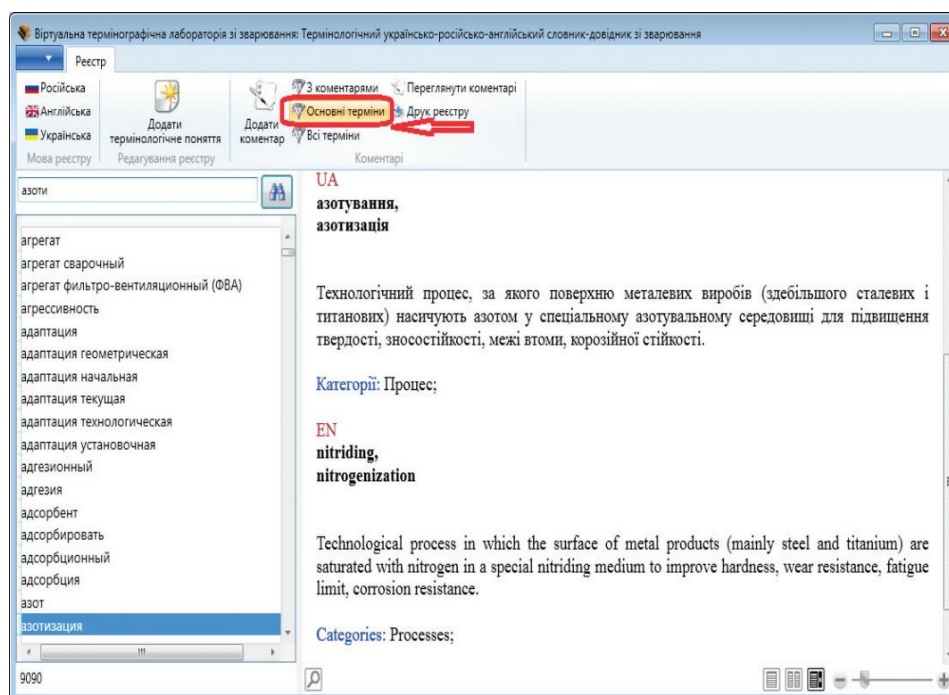


Рис. 6.3 – Зовнішній вигляд ВЛЛ багатомовного словника зі зварювання

⁴ Приклад взято з: Лінгвістично-інформаційні студії: праці Українського мовно-інформаційного фонду НАН У. країни: у 5 т. / В. А. Широков та ін. Т. 5: Віртуалізація лінгвістичних технологій. Київ. Український мовно-інформаційний фонд НАН України. 2018. 239 с. С. 187.

Усе це засвідчує, що сучасні електронні словники, зокрема у форматі віртуальних лабораторій, мають не лише функційну, а й виразну структурну специфіку, що безпосередньо впливає на організацію словникових даних та способи їхньої репрезентації – питання, яке розглянемо далі.

Структурна типологія відображає спосіб організації лексикографічного матеріалу в електронному словнику – тобто, як побудовано його макро- та мікроструктуру, які принципи упорядкування, зв'язків між статтями та навігації реалізовані в системі. Якщо для традиційних паперових словників типовим був лінійний, послідовний спосіб розташування слів, то для електронних систем характерна мережева (графова) модель, що дозволяє створювати динамічні зв'язки між статтями, поняттями та формами. Основні типи структур:

1. Абеткова структура – найпоширеніший тип, за яким реєстрові одиниці впорядковуються за алфавітним принципом. Така організація зручна для швидкого пошуку, але не відображає семантичних зв'язків між словами. *Приклади:* більшість тлумачних і перекладних словників («Словники України online», *Oxford English Dictionary*, *ABBYY Lingvo*).

2. Тезаурусна структура – базується на групуванні лексем за семантичними відношеннями (синонімія, антонімія, гіпонімія, меронімія тощо). У тезаурусах важливим є ієрархічний принцип, де кожна одиниця пов'язана з іншими у вигляді сітки понять. *Приклади:* *WordNet*, *EuroWordNet*, тезаурусна частина словника *Merriam-Webster*.

3. Онтологічна структура – орієнтована не лише на лексичні одиниці, а на поняттєві сутності та їхні відношення (родо-видові, частково-цілісні, причинно-наслідкові). Така структура використовується у семантичних мережах та спеціалізованих базах знань. *Приклади:* *SUMO (Suggested Upper Merged Ontology)*, *DBpedia Ontology*.

4. Мережева структура – найхарактерніша для інтерактивних багатомовних і багатофункційних ресурсів, де словникові статті пов'язані гіперпосиланнями, що дозволяють переходити від однієї одиниці до іншої. *Приклади:* *Wiktionary*, *Glosbe*, *Reverso Context*.

5. **Гібридна** – поєднує алфавітну, семантичну та онтологічну організацію, надаючи користувачеві кілька варіантів пошуку – за формою, значенням, контекстом або тематикою. *Приклади: IATE (InterActive Terminology for Europe), Merriam-Webster Online*, а також комерційні системи на кшталт *ABBY Lingvo*, які поєднують класичну словникову статтю з семантичними мережами та мультимедійними елементами.

Як приклад поєднання алфавітної та тезаурусної структур можна навести тлумачні словники *Merriam-Webster's Dictionary online, Cambridge Dictionary online*, а також багатомовний термінологічний словник *Turbolex*. Таке поєднання забезпечує як швидкий пошук за реєстровим словом, так і можливість тематичної навігації за смисловими полями. Як слушно зауважує Х. Касарес, «сукупність слів, упорядкованих відповідно до їх змісту, представляє в системі знань певну ієрархію, яка перевершує ту, що виникає в результаті послідовного розташування слів і яку очолює маловажливий, нестійкий, а іноді безглуздий елемент – початкова літера реєстрового слова»⁵. Дослідник розглядає **алфавітний реєстр** лише як допоміжний інструмент до **ідеографічного словника**, подібно до предметного покажчика.

У термінологічних словниках тезаурусна структура реєстру відображає систему термінологічної підсистеми як у цілому, так і на рівні її окремих складників – мікрополя, термінологічної групи чи підгрупи. При цьому вона фіксує системні відношення між термінами («рід – вид», «частина – ціле») і ґрунтується на результатах ідеографічної класифікації аналізованого термінологічного матеріалу. На рисунку 5.4 подано приклад тезаурусного реєстру в словнику *Turbolex*, який демонструє взаємозв'язки між термінами та принципи їх тематичного впорядкування.

Алфавітна структура передбачає лінійне розташування реєстрових слів у порядку зростання їх початкових літер. Такий тип є найпоширенішим, оскільки забезпечує простоту пошуку і зручність для користувача. Ідеографічна, або

⁵ Casares J. Nuevo concepto del diccionario de la lengua / J. Casares. – Madrid : G. Koehler., 1921. 118 p.

тезаурусна, структура ґрунтується на змістовій ієрархії понять і відображає семантичні зв'язки між одиницями – синонімічні, антонімічні, родо-видові тощо. Її основна мета — показати системність лексики певної мови або підсистеми. Комбінована структура поєднує переваги обох принципів і дає змогу здійснювати пошук як за алфавітом, так і за тематичними групами.

Технологічна типологія відображає способи реалізації електронного словника з погляду технічних засобів, платформи, архітектури, доступу до даних та рівня інтерактивності. Якщо у традиційних друкованих словниках головним був зміст і структура подання, то у комп'ютерних лексикографічних системах не менш важливими стають технологічні параметри: формат даних, спосіб обробки, платформа розгортання, засоби інтеграції з іншими ресурсами. За цією типологією вирізняють **локальні, онлайніві, гібридні, мобільні, краудсорсингові та засновані на технологіях**. Наведемо їхню характеристику.

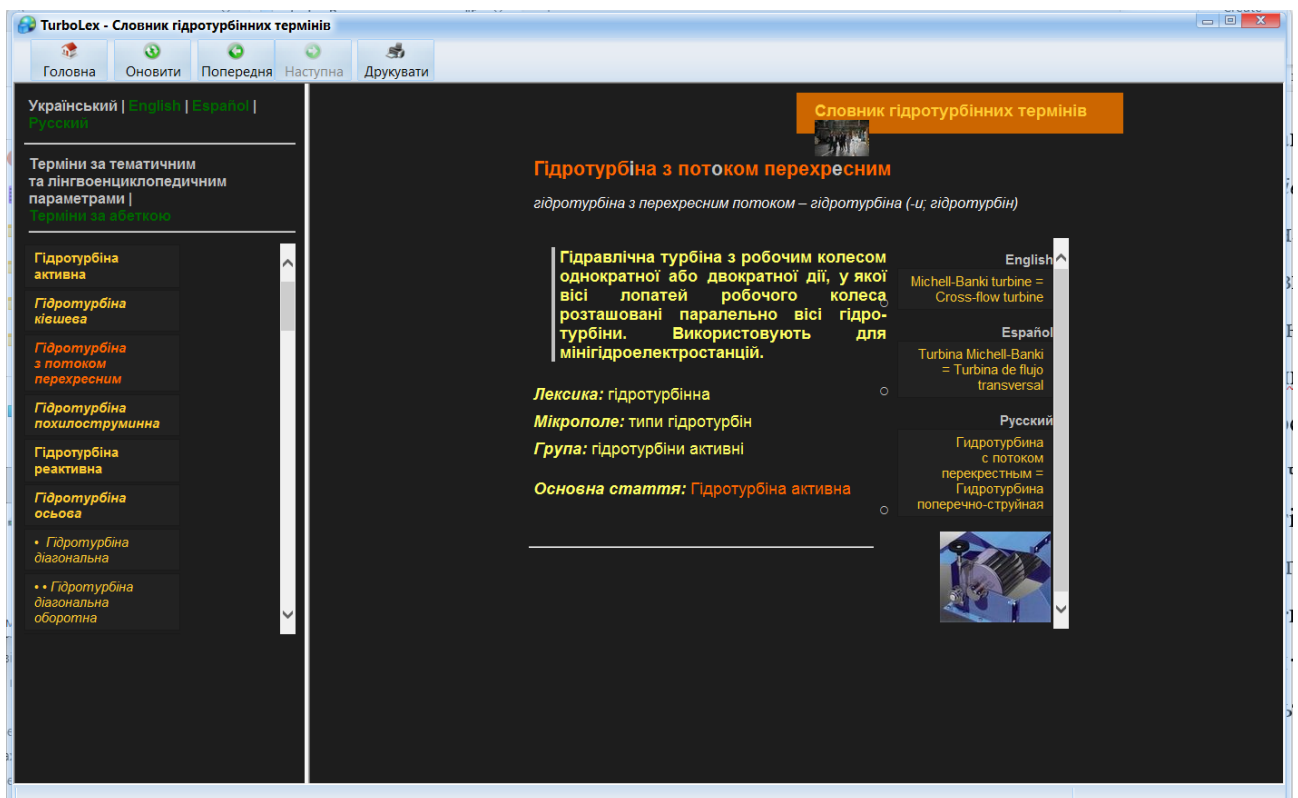


Рис. 6.4 – Тезаурусний реєстр словника *Turbolex*

Локальні словники — це словники, інстальовані на комп'ютер користувача або інтегровані в локальні програмні комплекси. Вони не потребують постійного доступу до мережі Інтернет. *Переваги:* висока швидкість роботи, стабільність функціонування, можливість використання без мережевого підключення.

Недоліки: складність оновлення бази даних, обмежена інтерактивність, відсутність автоматичної синхронізації з новими версіями.

До цього типу належать ранні версії *ABBYY Lingvo*, локальні бази *Multitran*, а також офлайн-словники для мобільних операційних систем (наприклад, *GoldenDict*). Такі програми використовували власні формати збереження лексикографічної інформації, проте поступово з'явилися універсальні програмні оболонки, здатні працювати з кількома форматами одночасно.

Так, *GoldenDict* (рис. 6.5) підтримує широкий спектр форматів: WordNet, XDXF (XML Dictionary eXchange Format), Babylon, StarDict, ABBYY Lingvo тощо. Програма дає змогу не лише відтворювати мультимедійні елементи (графіку, аудіо), а й здійснювати інтеграцію з вебресурсами: користувач може переглядати у вікні *GoldenDict* статті з Вікіпедії, Вікісловника та інших сайтів, що функціонують на рушії MediaWiki.



Рис. 6.5 – Зовнішній вигляд програми *GoldenDict*

Онлайн-словники: розгортаються на веб-серверах і доступні через браузер. Такі системи зазвичай реалізують **веб-інтерфейс**, **пошукові фільтри**, **API** для інтеграції з іншими застосунками. *Переваги:* постійне оновлення, доступ із будь-якої точки світу, можливість зворотного зв'язку з користувачами. *Недоліки:*

залежність від з'єднання з мережею. *Приклади:* «Словники України online», *Cambridge Dictionary Online*, *Lingvo Live*, *Glosbe*.

Онлайн-системи мають можливість не лише забезпечувати безпосередній доступ до словникових даних через веб-інтерфейс, а й об'єднуватися з іншими мовно-інформаційними ресурсами, утворюючи єдину лексикографічну мережу. Така інтеграція дає змогу користувачеві отримувати розширений контекст для кожної лексеми, переходити від словникової статті до суміжних ресурсів і порівнювати різні типи лінгвістичної інформації. Наприклад, *Wiktionary* у своїх словникових статтях подає гіперпосилання на зовнішні ресурси, серед яких портал «Словники України online», що надає словозмінні парадигми для українських лем, а також *Wikipedia*, де подано енциклопедичні відомості про відповідні поняття. Таким чином, формується мережевий принцип організації лексикографічного простору, у якому словник стає частиною динамічної інформаційної екосистеми.

Гібридні системи: поєднують локальну базу даних із можливістю синхронізації через Інтернет. У таких словниках частина інформації зберігається на пристрої, а частина – на сервері. *Переваги:* стабільна робота офлайн із можливістю отримання оновлень. *Приклади:* *Oxford Learner's Dictionaries App*, *Collins Cobuild Advanced Learner's*, пізні версії *ABBY Lingvo Live*. До гібридних систем також належать віртуальні лексикографічні лабораторії (ВЛЛ), розроблені в Українському мовно-інформаційному фонді НАН України. Ці лабораторії забезпечують доступ до баз даних різних словників – тлумачного (*СУМ-20*), етимологічного (*ЕСУМ*), граматичного, перекладних і термінологічних – що зберігаються на сервері з можливістю доступу через Інтернет. Користувач має змогу здійснювати аналітичний пошук, фільтрацію даних за параметрами та експортувати результати запитів, що розширює можливості дослідження лексикографічної інформації та сприяє формуванню єдиного лексикографічного простору. На рисунку 6.6 зображено головне вікно ВЛЛ ЕСУМ, що забезпечує корпоративний доступ через Інтернет за відповідним логіном та паролем.

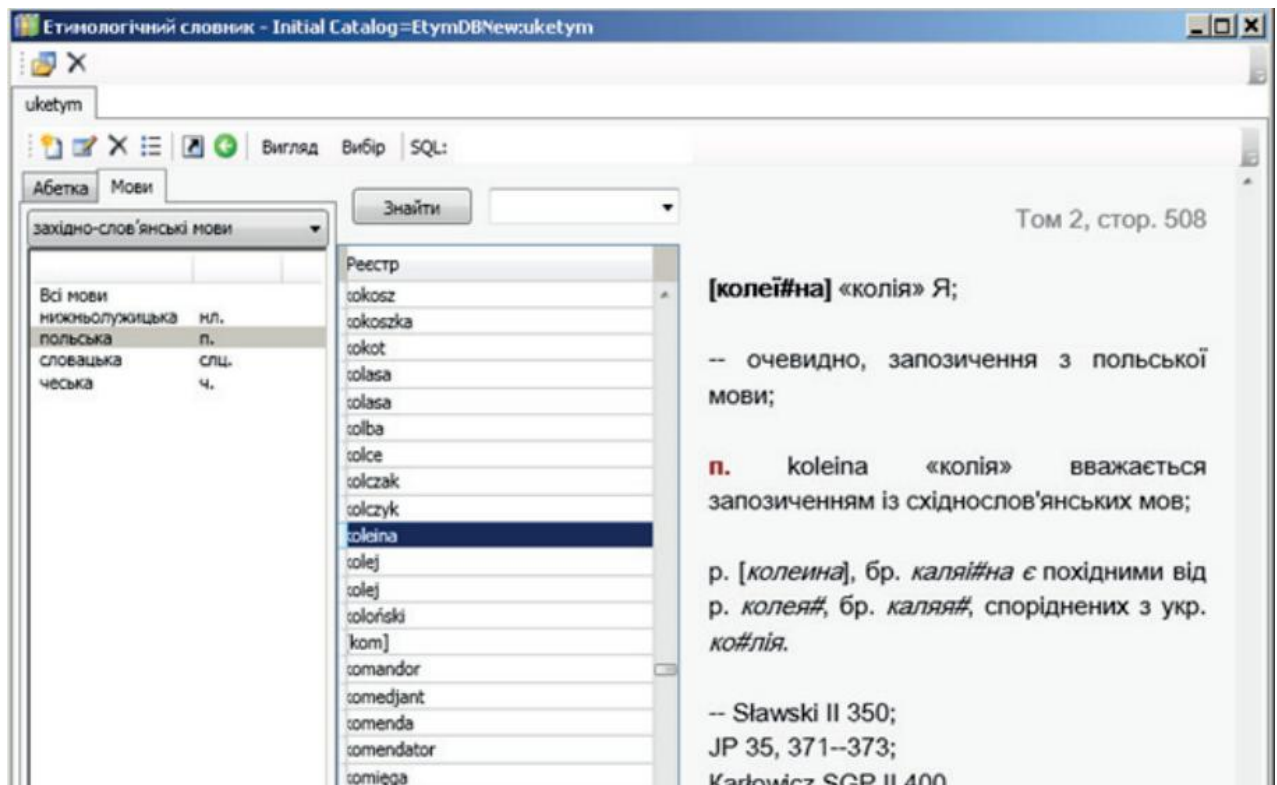


Рис. 6.6 – Головне вікно ЕСУМ.

Мобільні додатки (Mobile Dictionary Apps): реалізовані у вигляді програм для смартфонів і планшетів. *Переваги:* портативність, мультимедійність, швидка інтеграція з іншими додатками. *Приклади:* Merriam Webster, Oxford Dictionary, Oxford Learner's Dictionary, Power Thesaurus, WordDefiner English Dictionary, Diccionario de la lengua Española (RAE) та інші. З появою смартфонів провідні видавництва для просування своїх словників також розробили мобільні версії, які забезпечують користувачам швидкий пошук і доступ до лінгвістичної інформації у будь-якому місці та в будь-якій ситуації. Такі словники максимально адаптовані до технічних можливостей мобільного пристрою, зберігаючи при цьому повний функціонал. Крім базових пошукових функцій, мобільні додатки часто пропонують інтерактивні розділи – озвучення слів, збереження у «вибране», ведення особистих словників, тестові вправи для вивчення лексики. Деякі з них інтегровані з онлайн-платформами або хмарними сервісами, що дозволяє синхронізувати дані між різними пристроями користувача. Завдяки цьому мобільна лексикографія набуває статусу не лише довідкового, а й навчального інструмента, який поєднує функції словника, тренажера та персонального мовного помічника. Мобільні словники доступні для завантаження у сервісах

Google Play та App Store, що забезпечує їх широке поширення серед користувачів.

Краудсорсингові системи: укладання словника здійснюється спільнотою користувачів, які мають можливість додавати, редагувати й перевіряти матеріал. *Переваги:* швидке розширення, мовне розмаїття, актуальність. *Недоліки:* нерівномірність якості і потреба в модерації. *Приклади:* *Wiktionary*, *Urban Dictionary*, *OmegaWiki*. Такі краудсорсингові лексикографічні системи, як *Wiktionary*, поєднують принципи колективної участі з технологічною гнучкістю. Вони надають можливість не лише створення та редагування словникових статей, а й їх обговорення у спеціальних тематичних розділах. Кожна новостворена або змінена стаття проходить етап попередньої модерації, що забезпечує базовий контроль якості перед її публікацією.

На рисунку 6.7 подано вікно редагування словникової статті «вода». У лівій частині відображено вихідний код сторінки, представлений у форматі вікі-розмітки, тоді як у правій – попередній вигляд статті після застосування форматування. Така двопанельна структура редактора дозволяє користувачеві в реальному часі бачити зміни, що сприяє підвищенню точності та узгодженості контенту.

Важливу роль у підтриманні якості краудсорсингових лексикографічних ресурсів відіграє поєднання модерації та самоорганізації спільноти користувачів. Механізми попередньої перевірки, історії редагувань і системи репутаційних рейтингів сприяють дотриманню наукової точності та стабільності словникового контенту. Водночас активна участь користувачів у дискусіях і рецензуванні матеріалів формує внутрішні етичні норми й колективну відповідальність за достовірність поданих даних. Таким чином, модерація та спільнотна взаємодія виступають не лише технічними, а й соціолінгвістичними механізмами забезпечення якості сучасних електронних лексикографічних систем.

⚠ Увага: Ви не увійшли до системи. Коли ви зробите редагування, для вас буде створено тимчасовий обліковий запис. [Дізнатися більше](#). [Увійдіть в систему](#) або [створіть обліковий запис](#), щоб продовжити отримувати сповіщення після того, як виїде термін дії цього облікового запису, й отримати доступ до інших функцій.

Рис. 6.7 – Вікно редагування словникової статті

Використання штучного інтелекту у лексикографічних продуктах відкриває новий етап розвитку цифрових словників, де традиційні принципи опису лексики

поєднуються з адаптивними алгоритмами обробки мовних даних. Такі інструменти не лише автоматизують процес створення словникових статей, а й поступово формують інтелектуальне мовне середовище, у якому користувач стає співучасником лексикографічного процесу.

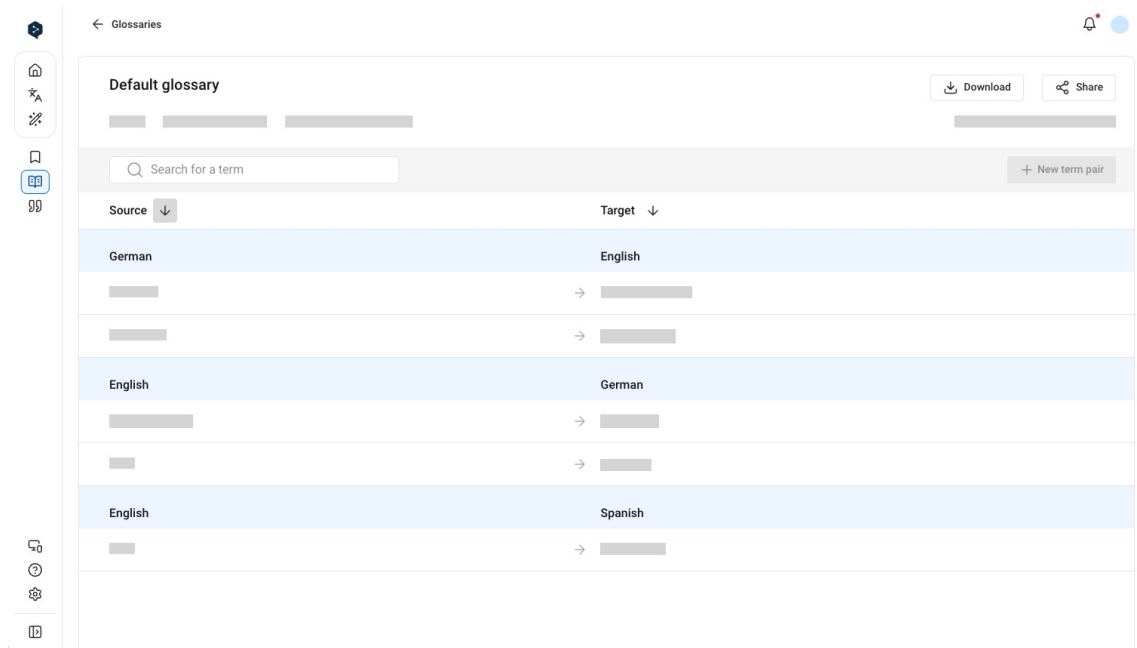


Рис. 6.8 – Вікно створення глосарію в *DeepL*

Узагальнюючи наведені типології, можна зазначити, що **функціональна, структурна та технологічна** класифікації перебувають у тісному взаємозв'язку і взаємозумовленості. Функціональна типологія визначає *мету та завдання* словника – тобто, що саме він має описувати й для кого створюється. Структурна типологія окреслює *спосіб організації даних* – як упорядковується матеріал і яким чином взаємодіють різні рівні (мікро-, медіо- та макро-). Технологічна типологія, своєю чергою, визначає *засоби реалізації* словника – де і яким способом забезпечується зберігання, пошук, оновлення та подання інформації користувачу. У сучасній електронній лексикографії всі три типи взаємодіють у межах єдиної концептуальної системи, де вибір формату даних, програмного середовища, інтерфейсу і навіть структури словникової статті безпосередньо пов'язаний із комунікативними функціями словника. Саме тому створення будь-якого електронного словника починається не з вибору технології, а з чіткого визначення його функцій, структури та користувацьких сценаріїв. Порівняльна характеристика розглянутих вище типологій класифікацій подана в додатку Д.16.

Питання для самоконтролю

1. У чому полягає основна мета класифікації електронних словників?
2. Які головні принципи покладено в основу типології електронних лексикографічних ресурсів?
3. За якими структурними ознаками розрізняють електронні словники?
4. Чим відрізняються абеткова, тезаурусна та змішана (гібридна) структури словникового реєстру?
5. Яке значення має тезаурусна структура для побудови термінологічних словників?
6. Як організовано інформацію у віртуальних лексикографічних лабораторіях (ВЛЛ)?
7. У чому полягають відмінності між локальними, онлайн та гібридними системами словників?
8. Які переваги та недоліки мають локальні електронні словники порівняно з онлайн-ресурсами?
9. Як функціонують краудсорсингові лексикографічні системи, і яку роль відіграють модерація та самоорганізація спільноти в забезпеченні якості даних?
10. Яким чином мобільні версії словників реалізують принципи адаптації лексикографічних ресурсів до сучасних технологічних умов?
11. Які інноваційні можливості відкривають системи зі штучним інтелектом (наприклад, DeepL Glossary) у сфері автоматизованої лексикографії?
12. Як, на вашу думку, поєднання традиційних лексикографічних підходів із технологіями штучного інтелекту впливає на майбутнє електронних словників?

Завдання для самоперевірки

1. Складіть схему класифікації електронних словників за трьома основними критеріями:
 - структурним,
 - функційним,
 - технологічним.Вкажіть приклади для кожного типу.

2. Порівняйте за трьома параметрами словники: «Словники України online», *Wiktionary*, *ABBYY Lingvo* (Параметри: тип створення, ступінь відкритості, рівень надійності лексикографічних даних).

3. Виберіть будь-який електронний словник із відкритим доступом (наприклад, *Cambridge Dictionary online*). Також опишіть, яку структуру має словниковий реєстр: абеткову, тезаурусну чи змішану. Обґрунтуйте відповідь прикладами.

4. Визначте, до якого типу за технологічною класифікацією належать такі ресурси:

- GoldenDict
- Multitran
- DeepL Glossary
- Віртуальна лексикографічна лабораторія (ВЛЛ)

Обґрунтуйте свою класифікацію.

5. Проаналізуйте, як саме функціонує модерація у краудсорсингових лексикографічних системах (на прикладі *Wiktionary*). Опишіть механізми контролю якості даних і роль спільноти користувачів.

6. Запропонуйте власну модель «ідеального» електронного словника, у якій поєднуються переваги академічного, комерційного та краудсорсингового підходів. Опишіть її основні функції, тип структури та способи перевірки достовірності матеріалу.

7. Знайдіть у Google Play будь-який мобільний словник української мови. Зробіть короткий опис його можливостей і визначте, які технічні рішення сприяють ефективності роботи з ним.

6.3. Інші типи електронних словників

Електронні словники також можна класифікувати за формою подання інформації, що безпосередньо пов'язано з мультимедійними можливостями цифрових технологій. За цим критерієм вирізняють **текстові, аудіо, відео та комбіновані словники**.

Текстові словники містять вербальну інформацію про слова, предмети, явища та поняття – це, як правило, семантичні, термінологічні чи енциклопедичні ресурси. Вони є основою більшості лексикографічних систем, зокрема академічних і навчальних. Приклади: *Словники України online*, *Oxford Learner's Dictionary*, *Cambridge Dictionary* тощо. За способом організації словникового тексту розрізняють **власне текстові** словники, у яких лінгвістичний матеріал подається у лінійній послідовності, та **гіпертекстові**, у яких словникові статті пов'язані системою гіперпосилань. Гіпертекст у таких словниках забезпечує не лише навігацію між статтями, а й візуалізує мережу взаємопов'язаних понять, дозволяючи користувачеві бачити лексичну систему як динамічну структуру. У деяких сучасних гіпертекстових словниках (наприклад, *WordNet* або *ConceptNet*) семантичні відношення між словами відображаються у вигляді **графів**, що поєднують ідеографічний та енциклопедичний принципи організації. Таким чином, межа між текстовими й комбінованими словниками поступово стирається, адже гіпертекст стає основою для створення **мультимедійних або візуалізованих лексикографічних систем**.

Аудіословники створюються для демонстрації вимови, інтонаційних особливостей, діалектних варіантів, а також для тренування слухового сприйняття під час вивчення іноземних мов. Такі ресурси широко застосовують у лінгводидактиці та фонетичних тренажерах (наприклад, *Forvo*, *Free Online English Pronunciation Dictionary* | *Howjsay*). У сучасних мультимедійних системах аудіофайли інтегруються безпосередньо у словникову статтю, що значно підвищує її функціональність. Варто зазначити, що для озвучення слів або фраз можуть використовуватися не лише попередньо записані аудіофайли, а й спеціальні онлайн сервіси, такі як *VOICE RSS* (<https://www.voicerss.org/api/>), *Google Text-to-Speech API*, *Text To Speech OpenAI API*, які перетворюють друкарський текст на мовлення за допомогою технології **Text-to-Speech** (синтез мовлення). Такі сервіси можуть озвучувати текст безкоштовно або за обраним тарифним планом.

Відеословники поєднують текстові й візуальні елементи, що дає змогу відобразити рухи, процеси, дії або невербальні форми спілкування, які складно описати словами. Найвідомішим прикладом є відеословники жестових мов (*SpreadTheSign, ASL Dictionary*), а також навчальні відеословники технічних спеціальностей, що демонструють принципи роботи механізмів або хід технологічних процесів. Такі типи словників також допомагають у навчанні мові, надаючи відео для пояснення семантики та особливостей використання того чи іншого слова. Серед них можна виокремити як загальномовні, так і спеціальні. До перших належать:

- *Viditionary*, що містить короткі відеоролики для ілюстрації значення слів з різних категорій, таких як тварини, природа та предмети.

- *Babel Video Dictionary*, що використовує відео для демонстрації значень фразових дієслів у реальних контекстах.

- *eJOY AI Dictionary*, розширення для браузера, яке надає відеоприклад для розвинення словникового запасу.

До числа других, спеціальних, можна зарахувати лексикографічні ресурси енциклопедичного характеру, такі як:

- *NASA's Picture Dictionary*, ілюстрована енциклопедія космосу для дітей 3-8 класів;

- *Math*, анімований та інтерактивний словник із простими поясненнями, визначеннями та прикладами математичних термінів.

Комбіновані словники об'єднують текст, звук і відео в єдиному інтерактивному інтерфейсі, створюючи комплексне мультимедійне середовище. Такі ресурси набувають особливої актуальності у сфері освітньої лексикографії, де важливе не лише пояснення, а й візуалізація та практичне застосування лексичних одиниць. Прикладом є мультимедійні навчальні системи на базі *ABBY Lingvo Live* або інтегровані словники у платформах *Duolingo* та *Babbel*. До комбінованих словників належать також онлайн енциклопедії, що органічно сполучають текстову, графічну та відеоінформацію про описувані предмети, явища та процеси. На рисунку 6.9 показано словникову статтю для слова *bridge*,

у якій текстовий опис (про саму конструкцію моста, інженерів, історію мостобудування) супроводжується ілюстраціями різних типів конструкцій та відео про мости.

Такі мультимедійні ресурси репрезентують новий етап у розвитку лексикографії, коли традиційна вербальна форма подання знань інтегрується з візуальними та аудіальними засобами. Це сприяє глибшому засвоєнню інформації, активізує асоціативне мислення й забезпечує кращу міждисциплінарну інтеграцію (мовознавства, педагогіки, інженерії, мистецтва). Комбіновані словники не лише виконують пізнавальну та навчальну функції, а й формують новий тип електронної культури знань, у якій користувач виступає не просто читачем, а активним учасником взаємодії зі словниковим контентом.

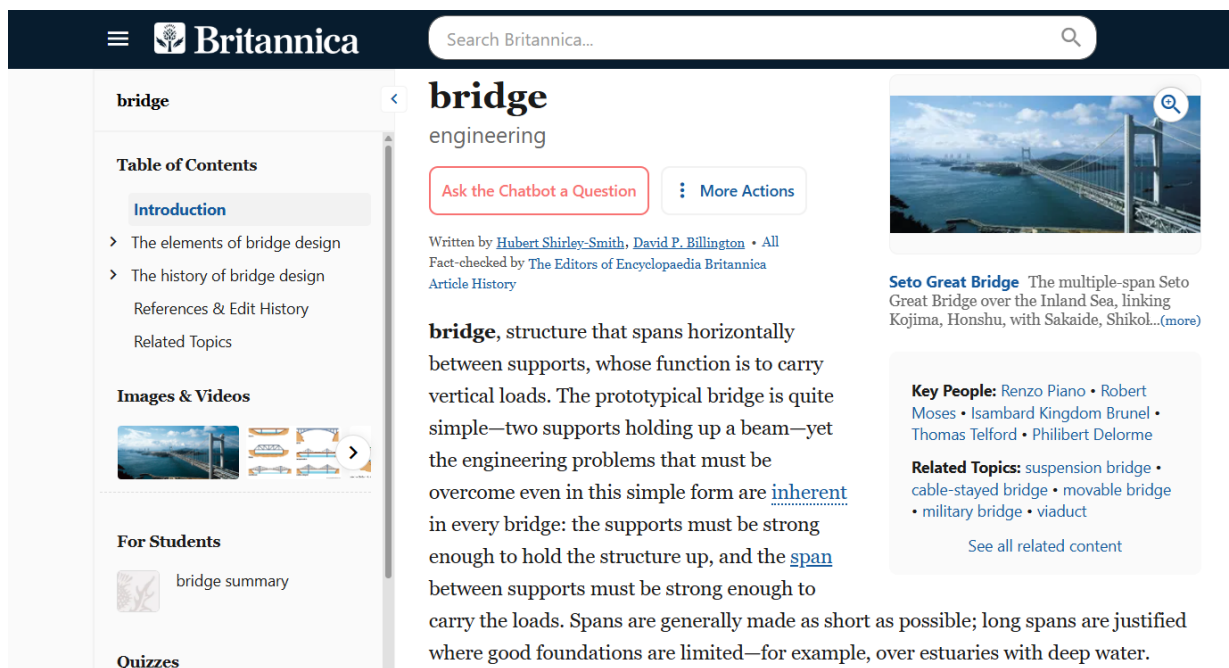


Рис. 6.9 – Словникова стаття енциклопедії *Britannica*

Як видно, розвиток мультимедійних технологій спричинив трансформацію самої концепції словника – від статичного сховища текстових даних до динамічної інтерактивної моделі мовного знання. Аудіо- та відеоеlementи, інтегровані у лексикографічну статтю, дозволяють подати мовну одиницю не лише як абстрактний знак, а як комунікативний феномен, який функціонує у реальному мовленні, жестах і діях. Завдяки цьому електронні словники стають інструментами міжмодальної репрезентації мови, поєднуючи семантичний, акустичний і візуальний рівні. Такий підхід формує нову парадигму цифрової

лексикографії, у якій знання подається в багатовимірній, мультимедійно збагаченій формі, що відповідає сучасним потребам навчання, перекладу й міжкультурної комунікації.

Питання для самоконтролю

1. За якими ознаками здійснюється класифікація електронних словників за формою подання інформації?
2. У чому полягає основна відмінність між текстовими, аудіо- та відеословниками?
3. Які функції виконують аудіословники у навчанні вимови та дослідженні діалектів?
4. У яких випадках використання відеословників є необхідним і чому?
5. Які можливості мультимедійні технології відкривають для репрезентації невербальних засобів комунікації?
6. Як поєднання текстових, аудіо- та відеоелементів у словнику впливає на повноту й достовірність мовного опису?
7. Яким чином мультимедійні словники сприяють формуванню нової парадигми цифрової лексикографії?
8. Яке значення має візуалізація системи понять у гіпертекстових словниках?
9. Чому відеословники є особливо корисними для вивчення невербальних засобів спілкування?
10. Які типи електронних словників, на твою думку, будуть найперспективнішими в майбутньому і чому?
11. Які завдання виконують аудіословники у навчанні мови та як реалізується технологія Text-to-Speech?
12. Які функції виконують відеословники у системі електронної лексикографії?
13. Які переваги комбінованих (мультимедійних) словників порівняно з традиційними текстовими?

14. Як поєднання тексту, зображення, звуку й відео впливає на ефективність засвоєння інформації користувачем?

Завдання для самоперевірки

1. Оберіть два мультимедійні електронні словники (наприклад, *Forvo*, *Cambridge Dictionary online*, *Spreadthesign*, *LinguaLeo* або будь-які інші за вашим вибором) та:

(1) визначте, до якого типу вони належать за формою подання інформації (текстові, аудіо-, відео- чи комбіновані).

(2) опишіть, які типи лексикографічної інформації в них подано (тлумачення, вимова, приклади, зображення, відео тощо).

(3) проаналізуйте, яким чином мультимедійні засоби підсилюють лексикографічний опис.

2. Розробіть структуру мультимедійної словникової статті для будь-якого слова. Укажіть, які елементи вона повинна містити, наприклад:

- реєстрове слово;
- аудіофайл із вимовою;
- відеофрагмент (для жестового чи невербального вираження);
- ілюстративне зображення;
- дефініцію;
- приклади вживання;
- посилання на суміжні статті.

3. Визначте, які типи інформації неможливо повноцінно відтворити у звичайному текстовому словнику, але можна — в аудіо чи відео форматі.

4. Запропонуйте концепцію електронного відеословника для однієї з галузей (наприклад, технічна термінологія, жестова мова, етнографічна лексика). Опишіть його функційне призначення, структуру бази даних та потенційну користь для лінгвістичних досліджень або освіти.

5. Запропонуйте концепцію власного комбінованого електронного словника (наприклад, для вивчення діалектів, термінів певної галузі тощо), описавши його структуру, функції та користь для лінгвістів або студентів.

6. Визначте, які функції аудіо- та відеокomпоненти виконують у процесі вивчення іноземної мови. Наведіть приклади словників або програм, що їх реалізують.

7. Дослідіть, як різні сервіси Text-to-Speech (Google TTS, VoiceRSS, OpenAI TTS) реалізують озвучення слів. Визначте, який із них забезпечує найвищу якість вимови та природність мовлення.

8. Створіть власну міні-словникову статтю у комбінованому форматі:

- текстове визначення;
- ілюстрація або фото;
- короткий аудіофайл із вимовою;
- посилання на відео або анімацію.

Опишіть, яку інформацію кожен елемент передає користувачеві.

Питання для поглибленого вивчення матеріалу

1. Принципи класифікації у паперовій та комп'ютерній лексикографії.
2. «Лексичні» та «нелексичні» словники.
3. Лексикографічні мережі у розвитку цифрової гуманітаристики.
4. Віртуальні лексикографічні лабораторії vs електронні словники.
5. Особливості використання графічної та аудіовізуальної інформації в електронних словниках.
6. Інтегровані лексикографічні системи.

РОЗДІЛ 3. ТЕХНОЛОГІЇ У КОМП'ЮТЕРНІЙ ЛЕКСИКОГРАФІЇ

7. СЕРЕДОВИЩА ТА ІНСТРУМЕНТИ ДЛЯ УКЛАДАННЯ ЕЛЕКТРОННИХ СЛОВНИКІВ

План заняття:

7.1. Середовища розроблення словників (DWS)

7.2. Віртуальні лексикографічні лабораторії (ВЛЛ)

7.3. Інструменти й сервіси для лексикографів

Процес укладання електронних словників охоплює широкий комплекс взаємопов'язаних етапів і завдань, серед яких: створення бази даних, розроблення алгоритмів обробки лексичного матеріалу, формування користувацького інтерфейсу та забезпечення стабільності функціонування системи. Разом із технічними аспектами важливе місце посідає лексикографічна робота, що передбачає добір, опис, систематизацію та редагування мовних одиниць, а також контроль якості словникового контенту. Для оптимізації цих процесів сучасні видавництва, науково-дослідні центри та університети використовують спеціалізовані **середовища розроблення словників** (*digital writing systems*, DWS) або **віртуальні лексикографічні лабораторії** (ВЛЛ). Такі платформи забезпечують інтегровану підтримку всіх етапів укладання словника – від опрацювання корпусних даних до фінальної верстки й публікації.

Поряд із професійними системами дедалі ширше застосовуються відкриті інструменти й онлайн-сервіси, що надають змогу створювати невеликі лексикографічні ресурси студентам, викладачам і дослідникам без потреби у складних технічних навичках.

7.1. Середовища розроблення словників (DWS)

Середовища розроблення словників (англ. *Dictionary Writing Systems* – DWS, або системи розроблення словників – СРС) – *це спеціалізовані програмні комплекси, призначені для компіляції, редагування та публікації словників та лексичних ресурсів різного типу.* Такі системи забезпечують комплексне

управління усіма етапами лексикографічного процесу – від створення реєстру до підготовки фінальної електронної або друкованої версії словника.

Оснoву СРС становить база даних, у якій зберігаються словникові статті, лексеми, приклади вживання, переклади, граматичні характеристики та інші відомості. Ці системи оснащені інструментами редагування, що дозволяють лексикографам працювати з матеріалом у зручному інтерфейсі, використовуючи шаблони словникових статей із фіксованою структурою. Це значно полегшує створення, редагування і уніфікацію словника. Ключові функції СРС включають:

- *укладання та редагування*: використання шаблонів словникових статей для збереження їхньої структурної єдності;
- *керування даними*: забезпечення довгострокового зберігання, повторного використання та узгодженості лексикографічних даних на всіх етапах роботи;
- *управління лексикографічними проєктами*: підтримка командної роботи, розподілу завдань між укладачами, контролю прогресу та версій словників;
- *публікація словників*: автоматичне формування вихідних файлів для друку чи електронного розповсюдження (зокрема у форматах XML, HTML, PDF);
- *автоматизована допомога*: створення попередніх версій словникових статей шляхом витягування лексичних одиниць і прикладів із корпусів текстів.

Отже, системи розроблення словників забезпечують повний цикл роботи з лексикографічним ресурсом, поєднуючи базу даних, редактор, засоби управління проєктом і розділі публікації в єдиному інтегрованому середовищі. Сьогодні існує ціла низка професійних систем розроблення словників, що використовують як у комерційних, так і в академічних проєктах. Вони різняться за архітектурою, функціональністю, ціною, підтримкою мов і рівнем автоматизації. Розглянемо деякі з них.

TLex (TshwaneLex) – це професійне багатофункціональне середовище для створення словників і термінологічних баз, що поєднує зручність користування з потужним набором лексикографічних інструментів (рис. 7.1). Програмний комплекс використовується провідними видавництвами, державними структурами, науковими установами та незалежними дослідниками в усьому світі. Однією з

головних ідей розробки TLex було створення інструменту, який не потребує від користувача високого рівня технічної підготовки: система інтуїтивно зрозуміла та орієнтована на лінгвістів, а не на програмістів. Інтерфейс TLex поділено на кілька функціональних зон:

- Список лем у лівій частині екрана, звідки користувач може швидко обрати будь-яке слово для перегляду або редагування;
- Область попереднього перегляду (WYSIWYG) праворуч, що відображає словникову статтю в її реальному вигляді. Попередній перегляд оновлюється миттєво після кожної зміни, дозволяючи бачити кінцевий результат у реальному часі;
- Click-to-edit – натискання на будь-який елемент у попередньому перегляді відкриває його у вікні редагування, що значно прискорює роботу.

Корисною особливістю системи є динамічне керування перехресними посиланнями: TLex автоматично оновлює посилання у разі зміни нумерації омонімів або значень, забезпечуючи цілісність словникової структури.

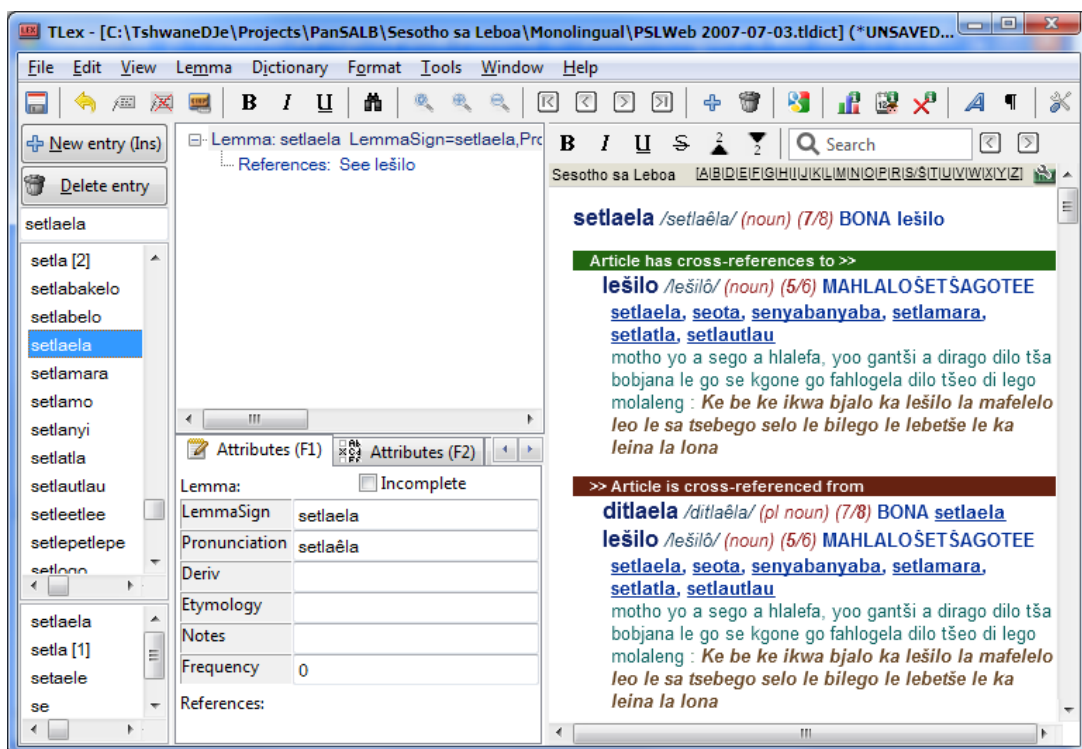


Рис. 7.1 – Головне вікно TLex

Вбудована система пошуку дозволяє здійснювати пошук по всьому корпусу словника з урахуванням регістру, цілих слів або із застосуванням регулярних виразів. Програма також підтримує мультимедійні елементи:

- до будь-якого поля можна додати звукові записи (для вимови лем, прикладів уживання тощо);

- можливе вставлення зображень і відеофайлів безпосередньо в статтю, що розширює дидактичний потенціал електронного словника.

Система стилів і форматування в TLex забезпечує гнучке налаштування візуального вигляду словника – кольори, шрифти, інтервали, автоматичні розділові знаки тощо. Зміни у стилях відображаються централізовано, тобто форматування оновлюється одночасно для всіх записів, що значно спрощує верстку та підготовку до публікації.

Для командної роботи передбачено розділ користувачів із розмежуванням доступу:

- кожен редактор має власний обліковий запис;
- система зберігає історію змін і дозволяє відстежувати, хто і коли редагував конкретну статтю чи її частину;

- фільтри дають змогу переглядати записи, створені або змінені окремими учасниками команди.

Особливу увагу приділено інтеграції корпусних технологій: вбудований корпусний інструмент дає змогу здійснювати пошук прикладів уживання безпосередньо під час редагування статті. Опція «автопошук» автоматично запускає пошук за поточним заголовним словом, а знайдений приклад може бути вставлений у запис одним кліком. Усе це робить TLex одним із найпотужніших середовищ для сучасної електронної лексикографії, придатним як для великих академічних проєктів, так і для навчальних або локальних словників.

IDM Dictionary Publishing System (DPS) – це професійне лексикографічне середовище (рис. 7.2), розроблене компанією IDM Group для створення, редагування та публікації словників у друкованому й електронному форматах. Система орієнтована на великі видавничі проєкти та забезпечує повний цикл роботи з лексикографічним контентом – від збирання матеріалу до багатоформатного випуску. DPS функціонує як централізована база даних, що підтримує багатокористувацький доступ. Це дозволяє декільком лексикографам

одночасно працювати над різними словниками або над різними розділами одного проекту, забезпечуючи узгодженість, контроль версій і високу продуктивність.

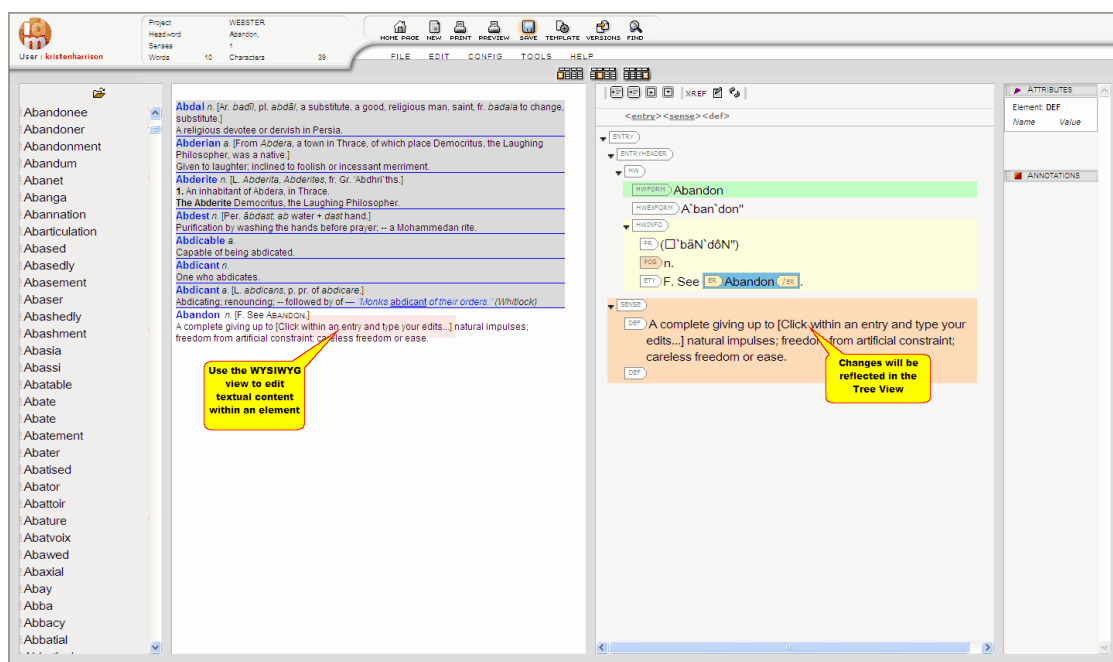


Рис. 7.2 – Основне вікно IDM DPS

Система активно використовується провідними світовими видавництвами, серед яких Oxford University Press, Cambridge University Press та Pearson Education, що свідчить про її професійний рівень і довіру з боку лексикографічної спільноти. Основними характеристиками DPS є:

- *централізоване управління контентом*: усі дані словників зберігаються у єдиній базі, що запобігає дублюванню та втраті інформації, а також гарантує узгодженість між різними версіями;

- *одночасна робота користувачів*: система дозволяє кільком редакторам, укладачам і коректорам одночасно працювати над спільним проектом, що особливо важливо для великих словників з великою кількістю статей;

- *керування користувачами та завданнями*: адміністратор може призначати завдання, визначати дедлайни, контролювати прогрес і розподіляти ролі в команді – від лексикографа до верстальника чи технічного редактора.

- *гнучке налаштування*: DPS адаптується до специфіки кожного словника: можна змінювати шаблони статей, структуру даних або схему XML без переривання роботи системи.

– *масштабованість*: платформа побудована як відкрита й розділна система, що дозволяє інтегрувати нові інструменти, розширювати сховища даних і підтримувати кілька паралельних проєктів.

– *підтримка кількох форматів виводу*: DPS генерує контент у різних форматах – XML, HTML, PDF, а також у спеціалізованих форматах для електронних і мобільних словників. Це забезпечує легку адаптацію до друкованих, онлайн і мультимедійних публікацій.

– *інтелектуальний контроль версій*: система відстежує історію змін у всіх конфігураційних файлах, зберігає попередні версії та дозволяє відновлювати їх у разі потреби. Це дає можливість повторно використовувати готові структури й адаптувати їх для нових проєктів.

Схема роботи в IDM Dictionary Publishing System (DPS) побудована за принципом централізованого управління контентом і поетапного виконання завдань. На початковому етапі адміністратор системи налаштовує параметри проєкту, створює облікові записи користувачів і визначає рівні доступу до бази даних. Далі лексикографам призначаються конкретні завдання, наприклад редагування певного набору словникових статей або додавання нових записів, із зазначенням термінів виконання.

Після отримання завдання користувач завантажує необхідний матеріал, опрацьовує його у своєму робочому середовищі та повертає результати до центральної бази даних. Система автоматично інтегрує оновлену інформацію, синхронізуючи зміни та зберігаючи попередні версії. У разі потреби DPS передає матеріали на наступний етап виробничого процесу – наприклад, до технічного редактора, коректора або публікаційного розділя.

Гнучкість архітектури DPS дозволяє створювати словники різного типу та призначення – від загальномовних і двомовних до галузевих спеціалізованих. Така організація забезпечує високий рівень узгодженості, оперативність оновлень та зменшення дублювання роботи у великих лексикографічних командах. Як видно, IDM DPS є комплексним рішенням для професійної лексикографії, яке поєднує функції управління даними, редакційного контролю,

автоматизації процесів і публікації у різних форматах. Завдяки своїй розділності та масштабованості система підходить як для академічних, так і для комерційних лексикографічних ініціатив. Порівняльна характеристика TLex та IDM DPS подана в таблиці 7.1.

Таблиця 7.1 – Порівняльна характеристика TLex та IDM DPS

Критерій / функція	TLex (TshwaneLex)	IDM Dictionary Publishing System (DPS)
Розробник	TshwaneDJe Software (ПАР)	IDM Group (Велика Британія)
Основне призначення	Створення словників і термінологічних баз будь-якого масштабу – від навчальних до академічних	Професійне середовище для великих видавничих проєктів і колективної роботи над словниками
Тип системи	Настільна програма (з можливістю інтеграції з корпусами та онлайн-середовищами)	Централізована серверна система з базою даних і багатокористувацьким доступом
Інтерфейс	Інтуїтивний, з попереднім переглядом (WYSIWYG), підтримкою «click-to-edit»	Веб- або серверний інтерфейс із керуванням користувачами, завданнями, версіями
Користувачі	Лінгвісти, укладачі словників, викладачі, студенти, дослідники	Великі видавництва, редакційні команди, технічні редактори, адміністратори проєктів
Підтримка мультимедіа	Так – звук, зображення, відео	Частково – для публікації в електронних форматах (XML, онлайн)
Інтеграція з корпусом	Вбудований корпусний пошук та автогенерація прикладів	Може інтегруватися з зовнішніми корпусами через API
Робота з кількома користувачами	Можлива (через окреме налаштування користувачів у проєкті)	Повноцінна колективна робота з централізованим управлінням і контролем версій
Контроль версій і змін	Відстеження змін користувачів на рівні записів	Інтелектуальна система контролю версій, журнал змін і можливість відновлення
Гнучкість і налаштування	Повне налаштування стилів, шаблонів, форматів, полів	Гнучка конфігурація словників без зупинки роботи системи
Масштабованість	Орієнтована переважно на малі та середні проєкти	Висока масштабованість, підтримка багатьох словників і команд
Формати виводу	HTML, PDF, XML, електронні формати для онлайн-публікацій	XML, HTML, PDF, цифрові формати для друку та мультимедіа
Особливі функції	Автоматичне оновлення перехресних посилань, стильова система, click-to-edit	Централізоване управління даними, командна робота, масштабованість
Тип користувача	Індивідуальний або невелика група	Велика редакційна команда, видавництво

Приклади користувачів	Наукові установи, університети, лексикографи	Oxford University Press, Cambridge University Press, Pearson Education
Основна перевага	Простота, швидкість, інтуїтивність	Колективність, масштабність, централізований контроль
Недолік / обмеження	Менша придатність для великих командних проєктів	Вимагає серверної інфраструктури та адміністрування

Отже, TLex найкраще підходить для індивідуальних або навчально-наукових лексикографічних проєктів, де важлива швидкість, наочність і простота. А DPS є оптимальним вибором для великих видавничих домів і академічних інституцій, що потребують командної роботи, централізованого управління та автоматизованого контролю всіх етапів укладання словника. Крім професійних систем існують інші платформи, розраховані на укладачів, які не є професійними лексикографами. До числа таких платформ належать Lexique Pro, Flex та Lexonomy.

Lexique Pro – десктопна комп’ютерна програма, яка за своїми функціями є переглядачем та редактором лексиконів (рис. 7.3). Її гнучкий та зрозумілий інтерфейс дає змогу укласти словники для студентів, лінгвістів і перекладачів. Lexique Pro формує базу даних і форматує її в інтерактивному переглядачі, роблячи роботу зі словником максимально зручною та наочною. Функціонал системи дає змогу створювати перехресні посилання за допомогою гіпертексту, формувати пошукові індекси, вставляти мультимедійні матеріали до словникових статей, генерувати алфавітні та тематичні реєстри, а також експортувати готовий словник у різних форматах – друкованому, десктопному або онлайнному.

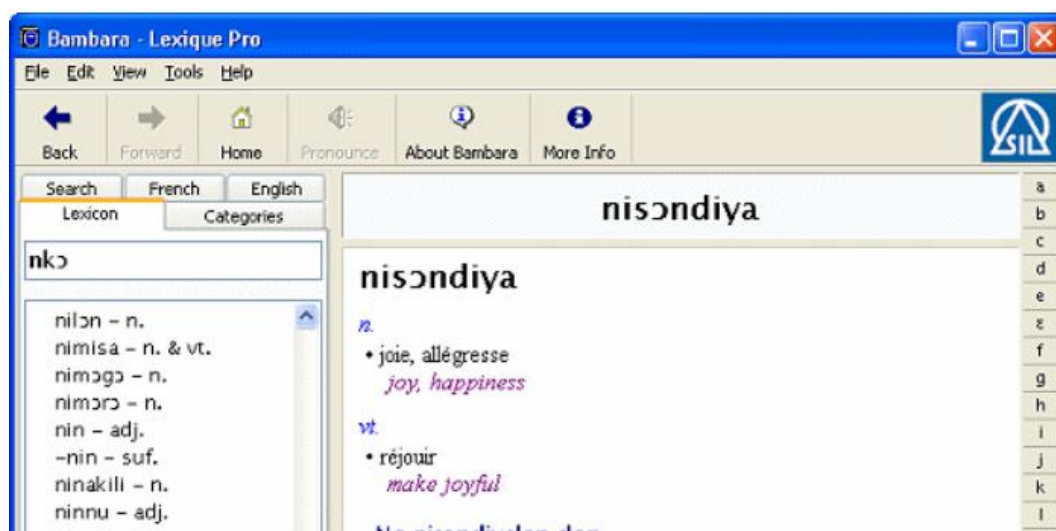


Рис. 7.3 – Головне вікно Lexique Pro

Flex (Full Linguistic EXpedition) – це безкоштовна програма з відкритим кодом, розроблена організацією SIL International для створення, редагування та керування мовними базами даних. Вона орієнтована насамперед на дослідників малодосліджених мов, польових лінгвістів і перекладачів Біблії, однак її потенціал дозволяє ефективно використовувати інструмент і для академічних або освітніх лексикографічних проєктів. Основна перевага Flex полягає у глибокій морфологічній та синтаксичній підтримці: користувач може будувати словотвірні парадигми, вказувати граматичні категорії, описувати фонологічні процеси та взаємозв'язки між словами. Програма також забезпечує експорт даних до інших форматів, включно з Lexique Pro, що дозволяє поєднувати ці системи в одному робочому процесі – від збору даних до публікації готового словника.

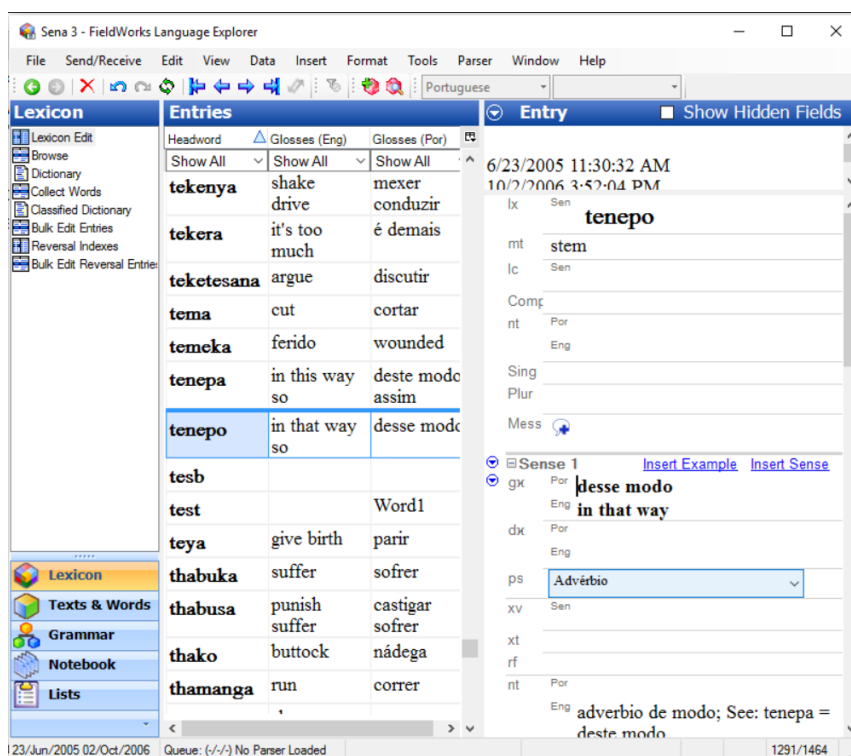
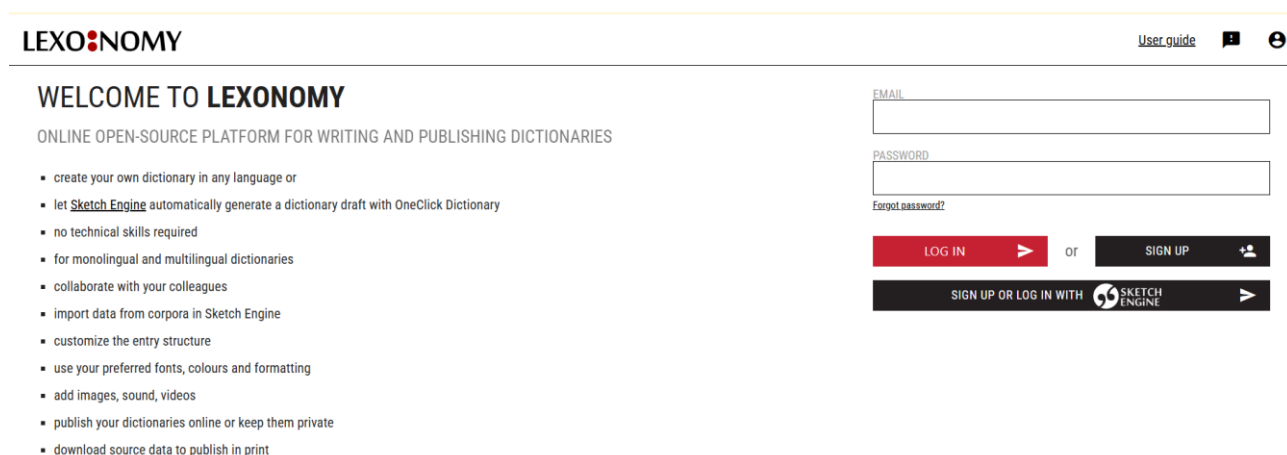


Рис. 7.4 – Головне вікно Flex

Lexonomy – сучасна онлайнова платформа для створення й редагування словників у спільному середовищі. Її головна ідея – зробити лексикографію доступною та колаборативною. Працюючи безпосередньо у браузері, користувач може створювати нові проєкти, редагувати статті, додавати теги, ілюстрації та приклади, а також керувати правами доступу для інших учасників. Lexonomy підтримує зберігання даних у форматі XML, що полегшує подальший імпорт або експорт у професійні системи. Інтерфейс платформи інтуїтивний і підтримує

багатомовність, що робить її зручною як для невеликих локальних словників, так і для масштабних міжнародних лексикографічних ініціатив.



LEXONOMY [User guide](#)

WELCOME TO LEXONOMY

ONLINE OPEN-SOURCE PLATFORM FOR WRITING AND PUBLISHING DICTIONARIES

- create your own dictionary in any language or
- let [Sketch Engine](#) automatically generate a dictionary draft with OneClick Dictionary
- no technical skills required
- for monolingual and multilingual dictionaries
- collaborate with your colleagues
- import data from corpora in Sketch Engine
- customize the entry structure
- use your preferred fonts, colours and formatting
- add images, sound, videos
- publish your dictionaries online or keep them private
- download source data to publish in print

EMAIL

PASSWORD

[Forgot password?](#)

LOG IN or SIGN UP

SIGN UP OR LOG IN WITH  SKETCH ENGINE

Рис. 7.5 – Головна сторінка Lexonomy

Отже, Lexique Pro, Flex і Lexonomy представляють три різні, але взаємодоповнювальні підходи до укладання електронних словників:

- Lexique Pro – для візуалізації та публікації;
- Flex – для глибинного лінгвістичного аналізу;
- Lexonomy – для спільної онлайн-роботи та гнучкого редагування.

Кожна з цих платформ відіграє свою роль у сучасній типології комп'ютерних словників, демонструючи, як цифрові інструменти поступово змінюють саму практику лексикографії – від індивідуального укладача до колективної, інтерактивної діяльності.

Порівнюючи ці три системи, можна відзначити, що з точки зору користувача найбільш доступною є Lexonomy, оскільки вона не потребує встановлення програмного забезпечення, працює в інтернет-браузері й забезпечує миттєвий спільний доступ до словника. Її інтерфейс інтуїтивно зрозумілий навіть для початківців, що робить платформу зручною для навчальних і громадських проєктів. Lexique Pro, натомість, надає більше можливостей для візуальної презентації словника – зручної навігації, гіпертекстових посилань та мультимедійних вставок, що особливо корисно для демонстраційних або популярних електронних видань. Flex має найбільшу дослідницьку глибину, адже дозволяє моделювати мовну систему на рівні фонології, морфології й синтаксису, що робить його незамінним для академічної лексикографії,

польових досліджень і розроблення корпусів мов. З погляду лексикографічної типології, ці три інструменти ілюструють різні напрями розвитку сучасної комп'ютерної лексикографії: від спеціалізованих систем для професіоналів – до універсальних, відкритих платформ, орієнтованих на колективну взаємодію, багатомовність і мультимедійність. Саме завдяки таким інструментам лексикографія стає дедалі більш інтерактивною, динамічною й доступною для ширшого кола користувачів.

Сучасні середовища укладання електронних словників демонструють широкий спектр підходів – від потужних професійних систем типу *IDM Dictionary Publishing System*, орієнтованих на колективну роботу великих видавництв, до доступних і відкритих платформ, як-от *Lexique Pro*, *Flex* та *Lexonomy*, що розширюють можливості індивідуальних дослідників і невеликих лексикографічних спільнот. Усі вони сприяють демократизації лексикографічної діяльності, роблять процес створення словників більш гнучким, інтерактивним і технологічно інтегрованим у цифрове середовище. Завдяки цим інструментам лексикографія поступово перетворюється на комунікативну, колаборативну й мультимедійну практику, де на перший план виходить не лише точність опису, а й зручність користування, оновлюваність і взаємодія між укладачем і користувачем.

Питання для самоконтролю

1. Що таке середовище розроблення словників (DWS / CPC) і які основні завдання воно виконує?
2. Які ключові функції мають сучасні DWS-системи (укладання, керування даними, публікація тощо)?
3. У чому полягає відмінність між професійними лексикографічними системами та платформами для непрофесійних укладачів?
4. Які переваги використання централізованої бази даних у системі *IDM Dictionary Publishing System*?
5. Як у DPS реалізовано функцію спільної роботи кількох користувачів над одним проєктом?

6. Які типи форматів вихідних даних підтримує DPS і чому це важливо для сучасної лексикографії?
7. У чому полягають особливості та переваги роботи з *Lexique Pro*?
8. Які можливості щодо мультимедіа й гіпертексту надає *Lexique Pro*?
9. Які спільні риси мають платформи *Flex*, *Lexonomy* і *Lexique Pro*?
10. Як, на вашу думку, середовище DWS впливає на якість і швидкість створення словника?

Завдання для самоперевірки

1. Знайдіть на офіційних сайтах TLex, Lexonomy та Lexique Pro опис програм і порівняйте їхні інтерфейси. Коротко занотуйте, які функції здалися найзручнішими в кожній системі. Опис може бути в текстовій або табличній формі.
2. Перегляньте демо-версію або скриншоти TLex і Lexonomy. Які типи полів словникової статті ти бачиш? Чи можна відрізнити структуру «лема – значення – приклад»?
3. Складіть порівняльну таблицю для чотирьох середовищ: TLex, IDM DPS, Lexique Pro, Lexonomy. Укажіть параметри:
- тип системи (професійна / напівпрофесійна);
 - архітектура (десктопна / онлайн / гібридна);
 - підтримка мультимедіа;
 - підтримка спільного редагування;
 - формати експорту;
 - переваги;
 - обмеження.
4. Проаналізуйте, яка система (з розглянутих) найкраще підійде для створення:
- а) студентського навчального словника новітніх слів;
 - б) академічного термінологічного словника;
 - в) мультимедійного словника сленгу.
- Обґрунтуйте свою відповідь.

5. Створіть модель словникової статті у вигляді таблиці (наприклад, у Word або Excel). Поля:

- лема
- частина мови
- тлумачення
- приклад уживання
- переклад (якщо двомовний)
- примітка
- гіперпосилання на споріднені слова

Позначте, яке поле могло б бути мультимедійним (звуковий запис, ілюстрація).

6. Зареєструйтеся в Lexonomy (<https://www.lexonomy.eu>) та створіть міні-словник (3–5 статей). Оберіть тему, наприклад, *неологізми українського інтернет-сленгу* або *словник звуків природи*. Коротко опишіть процес (які поля створено, що було зручно, що ні).

7. Розробіть концепцію власного електронного словника (у будь-якій системі):

- назва словника;
- цільова аудиторія;
- тип контенту (мовний, тематичний, енциклопедичний);
- платформа, яку тиобираєш (TLex, Lexonomy, Lexique Pro тощо);
- структура статті (5–7 полів);
- приклад 1–2 статей.

7.2. Віртуальні лексикографічні лабораторії (ВЛЛ)

Віртуальна лексикографічна лабораторія (ВЛЛ) – це інноваційна модель організації лексикографічної діяльності, що поєднує технології віддаленого доступу, спільної праці над базами даних і комплексного лексикографічного аналізу. Термін «віртуальна лексикографічна лабораторія» увів академік Володимир Широков, директор Українського мовно-інформаційного фонду НАН України (УМІФ), для позначення цифрової інфраструктури, що забезпечує

дистанційне редагування, оновлення та використання лексикографічних ресурсів.

Основна ідея полягає у створенні єдиного цифрового середовища, у якому науковці, редактори, мовознавці, програмісти й перекладачі можуть працювати з **лексикографічними базами** незалежно від свого місця перебування. Завдяки цьому концепт ВЛЛ виходить за межі традиційного уявлення про «редакцію словника» і набуває ознак відкритої, інтегративної, колективної платформи. У межах ВЛЛ реалізовано доступ до провідних лексикографічних ресурсів УМІФ, зокрема (рис. 7.6):

- «Словника української мови у 20 томах» (СУМ-20) – сучасної академічної лексикографічної бази, що оновлюється в режимі реального часу;
- «Етимологічного словника української мови» (ЕСУМ) – наукового видання, що відображає походження й історичний розвиток української лексики;
- **граматичних словників**, які автоматично генерують словозміну слів української та інших мов;
- **термінологічних баз даних**, створених на основі даних спеціальних словників.

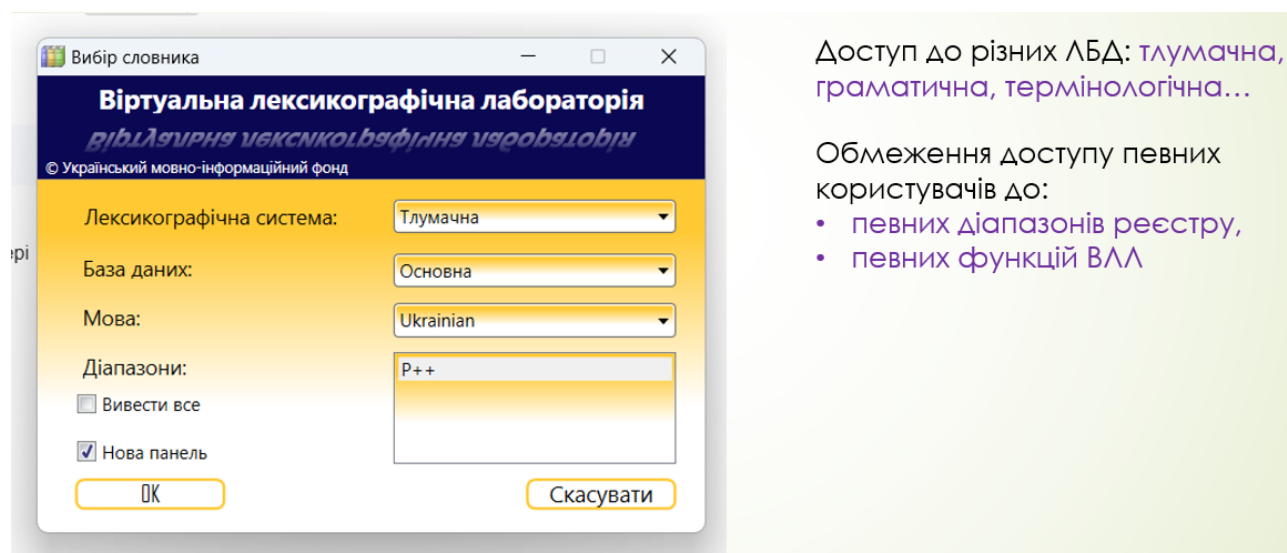


Рис. 7.6 – Вікно доступу до різних ЛБД ВЛЛ

Програмний інструментарій ВЛЛ (рис. 6.7) забезпечує повний цикл лексикографічної обробки даних – від укладання нових **словникових статей** до перевірки, актуалізації й пошуку інформації. Ключові розділи системи охоплюють:

(1) редактор словникових статей, що дозволяє вводити, змінювати та рецензувати записи;

(2) розділ режимів перегляду, що надає вибір різних режимів відображення словникових статей у головному вікні та під час збереження у файлі;

(3) інструменти для здійснення контролю за ходом робіт над словником, а також ведення показників виконання, що забезпечують виконання проєкту у встановлені терміни;

(4) інструменти для формування вибірок, за різними параметрами: лема, значення, частина мови, джерела цитування тощо.

Завдяки хмарним технологіям та централізованій архітектурі ВЛЛ, науковці отримують можливість працювати з базами у **синхронному режимі**, що забезпечує оперативність і узгодженість змін. Кожен учасник має визначену роль від укладача до рецензента, редактора чи адміністратора системи. Отже, ВЛЛ є цифровим ядром сучасної української лексикографії, де поєднано **лінгвістичну експертизу**, інформаційні технології та принципи відкритої науки. Вона сприяє не лише оновленню словникового корпусу української мови, а й також інтернаціоналізації української лексикографії, відкриваючи доступ до даних для міжнародних наукових спільнот.

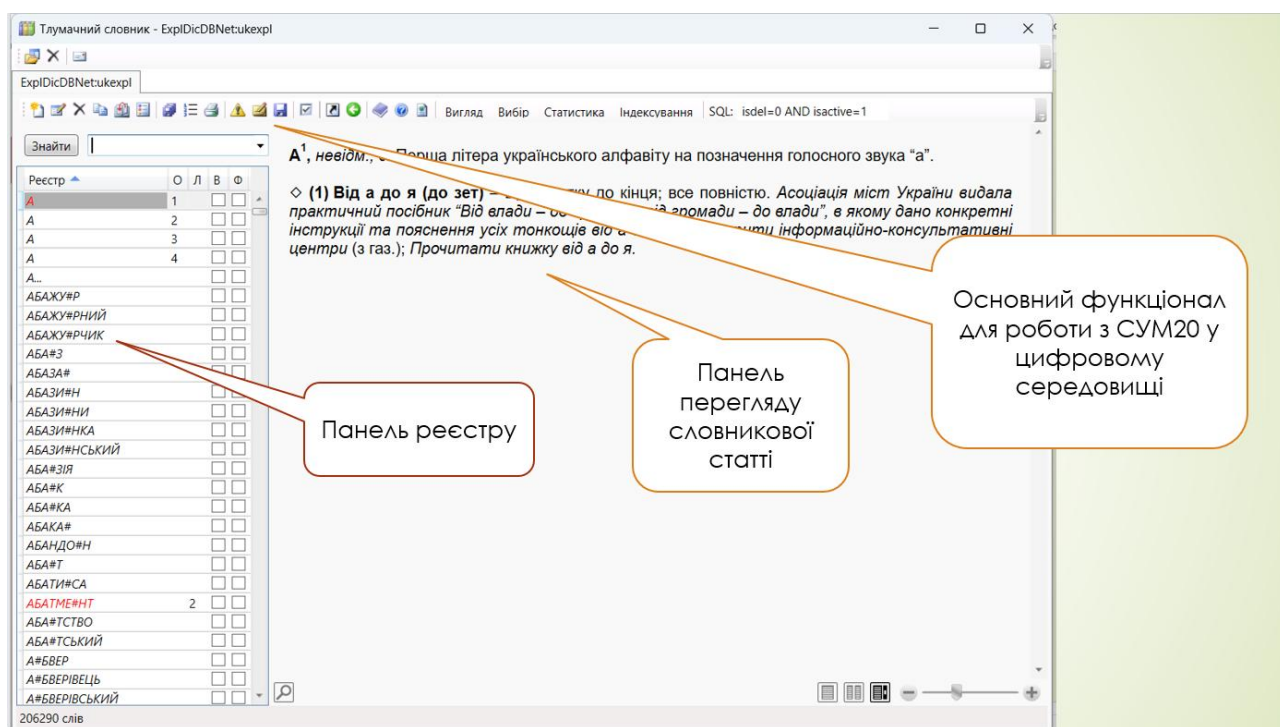


Рис. 7.7 – Головне вікно ВЛЛ СУМ-20

Технологічна архітектура ВЛЛ побудована на принципах розділності, відкритості та взаємодії між компонентами, що забезпечують як автономне функціонування окремих словникових проєктів, так і їхню інтеграцію в єдину інформаційну екосистему. Умовно структура ВЛЛ складається з трьох рівнів:

1. Базовий рівень (серверна інфраструктура) включає сховище даних, сервери опрацювання запитів та резервні копії. Саме тут розміщуються **лексикографічні бази даних (ЛБД)**, які зберігають усю структуровану інформацію: лема, словникові статті, граматичні параметри, семантичні мітки, цитати та приклади.

2. Рівень прикладного програмного забезпечення – це веб- і десктопні клієнти, через які користувачі взаємодіють із системою. На цьому рівні реалізовано функції редагування, перевірки, коментування, лінгвістичного аналізу та експорту даних у різні формати (рис. 7.8).

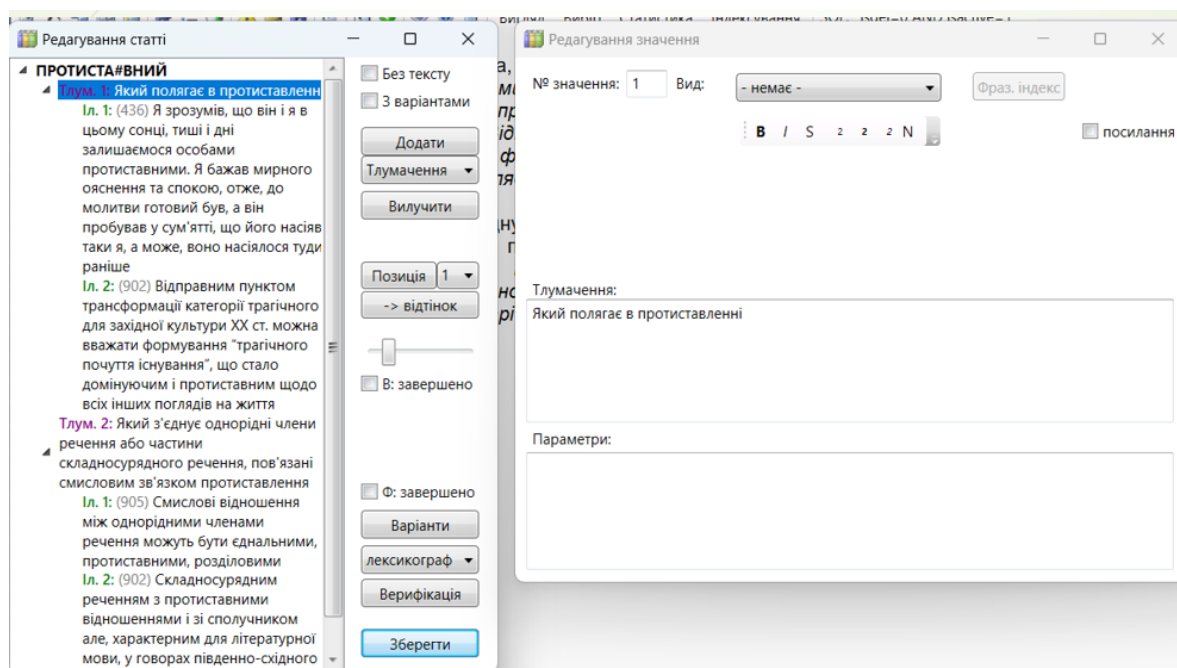


Рис. 7.8 – Вікна редагування словникової статті

3. Інтерфейс користувача (UI/UX-рівень), що забезпечує зручність доступу до словникових матеріалів. Через інтерфейс користувач може виконувати пошук, переглядати, створювати й редагувати словникові статті, створювати звіти, вносити зміни або пропозиції. Архітектура взаємодії користувачів та системи побудована на принципі рольового доступу (рис. 7.9). Учасники проєкту поділяються на кілька категорій:

- *укладачі*: створюють і редагують словникові статті, додають приклади, джерела, семантичні зв'язки.
- *рецензенти та наукові редактори*: перевіряють лінгвістичну достовірність і стиль оформлення матеріалу.
- *адміністратори*: керують доступами, контролюють версії, здійснюють технічний супровід і резервування даних.
- *користувачі (читачі)*: отримують обмежений доступ до перегляду публічних або тестових версій словників.

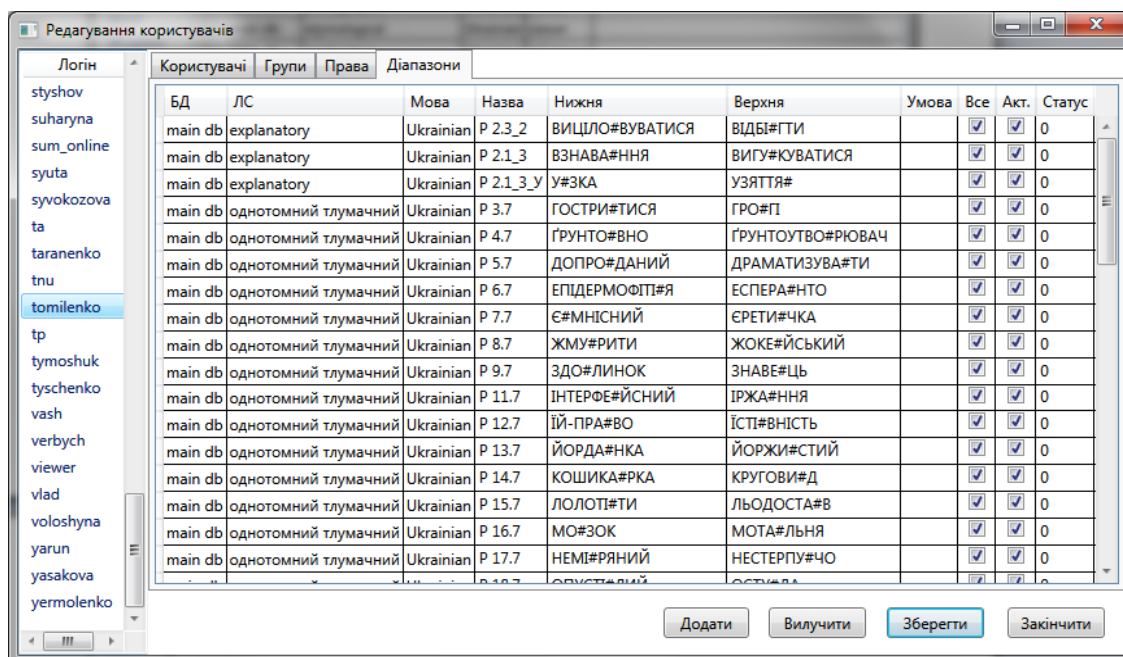


Рис. 7.9 – Вікно редагування користувачів СУМ-20

Система підтримує механізм версіонування, що дозволяє відстежувати історію змін кожної словникової статті. Кожна модифікація фіксується із зазначенням дати, автора та типу зміни (додавання, редагування, вилучення). Важливою складовою архітектури є **семантична інтеграція**, тобто узгодження структури даних між різними словниковими базами. Це дає змогу створювати міжсловникові перехресні посилання – наприклад, від леми у СУМ-20 до відповідної статті в граматичному словнику. Для зберігання та обробки інформації в УМІФ використовується гнучка **модель даних**, яка базується на **XML-структурах і стандартах TEI (Text Encoding Initiative)**. Така модель забезпечує сумісність українських електронних словників із міжнародними лексикографічними проєктами та дозволяє автоматично експортувати дані

в універсальні формати для наукових корпусів. Крім того, ВЛЛ передбачає інтеграцію з інструментами штучного інтелекту: автоматичне розпізнавання частин мови, генерацію словникових прикладів, пошук контекстів у корпусах текстів, а також попередню перевірку узгодженості термінів.

Віртуальні лексикографічні лабораторії (ВЛЛ) стали ключовим інструментом сучасної цифрової лексикографії, оскільки поєднують наукову глибину традиційної лексикографії з оперативністю та гнучкістю цифрових технологій. Їхня поява змінила підхід до створення, перевірки, публікації та підтримки словникових баз, зробивши процеси більш прозорими, колаборативними й довготривалими. Нижче детально висвітлено їх переваги.

1. Ефективна колективна робота. ВЛЛ забезпечує синхронну взаємодію лексикографів, редакторів, технічних спеціалістів і консультантів, які можуть працювати над одним словником із різних місць. Система автоматично координує їхні дії, запобігає конфліктам версій і надає інструменти для контролю якості. Завдяки централізованому сховищу даних кожен учасник бачить актуальний стан словника, має доступ до історії змін. Також система може генерувати аналітичні звіти щодо виконаних завдань (рис. 7.10). Для укладача це означає можливість працювати з будь-якого комп'ютера або навіть планшета, не турбуючись про локальні файли чи версії документів. Для адміністратора – це гарантія узгодженості даних і повного контролю над процесом редагування.

	Статті	Значення	Відтинки	Ілюстрації	Словосполучення	Значення CCC	Ілюстрації CCC	Коментарі
Всього:	206295	238094	50090	1	20491	24020	1	251656
Додано:	37	57	28	484	1	2	26	568
Змінено:	773	346	80	1322	213	21	77	
Вилучено:	0	10	7	199	4	4	11	
Відпрацьовано:	55672				7693			0

Рис. 7.10 – Звіт щодо виконаної роботи над СУМ-20

2. Підвищення якості та актуальності словникових матеріалів. Завдяки вбудованим інструментам валідації та автоматичної перевірки ВЛЛ дозволяє мінімізувати людські помилки. Система може автоматично виявляти неповні

поля, дублікати, невідповідності між граматичними параметрами або семантичними тегами. Вона також підтримує функцію контекстного аналізу, що дає змогу швидко знаходити приклади вживання слова в корпусі української мови. Крім того, регулярне оновлення бази даних дозволяє оперативно реагувати на мовні зміни – появу неологізмів, зміну частотності вживання, нові запозичення тощо. Таким чином, словник у ВЛЛ перестає бути статичним – він стає динамічною системою знань.

3. Прозорість і відтворюваність наукових процедур. Кожна дія користувача у ВЛЛ фіксується – дата, автор, тип операції. Це забезпечує повну наукову трасованість, тобто можливість відстежити, як і чому було внесено певну зміну до словникової статті. Завдяки цьому редактори та науковці можуть реконструювати логіку укладання словникової одиниці, перевірити її джерела або уточнити дані для майбутніх проєктів. Така відтворюваність є надзвичайно важливою для академічних словників, що формуються протягом багатьох років і проходять кілька етапів редагування та публікації.

4. Можливість інтеграції з іншими ресурсами. Сучасні ВЛЛ розробляються з урахуванням принципів відкритої архітектури та інтероперабельності. Це означає, що дані з ВЛЛ можна обмінювати з іншими електронними корпусами, термінологічними базами чи системами автоматичного перекладу. Наприклад, дані зі СУМ-20 можна використовувати у навчальних лінгвістичних додатках або у проєктах цифрової гуманітаристики. Крім того, завдяки стандартам XML і TEI інформація з ВЛЛ легко конвертується у формати, придатні для онлайнових публікацій, інтерактивних словників чи мобільних застосунків.

5. Довготривалість і збереження результатів. Віртуальна лабораторія виконує також функцію архіву наукових даних. Усі версії словника, у тому числі проміжні, залишаються збереженими, що гарантує відновлення інформації навіть через десятиліття. Це особливо важливо для академічних словників великого обсягу, створення яких триває десятиліттями – наприклад, «Словника української мови» чи «Етимологічного словника української мови».

6. *Публічна відкритість та демократизація лексикографії.* Ще однією перевагою ВЛЛ є можливість публічного доступу до частини словникових баз. Користувачі, студенти, перекладачі або дослідники можуть не лише переглядати словникові статті, але й пропонувати доповнення або виправлення (рис. 7.11). Такий підхід наближає лексикографію до **краудсорсингової моделі**, де спільнота стає активним учасником оновлення мовних ресурсів.

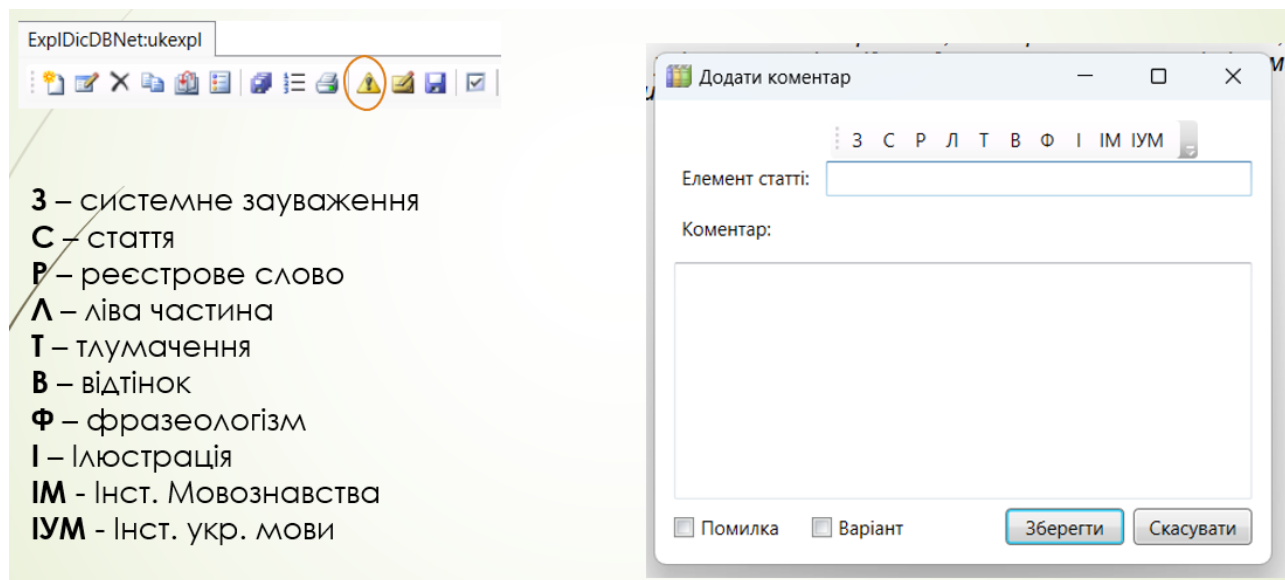


Рис. 7.11 – Вікно додавання коментаря до словникової статті

Віртуальні лабораторії Українського мовно-інформаційного фонду (УМІФ) становлять унікальну технологічну та наукову платформу, що забезпечує інтеграцію національних лексикографічних ресурсів у єдину цифрову екосистему. У межах цієї моделі об'єднано роботу над кількома базовими лексикографічними джерелами – Словником української мови у 20 томах (СУМ-20), Етимологічним словником української мови (ЕСУМ) та низкою граматичних баз даних, зокрема Граматичним словником української мови. Інтеграційна модель УМІФ реалізує принцип взаємопов'язаності лексичних, граматичних та етимологічних рівнів, що робить можливим створення динамічного багатовимірного ресурсу української мови. Зокрема:

- у процесі редагування статей СУМ-20 лексикограф може одразу звернутися до етимологічних коментарів з ЕСУМ або перевірити граматичну характеристику слова в базі морфологічних форм;

- при оновленні граматичних словників дані синхронізуються із лексичним і семантичним описом у СУМ-20;
- система автоматично виявляє невідповідності між різними рівнями опису (лексичним, граматичним, етимологічним) і пропонує корекцію.

Завдяки віртуальним лабораторіям УМІФ забезпечується єдність концепцій і стандартів опису, узгодження термінології, формальних структур та посилань. Цифрова інтеграція дозволяє створювати нові типи аналітичних інструментів – наприклад, автоматичний морфолого-етимологічний пошук або побудову динамічних зв'язків між словниковими статтями різних баз. У перспективі така модель є основою для побудови і розширення Національного лексикографічного порталу української мови, який поєднує словники, корпуси, граматичні ресурси, ономастичні бази й термінологічні системи. Віртуальні лабораторії, що функціонують на базі УМІФ, фактично є прототипом цього порталу та демонструють, як сучасна цифрова лексикографія може забезпечити системність, відкритість і наукову точність у формуванні мовного знання.

Віртуальні лабораторії Українського мовно-інформаційного фонду демонструють приклад сучасної інституційної інтеграції лексикографічних ресурсів, яка охоплює всі рівні мовного опису від лексичного та граматичного до етимологічного й корпусного. Такий підхід створює якісно нову модель національної цифрової лексикографії, що має низку стратегічних переваг, зокрема належать наукова цілісність опису, єдність стандартів і форматів, розширення аналітичних можливостей, відкритість та кооперація. Розглянемо детально наведені переваги.

Наукова цілісність опису. Об'єднання СУМ-20, ЕСУМ і граматичних словників у межах єдиної бази даних забезпечує системність і узгодженість опису української лексики. Лексема розглядається не ізольовано, а в контексті свого походження, морфологічної структури, частотності й функціонування в сучасній мові. Це дозволяє створювати комплексні мовні профілі слів, що мають значення для лінгвістичних досліджень, машинного перекладу, семантичного аналізу й освітніх ресурсів.

Єдність стандартів і форматів. Завдяки спільному інструментарію та стандартизованим форматам подання даних (зокрема XML-структурам) забезпечується взаємна сумісність між різними словниковими базами. Це відкриває шлях до автоматичного обміну інформацією та консолідації мовних даних у рамках національної інфраструктури відкритих даних.

Розширення аналітичних можливостей. Інтегрована модель дозволяє створювати інноваційні інструменти для аналізу – від автоматичного морфемного розбору до виявлення етимологічних зв'язків і побудови динамічних семантичних мереж. Такі механізми сприяють глибшому осмисленню лексичної системи української мови та її розвитку в діяхронічній перспективі.

Відкритість і кооперація. Віртуальні лабораторії створюють умови для дистанційної співпраці лексикографів, редколегій і дослідників. Це дозволяє залучати ширше коло фахівців, підтримувати міжінституційні проекти, а також сприяти публічній науковій комунікації через інтерактивні платформи та електронні ресурси.

Але разом із цим аналізована модель несе у собі певні виклики, до числа яких належать проблема сумісності даних, забезпечення стабільності та якості, технічна підтримка та довготривала сталість та питання доступу й авторських прав.

Однак, разом із цим аналізована модель несе у собі й низку викликів, що постають як на технічному, так і на організаційному рівнях. До числа таких викликів належать, насамперед, *проблема сумісності даних* – необхідність уніфікації форматів, структур і тегів, які використовуються в різних словникових базах, створених у різний час і за різними принципами. Не менш важливим є *забезпечення стабільності та якості* даних у процесі постійного оновлення та інтеграції – це потребує надійних алгоритмів верифікації, редакторського контролю та узгодження між фахівцями різних напрямів. Окрему складність становлять питання *технічної підтримки й довготривалої сталості*: підтримка інфраструктури в актуальному стані, оновлення програмного забезпечення, збереження сумісності з новими версіями баз і платформ. Також не можна

оминути *аспект доступу та авторських прав*, який визначає, хто, у якій формі та на яких умовах може використовувати напрацьовані ресурси – як для наукових, так і для освітніх чи комерційних цілей. Усі ці фактори зумовлюють потребу у виробленні комплексної політики управління цифровими лексикографічними ресурсами та їх інтеграції в межах єдиної системи.

Віртуальні лабораторії УМІФ можуть стати ядром Національної платформи української лексикографії, де об'єднуватимуться корпуси текстів, словники різних типів (тлумачні, перекладні, термінологічні), граматичні й ономастичні бази. Подальший розвиток передбачає інтеграцію з міжнародними проєктами відкритих лексикографічних даних (наприклад, *Global WordNet*, *CLARIN*), що дозволить українській лексикографії стати повноправною частиною світової цифрової лінгвістичної інфраструктури.

Питання для самоконтролю

1. Поясніть, що таке віртуальна лексикографічна лабораторія (ВЛЛ).
2. Хто запровадив цей термін і з якою метою?
3. Які словники Українського мовно-інформаційного фонду реалізовано у форматі ВЛЛ?
4. Назвіть основні функції віртуальної лексикографічної лабораторії.
5. У чому полягає відмінність ВЛЛ від традиційних комп'ютерних словникових систем?
6. Поясніть, як ВЛЛ сприяє підвищенню якості словникових матеріалів.
7. Проаналізуйте схему взаємодії лексикографів у ВЛЛ. Які переваги й ризики вона має?
8. Розкрийте поняття наукової трасованості в контексті лексикографії.
9. Як у ВЛЛ реалізовано контроль версій і відстеження змін у словниковій базі?
10. У чому полягає інтероперабельність ВЛЛ і чому вона важлива для сучасних лінгвістичних досліджень?
11. Чи може ВЛЛ повністю замінити традиційне видавництво словників?

12. Які загрози для збереження мовної достовірності може створювати краудсорсинг у ВЛЛ?
13. Чи варто відкривати частину функцій ВЛЛ для публічного редагування, як у Wikipedia?
14. Як використання ВЛЛ змінює роль лексикографа у XXI столітті?
15. У чому полягає значення ВЛЛ для розвитку української наукової лексикографії?

Завдання для самоперевірки

1. Складіть короткий опис (до 10 речень) уявної віртуальної лабораторії для студентського лексикографічного проєкту. Вкажіть її цілі, учасників, типи даних і очікувані результати.
2. Розробіть порівняльну таблицю: традиційна лексикографічна робота vs робота у ВЛЛ.
3. Опишіть, як ВЛЛ може бути інтегрована з корпусом української мови або з навчальною платформою Moodle.
4. Створіть схему життєвого циклу словникової статті у ВЛЛ: від створення до публікації.
5. Підготуйте коротке есе (1–1,5 стор.) на тему: «Віртуальні лексикографічні лабораторії як крок до відкритої науки і цифрової гуманітаристики».

7.3. Інструменти й сервіси для лексикографів

Не менш важливу роль у створенні словників, розвитку лексикографічної думки та вдосконаленні технологій комп'ютерної лексикографії відіграють спеціалізовані сервіси та інструменти, які можуть бути корисними на різних етапах роботи над лексикографічним проєктом – від збирання корпусних даних до оцінювання якості готових словників. У цьому підрозділі розглянуто інструменти й сервіси, що складають Європейську лексикографічну інфраструктуру ELEXIS (European Lexicographic Infrastructure, <https://elex.is/>) – масштабний проєкт, який фінансується Європейським Союзом із метою створення єдиної цифрової платформи для лексикографії.

Метою ELEXIS є розроблення комплексних цифрових рішень для об'єднання словникових ресурсів, забезпечення їхньої взаємодії, сумісності та доступу через загальну платформу з відкритим інтерфейсом. Усі доступні сервіси й інструменти зібрано на офіційному сайті проєкту (<https://elex.is/tools-and-services/>), де користувач може ознайомитися з їхнім функціоналом. Залежно від типу облікового запису, деякі інструменти пропонують повний набір можливостей (для передплатників), а інші — лише демонстраційний режим. Основною платформою екосистеми ELEXIS є система Sketch Engine, до якої інтегровані всі інші сервіси. Наявність акаунта в Sketch Engine автоматично забезпечує доступ до всіх пов'язаних ресурсів і розділів, що значно спрощує роботу дослідника або укладача словника.

Sketch Engine (<https://www.sketchengine.eu/>). Система запитів, побудови та управління корпусами Sketch Engine дає змогу користувачам створювати та працювати з понад 550 текстовими корпусами на більш ніж 90 мовах і 30 алфавітах. Вона містить низку унікальних інструментів для аналізу великих корпусів обсягом до 60 мільярдів слів, що відкриває широкі можливості для автоматичного виявлення лексичних зв'язків, сполучуваності та типових контекстів. Крім того, система пропонує повністю автоматизовану функцію побудови словників на основі корпусних даних, що суттєво скорочує час попередньої аналітичної роботи (рис. 7.12).

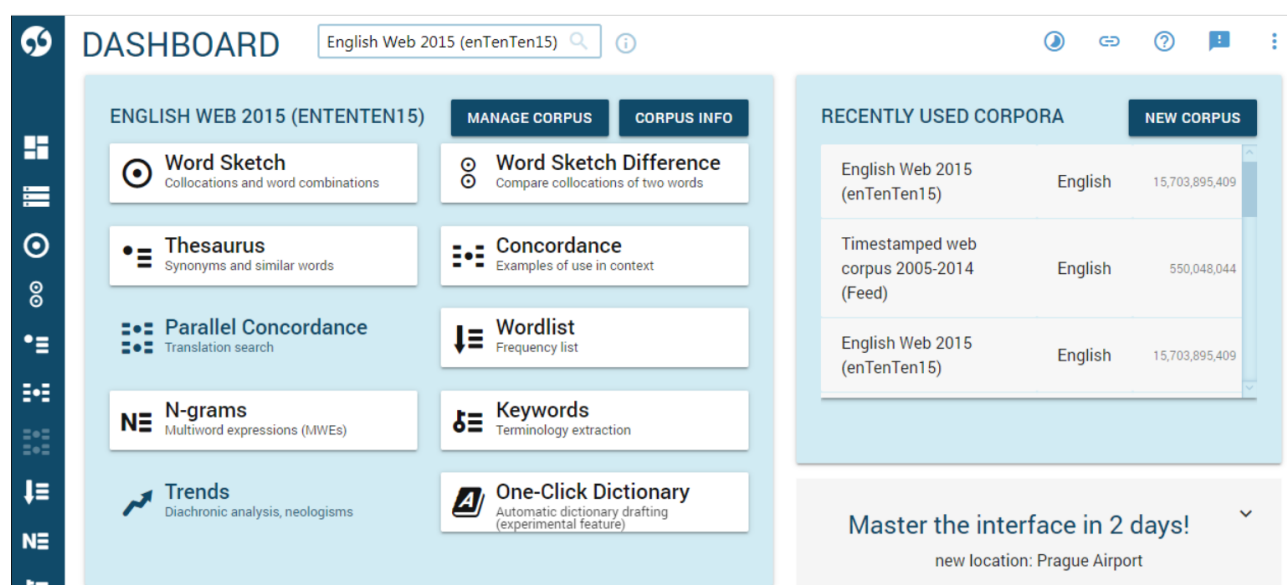


Рис. 7.12 – Головна панель Sketch Engine

Будь-яка лексикографічна робота починається з відбору матеріалу, і в цьому аспекті Sketch Engine надає доступ до великої кількості корпусів різного призначення – загальномовних, спеціалізованих, одномовних або паралельних. Із них можна вилучати як окремі слова, так і словосполучення чи колокації. Будь-який користувач, що має акаунт у системі, може створювати власні корпуси, завантажуючи електронні документи. Система автоматично індексує слова, проводить частотний аналіз і створює конкорданс – список усіх вживань слова в контексті. На основі корпусу користувач може отримати:

а) словниковий реєстр із можливістю налаштування параметрів (обсяг вибірки, частотність, стоп-слова тощо);

б) контексти вживання слів із можливістю фільтрації для добору найрепрезентативніших прикладів, що ілюструють значення слова у типовому мовному середовищі;

в) колокації, тобто поєднання слова з іншими лексемами різних граматичних класів, що дозволяє простежити закономірності сполучуваності.

Системи Sketch Engine та Lexonomy мають спільну інтеграційну платформу: користувач може безпосередньо підключати створені корпуси до Lexonomy та працювати з ними під час створення чи редагування словникових статей. Така взаємодія дозволяє поєднати корпусну аналітику з процесом укладання словників у межах єдиного робочого середовища. Одним із корисних розділів є OneClick Dictionary, про який детально викладено нижче.

OneClick Dictionary (OCD). Це розділ для автоматизованого створення словників, який поєднує систему управління корпусом (наприклад, Sketch Engine, noSketch Engine) або навіть звичайні таблиці Excel із платформою для створення та онлайн-публікації словників Lexonomy. OCD дає змогу автоматично сформувати початковий проєкт словника – із заголовними словами, словоформами, словосполученнями та прикладами вживання (рис. 7.13), який згодом імпортується та редагується лексикографом у Lexonomy.

Завдяки цьому інструменту лексикограф має можливість перенести свій основний інтелектуальний ресурс на етап аналітичного редагування, уникаючи

рутинної ручної обробки даних перед створенням чернетки. Таким чином, OCD придатний не лише для професійної лексикографії, а й для так званої *спонтанної лексикографії* – невеликих проєктів глосаріїв, спеціалізованих чи навчальних словників, що їх часто укладають викладачі або фахівці без спеціальної підготовки.

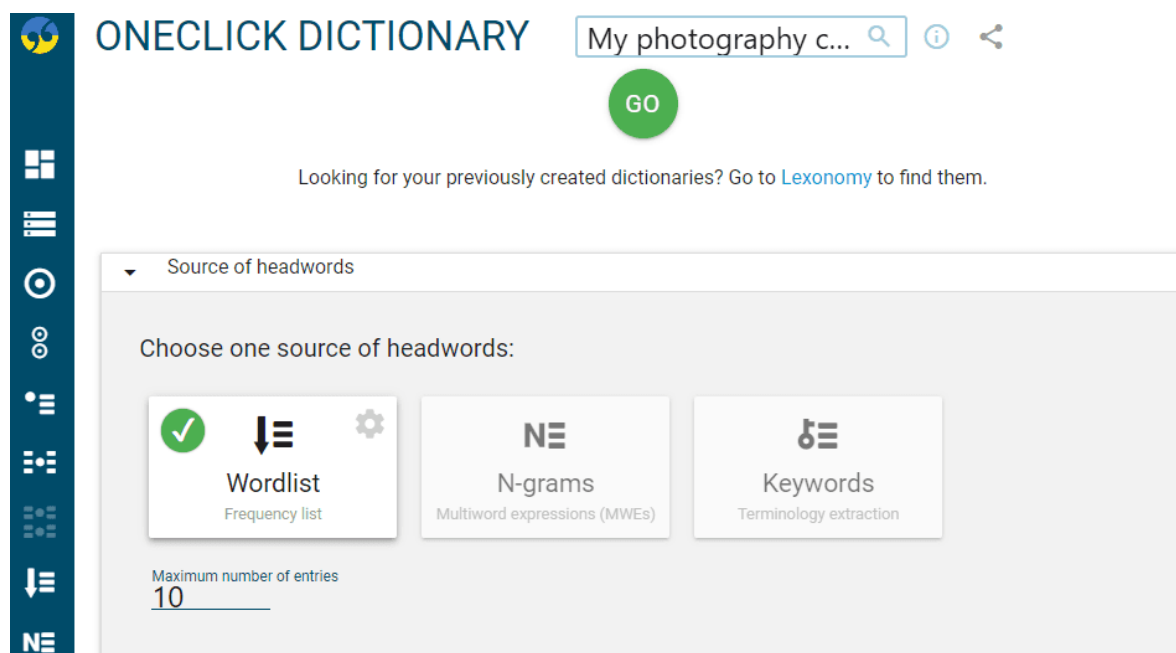


Рис. 7.13 – Головне вікно розділя OneClick Dictionary

Цей розділ створено як додатковий інструмент для автоматизованого укладання словникових статей на основі корпусних даних, отриманих із системи Sketch Engine. Він дає змогу з мінімальними зусиллями сформувати структуровану словникову базу, що включає ключові лексичні одиниці, їхні граматичні характеристики, частотні показники, типові контексти та колокації. Розділ дає можливість користувачеві обрати інформаційні елементи, на основі яких будуть створені словникові статті. Проте він не має функції автоматичного створення дефініцій, тому етап семантичного опрацювання потребує участі редактора або лексикографа. Така побудова системи забезпечує баланс між автоматизацією процесів і контролем за якістю лексикографічного опису.

Elexifier (<https://elexifier.elex.is/>) – це хмарна служба для перетворення словників у стандартизований формат, придатний для машинного зчитування та подальшої інтеграції в інші лексикографічні середовища. Вона використовує передові технології XML-аналізу та машинного навчання, що дозволяє

автоматизувати процес конвертації навіть складних словникових структур. Користувачі можуть завантажувати свої PDF- або XML-словники до Elexifier, визначати правила відображення для XML-перетворення або створювати навчальний набір для системи машинного навчання, яка виконує автоматичне розпізнавання структур у PDF-файлах. У результаті користувач отримує конвертований словник у форматі, сумісному зі стандартом TEI, побудований на основі моделі даних ELEXIS (рис. 7.14).

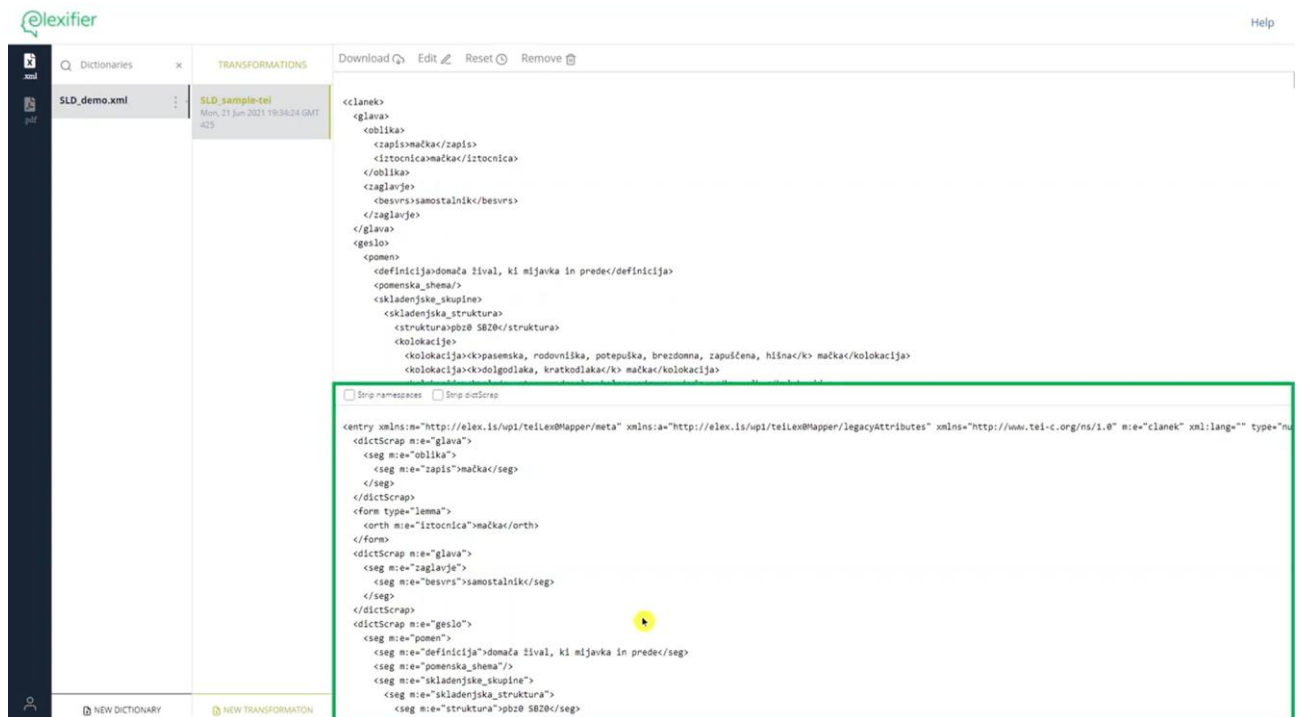


Рис. 7.14 – Вікно інструмента Elexifier

Аналізований інструмент виконує парсинг словника, завдяки чому отриманий ресурс можна у подальшому використовувати у системах обробки природної мови (NLP) або на платформах, подібних до Lexopomtu (чи інших систем редагування словників, наприклад СРС), для створення його повноцінного електронного аналога. Під час здійснення конвертації словника користувачеві необхідно чітко вказати структурні елементи словникової статті та їхні формальні маркери – це забезпечує коректність розпізнавання ієрархії словникових компонентів і підвищує точність отриманої структури. Таким чином, Elexifier не лише спрощує технічний процес конвертації, але й забезпечує стандартизацію словникових ресурсів, що є необхідною умовою для подальшої їхньої інтеграції в цифрові інфраструктури лексикографії.

EDiE (ELEXIS Dictionary Evaluator) – це інструмент для оцінювання якості, доступності та зручності використання лексикографічних ресурсів, опублікованих через API ELEXIS (<https://elexis-eu.github.io/elexis-rest/elexis.html>). Його основне завдання – забезпечити порівняльний аналіз словників, інтегрованих у систему ELEXIS, і допомогти користувачам (дослідникам, лексикографам, викладачам) у виборі найвідповідніших ресурсів для конкретних дослідницьких або практичних цілей. EDiE надає користувачам такі можливості:

– *Оцінка та порівняння*: система аналізує різні параметри словників, включно з їхніми метаданими та структурою словникових записів. Створюючи профілі для кожного ресурсу, EDiE забезпечує об’єктивне порівняння словників за низкою показників.

Ідентифікація пов’язаних даних: на основі створених профілів EDiE може виявляти потенційні точки зв’язку між словниками, сприяючи їх інтеграції та формуванню єдиного взаємопов’язаного лексичного простору.

Агреговані метрики: інструмент обчислює узагальнені статистичні показники (обсяг, оновлюваність, тип джерела тощо), що дає змогу порівнювати словники між собою та визначати, які з них найповніше відповідають потребам користувача.

Оцінка доступності та зручності використання: окремий розділ EDiE аналізує, наскільки легко користувачеві отримати доступ до словникових даних, зрозуміти їх структуру та використовувати їх у власних дослідженнях.

Таким чином, EDiE виступає не лише інструментом контролю якості, а й аналітичною платформою, яка підтримує стандартизацію, прозорість та порівнюваність лексикографічних ресурсів у межах екосистеми ELEXIS. Його використання сприяє підвищенню якості цифрових словників і створенню спільного європейського лексикографічного простору.

Elexifinder (<https://finder.elex.is/>) – це спеціалізований пошуковий інструмент, створений у межах інфраструктури ELEXIS для пошуку та навігації в наукових публікаціях, присвячених лексикографії, корпусній лінгвістиці та суміжним дисциплінам (рис. 7.15). Його головна мета – забезпечити лексикографів,

дослідників і розробників словників швидким доступом до актуальних теоретичних і практичних матеріалів, що стосуються сучасних підходів до створення й аналізу словників. Основні функціональні можливості ELEXIFINDER:

– *Пошук за поняттями та контекстами*: користувач може здійснювати пошук за ключовими словами, концептами або термінами, що мають відповідні сторінки у Вікіпедії, а також за іншими параметрами (автор, мова, джерело, конференція тощо).

– *Розширений контент*: база ELEXIFINDER охоплює наукові статті, відео, конференційні матеріали, звіти та презентації, що робить його універсальним інформаційним інструментом для дослідників.

– *Фільтрування та категоризація результатів*: система дозволяє сортувати результати за різними критеріями, що сприяє швидкому знаходженню матеріалів, релевантних до конкретної теми або дослідницького напрямку.

– *Інтеграція з іншими інструментами ELEXIS*: знайдені джерела можуть бути використані для формування корпусів, поповнення баз знань або для безпосереднього цитування в рамках проєктів, пов'язаних з Sketch Engine, Lexonomy чи EDiE.

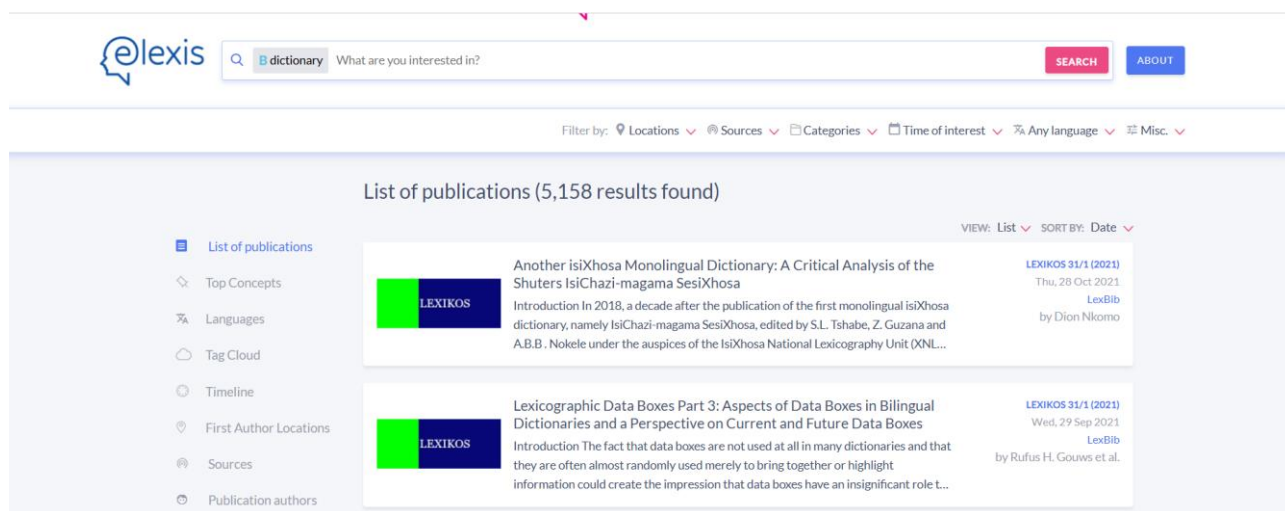


Рис. 7.15 – Вікно інструмента Elexifinder

Отже, ELEXIFINDER є не просто бібліографічним пошуком, а й також інтелектуальним дослідницьким навігатором, що допомагає фахівцям орієнтуватися в актуальних наукових тенденціях, відстежувати нові публікації та

підтримувати академічну взаємодію в межах міжнародної лексикографічної спільноти.

Новинний канал ELEXIS (<https://elex.is/tools-and-services/lexicographic-news/>) – це інформаційно-аналітичний сервіс, який забезпечує автоматичне збирання, оброблення та подання новин, пов'язаних із лексикографією, мовознавством і цифровими гуманітарними науками. Його основою є API Event Registry, що здійснює моніторинг близько 30 000 джерел новин у понад 35 мовах світу. Ключові функції та особливості сервісу:

- *Автоматизоване вилучення новин*: система щоденно сканує інформаційний простір і відбирає матеріали, що мають відношення до тематики лексикографії, корпусної лінгвістики, цифрових словників і мовних технологій.

- *Мультимовна підтримка*: завдяки багатомовній архітектурі Event Registry новинний канал ELEXIS забезпечує доступ до інформації з різних регіонів світу, дозволяючи відстежувати глобальні тенденції розвитку лексикографії.

- *Фільтрація та персоналізація контенту*: користувачі можуть відбирати матеріали за тематичними категоріями, мовами або типами джерел, формуючи власну стрічку новин відповідно до професійних інтересів.

- *Інтеграція з платформами ELEXIS*: новини можуть бути пов'язані з іншими інструментами системи (ELEXIFINDER, EDiE, Sketch Engine), що дозволяє відразу переходити від новинних повідомлень до відповідних наукових джерел або корпусних даних.

У такий спосіб Новинний канал ELEXIS виконує функцію інформаційного мосту між науковими дослідженнями, практикою лексикографії та професійною спільнотою, підтримуючи безперервний обіг знань і формуючи спільний інформаційний простір для розвитку європейської та світової лексикографії.

Отже, інфраструктура ELEXIS утворює єдину інтегровану екосистему цифрових інструментів, що охоплює всі етапи лексикографічного процесу – від збирання корпусних даних до їх аналітичної обробки, автоматизованого створення словникових статей, оцінювання якості лексичних ресурсів і поширення наукових знань. Sketch Engine забезпечує корпусну основу для

досліджень і побудови словників; OneClick Dictionary створює первинні проекти словників із корпусів; Elexifier стандартизує структури словників і готує їх до подальшої обробки; EDiE аналізує та оцінює якість лексикографічних ресурсів; Elexifinder допомагає орієнтуватися в наукових публікаціях і новітніх тенденціях, а Новинний канал ELEXIS підтримує актуальний інформаційний простір лексикографічної спільноти. У взаємодії ці інструменти формують нову парадигму цифрової лексикографії, де головну роль відіграють взаємозв'язок даних, стандартизація форматів, автоматизація аналітичних процесів і відкритість доступу. Саме завдяки таким екосистемам, як ELEXIS, сучасна лексикографія поступово переходить від локальних словникових проєктів до глобальної мережі взаємопов'язаних лексичних знань.

Питання для самоконтролю

1. Що таке ELEXIS і яка його головна мета?
2. Які типи ресурсів об'єднує інфраструктура ELEXIS?
3. Яке призначення має система Sketch Engine?
4. У чому полягає відмінність між корпусом текстів і словником?
5. Які функції виконує Lexonomy?
6. Як пов'язані між собою Sketch Engine і Lexonomy?
7. Що таке Elexifier, і яку роль відіграє парсинг у його роботі?
8. Для чого використовується інструмент EDiE?
9. Які типи оцінювання здійснює EDiE (метадані, доступність, порівняльний аналіз)?
10. Які можливості має OneClick Dictionary, і які обмеження у нього є?
11. Як автоматизація процесів у Lexonomy змінює роль лексикографа?
12. Яким чином інтеграція корпусних інструментів полегшує укладання словників нового покоління?
13. Які переваги й ризики має використання відкритих API у лексикографічних системах?
14. Чому функція автоматичного створення дефініцій поки не реалізована у більшості систем?

Завдання для самоперевірки

1. Зареєструйтесь у системі Sketch Engine (демоверсія) і створіть власний корпус українських текстів. Визначте найчастотніше слово та побудуйте список його колокацій. Зробіть скриншот або короткий опис результатів.

2. Використайте Lexonomy для створення пробного словника. Створіть 3 словникові статті з прикладами, перекладом і граматичною інформацією. Порівняйте структуру вашого словника зі СУМ-20.

3. Спробуйте за допомогою Elexifier (демоверсії) конвертувати невеликий XML-файл словника у формат TEI. Які елементи розпізнає система автоматично, а які вимагають ручної верифікації?

4. Розробіть прототип інтерфейсу українського електронного словника нового покоління (опис або схема).

5. Розробіть схему (у вигляді таблиці) взаємозв'язків між основними сервісами ELEXIS:

- Sketch Engine
- Lexonomy
- Elexifier
- EDiE
- OneClick Dictionary
- Elexifinder.

Питання для поглибленого вивчення матеріалу

1. Віртуалізація лексикографічних систем. Віртуальні лексикографічні лабораторії.

2. Віртуальна лексикографічна лабораторія ВЛЛ СУМ-20 та Український національний лінгвістичний корпус (УНЛК).

3. Віртуальні лексикографічні лабораторії у сучасній лексикографії. Досвід Українського мовно-інформаційного фонду НАН України.

4. Лексикографічні інструменти нового покоління: від словникової спадщини до інноваційної дидактичної практики.

5. Інструмент OneClickTerms для створення одномовних та багатомовних глосаріїв.
6. Використання можливостей платформ SketchEngine та NoSketchEngine в електронному словникарстві.
7. Лексикографічні аддони до Інтернет-браузерів (Dictionaryz, eJOY AI Dictionary, Google Dictionary, DictZone English dictionaries, Instant Dictionary by GoodWordGuide.com та інші).
8. Автоматизація процесу укладання словника: створення та постредагування.

8. КОРПУСНІ ТЕХНОЛОГІЇ У КОМП'ЮТЕРНІЙ ЛЕКСИКОГРАФІЇ

План заняття:

8.1. Текстові корпуси як основа для створення словників

8.2. Корпусні інструменти для створення словників

8.3. Електронні словники в корпусній лінгвістиці

Текстовий корпус – це спеціально сформоване, структуроване та комп'ютерно оброблене зібрання писемних або усних текстів певної мови, що відображає її сучасний стан або історичний розвиток. Корпуси формуються з дотриманням низки обов'язкових параметрів – репрезентативності, розмітки, збалансованості та пошукової доступності. Їхнє головне призначення – забезпечення можливості наукового аналізу та практичного вивчення мови.

Для лексикографії корпуси є незамінним джерелом фактичного матеріалу, на основі якого укладачі словників описують лексику, простежують її уживання, стилістичні відтінки, частотність і сполучуваність. Саме корпуси дозволяють виявляти закономірності функціонування слів у реальних мовних контекстах, що унеможливорює суб'єктивізм і забезпечує об'єктивність словникових даних. Корпусні приклади допомагають формулювати точні дефініції, добирати ілюстрації, фіксувати сталі вирази, колокації, терміни та тематичні поля.

Інакше кажучи, жоден сучасний словник не обходиться без корпусної бази – загальномовної або спеціалізованої. Застосування корпусів дозволяє не лише відтворити природне мовне середовище, а й забезпечити автоматизацію багатьох етапів лексикографічного процесу. Сучасні корпусні інструменти дають змогу не лише здійснювати пошук, аналіз і статистичну обробку лексичних одиниць, а й автоматично генерувати електронні словники на основі корпусних даних. У подальших підрозділах розглядатимуться текстові корпуси як основа для створення словників, використання корпусних даних, а також основні програмні платформи, що використовуються в сучасній комп'ютерній лексикографії.

8.1. Текстові корпуси як основа для створення словників

Текстові корпуси становлять фундамент сучасної лексикографії, оскільки забезпечують достовірні статистичні дані про функціонування мовних одиниць

у реальному мовленні. Вони дають змогу визначати частотність уживання слів, аналізувати їх семантичні відтінки, синтаксичні ролі та контекстні зв'язки, що дозволяє об'єктивно фіксувати мовні явища. Завдяки корпусам лексикографи мають можливість описувати лексику не інтуїтивно, а спираючись на фактичні мовні спостереження.

Корпусні дані використовуються на всіх етапах створення словників: від відбору реєстрових одиниць до написання дефініцій і добору ілюстрацій. На основі корпусів укладаються частотні та колокаційні словники, тлумачні й перекладні ресурси, а також галузеві термінологічні бази. Крім того, корпуси дають змогу перевіряти гіпотези щодо норми й варіантності, простежувати динаміку мовних змін, виявляти нові слова (неологізми) та тенденції в уживанні лексики. Отже, корпуси виступають не лише джерелом прикладів для словників, а й інструментом формування сучасних лексикографічних принципів, у яких переважає доказовість, системність і відкритість даних.

Головною характеристикою корпусу, що відрізняє його від простої вибірки текстів, є наявність **розмітки**. Розмітка полягає у *приписуванні текстам або їхнім компонентам спеціальних міток, які позначають певні види лінгвістичної інформації чи метаданих*. У корпусах зазвичай використовують три основні типи розмітки: зовнішню (екстралінгвістичну), структурну та лінгвістичну.

Екстралінгвістична розмітка надає зовнішні характеристики тексту, що описують обставини його створення і не належать до самого змісту. До таких характеристик належать:

- відомості про автора: ім'я, вік, стать, роки життя тощо;
- відомості про текст: назва, рік і місце видання, а також жанр і тематика.

Таке кодування інформації має назву метарозмітки. Нижче на малюнку 8.1 показано приклад зовнішньої розмітки. Теги <creation>, <textClass>, <textDesc>, <particDesc> та <settingDesc> позначають, відповідно, інформацію про місце і час створення тексту, класифікація тексту, опис тексту, опис автора і учасників комунікації та умови здійснення акту комунікації.

```

<teiHeader id="TRIF1" target="CRONIN/dom.txt" type="text" lang="en">
  <fileDesc>
    <titleStmt>
      <title>The Citadel</title><author>J. A. Cronin</author>
      <extent type="w">45238</extent>
      <extent type="u">4</extent>
      <extent type="s">4132</extent>
    </titleStmt>
    <sourceDesc>
      <respStmt><resp>Little, Brown and Co.</resp></respStmt>
      <address>https://lingualeo.com/en/jungle/croninaj-the-citadel-143681</address>
    </sourceDesc>
  </fileDesc>

```

Рис. 8.1 – Метарозмітка: дані про автора

Структурна розмітка відображає внутрішню організацію тексту та його ієрархію. Вона визначає межі основних структурних одиниць – абзаців, речень, реплік діалогу, віршових рядків, а також форматні елементи (заголовки, підписи, списки тощо). Така розмітка дає змогу системі правильно відтворювати текст, проводити пошук на рівні структурних компонентів і забезпечує зручність подальшої обробки корпусу. Приклад структурної розмітки в XML-форматі може мати вигляд, як це показано на рисунку 8.2.

```

<text id="001" title="Місто">
  <paragraph id="p1">
    <sentence id="s1">Київ – столиця України.</sentence>
    <sentence id="s2">Він розташований на берегах Дніпра.</sentence>
  </paragraph>
</text>

```

Рис. 8.2 – Приклад структурної розмітки

У цьому прикладі теги <text>, <paragraph> і <sentence> позначають відповідно текст, абзац і речення. Така ієрархічна структура допомагає системі точно ідентифікувати межі мовних одиниць і використовувати їх для подальшого лінгвістичного аналізу. Наступним етапом є **лінгвістична розмітка**, яка

поглиблює опис тексту, додаючи морфологічну, синтаксичну або семантичну інформацію.

Лінгвістична розмітка – це найбільш змістовний рівень кодування корпусу, що забезпечує глибокий мовний аналіз тексту. Вона охоплює кілька рівнів: морфологічний, синтаксичний та семантичний. **Морфологічна розмітка** додає до кожного слова інформацію про його частину мови, граматичні категорії (рід, число, відмінок, час, аспект тощо) та базову форму (лему). Це дозволяє будувати частотні списки лем, аналізувати граматичні тенденції та створювати автоматизовані словникові реєстри. **Синтаксична розмітка** визначає граматичні зв'язки між словами (підмет, присудок, додаток тощо) та структуру речень. Вона використовується для аналізу моделей керування, побудови синтаксичних дерев і дослідження закономірностей поєднання слів. **Семантична розмітка** позначає значення слів або фраз, тематичні ролі, кореференції, типи предикатів і може бути пов'язана зі словниками тезаурусного типу або онтологіями (наприклад, WordNet). Приклад фрагмента морфологічно розміченого тексту подано на рисунку 8.3.

```
<w lemma="Київ" pos="noun" case="nom" gender="masc">Київ</w>
<w lemma="бути" pos="verb" tense="pres" person="3">є</w>
<w lemma="столиця" pos="noun" case="nom" gender="fem">столиця</w>
<w lemma="Україна" pos="noun" case="gen" gender="fem">України</w>
```

Рис. 8.3 – Приклад морфологічної розмітки

У такому представленні кожне слово позначене тегом <w> із набором атрибутів, які кодують його граматичні характеристики. Це дозволяє системі здійснювати складний пошук – наприклад, знаходити всі прикметники в орудному відмінку або всі вживання певної леми незалежно від форми. Лінгвістична розмітка є основою корпусної лінгвістики, оскільки саме вона робить тексти «читабельними» для машинного аналізу та забезпечує інтеграцію корпусів із лексикографічними платформами, такими як **Sketch Engine**, **NoSketch Engine** або **UDPipe**. Окрім них, існують й інші системи, що розробляються або використовуються науковими установами та організаціями, які працюють у галузі комп'ютерної лінгвістики та прикладної лексикографії. **UDPipe**

призначена насамперед для розробників програмного забезпечення та професійних користувачів, тоді як **Sketch Engine** і **NoSketch Engine** орієнтовані на ширше коло лінгвістів, дослідників і викладачів. Обидві ці платформи мають подібну архітектуру, проте відрізняються за функціональністю та рівнем доступу до інструментів аналізу. Обидві платформи походять від одного розробницького середовища (Lexical Computing Ltd) і використовують однакові принципи корпусної архітектури, однак різняться за рівнем функціональності, доступу та цільового призначення.

Sketch Engine – комерційна онлайн-платформа, призначена для створення, обробки та аналізу великих текстових корпусів. Вона забезпечує повноцінний набір інструментів для корпусного та лексикографічного аналізу, зокрема:

- **Word Sketch**: автоматичне створення словникових статей на основі частотних колокацій;
- **Thesaurus**: побудова переліку синонімів або групи слів, що належать до спільної категорії (семантичної);
- **Word Sketch Difference**: порівняння лексичних профілів двох слів (спільні та диференційні риси);
- **Keywords**: автоматичне вилучення термінів із текстів під час роботи зі спеціалізованими (галузевими) корпусами;
- **Wordlist**: генерація словникового реєстру на основі частот уживання в текстах;
- **Trends**: аналіз слів (неологізмів або архаїзмів), частота вживання яких змінюється з часом (діахронічний аналіз);
- **Concordance** та **Parallel Concordance**: робота з одномовними та паралельними корпусами.

Платформа має зручний веб-інтерфейс, підтримує понад 90 мов і забезпечує інтеграцію з національними корпусами та словниковими базами. Доступ до неї здійснюється за передплатою або через університетські ліцензії.

NoSketch Engine – безкоштовна, локально встановлювана версія тієї ж технології, орієнтована на навчальні та дослідницькі потреби. Вона надає базовий набір функцій корпусного пошуку та аналізу, зокрема:

- пошук за допомогою розширеного формату запитів **CQL (Corpus Query Language)**;
- формування конкордансів, частотних списків і колокацій;
- можливість самостійного додавання та розмітки корпусів;
- підтримку морфологічного тегування через **UDPipe** або **TreeTagger**.

На відміну від **Sketch Engine**, ця система не має розширених інструментів автоматичного створення словникових нарисів і статистичного моделювання, але дозволяє повністю контролювати структуру корпусу та його розмітку.

Виходячи з вищесказаного, **Sketch Engine** доцільно використовувати для комплексних лексикографічних проєктів, у яких потрібен автоматизований аналіз і візуалізація даних, тоді як **NoSketch Engine** є оптимальним для навчальних установ та дослідницьких лабораторій, що працюють із невеликими або експериментальними корпусами. Порівняльна характеристика обох платформ наведена в таблиці 8.1.

Таблиця 8.1 – Sketch Engine vs NoSketch Engine

Параметр	Sketch Engine	NoSketch Engine
Тип платформи	Комерційна онлайн-система	Безкоштовне локальне програмне забезпечення
Розробник	Lexical Computing Ltd (Велика Британія)	Lexical Computing Ltd (Велика Британія)
Призначення	Професійна лексикографія, корпусна лінгвістика, автоматичне створення словників	Академічні, дослідницькі та навчальні цілі
Доступ	Онлайн через передплату або університетську ліцензію	Локальна установка, вільний доступ
Підтримувані мови	Понад 90 мов	Будь-які (за умови власного корпусу)
Інтерфейс користувача	Вебінтерфейс із розширеними інструментами візуалізації	Вебінтерфейс із базовими функціями
Основні інструменти	Word Sketch, Sketch Diff, Thesaurus, Trends, Parallel Concordance, Frequency Lists	Concordance, Word List, Collocations, CQL Search
Мовний запит	CQL (Corpus Query Language)	CQL (Corpus Query Language)
Типи корпусів	Загальномовні, галузеві, паралельні, навчальні, користувацькі	Будь-які, створені користувачем

Розмітка	Морфологічна, синтаксична, частково семантична	Морфологічна (UDPipe, TreeTagger)
Інтеграція	Підтримка словникових проєктів, API для розробників, зв'язок із корпусами ЄС, Oxford, TenTen	Можливість інтеграції з локальними корпусами
Формати експорту	CSV, XML, JSON, XLSX	CSV, TXT
Можливість створення власного корпусу	Так, через інтерфейс або API	Так, вручну або через командний рядок
Переваги	Потужний аналітичний інструментарій, висока швидкість, зручна візуалізація	Безкоштовність, відкритість, гнучкість налаштувань
Обмеження	Платна ліцензія, залежність від серверів	Відсутність автоматичних лексикографічних розділів
Оптимальна сфера використання	Створення великих словників, лінгвістичні дослідження, видавничі проєкти	Локальні корпуси, навчальні курси, експериментальні дослідження

Обидві платформи дають змогу працювати з різними корпусами. Так, у лексикографічній практиці використовується кілька основних типів корпусів, кожен із яких має власне призначення, структуру та цільову аудиторію. До них належать загальномовні, спеціалізовані (галузеві), паралельні, навчальні, та історичні, діалектні та інші корпуси. Деякі можуть поєднувати ознаки кількох типів, наприклад паралельний корпус галузевих текстів *The JRC-Acquis Multilingual Parallel Corpus*, що містить тексти законодавчих актів ЄС і охоплює 231 мовну пару. Ще одним показовим прикладом є *Parallel English-Croatian Corpus in the Domain of Air Traffic*, який включає 178 англомовних і хорватських текстів, пов'язаних із законодавством у сфері авіаційного руху. Розглянемо детальніше особливості загальномовних, галузевих і паралельних корпусів та специфіку роботи з ними під час укладання словників.

Загальномовні корпуси – це корпуси, що відображають сучасну мову в усіх її функціональних різновидах – художньому, публіцистичному, науковому, розмовному тощо. Їх основне призначення полягає в аналізі лексики, граматики, стилістики та інших аспектів загальнонаціональної мови. Такі корпуси забезпечують дослідників матеріалом для створення нормативних словників, граматик, частотних списків і словників колокацій (див. порівняльну характеристику в додатку Д.19). Серед найвідоміших варто виокремити:

– українські: *Український національний лінгвістичний корпус, Генеральний регіональний анований корпус (ГРАК), Корпус текстів української мови, Український веб-корпус Лейпцизького університету;*

– англійські: *British National Corpus (BNC), Corpus of Contemporary American English (COCA), Open American National Corpus (ANC), Bank of English (BoE), Brown Corpus, Oxford English Corpus;*

– німецькі: *DWDS Core Corpus, Deutscher Wortschatz Project, IDS-Corpora, German Wikipedia Text Corpus, Open Parallel Corpus (OPUS);*

– французькі: *ELEXIS French Web 2020, Frantext, French Trends Corpus, Corpus for the Study of Contemporary French, Collection de corpus oraux numériques (COCOON);*

– іспанські: *Corpus del Español del Siglo XXI (CORPES XXI), Corpus del Español Actual, Corpus Escrito del Español (CEDEL 2), Corpus Diacrónico del Español (CORDE).*

Одним із сучасних потужних інструментів української лексикографії та лінгвістичних досліджень є Український національний лінгвістичний корпус (УНЛК), створений в Українському мовно-інформаційному фонді НАН України в межах робіт із формування та вдосконалення Національної словникової бази України. Корпус було укладено передусім для потреб створення Словника української мови у 20 томах (СУМ-20). В УНЛК зберігаються тексти різних стилів – художнього, наукового, офіційно-ділового, публіцистичного та інших, що забезпечує репрезентативність мовного матеріалу.

Як показано на рисунку 8.4, інтерфейс системи включає панель пошуку, панель бібліографічного опису джерела та панель контекстів, що дає змогу оперативно переглядати результати й аналізувати їх у структурованому вигляді. Доступ до корпусних матеріалів здійснюється через пошукову підсистему, у межах якої в УНЛК реалізовано два основні види пошуку: за бібліографічним описом та повнотекстовий пошук (за лінгвістичними параметрами). Пошук за бібліографічним описом здійснюється через роботу з метаданими, серед яких:

- назва видання, прізвище автора (редактора, укладача або колективного автора);
- стиль, серія, жанр, рік видання (або часовий проміжок), анотація, примітки,

– бібліографічні відомості, ISBN чи ISSN та інші характеристики джерела.

Пошук за лінгвістичними параметрами забезпечується повнотекстовим індексом, що дозволяє:

- враховувати порядок слів у запиті;
- здійснювати пошук у межах визначеної підмножини документів;
- застосовувати лематизацію (пошук усіх словоформ певної лем);
- враховувати синонімічні ряди та розширювати запит за рахунок семантично споріднених одиниць.

Отже, призначення УНЛК полягає у забезпеченні лексикографічної бази для укладання великих академічних словників, тому в ньому особливо ретельно збалансовано жанрову, стилістичну та хронологічну репрезентативність текстів. УНЛК також містить значно ширшу бібліографічну інформацію про джерела, що є надзвичайно важливим для укладачів словникових статей, адже дозволяє відстежувати походження прикладів, їхній стиль, час створення та авторську специфіку.

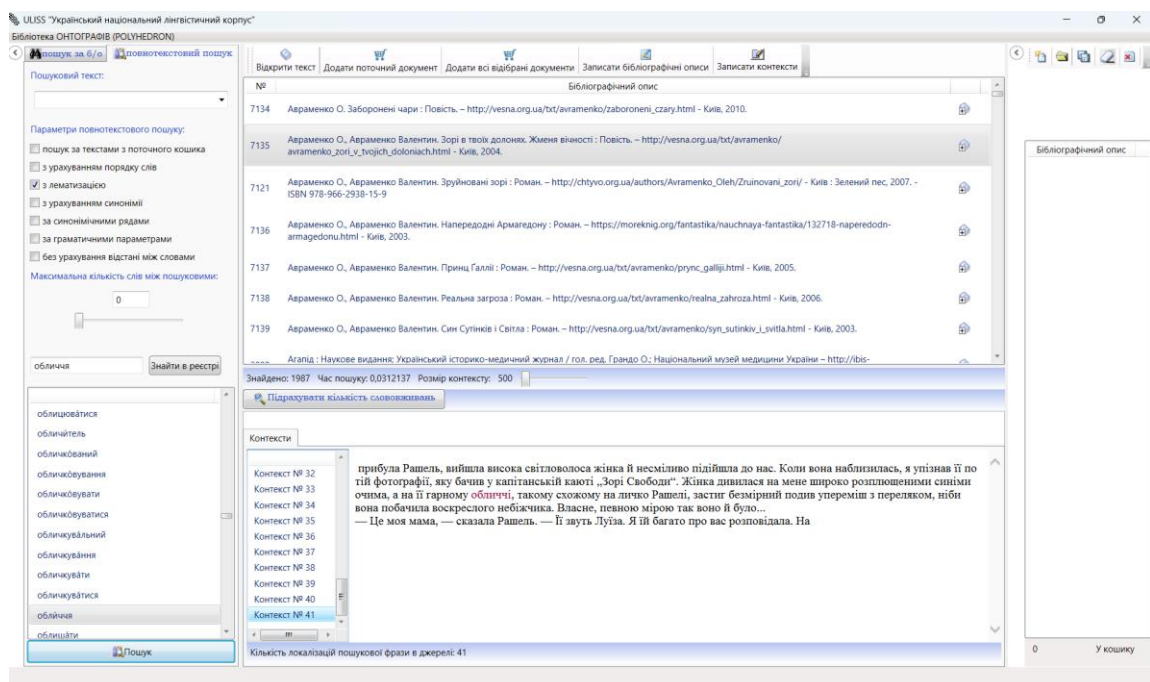


Рис. 8.4 – Інтерфейс УНЛК

Розглянемо інший приклад – Генеральний регіональний анотований корпус (ГРАК), орієнтований на детальну морфологічну та семантичну розмітку. Для роботи з ним широко використовується мова CQL (Corpus Query Language) – спеціалізована мова запитів, призначена для пошуку та аналізу лінгвістичних моделей у корпусних даних. Вона дає змогу формулювати складні пошукові конструкції, що виходять за межі звичайного пошуку за ключовими словами, і забезпечує можливість знаходити конкретні граматичні, лексичні або морфосинтаксичні структури. Водночас слід враховувати, що система розмітки різних корпусів може суттєво відрізнятися. Це впливає на спосіб побудови запитів мовою CQL і потребує уважного врахування специфіки конкретного корпусу. Знання цієї мови та вміння будувати запити є важливими для лексикографів, оскільки вони дають змогу ефективно вилучати потрібну інформацію для словникової статті. У таблиці 8.2 наведено приклади запитів до корпусу ГРАК-19а, сформульованих для пошуку словосполучень типу «прикметник + іменник», семантика яких відображає різні характеристики обличчя. Такі запити були використані під час укладання словника епітетів.

Таблиця 8.2 – Приклади словосполучень із ГРАК-19а

CQL запит	Семантика словосполучення	Приклади словосполучень
[tag="adj.*"] [lemma="обличчя"]	Всі словосполучення прикметників з лемою обличчя	кругле обличчя, бліде обличчя, пооране обличчя, виголене обличчя тощо
[semtag="*.color.*"] [lemma="обличчя"]	Колір обличчя	біле обличчя, червоне обличчя, чорне обличчя тощо
[semtag="*.abst:hum.*"] [lemma="обличчя"]	Абстрактна характеристика обличчя	людське обличчя, добре обличчя, чисте обличчя, святе обличчя, просте обличчя, дороге обличчя
[semtag="*.age:min.*"] [lemma="обличчя"]	Вікова характеристика обличчя	маленьке обличчя
[semtag="*.physio.*"] [lemma="обличчя"]	Фізичний стан обличчя	хворе обличчя
[semtag="*.size:max.*"] [lemma="обличчя"]	Розмір обличчя	величезне обличчя
[semtag="*.posit.*"] [lemma="обличчя"]	Позитивна оцінка	добре обличчя, гарне обличчя, дороге обличчя, справжнє обличчя
[semtag="*.negat.*"] [lemma="обличчя"]	Негативна оцінка	люте обличчя

У поданій вище таблиці продемонстровано застосування семантичної розмітки для прикметників, що сполучаються з аналізованим словом. У корпусі ГРАК для

кожної частини мови передбачено окремий тип семантичної розмітки, що дає змогу точніше класифікувати та вилучати відповідні лексичні одиниці. Наступні приклади (таблиця 8.3) показують, як за допомогою відповідних запитів можна вилучити словосполучення з прийменниковими та безприйменниковими конструкціями зі словом *обличчя*.

Таблиця 8.3 – Багатокомпонентні словосполучення з ГРАК-19а

SQL запит	Приклади словосполучень
[tag="adj.*"] [tag="prep.*"] [lemma="страх"] [lemma="обличчя"]	<i>перекошене від страху обличчя, посіріле від страху обличчя, побіліле від страху обличчя, викривлене від страху обличчя, спотворене від страху обличчя, охололе від страху обличчя, зведене від страху обличчя</i>
[tag="adj.*"] [tag="prep.*"] [lemma="холод"] [lemma="обличчя"]	<i>посиніле від холоду обличчя, буре від холоду обличчя, посіріле від холоду обличчя, почервоніле на холоді обличчя, рум'яне від холоду обличчя, зчорніле від холоду обличчя</i>
[lemma="обличчя"] [lemma="колір"] [tag="adj.*"] [tag="noun.*"]	<i>обличчя кольору мертвого місяця, обличчя кольору пергаментного паперу, обличчя кольору смаженої кави, обличчя кольору синьої кістки, обличчя кольору сірої замазки</i>

Для лексикографів деталізована морфологічна та семантична розмітка має ключове значення, адже вона забезпечує можливість точного вилучення словоформ, сполучень і контекстів, необхідних для створення повноцінних словникових статей. Чим багатша й тонша розміченість корпусу – тим ширшими є аналітичні можливості дослідника: від розмежування значень і стилістичних відтінків до виявлення типових моделей сполучуваності та колокацій. Така точність даних особливо важлива під час укладання галузевих, тлумачних і семантичних словників, де якість результату безпосередньо залежить від якості вихідної розмітки.

Спеціалізовані корпуси – це корпуси, що охоплюють тексти певної галузі знань, професійної сфери або комунікативної ситуації (див. порівняльну характеристику в додатку Д.20). Вони призначені для дослідження термінології, жанрових особливостей і мовних закономірностей у межах конкретної дисципліни. Основними ознаками спеціалізованих корпусів є:

- (1) обмежена предметна сфера (медицина, право, економіка, техніка тощо);

- (2) однорідність жанрів (наукові статті, нормативні документи, інструкції, договори);
- (3) висока термінологічна насиченість;
- (4) наявність тематичної розмітки, що дозволяє виділяти терміни, скорочення, формули, стандартизовані словосполучення.

Такі корпуси широко застосовують у термінографії, галузевій лексикографії, машинному перекладі, а також при створенні електронних словників фахової лексики. Серед прикладів спеціалізованих корпусів можна назвати:

- українські: *Український парламентський корпус, Корпус законів та правових актів, Підкорпус текстів із програмування (ГРАК-19a)*;
- англійські: *British Academic Written English (BAWE), Medical English Corpus, British Law Report Corpus*;
- німецькі: *Huge German Corpus (HGC), Jurakorpus, German Parliamentary Corpus (GerParCor)*;
- французькі: *Collection Astrophysique, Collection Food S&T, Écrits scientifiques en français (Scientext)*;
- іспанські: *El Corpus Científico-Técnico de la Real Academia Española (CORPES XXI), Corpus jurídico (Universidad del País Vasco), Spanish Biomedical Crawled Corpus*.

За допомогою таких корпусів можна створити словниковий реєстр. На рисунку 8.5 показано приклад термінологічного реєстру, сгенерованого за допомогою інструмента **Keywords** для укладання глосарія з програмування на основі підкорпусу текстів ГРАК-19a.

KEYWORDS

GRAC19a_articles_pro...

SINGLE-WORDS

reference corpus: Grac v.19 (items: 3,480)

Lemma	Lemma	Lemma
1 sdn	11 маршрутизація	21 кодер
2 spice	12 хеммінг	22 azure
3 nfv	13 інформаційно-технологічний	23 кодування
4 мережний	14 градієнтний	24 зрілість
5 cvss	15 завадостійкий	25 мрэнм
6 підмножина	16 еталонний	26 градієнт
7 парність	17 каузальний	27 топологічний
8 підпроцес	18 програмно-конфігурований	28 біт
9 вразливість	19 діпдо	29 попарний
10 mpc-регулятор	20 каскадний	30 aws

Rows per page: 50 1–30 of 30 < > 1 / 1

Рис. 8.5 – Термінологічний реєстр на основі підкорпусу ГРАК-19a

За допомогою таких корпусів можна сформувати словниковий реєстр. На рисунку 8.4 наведено приклад термінологічного реєстру, згенерованого за допомогою інструмента **Keywords** для укладання глосарія з програмування на основі підкорпусу текстів ГРАК-19а. Для кожного терміна можна отримати **конкорданс**, який демонструє приклади його вживання та можливі його сполучення, з іншими словами. У таблиці 8.4 подано приклади термінологічних сполучень (колокацій), отриманих із конкордансу для терміна *маршрутизація*. Для побудови таких сполучень та/або колокацій можна сформулювати відповідні запити мовою **CQL**.

Таблиця 8.4 – Термінологічні сполучення і колокації

Запит мовою CQL	Тип сполучення	Приклади сполучень
[tag=>adj.*>] [lemma=>»маршрутизація»]	ADJ + маршрутизація	безпечна маршрутизація, багатошляхова маршрутизація
[tag=>noun.*>] [tag=>adj.*>] [lemma=>»маршрутизація»]	N+ ADJ + маршрутизація	Засоби безпечної маршрутизації, завдання безпечної маршрутизації, методи безпечної маршрутизації
[tag=>noun.*>] {0,2} [tag=>adj.*>] [lemma=>»маршрутизація»]	N + [Ø, N] + ADJ + маршрутизація	Дослідження методів безпечної маршрутизації, використання засобів безпечної маршрутизації

Завдяки таким корпусам лексикографи отримують змогу автоматично виділяти фахові терміни, визначати їхню сполучуваність і контекстуальні варіанти вживання, що значно полегшує укладання спеціалізованих словників, наприклад: медичних, технічних юридичних чи економічних.

Паралельні корпуси – це корпуси, що містять тексти двома або кількома мовами, вирівняні за реченнями або фрагментами для зручності зіставлення (див. порівняльну характеристику в додатку Д.21). Ключовими особливостями паралельних корпусів є:

- (1) наявність текстів-оригіналів і перекладів, що точно співвіднесені між собою;
- (2) вирівнювання за реченнями або сегментами, яке забезпечує можливість пошуку еквівалентів;
- (3) багатомовна структура, що дозволяє охоплювати десятки мов;
- (4) використання у лексикографії для автоматичного виявлення перекладних відповідників, міжмовних колокацій і типових структур.

Їхнє основне призначення полягає у забезпеченні можливості автоматичного або напіваавтоматичного аналізу перекладів, дослідження міжмовних відповідників, а також у створенні двомовних і багатомовних словників. Такі корпуси є незамінними в перекладознавстві, зіставній лінгвістиці, машинному перекладі та комп'ютерній лексикографії. До найвідоміших паралельних корпусів належать:

Іншомовні проєкти

JRC-Acquis Multilingual Parallel Corpus

OPUS (Open Parallel Corpus)

ParaCrawl

EuroParl Parallel Corpus

ParlaMint Corpora

Українські проєкти

ParaRook Corpora

English-Ukrainian Parallel Corpus of IT Texts

Polish-Ukrainian Parallel Corpus

На рисунку 7.6 наведено приклад паралельного конкордансу, англійська частина якого містить тексти зі словом *humanity*, а українська – переклад цих текстів разом із еквівалентом до аналізованого слова (*людство, люди, людяність, гуманний, народ*).

PARALLEL CONCORDANCE		ParaRook EN→UK	
lemma humanity • 50 16.95 per million tokens • 0.0017%		Sort GDEX	
			ParaRook UK→EN
① doc#26	<s> It is the mass for which the true Humanity cares. </s>	<s> Загал - ось про кого дбає в нас істинна людяність. </s>	
① doc#62	<s> With one last touch of humanity I turned the blade back and struck him with the butt. </s>	<s> Остання крапля людяності озвалась у мені, - я обернув сікача до себе лезом і вдарив священика тільки рукою. </s>	
① doc#91	<s> They saw humanity in its darkest, most painful hours. </s>	<s> Вони бачили людей в найтемніші і найкошмарніші моменти їхнього життя. </s>	
① doc#92	<s> Meyer tried to speak, to reaffirm his humanity . </s>	<s> Меєр спробував заговорити, підтвердити свою людську сутність, але слова давалися через силу. </s>	
① doc#9	<s> And now was I indeed wretched beyond the wretchedness of mere humanity . </s>	<s> Тепер-таки я справді зробився найнещасливіший із смертних. </s>	
① doc#32	<s> I felt that this small boy had far more humanity than I did. </s>	<s> я відчув, що цей маленький хлопчик набагато гуманніший за мене. </s>	
① doc#7	<s> Oh, it would be very nice to fight Armageddon with the mass of humanity . </s>	<s> Певна річ, було б добре перемогти в Армагедоні силами людства. </s>	
① doc#62	<s> It was the first time I realised that the Martians might have any other purpose than destruction with defeated humanity . </s>	<s> Тільки зараз я вперше подумав, що, може, марсіяни мають якісь інші наміри щодо людства, аніж просто знищити його. </s>	
① doc#62	<s> And besides——" He looked at me. "Aren't you satisfied it is up with humanity ? </s>	<s> Але... - він глянув на мене, - гадаєте, що з людством ще не покінчено? </s>	
Rows per page: 10		11–20 of 50	1< < 2 / 5 > >1

Рис. 8.6 – Паралельні тексти для слова *humanity* в *ParaRook*

Так само, як і в одномовних корпусах, у паралельних корпусах можна формулювати запити мовою **CQL**. У таблиці 8.5 наведено приклади таких запитів та результати їх опрацювання, що демонструють можливості пошуку й аналізу паралельних відповідників у двох мовах. Під час роботи з паралельними корпусами CQL-запити мають певну специфіку, зумовлену вирівнюванням

сегментів між двома мовами. На відміну від одномовного пошуку, де аналізується суто внутрішня структура речення, у паралельному корпусі запит може охоплювати одразу дві площини: форму або структуру вихідного мовного елемента та його перекладний відповідник. Це дає змогу здійснювати пошук асиметричних відповідників, аналізувати різні способи перекладу одного й того самого слова чи конструкції, виявляти сталі перекладні моделі та розбіжності між мовами. Таким чином, CQL-запити в паралельних корпусах забезпечують не лише вилучення лексичних одиниць, а й глибший контрастивний аналіз.

Таблиця 8.5 – Англійські колокації та їхні відповідники

Запит мовою CQL	Англійські колокації	Українські колокації
[upos="ADJ"] [lemma="water"]	<i>cold water, filthy water, clear water, shallow water</i>	<i>холодна вода, ставок, прозорий плес, мілководдя</i>
[upos="NOUN"] [upos="ADP"] [lemma="water"]	<i>cup of water, pailful of water, jet of water</i>	<i>склянка води, відро води, струмінь води</i>
[lemma="water"] [upos="NOUN"]	<i>water level, water pistol, water journey, water colors</i>	<i>рівень води, лавандовий пістолет, подорож водою, акварельні фарби</i>
[lemma="water"] [upos="VERB"]	<i>water runs, water gurgles, waters advance</i>	<i>вода тече, вода булькоче, вода прибуває</i>

Отже, паралельні корпуси істотно підвищують якість двомовних словників, оскільки надають не лише переклад окремого слова, а й типові контексти його вживання в обох мовах. Це дозволяє точніше встановлювати семантичні відповідники, уникати хибних чи надто широких перекладних рішень, а також забезпечує узгодженість термінології між мовами.

Питання для самоконтролю

1. Що таке текстовий корпус? Які обов'язкові характеристики відрізняють його від простої вибірки текстів?
2. Які функції виконують корпуси в сучасній лексикографії?
3. Які типи корпусів використовуються у лексикографічній практиці?
4. Чому корпуси є основою для створення сучасних словників?
5. Чим загальномовні корпуси відрізняються від спеціалізованих?
6. Які особливості паралельних корпусів і яку роль вони відіграють у дво- та багатомовній лексикографії?

7. Які сфери лінгвістичних і прикладних досліджень потребують спеціалізованих корпусів?
8. У чому полягає відмінність між зовнішньою (екстралінгвістичною), структурною та лінгвістичною розміткою?
9. Які види лінгвістичної розмітки найчастіше застосовують у корпусній лінгвістиці?
10. Що таке токенізація, лематизація, морфологічне тегування? Чому лінгвістична розмітка є ключовою для автоматизованої обробки корпусів?
11. Які можливості надає Sketch Engine для лексикографічного аналізу? Які функції має NoSketch Engine?
12. Як корпуси допомагають у створенні словникових статей? Які типи інформації про слово можна отримати з корпусу?
13. Яку роль відіграють колокації в сучасних словниках?
14. Чому частотні дані важливі для укладання словників?
15. Які переваги надає використання корпусів для термінографії?

Завдання для самоперевірки

1. Для кожного опису оберіть один із типів корпусу: *загальномовний, спеціалізований, паралельний, навчальний, історичний*. Приклади описів (можна використовувати або додати свої):

– Містить підкорпус художньої літератури, публіцистики, документів і має понад 100 млн слововживань.

– Складається з медичних протоколів і фахових статей.

– Містить тексти 18–19 століття з повною морфологічною розміткою.

– Об'єднує українські та польські переклади наукових текстів.

– Містить студентські есе з позначенням граматичних помилок.

2. Для наведених прикладів визначте тип розмітки: (а) екстралінгвістична, (б) структурна, (в) лінгвістична. Приклади:

Автор: Ліна Костенко, Рік видання: 1979.

<p> ... </p>, <title> ... </title>, поділ на розділи.

VERB, ADJ, NOUN, лематизація слова «йшов → йти».

3. Порівняйте корпусні платформи за їхніми головними характеристиками та заповніть таблицю:

Платформа	Основне призначення	Тип користувачів	Переваги	Обмеження
Sketch Engine				
NoSketch Engine				
UDPipe				

4. Уявіть, що ви досліджуєте слово «*ефективність*». Опишіть, яку інформацію ви могли б отримати з корпусу:

- частотність;
- типові контексти;
- жанрові відмінності;
- ключові колокації;
- приклади з носіями стилю.

5. Опишіть кроки створення невеликого тематичного корпусу, наприклад: корпус юридичних текстів, корпус спортивної журналістики, корпус технічних інструкцій. Необхідно врахувати:

- джерела текстів;
- етапи збирання;
- очищення й нормалізації;
- розмітки;
- формат збереження;
- можливості інтеграції в лексикографічний проєкт.

6. Наведено три фрагменти (умовні), отриманих із корпусу текстів за запитом користувача:

*Система демонструє високу **ефективність** при низьких затратах енергії.*

*Уряд обговорює **ефективність** реформ.*

***Ефективність** у бізнесі визначається правильністю управлінських рішень.*

З'ясуйте: (1) чи належать фрагменти до одного домену; (2) які колокації повторюються; (3) чи зміщується значення слова залежно від тематики.

8.2. Корпусні інструменти для створення словників

Платформи, що працюють із корпусами, надають широкий спектр інструментів для створення словників на основі корпусних даних. Однією з найвідоміших і найбільш функціональних таких платформ є Sketch Engine, яка містить засоби для автоматичної генерації не лише словникового реєстру, а й повноцінних словникових статей. До структури таких статей можуть входити ілюстративні приклади, колокації, типові граматичні та лексичні профілі, а також синонімічні ряди. До інструментів Sketch Engine, що безпосередньо підтримують лексикографічну роботу, належать:

- OneClick Dictionary;
- One Click Terms;
- Good Dictionary Examples (GDEX);
- TickBox Lexicography (TBL);
- Sketch Engine i Lexonomy.

OneClick Dictionary – це інструмент для автоматичної генерації чернеток словникових статей на основі корпусних даних. Алгоритм аналізує лексеми в корпусі, добирає для них дефініції, типові колокації, приклади вживання, синоніми та навіть графічні зображення. Сформовані таким чином статті виступають стартовою основою для подальшої ручної лексикографічної обробки, що істотно прискорює укладання термінологічних і тлумачних словників.

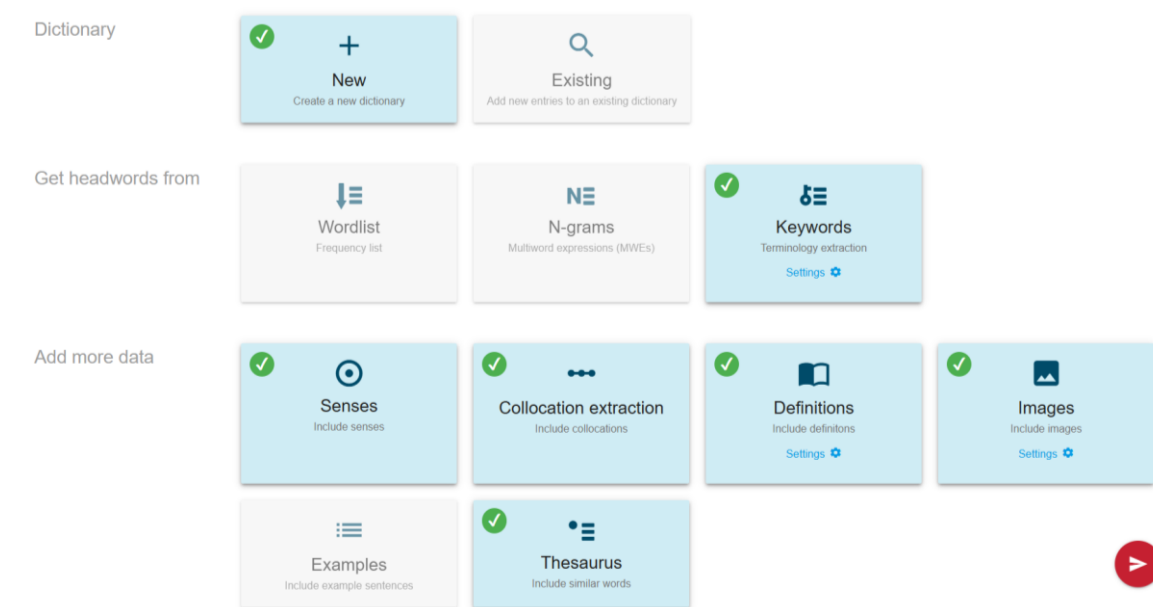


Рис. 8.7 – Інтерфейс інструмента OneClick Dictionary

Як показано на рисунку 8.7, для автоматичної генерації словника користувач має обрати опцію «Створити новий» (*Create a new dictionary*) або «Доповнити вже існуючий» (*Add new entries to an existing dictionary*). Далі необхідно визначити джерело словникового реєстру, яким може бути:

- *Wordlist* – список слів на основі частотності;
- *n-grams* – багатослівні сполучення;
- *Keywords* – ключові слова галузевого корпусу.

На завершальному етапі користувач налаштовує структуру словникової статті, додаючи необхідні інформаційні елементи:

- *Senses* – значення;
- *Collocation Extraction* – добір колокацій;
- *Include Definitions* – автоматичні дефініції;
- *Include images* – ілюстрації;
- *Include Examples* – приклади вживання;
- *Thesaurus* – синоніми та дотичні терміни.

Одним із важливих нововведень є автоматична генерація дефініцій (тлумачень) засобами штучного інтелекту. Система дозволяє обрати одну з моделей великомасштабних мовних моделей (Chat GPT-5, Claude Sonnet 4, Gemini 2.5 Flash, Llama 4 Maverick, Mistral Nemo та інші), кожна з яких має власні підходи до узагальнення значень, стилістичної нейтральності та точності формулювань. Використання таких інструментів не лише прискорює процес формування тлумачних блоків, а й дає змогу отримати кілька варіантів дефініцій для подальшого редакторського добору, що значно підвищує ефективність роботи лексикографа.

Результатом роботи інструменту OneClick Dictionary є автоматично сформований у Лехопоту словник, структура та зміст якого можуть надалі редагуватися й удосконалюватися лексикографом. Обсяг необхідних доповнень та доопрацювань значною мірою залежить від якості та репрезентативності корпусу: чим повніше й збалансованіше корпус охоплює різні типи текстів, тим точнішими будуть автоматично підібрані дефініції, колокації та приклади.

У випадку ж менш збалансованих або галузевих корпусів лексикографові може знадобитися суттєвіший ручний аналіз і корекція матеріалу.

Після автоматичної генерації словникових статей лексикограф зазвичай здійснює кілька етапів ручного опрацювання матеріалу. Насамперед перевіряється відповідність дефініцій нормам словникової подачі: усуваються надмірні фрагменти, уточнюється зміст, дотримуються стилістичні вимоги до тлумачення. Далі здійснюється відбір і редагування колокацій, оскільки автоматичні списки можуть містити повтори, випадкові збіги або нехарактерні для норми сполучення. Також лексикограф перевіряє ілюстративні приклади, добираючи найбільш репрезентативні з погляду стилю, граматики та семантики. За потреби допрацьовуються синонімічні ряди та пов'язані терміни, уточнюється обсяг значення і його місце у структурі словникової статті. У підсумку автоматично згенерований словник перетворюється на опрацьований ресурс, придатний для публікації або інтеграції у великий лексикографічний проєкт. На рисунку 8.8 представлено результат роботи інструменту OneClick Dictionary – фрагмент словника авіаційних термінів, укладеного на основі спеціалізованого корпусу технічних інструкцій корпорації *Boeing*. Приклад демонструє, як автоматично згенеровані словникові статті можуть відображати термінологію певної галузі з урахуванням типової фахової лексики, характерних колокацій і контекстів уживання.

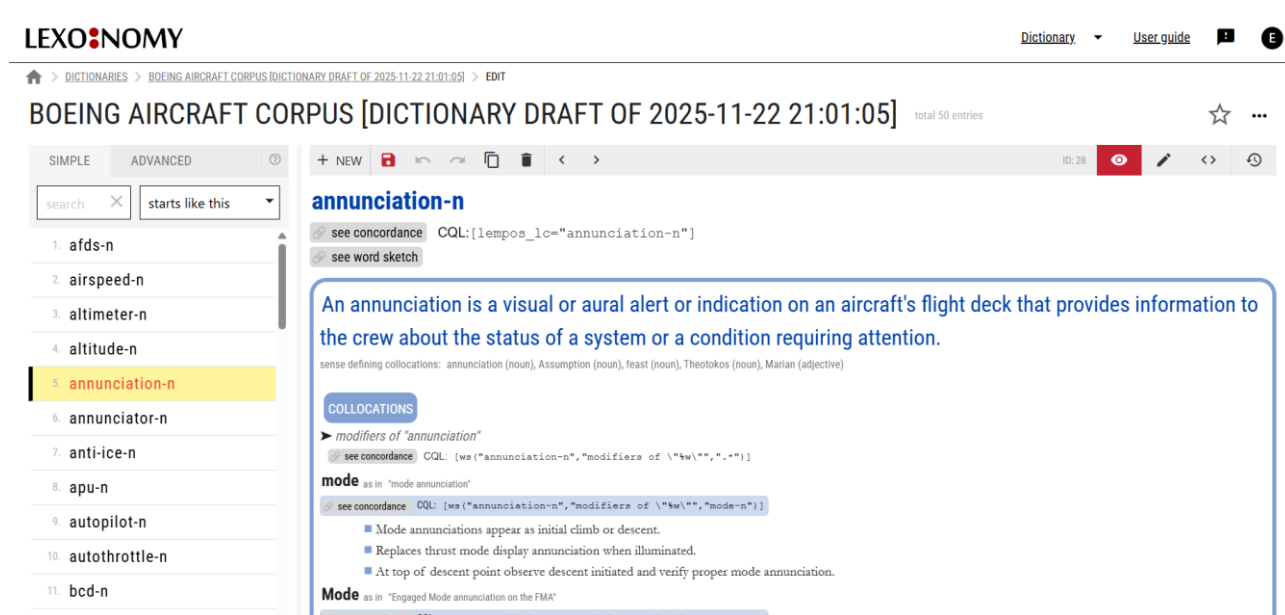


Рис. 8.8 – Авіаційний словник, створений за допомогою OneClick Dictionary

Отже, такий інструмент дає можливість швидко згенерувати попередню версію словника, яка надалі може бути доопрацьована фахівцем: уточнена, доповнена або структурована відповідно до типу видання. OneClick Terms є ще одним спеціалізованим інструментом корпусної лексикографії, що призначений для укладання галузевих словників – як одномовних, так і багатомовних. На відміну від OneClick Dictionary, який працює з широким спектром лексики, OneClick Terms орієнтований саме на виділення термінів, визначення їхніх ключових перекладних відповідників і побудову структурованих термінологічних реєстрів на основі корпусного аналізу.

OneClick Terms (рис. 8.9) використовується для автоматизованого виділення термінів та терміносполучень у спеціалізованих корпусах. Система застосовує частотні, статистичні та контрастивні моделі (зокрема порівняння спеціалізованого й загального корпусів), щоб виокремити ключові терміни галузі. Інструмент можна використовувати під час укладання термінологічних словників, тезаурусів та глосаріїв технічних або наукових дисциплін.

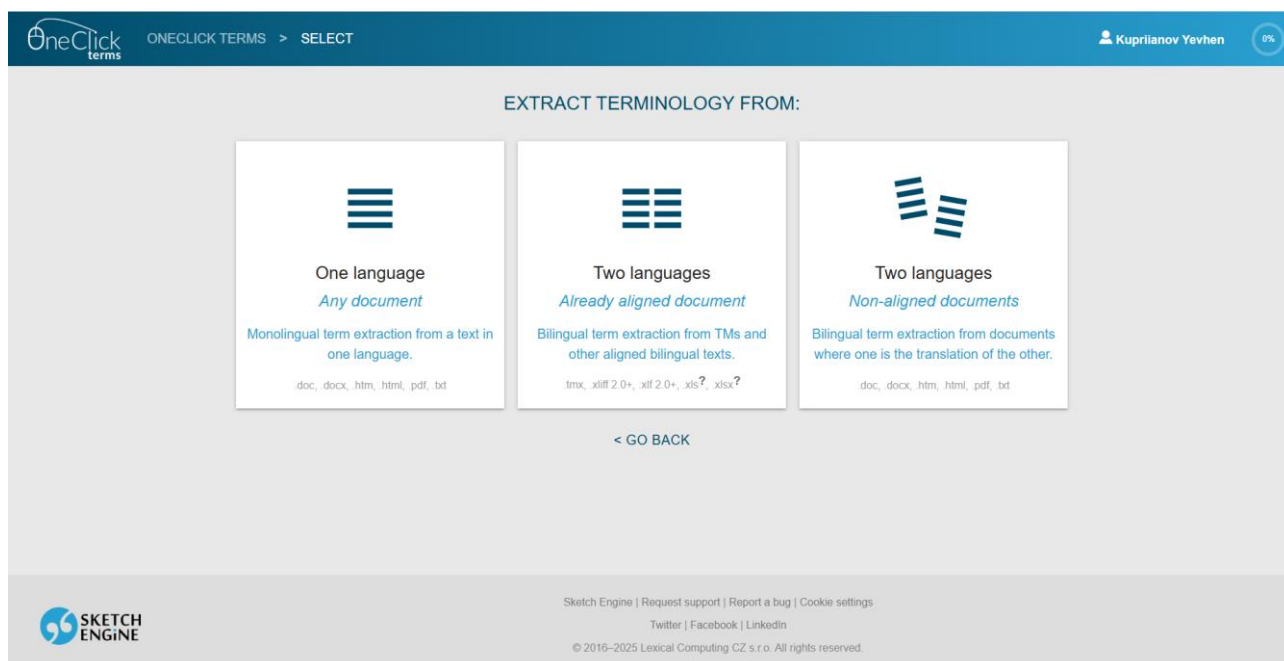


Рис. 8.9 – Головне вікно OneClickTerms

Для роботи з інструментом OneClick Terms користувачеві необхідно створити акаунт у Sketch Engine і увійти до системи під своїми даними. Інтерфейс інструмента є максимально спрощеним та інтуїтивно зрозумілим, що дозволяє здійснювати терміновиділення без попередньої спеціальної підготовки. Після

авторизації відкривається основне робоче вікно OneClick Terms, у якому пропонується вибрати один із трьох режимів вилучення термінів і термінологічних колокацій:

(1) **One language (Any document)** – вилучення термінів з одного або кількох одномовних документів у форматах *DOC, DOCX, HTM, HTML, TXT*. Це режим для створення одномовних термінологічних словників або попередніх термінологічних реєстрів на основі спеціалізованих текстів;

(2) **Two languages (Already aligned documents)** – обробка двомовних документів, які вже мають вирівнювання сегментів (формати *TMX, XLIFF 2.0+, XLF 2.0+, XLS, XLSX*). Цей режим дає змогу автоматично зіставляти терміни та їхні перекладні відповідники в паралельних документах, що значно прискорює укладання двомовних і багатомовних термінологічних словників;

(3) **Two languages (Non-aligned documents)** – вилучення термінів із двох незіставних текстів, один із яких є перекладом іншого (формати *DOC, DOCX, HTM, HTML, TXT*). У цьому випадку вирівнювання здійснює сама система, що робить можливим опрацювання великих масивів текстів навіть без попередньої підготовки перекладних корпусів.

Обравши режим, користувач завантажує файли, після чого OneClick Terms автоматично проводить аналіз: визначає спеціальну лексику, витягує ключові терміни, частотні термінологічні сполучення, а у двомовних режимах також формує *міжмовні відповідники* та базові термінологічні пари. Після завантаження документів у вибраному режимі OneClick Terms переходить до етапу автоматичного аналізу тексту, у межах якого виконується кілька послідовних операцій. Система:

- ідентифікує лексеми, що мають ознаки термінів (висока інформативність, доменна специфічність, контекстна стійкість);
- визначає частоту їхнього вживання в корпусі завантажених текстів;
- кількість термінів та терміносполучень для відображення в списку результатів;

– виявляє словосполучення термінологічного типу (наприклад, *основна плата, авіонічний розділ, крива тяги*);

– у двомовних режимах автоматично встановлює відповідники між мовами. Для точнішого добору термінів інструмент пропонує низку налаштувань фільтрації. Користувач може:

- задати *мінімальну частоту* появи терміна;
- визначити, чи враховувати *морфологічні варіанти* (лематизація);
- застосувати *стоп-списки*, щоб вилучити загальноновживані слова.

У разі роботи з двомовними документами інструмент також дозволяє налаштувати ступінь збігу сегментів, що впливає на якість вирівнювання та подальшого зіставлення термінів. Результатом роботи OneClick Terms є структурований термінологічний перелік, який включає:

- ключові терміни;
- частотність їхнього вживання;
- типові термінологічні колокації;
- у двомовних режимах – перекладні відповідники та їхню частоту;
- приклади контекстів для кожного терміна (за потреби).

Усі отримані дані можна завантажити у вигляді таблиці у форматі **XLSX** або **CSV**, що дає змогу використовувати їх для подальшого укладання галузевого термінологічного словника чи інтеграції в інші лексикографічні інструменти. Крім того, результати терміновиділення можуть бути експортовані до власного корпусу користувача на платформі **Sketch Engine**, що забезпечує зручну подальшу роботу з термінами в межах одного середовища. На рисунку 8.10 показано результат роботи One Click Terms – перелік усіх термінів із технічних інструкцій з ремонту електричних насосів.

Приклади, автоматично вилучені системою з корпусів, не завжди є достатньо зрозумілими або репрезентативними. Тому під час їх відбору та редагування доцільно використовувати інструмент **GDEX**, який допомагає визначити найбільш якісні та інформативні ілюстративні приклади для словникових статей.

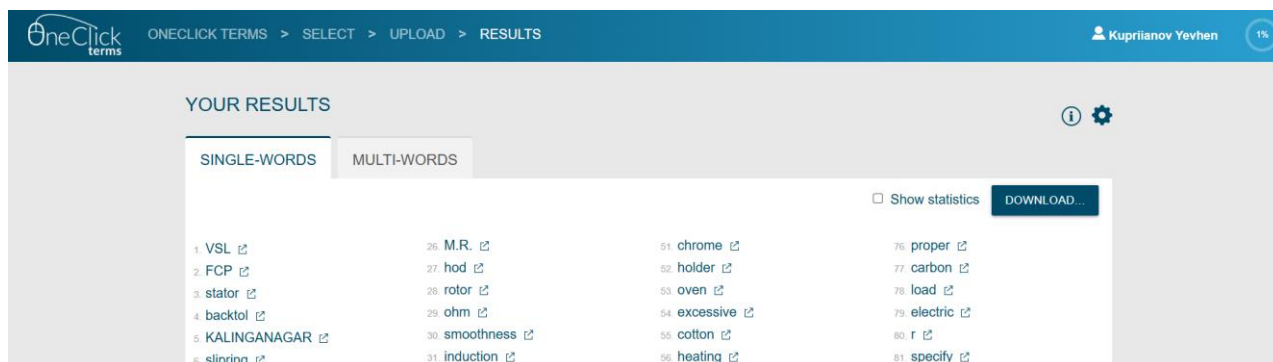


Рис. 8.10 – Результат роботи OneClick Terms

GDEX – інструмент для автоматичного добору найкращих ілюстративних прикладів. Він оцінює корпусні речення за низкою критеріїв: ясність синтаксичної структури, відсутність зайвих займенників і анафор, нейтральність стилю, адекватна довжина, відсутність лексичних або граматичних аномалій.

Без застосування GDEX система часто подає приклади у довгих, перевантажених технічних фрагментах, які не завжди підходять для словникової статті. Наприклад, це можуть бути складні службові повідомлення, інструкції або суцільні блоки тексту з надлишком цифр і параметрів. Такі приклади важко прочитати та неможливо використати як ілюстративний матеріал без значного редагування (рис. 8.11а). У результаті GDEX пропонує речення, які можна одразу використовувати як ілюстративні приклади у словнику без значної редакторської правки (рис. 8.11б).

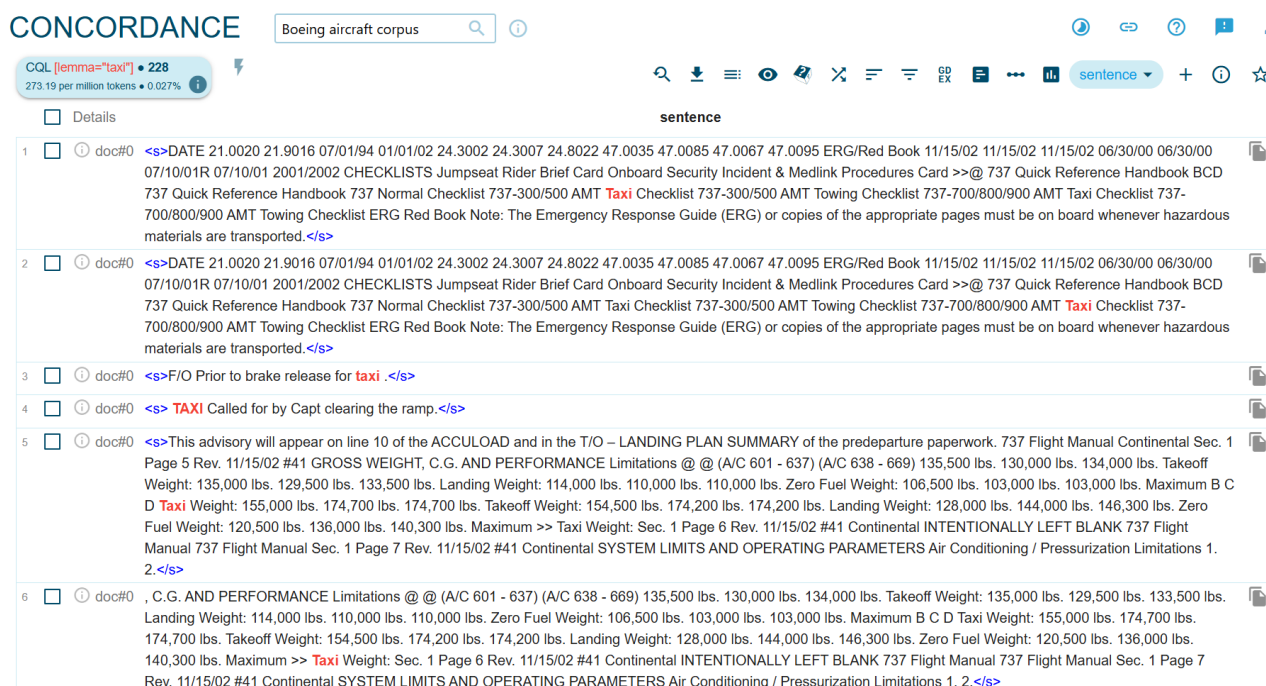


Рис. 8.11а – Ілюстративні приклади без застосування GDEX

Алгоритм GDEX аналізує кожен потенційний приклад за набором критеріїв і нараховує йому оцінку, що відображає його придатність для словникової статті. Система враховує такі параметри:

- довжина речення (оптимальна – середня);
- відсутність або мінімальна кількість власних назв;
- відсутність складних або маловживаних лексем;
- чітка синтаксична структура;
- близькість слововживання до базових, найтипівіших контекстів;
- мінімальна фрагментація (без занадто коротких чи вирваних фрагментів).

Приклади з найвищим рейтингом подаються у верхній частині списку. На рисунку 8.11б видно, що для кожного прикладу система автоматично присвоює **GDEX-score**, який відображає придатність речення для словникової статті. Чим вищий показник (наближений до 1.0), тим кращим вважається приклад за сукупністю критеріїв – зрозумілістю, граматичною правильністю, нейтральністю стилю та репрезентативністю значення слова.

CONCORDANCE Boeing aircraft corpus

CQL [lemma="taxi"] • 228
273.19 per million tokens • 0.027%

Sort GDEX x

search download print expand zoom in zoom out list view settings sentence + info star

	Details	sentence	GDEX score
1	☐ ⓘ doc#0	<s>A slightly higher thrust setting is required to start taxiing .</s>	0.960
2	☐ ⓘ doc#0	<s>The tendency is to taxi faster than desired.</s>	0.960
3	☐ ⓘ doc#0	<s>Rudder pedal differential will also result in abnormal nose wheel steering control displacement during taxi and takeoff.</s>	0.960
4	☐ ⓘ doc#0	<s>Any significant rudder trim application on the ground should be apparent to the crew during taxi and takeoff roll.</s>	0.960
5	☐ ⓘ doc#1	<s>ON – nose wheel well taxi light illuminated.</s>	0.960
6	☐ ⓘ doc#0	<s>Call the station on company radio and have the marshaller return and give taxi clearance.</s>	0.95
7	☐ ⓘ doc#0	<s>Idle thrust is adequate for taxiing under most conditions.</s>	0.95
8	☐ ⓘ doc#0	<s>Single engine taxi is authorized after landing.</s>	0.95
9	☐ ⓘ doc#0	<s> Taxi as close as possible on the centerline and avoid large nose steering inputs.</s>	0.95
10	☐ ⓘ doc#0	<s>Use caution when taxiing onto the runway for takeoff.</s>	0.95
11	☐ ⓘ doc#0	<s>A manual update cannot be done during taxi or flight.</s>	0.95
12	☐ ⓘ doc#0	<s>This checklist will be completed prior to releasing the brakes for taxi .</s>	0.949
13	☐ ⓘ doc#0	<s>The appropriate taxi speed will depend on turn radius and surface condition.</s>	0.949
14	☐ ⓘ doc#0	<s>The normal straight away taxi speed should not exceed approximately 20 knots.</s>	0.949
15	☐ ⓘ doc#1	<s> Taxiing with flaps extended subjects the flaps and flap drives to snow and slush accumulations from the main gear wheels.</s>	0.927
16	☐ ⓘ doc#0	<s>The crew should not hesitate to discontinue taxiing if any doubt exists about the aircraft's location on the field.</s>	0.926

Рис. 8.11б – Ілюстративні приклади після застосування GDEX

Наприклад, більшість наведених речень мають рейтинг 0.95–0.96, що свідчить про їх високу якість і лексикографічну придатність. Приклади з нижчим рейтингом (наприклад, 0.85 або менше) можуть містити надто спеціалізовану лексику, складні конструкції, зайві елементи або менш типові контексти. Саме

тому GDEX дає змогу лексикографу швидко зорієнтуватися, які речення варто включати до словникової статті, а які – проігнорувати або замінити.

Завдяки GDEX лексикограф може одразу вибрати найкращі варіанти або переглянути ширший діапазон. У процесі формування словникової статті важливо подати не просто випадковий приклад, а типовий, зрозумілий та лексикографічно придатний. Саме цей інструмент:

- значно скорочує час на ручний перегляд десятків випадків;
- допомагає уникнути прикладів із надто спеціалізованого або нетипового стилю;
- відсіює речення, де слово має інше, нецікаве для словникової статті значення;
- автоматично формує добірку речень, придатних навіть для двомовних словників.

Користувач може використовувати стандартний набір правил GDEX або створювати власні фільтри, адаптовані до типу словника (наприклад, для дитячих, шкільних чи технічних видань). Крім того, у фільтрах можна задати окремі параметри для відображення саме тих контекстів, у яких аналізоване слово утворює релевантні колокації, що істотно полегшує добір якісних прикладів.

CONCORDANCE Boeing aircraft corpus

QCL [lemma="taxi"] • 228
273.19 per million tokens • 0.027%

filter [lemma="taxi"][tag="NN.*"] (-3..3,+KWIC) • 89
106.64 per million tokens • 0.011%

Sort GDEX

Details sentence GDEX score

1	doc#1	<s>ON – nose wheel well ,taxi ,light illuminated.</s>	0.950
2	doc#0	<s>Call the station on company radio and have the marshaller return and give ,taxi ,clearance .</s>	0.95
3	doc#0	<s>The appropriate ,taxi ,speed will depend on turn radius and surface condition.</s>	0.949
4	doc#0	<s>The normal straight away ,taxi ,speed should not exceed approximately 20 knots.</s>	0.949
5	doc#0	<s>Generally, single engine taxi will result in less overall fuel consumption and is appropriate in most ,taxi ,situations .</s>	0.925
6	doc#0	<s>Also, evidence of any ice formation that may have been induced by ,taxi ,operations is sought.</s>	0.925
7	doc#0	<s>Allow the aircraft to accelerate, then brake to a very slow ,taxi ,speed and release the brakes completely.</s>	0.925
8	doc#0	<s>ON – Nose wheel ,taxi ,light illuminated.</s>	0.925
9	doc#1	<s>Prior to ,taxi ,speed , disarm the autobrake and continue manual braking as required.</s>	0.925
10	doc#0	<s>For directional control, use the rudder pedals down to ,taxi ,speed .</s>	0.924
11	doc#0	<s>The transition speed should be closer to a safe ,taxi ,speed on very slippery runways or when runway length is limited.</s>	0.904
12	doc#0	<s>It is essential that the ,taxi ,speed be kept low enough that the aircraft can be stopped in the space available.</s>	0.904
13	doc#0	<s>Speeds in excess of this, when combined with long ,taxi ,distances , cause heat buildup in the tires.</s>	0.900

Рис. 8.12 – Пошук колокацій з використанням GDEX

На рисунку 8.12 продемонстровано, як інструмент GDEX у поєднанні з фільтрами для колокацій дозволяє отримати не лише «хороші» приклади речень, а й побачити типові комбінаторні зв'язки леми *taxi*. У цьому списку контексти доповнені виділеними колокаціями – *taxi light*, *taxi clearance*, *taxi speed*, *taxi operations*, *taxi situations*, – що полегшує розуміння типових сполучень у технічних текстах. Завдяки цьому укладач може одразу бачити, які саме словосполучення є найбільш частотними та найбільш характерними для авіаційної термінології, а відповідно – включати їх до словникової статті як окремі гнізда або приклади вживання. Таке візуальне підсвічування колокацій значно прискорює аналітичний етап укладання словника, адже користувач не просто переглядає речення, а бачить структуру терміносистеми у «живому» контексті.

Окрім підсвічування колокацій, інтерфейс також використовує стандартне для корпусної лінгвістики представлення **KWIC (Key Word In Context)** – ключове слово в оточенні контексту. Такий формат дозволяє миттєво побачити, які слова найчастіше з'являються зліва чи справа від терміна *taxi*, а також оцінити синтаксичні патерни, у яких він функціонує. Впорядкування списку за лівим або правим контекстом допомагає виявити стійкі шаблони вживання, що особливо корисно для термінів, які утворюють розгалужені словотвірні або комбінаторні гнізда.

Для точнішого аналізу користувач може застосовувати частиномовні фільтри, обмежуючи пошук певними частинами мови – наприклад, виділити лише випадки, де *taxi* сполучається з іменниками (*taxi clearance*, *taxi operations*) або прикметниками (*normal taxi speed*). Це дозволяє сконцентруватися на тих типах колокацій, які планується включити в словникову статтю: терміносполучення, кліше, назви параметрів чи операцій. Сформовані у такий спосіб групи колокацій безпосередньо переносяться у словниковий реєстр або статтю. При цьому укладач може:

- створити **гнізда** (наприклад, *taxi speed*, *taxi distance*, *taxi clearance*),
- додати типові приклади використання з високими GDEX-оцінками,

- подати **тлумачення** з опорою на реальні технічні контексти,
- позначити варіанти перекладу у двомовних словниках.

Завдяки такій інтеграції корпусних даних у процес укладання статті кінцевий словник не лише відображає термін у його «чистому» вигляді, а й демонструє його реальне функціонування в авіаційних інструкціях, що суттєво підвищує точність і практичну цінність видання. На основі отриманих контекстів укладач може створити окремий словник колокацій із типовими прикладами їхнього уживання; для цього передбачено спеціальний інструмент **TickBox Lexicography**, детальний огляд якого подано нижче.

TickBox Lexicography – це функціональність Sketch Engine, розроблена для ефективного експорту релевантних даних корпусу у форматі, зручному для систем лексикографічного редагування. Завдяки TBL лексикограф може вибирати колокації та приклади безпосередньо з Word Sketch, позначаючи їх за допомогою «галочок» (tick-box). Інструмент поєднує можливості Word Sketch та GDEX, автоматично підбираючи до кожної обраної колокації набори прикладів, які вже сортовані за GDEX-оцінками. Після вибору колокацій і прикладів лексикограф може зберегти їх разом в XML-файлі, який підходить для імпорту в систему створення словників (DWS).

Формат цього XML-файлу можна налаштовувати відповідно до специфічних вимог словникової системи (наприклад, з позначенням граматичних міток, тезауруса тощо). Це суттєво скорочує час перенесення корпусних даних у словниковий редактор, адже функція автоматизує і структурує процес лексикографічного добору даних. На рисунку 8.13 подано приклад структури XML-файлу, що може згенерувати TickBox Lexicography та Word Sketch. Значення тегів, що застосовуються в файлі, показано у таблиці 8.6.

Таблиця 8.6 – Значення тегів

Тег	Значення
<entry>	словникова стаття
lemma i pos	параметри, які TBL може додавати автоматично
<collocations>	блок колокацій, відібраних у Word Sketch
<collocation>	підблок для окремої колокації
<examples>	приклади до колокації


```

<entry>
<keyword>обличчя</keyword>
<gramrel>
  <grname>modifiers of "обличчя"</grname>
  <collocation>
    <collo>блідий</collo>
    <example>На її блідому <b>обличчі</b> залишилися живими тільки очі.</example>
    <example>На блідому <b>обличчі</b> від гніву не залишилось і сліду.</example>
    <example>Раптом легкий вітерець проповз по блідому <b>обличчю</b>.</example>
    <example>Світло зірок лягає на бліде витягнуте <b>обличчя</b>.</example>
    <example>Бліде <b>обличчя</b> кольору молодого меду було незворушним.</example>
    <example>З її блакитних очей по блідому <b>обличчі</b> продовжували збігати сльози.</example>
  </collocation>
  <collocation>
    <collo>жіночий</collo>
    <example>Жіночі <b>обличчя</b> найчастіше прикрашали обкладинки радянських журналів про кіно.</example>
    <example>У війни не повинно бути жіночого <b>обличчя</b>.</example>
    <example>Кажуть – у війни не жіноче <b>обличчя</b>.</example>
    <example>Ознаки вікових змін в першу чергу проявляються на жіночому <b>обличчі</b>.</example>
    <example>Тепер прийшла черга поглянути і на жіноче <b>обличчя</b> святості.</example>
    <example>Чоловічий світ маркетингу знайшов бінарність і жіноче <b>обличчя</b>.</example>
  </collocation>
</collocation>

```

Рис. 8.13 – Фрагмент структури в XML-файлі

Після імпорту XML-файлу в Lexonomy (див. підрозділ 6.1) укладач отримує повноцінний доступ до структури кожної статті. На цьому етапі можна не лише переглядати автоматично згенеровані елементи, а й адаптувати їх відповідно до редакційних вимог: додавати дефініції або кілька значень, упорядковувати колокації за типами, вносити уточнення щодо граматичних чи стилістичних характеристик, додавати переклади або посилання на споріднені лексеми. Крім того, у Lexonomy можна налаштувати шаблон відображення статті – створити власну структуру статей, додати нові поля (наприклад, «Сфера вживання», «Джерело прикладу», «Тип колокації», «Етимологія»), а також керувати тим, що саме буде відображатися для користувача у фінальному словнику.

Інтеграція **Sketch Engine** з **Lexonomy** забезпечує безпосередній імпорт корпусних даних у словник: лексеми, колокації, приклади використання, частотні профілі та інші елементи, згенеровані в Sketch Engine, можуть автоматично переноситися у структуру словникової статті в Lexonomy. Такий механізм суттєво прискорює роботу лексикографа, мінімізує потребу в ручному копіюванні та гарантує узгодженість між корпусними даними й словниковими записами. Більше того, окремі елементи словникової статті — наприклад, «Колокації», «Контексти», «Синоніми», «Приклади» — можуть бути інтегровані з конкретним корпусом, що дозволяє автоматично вилучати з нього відповідні дані для подальшого редагування та включення в статтю. Усе це робить процес укладання словника гнучкішим та точнішим (див. рис. 8.14).

З технічного погляду інтеграція між Sketch Engine і Lexonomu працює через спеціальні API-механізми, які дозволяють Lexonomu отримувати доступ до корпусних результатів у напівавтоматичному режимі. Після налаштування зв'язку проєкт словника в Lexonomu «пам'ятає» обраний корпус, тому під час оновлення корпусних даних або їх повторного аналізу (наприклад, після додавання нових документів) лексикограф може автоматично оновити відповідні частини словникових статей. Це гарантує актуальність інформації й запобігає дублюванню роботи.

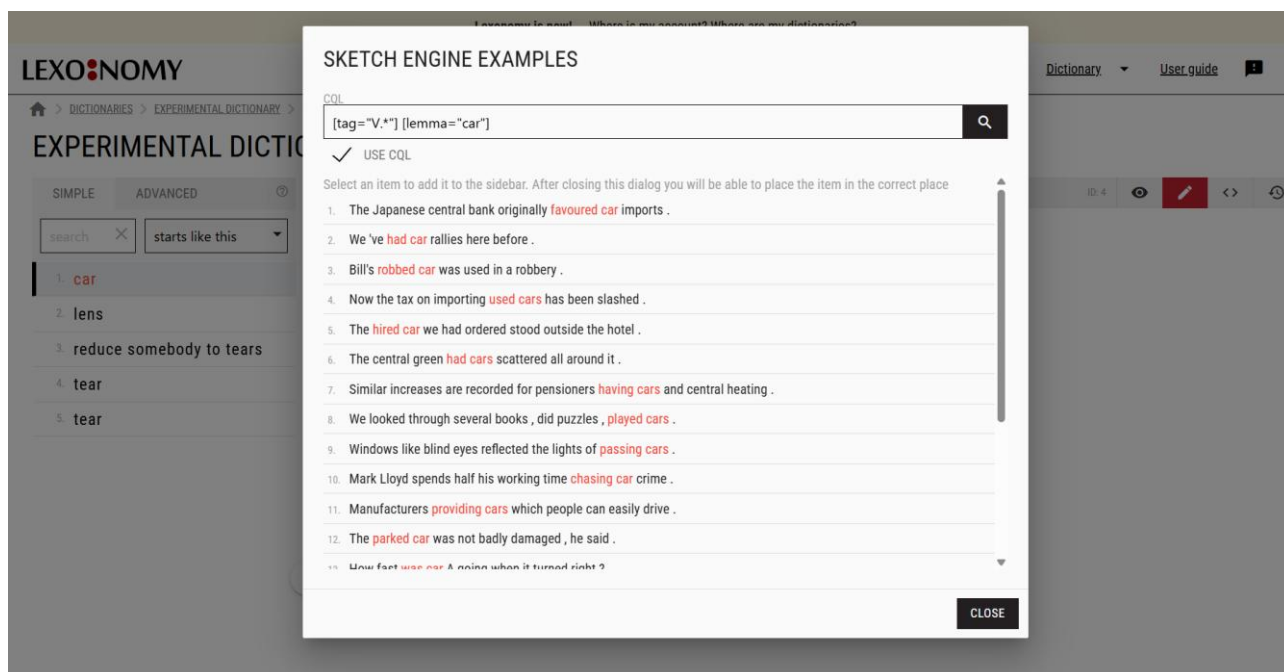


Рис. 8.14 – Робота з корпусом у Lexonomu

Інтеграція також підтримує кілька типових сценаріїв використання. Перший – початкове формування статей: лексикограф імпортує частотні словники, колокації та контекстні приклади для створення «чернеток» статей, які згодом вручну опрацьовує. Другий сценарій – поглиблене редагування: працюючи над окремою статтею, можна здійснювати точкові запити до корпусу прямо з інтерфейсу Lexonomu та вставляти нові приклади або колокації. Третій сценарій – масове оновлення: після розширення корпусу або зміни фільтрів добору прикладів GDEX усі статті, що містять пов'язані елементи, можуть бути автоматично оновлені відповідно до нових даних. Такий підхід забезпечує не лише ефективність, а й динамічність словника, адже його зміст може постійно оновлюватися разом із розвитком корпусу, не потребуючи створення нової версії з нуля.

Працюючи з платформою Sketch Engine, слід враховувати її переваги та обмеження. До переваг належать автоматизація ключових етапів укладання словника (генерація реєстру, статей, прикладів), інтегрованість з Lexonomy та іншими платформами, гнучку систему запитів (CQL), розвинений інструментарій лінгвістичного аналізу, можливість роботи з одномовними та багатомовними корпусами та створення користувацьких корпусів, а також високу швидкість опрацювання даних. Як обмеження слід зазначити потребу знати мову CQL для повноцінного використання корпусів, доступність багатьох функцій лише за передплатою, не завжди коректну розмітку особливо для нових термінів або нечастих конструкцій та необхідність додаткового редагування згенерованих статей. Порівняльну характеристику інструментів Sketch Engine наведено у таблиці 8.7.

Таблиця 8.7 – Характеристика інструментів Sketch Engine

Інструмент	Основна функція	Тип даних	Для яких словників	Переваги	Обмеження
OneClick Dictionary	Генерація чернеток словникових статей	Колокації, приклади, профілі	Тлумачні, навчальні	Швидкість, цілісна стаття	Потребує редагування
OneClick Terms	Виявлення термінів	Частотність, приклади	Термінологічні	Точна ідентифікація термінів	Нерелевантні одиниці можливі
GDEX	Автоматичний добір прикладів	Корпусні речення	Усі типи словників	Якісні приклади	Не завжди враховує контекст
TBL	Формування структури статті через позначки	Обрані лексикографом дані	Тлумачні, навчальні	Контроль редактора	Потребує ручного вибору
Sketch Engine + Lexonomy	Інтеграція даних у словник	Увесь набір корпусних даних	Усі типи	Прямий імпорт, економія часу	Залежність від двох платформ

Отже, Sketch Engine пропонує потужний інструментарій для автоматизованого опрацювання мовного матеріалу, який може слугувати надійною основою для укладання різних типів словників. Водночас навіть найякісніші автоматично згенеровані дані потребують певного редагування та коригування з боку фахівця, аби забезпечити точність, відповідність нормам і повну відповідність концепції конкретного словника.

Питання для самоконтролю

1. Які основні інструменти Sketch Engine призначені для укладання словників? Назвіть їх і коротко поясніть призначення кожного.
2. У чому полягає принцип роботи OneClick Dictionary? Як система формує словникові статті та які типи даних можуть бути додані автоматично?
3. Чим OneClick Terms відрізняється від OneClick Dictionary? У яких випадках доцільно використовувати кожен із цих інструментів?
4. Які формати документів підтримує OneClick Terms у режимах одномовного та двомовного опрацювання?
5. Що таке GDEX і яку роль він відіграє в лексикографії? Чому не рекомендується включати в словникову статтю приклади без попередньої фільтрації?
6. Які критерії лежать в основі оцінювання прикладів GDEX? Як інтерпретувати високий або низький рейтинг пропозиції?
7. Які переваги дає використання TickBox Lexicography? У чому особливість його роботи порівняно з традиційними методами добору колокацій?
8. Як працює інтеграція Sketch Engine з Lexonomy? Які дані можуть передаватися автоматично й що це дає лексикографу?
9. Чому автоматично згенеровані корпусні дані потребують додаткового редагування? Наведіть приклади потенційних проблем.
10. Які можливості щодо експорту даних передбачені в OneClick Terms та OneClick Dictionary? У яких форматах користувач може зберегти результати?

Завдання для самоперевірки

1. Складіть список з 5–7 термінів з будь-якої галузевої теми (наприклад, авіація, медицина, інформаційні технології) та опишіть, який інструмент Sketch Engine був би найбільш ефективним для їх вилучення і чому.
2. Проаналізуйте два приклади: один до фільтрації GDEX, інший – після. Опишіть, які саме параметри GDEX покращили якість прикладу.

3. Сформуйте запит для вилучення колокацій типу «прикметник + іменник» у будь-якому корпусі Sketch Engine. Вкажіть, які особливості корпусу потрібно врахувати (частини мови, теги, морфологічна розмітка).

4. Створіть прототип словникової статті для одного терміна, додавши:

- дефініцію,
- колокації,
- щонайменше 2 приклади GDEX,
- можливі синоніми або співвіднесенні терміни. Вкажіть, які саме дані ви б імпортували зі Sketch Engine.

5. Опишіть кроки імпорту XML-файлу до Lexonomy. Що можна робити зі статтею після імпорту? Які елементи зазвичай редагують вручну?

6. Порівняйте три способи роботи інструмента OneClick Terms (one language, aligned texts, non-aligned texts). Вкажіть переваги та обмеження кожного підходу.

7. Створіть таблицю порівняння «Було/Стало» для кількох прикладів корпусного матеріалу до та після використання GDEX (3–5 рядків).

8. Проаналізуйте мінімум три колокації терміна, що вас цікавить (наприклад, *engine failure*, *hydraulic pressure*, *flight crew*). Для кожної опишіть частоту, тип граматичного зв'язку та приклад уживання.

8.3. Електронні словники в корпусній лінгвістиці

Електронні словники, призначені для автоматичного морфологічного аналізу та розмітки (pos-tagging), становлять критично важливу ланку сучасних корпусних технологій. Вони не просто містять перелік слів – вони формалізують правила мови: подають парадигми, частини мови, граматичні категорії та допустимі словоформи. На основі таких ресурсів алгоритми маркування корпусу можуть виконувати розпізнавання частин мови, лематизацію, побудову морфологічних дерев та зняття полісемії. POS-tagger використовує словник у кілька способів:

- *лематизація*: визначення базової форми введеної словоформи.
- *визначення можливих тегів*: словник подає всі варіанти граматичних тлумачень.

– *зняття неоднозначності*: попередній список можливих тегів комбінується з контекстною моделлю (статистичною або нейронною).

– *перевірка відповідності*: недійсні комбінації (наприклад, «пити + імперфект + родовий множини» у якомусь формальному форматі) відкидаються.

Одним із центральних електронних морфологічних ресурсів для української мови є *Великий словник української мови* (ВЕСУМ), що використовують у корпусі ГРАК та в низці інструментів морфологічного аналізу й автоматичної розмітки (POS-tagging, лематизація). Його характеристиками є:

- Обсяг близько 430 000 лем (список лем постійно збільшується).
- Структура: кожна лема представлена набором словоформ, морфологічними тегами та додатковими атрибутами.
- Гнучкість: система тегів дозволяє описувати широкий набір морфологічних груп і граматичних ознак.
- Обробка неоднозначностей: омоніми розділено, якщо вони відрізняються морфологічно (різні частини мови або різні набори тегів).
- Ненормативні елементи: у словнику враховано сленгові, вульгарні, архаїчні та інші позанормативні одиниці, а також ненормативні форми.
- Історичні/правописні мітки: присутні теги, що позначають належність форми або написання до варіантів правопису (наприклад, правописи 1992–2019 років).
- Суржикові (покручі): містить інформацію про приблизно 8 000 покрусів, їхні варіанти заміни.

Кожна мовна одиниця у ВЕСУМ має розгорнутий набір тегів для опису її морфологічних властивостей: від базової належності до граматичного класу та зазначення ключових граматичних категорій (роду, числа, відмінка, виду, особи тощо) до фіксації її словозмінних форм і морфологічних варіантів. До таких варіантів належать як нормативні, так і позанормативні форми – зокрема сленгові, розмовні та покрусі (суржикові одиниці), для яких словник додатково подає рекомендовані нормативні відповідники. Завдяки цьому ВЕСУМ може

відображати повний спектр реального мовного вжитку та забезпечувати точнішу автоматичну розмітку. Приклади розмічення одиниць наведено в таблиці 8.8.

Таблиця 8.8 – Приклади розмічення одиниць тегами ВЕСУМ

Лема	Розмічення	Коментар
вода	noun:inanim:f:v_naz	іменник, неістота, ж.р. у формі род. відм.
водо	noun:inanim:f:v_kly	іменник, неістота, ж.р. у формі клич. відм.
синій	adj:m:v_naz:compb	прикм., ч.р., наз. відм., базова форма
синєє	adj:n:v_naz:compb:long	прикм., ч.р., наз. відм., базова форма, нестягнена форма
бути	verb:imperf:inf	дієслово недок. виду у формі інфінітива
будь	verb:imperf:inf:short	дієслово недок. виду в короткій формі інф.
будь	verb:imperf:impr:s:2	дієслово недок. виду наказового сп. 2 ос. одн.
єсть	verb:imperf:pres:s:2:predic:rare	дієслово недок. виду, теп.ч., 2 ос. одн., рідк.

ВЕСУМ працює як додаток в офісних пакетах LibreOffice та Microsoft Word, інтегрується в онлайн-редактори на кшталт Google Docs, а також підтримується браузерами Firefox та Chrome. Крім того, словник «вмонтовано» в пошуковий двигун Apache Lucene, що дає змогу застосовувати його морфологічні ресурси у великих інформаційно-пошукових системах. Усі ці інтеграції роблять ВЕСУМ універсальним інструментом для морфологічної розмітки, редагування та автоматичної перевірки текстів. Нижче на рисунку 8.15 показано результат розмічення тексту у форматі XML парсером TagText із використанням словника.

```

<sentence>
  <tokenReading>
    <token value='Україна' lemma='Україна' tags='noun:inanim:f:v_naz:prop:geo' />
  </tokenReading>
  <tokenReading>
    <token value='е' lemma='бути' tags='verb:imperf:pres:p:1' />
    <token value='е' lemma='бути' tags='verb:imperf:pres:p:2' />
    <token value='е' lemma='бути' tags='verb:imperf:pres:p:3' />
    <token value='е' lemma='бути' tags='verb:imperf:pres:s:1' />
    <token value='е' lemma='бути' tags='verb:imperf:pres:s:2' />
    <token value='е' lemma='бути' tags='verb:imperf:pres:s:3' />
  </tokenReading>
  <tokenReading>
    <token value='суверенна' lemma='суверенний' tags='adj:f:v_kly' />
    <token value='суверенна' lemma='суверенний' tags='adj:f:v_naz' />
  </tokenReading>
  <tokenReading>
    <token value='і' lemma='і' tags='conj:coord' />
    <token value='і' lemma='і' tags='part' />
  </tokenReading>
  <tokenReading>
    <token value='незалежна' lemma='незалежний' tags='adj:f:v_kly:compb' />
    <token value='незалежна' lemma='незалежний' tags='adj:f:v_naz:compb' />
  </tokenReading>

```

Рис. 8.15 – Результат роботи парсера TagText у форматі XML

Як було зазначено вище, словник ВЕСУМ використовується для морфологічної розмітки текстів у корпусі ГРАК, що продемонстровано на

рисунку 8.16. Під час перегляду результатів розмічення морфологічні теги можуть бути відображені як для ключового слова в контексті (KWIC), так і для всіх слів у тексті, що забезпечує повноцінний аналіз граматичної структури та полегшує подальшу лінгвістичну інтерпретацію. Завдяки такій деталізації користувач може виконувати складні морфологічні запити — наприклад, шукати форми певного відмінка, виду чи часу — та отримувати точніші результати порівняно з пошуком лише за поверхневими формами.

The screenshot shows the CONCORDANCE web interface. At the top, there's a search bar with 'Grac v.19a' and a search icon. Below it, a red box highlights the lemma 'комунікувати' with 4,361 occurrences and a frequency of 2.15 per million tokens. The interface has tabs for 'Details', 'KWIC', and 'Right context'. The KWIC tab is active, showing a table of results. Each row contains a list item number, a checkbox, a left context snippet, the lemma with its morphological tag, and a right context snippet. The lemma 'комунікувати' is shown with various tags like 'verb imperf past p', 'verb imperf pres s 3', 'verb imperf pres p 3', 'verb imperf pres s 3', 'verb imperf pres p 3', 'verb imperf pres s 3', 'verb imperf pres p 3', 'verb imperf pres s 3', 'verb imperf pres p 3', and 'verb imperf pres s 3'.

	Left context	KWIC	Right context
1	и в Донецькому басейні збудувати коло 600 верст залізниці, котрі б краще	комунікували verb imperf past p	шахти поміж собою та доставляли продукти, — то легко можна було б знач
2	/s>s>Воєнне положення в Галичині. /s>s>Виділ II. /s>s>Штабу Д. О. Г.	комунікує verb imperf pres s 3	: 30. серпня в полудне: Акція ген. /s>s>Павленка розвивається успішно. <
3	іл тонаж. /s>s>Перед переговорами в Ризі. /s>s>Пресовий Виділ МСЗ.	комунікує verb imperf pres s 3	: Дня 7, ц. м. відібрано з Москви таку телеграму адресовану до мін. загр. сп
4	/s>s>Японці посуваються у глибину суші. /s>s>Головна японська квартира	комунікує verb imperf pres s 3	, що частини, які причалили в заливі Біяс, осягнули лінію Таншуні. /s>s>Віс
5	Ваневі Себенів та Пивовщині читальні " Просвіти ". /s>s>Це сталося —	комунікує verb imperf pres s 3	ПАТ — внаслідок " вибриків української молоді проти спокою мешканців по
6	ис, що так, як було, не буде ніколи. /s>s>А як буде? /s>s>А буде, дійсно,	комуніковано verb perf impers	всі функції унікальної перлини оздоровлення дітей у Міністерстві сім'ї і мол
7	/ та в Берліні. /s>s>Через очолюваний ним інтернет-часопис e-flux щодня	комунікують verb imperf pres p 3	близько 70 тисяч художників з усього світу. /s>s>В Арсеналі Антон Відокл
8	фахових знань, людина стає випускником школи життя, де заново вчиться	комунікувати verb imperf inf	, розгадувати певні коди на рівні національностей і розуміти, що люди всюд
9	зтів би, щоб наш близький контакт не припинився через Ваш виїзд, щоб ми	комунікували verb imperf past p	навзаєм, часто й істотно. /s>s>Можливо, навіть, що такий контакт на відс
10	лись нові можливості, і ми не маємо їх опустити. /s>s>Роль бізнесу зараз	комунікувати verb imperf inf	західним інвесторам, що тут є ті компанії, які вміють добре працювати і з як

Рис. 8.16 – Відображення морфологічної розмітки KWIC

Отже, ВЕСУМ посідає ключове місце в українській корпусній лінгвістиці як один із найповніших і найстандартизованиших морфологічних словників. Його інтеграція в інструменти розмічення текстів — від TagText до тагера ГРАК — робить його фактичним еталоном для автоматичного морфологічного аналізу. Завдяки уніфікованій системі тегів, широкій охопності лексики та підтримці різних платформ, ВЕСУМ забезпечує узгодженість, точність і відтворюваність корпусних даних. У сучасних корпусах він не просто словник-довідник, а інфраструктурний компонент, що гарантує якісне анотоване мовне середовище для досліджень і лінгвістичних застосунків. Аналогами ВЕСУМ у світовій практиці є MULTEXT-East Morphosyntactic Lexicons, CELEX Lexical Database, BNC XML Lexicon, UD Lexicons та Apertium dictionaries. Усі ці ресурси виконують подібну функцію — забезпечують стандартизовану морфологічну інформацію для корпусів і мовних технологій. На цьому тлі ВЕСУМ упевнено вписується в міжнародний лексикографічний ландшафт, пропонуючи українській

мові той самий рівень структурованості й формалізації, який давно є нормою для провідних європейських мов.

Морфосинтаксичні словники MULTEXT-East – багатомовний ресурс, що містить словоформи, базові форми (леми) та детальний морфосинтаксичний опис (MSD) для широкого кола мов. Створені в межах проєкту MULTEXT-East, ці лексикони стали ключовим інструментом для задач автоматичної обробки природної мови. З часом ресурс було розширено, і він охопив понад десяток мов – переважно слов'янських та інших мов Центральної й Східної Європи. До них належать болгарська, чеська, англійська, естонська, угорська, македонська, перська, польська, румунська, російська, сербська, словацька та українська.

Кожна словникова стаття містить три основні поля: **словоформу** (з відповідним відмінюванням), **лему** (основну форму) та **морфосинтаксичний опис (MSD)**. Поле MSD подає детальну інформацію про граматичні властивості одиниці відповідно до узгодженої специфікації, створеної на основі схеми **EAGLES**. Лексикони призначено для підтримки різноманітних NLP-застосунків та для створення багатомовних анотованих корпусів. Дані звичайно кодуються у форматі **XML** із використанням стандарту **Text Encoding Initiative (TEI)**, що забезпечує їхню сумісність і можливість широкого використання. У таблиці 8.9 подано приклади MSD розмітки для слів із різних європейських мов. Додаткові приклади MSD розмітки подано в додатку Д.22.

Таблиця 8.9 – Приклади MSD розмітки

Словоформа	Лема	MSD	Розшифрування
cats (англ.)	cat	Ncmpl	іменник, загальний, чол. рід*, множина, неістота
гарні (укр.)	гарний	Afp-pn	прикметник, повна форма, множина, позитивний ступінь, називний
четях (болг.)	чета	Viptls	дієслово, імперфект, дійсний спосіб, перехідне, 1 ос., одн.

Як видно, ключовою рисою проєкту MULTEXT-East є розроблення уніфікованої моделі опису морфосинтаксичних характеристик, яка забезпечує порівнюваність та узгодженість мовних даних для широкого спектра європейських мов. Такий підхід дає змогу застосовувати єдині принципи тегування до мов із різними морфологічними системами, що особливо важливо для багатомовних корпусів та інструментів автоматичної обробки мови. Завдяки

цьому стандартизованому підходу лексикони MULTEXT-East можуть слугувати основою для узгодженого морфосинтаксичного тегування, багатоцентричних корпусних проєктів і алгоритмів машинного навчання, що працюють із паралельними та порівнюваними текстами різних мов.

CELEX Lexical Database охоплює ASCII-версії лексичних баз даних англійської (версія 2.5), голландської (версія 3.1) та німецької (версія 2.0) мов. Цей масштабний ресурс є результатом колаборації кількох провідних наукових установ: Університету Неймегена, Інституту нідерландської лексикології в Лейдені, Інституту психолінгвістики імені Макса Планка та Інституту досліджень сприйняття в Ейндговені. Попередня обробка й фінальне виробництво наборів даних були виконані центром LDC (Linguistic Data Consortium). CELEX надає винятково детальний лексичний опис, охоплюючи кілька рівнів мовної структури:

- *орфографія*: альтернативні написання, побуквений склад слова, поділ на склади;
- *фонологія*: фонетичні транскрипції, варіативність вимови, структура складів, розміщення основного наголосу;
- *морфологія*: дериваційна та композиційна структура, повні парадигми словозміни;
- *синтаксис*: словотвірні класи, синтаксичні підкатегорії, аргументні структури;
- *частотність*: частоти словоформ та лем, отримані з репрезентативних корпусів відповідних мов.

На відміну від багатьох сучасних електронних лексиконів, CELEX не прив'язаний до певної системи керування базами даних. Усі дані подано в ASCII-файлах, організованих у дереві каталогів UNIX, що робить ресурс максимально універсальним, але водночас вимагає від користувача базових технічних навичок. Доступ і пошук інформації здійснюються через інструменти на кшталт **AWK** або **ICON**, а унікальні ідентифікаційні номери дозволяють зв'язувати між собою дані з різних таблиць. Частина відомостей не зберігається у вигляді готових значень і

генерується «на льоту» – для цього надається набір вбудованих функцій AWK, що відтворюють необхідні обчислення.

У підсумку CELEX відіграє роль універсального багаторівневого лексичного ресурсу, що поєднує орфографічні, фонологічні, морфологічні, синтаксичні та частотні дані в єдиній структурі. Завдяки такій глибині й стандартизованості опису CELEX широко використовується для корпусної розмітки, створення мовних моделей, лінгвістичних експериментів та розробки інструментів автоматичної обробки мовлення і тексту. Його багаторівневий формат дозволяє інтегрувати базу як у класичні корпусні системи, так і в сучасні NLP-платформи – від морфологічних тегерів до систем синтезу та розпізнавання мовлення.

BNC XML Lexicon – це лексичний ресурс, створений на основі British National Corpus (BNC), одного з найбільших і найвпливовіших корпусів англійської мови. Лексикон являє собою структурований у форматі XML набір словникових одиниць, який містить інформацію про лема, частини мови, граматичні характеристики та частотні показники, отримані з усього корпусу. Його мета – надати уніфікований лексичний опис, який може використовуватися для корпусного пошуку, морфосинтаксичної розмітки, автоматичного тегування нових текстів, а також для лексичної підтримки різноманітних NLP-інструментів. У кожній словниковій статті зазвичай подано:

- лему, яка слугує базовою формою;
- *частиномовний тег* згідно з системою POS-тегів BNC;
- *граматичні варіанти* (наприклад, форми однини/множини, часи дієслів);
- *частотні дані*, отримані з BNC у цілому та його підкорпусів;
- *варіативність* (орфографічні чи морфологічні варіанти).

XML-формат забезпечує зручну інтеграцію лексикону з корпусними менеджерами, тегерами та програмами узгодженого пошуку. Завдяки суворій структурі XML дані легко конвертуються в інші моделі подання – TEI Lex-0, JSON, TSV, словникові бази тощо. BNC XML Lexicon став одним із перших великих лексичних ресурсів англійської, які поєднували корпусну частотність із класичним лексикографічним описом, тому його часто використовують як

«еталонний» набір для тестування нових POS-тегерів і лематизаторів. У лінгвістичних дослідженнях він виконує роль стабільного опорного словника для порівняльних корпусів і для верифікації морфологічних моделей.

UD Lexicons – це лексичні ресурси, створені на основі проєкту Universal Dependencies (UD), глобальної ініціативи зі стандартизації морфологічної та синтаксичної анотації для сотень мов. На відміну від традиційних словників, UD Lexicons виникають як «похідні» від анотованих корпусів UD: вони автоматично або напівавтоматично генеруються з розмічених текстів, а потім доповнюються мовними спільнотами. Мета UD Lexicons – забезпечити уніфіковану морфологічну інформацію, яку можна однаково інтерпретувати для різних мов, що особливо важливо у багатомовних NLP-застосунках. Кожна словникова стаття містить:

- Lemma – базову форму слова;
- UPOS – універсальну частину мови відповідно до Universal POS tags;
- XPOS – мовоспецифічний тег, що зберігає локальну традицію розмічення;
- FEATS – набір універсальних граматичних ознак (число, рід, відмінок, час, аспект тощо);
- Lexeme-specific features – нестандартні або додаткові категорії, характерні лише для конкретної мови;
- Frequency data – частотні показники на основі корпусу UD.

Оскільки UD моделі розмітки побудовані на принципах максимальної універсальності, UD Lexicons працюють як інструмент «вирівнювання» морфологічних систем різних мов. Завдяки цьому вони широко застосовуються у багатомовному тегуванні, машинному перекладі, розробці словозмінних і морфологічних моделей, семантичному аналізі та синтаксичних парсерах (Stanza, spaCy UD-pipelines, UDPipe).

Більшість UD Lexicons доступні у відкритому доступі на GitHub та дистрибутивах UD. Їх структура стандартизована, що спрощує імпорт у корпусні менеджери, термінологічні бази та навчання мовних моделей. Особливістю UD Lexicons є те, що вони постійно оновлюються: щороку виходять нові версії разом

з оновленнями корпусів. Це робить їх надзвичайно релевантним ресурсом, адаптованим до сучасного мовного вжитку, на відміну від традиційних статичних морфологічних словників.

Apertium dictionaries – це набір морфологічних та перекладних лексиконів, що входять до екосистеми Apertium, відкритої платформи для машинного перекладу та обробки природної мови. На відміну від UD Lexicons, які впливають із корпусної анотації, словники Apertium створюються насамперед для моделювання словозміни та автоматичного перекладу, а тому мають багатокomпонентну структуру та високий ступінь формальної чіткості. Apertium використовує формалізм **LTtoolbox** – набір правил і структур для зберігання: лексем (лем і словоформ), морфологічних ознак, словозмінних парадигм, флективних шаблонів, перекладних відповідників (у випадку двомовних словників). Кожен словник Apertium являє собою XML-файл з чітко визначеними блоками:

1. *Моно-лексикон (monodix)*, який описує морфологію конкретної мови: лем, частини мови, граматичні категорії, словозмінні парадигми.

2. *Білексикон (bidix)* – двомовний словник для перекладу між двома мовами: відповідники лем, вказівки на омонімію, контекстні обмеження (правила вибору значення).

3. *Правила обмежень (transfer rules)* – шаблони, які керують перекладом словосполучень і конструкцій.

4. *Фінальна компіляція* – словник перетворюється на високоефективний бінарний формат, який забезпечує швидку обробку текстів.

Apertium підтримує понад 40 мов та 70+ мовних пар, серед яких іспанська, каталонська, галісійська, окситанська; англійська, французька, німецька; слов'янські мови (болгарська, словенська, македонська, російська, українська – частково); тюркські, угро-фінські, індійські мови. Особливість Apertium у тому, що воно активно заохочує мовні спільноти до створення власних словників – тому для малих і міноритарних мов (баскська, румунська, окситанська) Apertium часто є єдиним відкритим джерелом структурованих морфологічних ресурсів.

Попри те, що первинно Apertium проєктувався для перекладу, його словники часто застосовують у морфологічному розборі, лематизації корпусів, – розміченні словоформ, – автоматичному створенні словозмінних парадигм, та побудові навчальних даних для NLP-моделей.

Усі розглянуті ресурси – лексикони, які виконують подібну функцію (морфологічний опис слів), але створювалися з різними цільовими моделями, обсягами та історичними контекстами. Відповідно, їхні підходи до структурування даних, репрезентації граматичної інформації та інтеграції з корпусними технологіями суттєво відрізняються. Нижче у таблиці 8.10 подано порівняльну характеристику, що допомагає побачити спільні риси та ключові відмінності між п'ятьма основними системами.

Таблиця 8.10 – Порівняльна характеристика словників

Ресурс	Основна сфера	Формат тегів	Мовне покриття	Сильні сторони	Обмеження
VESUM	Корпуси, морфологічний аналіз	Національний	1	Точність, адаптованість	Менш універсальний
MULTEX T-East	Мультимовні корпуси	Позиційні коди	~15	Стандартизованість	Старіння стандарту
CELEX	Психолінгвістика, морфологія	Спеціальний	3	Дуже глибока деталізація	Обмежена кількість мов
UD Lexicons	NLP, корпуси	UD-tags	100+	Інтероперабельність, сучасність	Іноді спрощення
Apertium	Машинний переклад	LEXC/HFST	40+	Практичність, гнучкість	Нефокусованість на корпусах

Кожен проаналізований ресурс відображає свій “культурний код” у мовних технологіях. VESUM є прикладом національного, глибоко мовно-орієнтованого підходу. MULTEXT-East – це перше покоління універсальних мультимовних морфологічних стандартів. CELEX являє собою академічний фонологічно-морфологічний ресурс. UD Lexicons створювався як сучасний глобальний стандарт корпусної морфології та синтаксису. Apertium – практична морфологія, зрошена потребами перекладу. У підсумку, українська корпусна лінгвістика стоїть у діалозі з усіма цими підходами, а VESUM та UD Lexicons сьогодні утворюють найдієвіший тандем для сучасних NLP-інструментів і моделювання мови.

Питання для самоконтролю

1. Що таке VESUM і яку роль він відіграє в українській корпусній лінгвістиці?
2. Які відмінності існують між національними морфологічними тегсетами (типу VESUM) та міжнародними (UD, MULTEXT-East)?
3. Чим відрізняється підхід MULTEXT-East до морфологічного опису мов від підходу UD Lexicons?
4. Які типи лінгвістичної інформації містить CELEX, і чому він корисний у психолінгвістиці?
5. У чому полягає практичний фокус Apertium dictionaries, і чому їхня морфологія відрізняється від корпусних стандартів?
6. Що означає “інтероперабельність” у контексті морфологічних тегсетів і які ресурси є найбільш інтероперабельними?
7. Які типи лінгвістичних завдань найчастіше розв’язуються за допомогою UD Lexicons?
8. Чому MULTEXT-East вважається історично важливим проєктом?
9. У чому сильні та слабкі сторони Apertium у порівнянні з ресурсами VESUM або UD?
10. Які критерії варто враховувати під час вибору морфологічного лексикону для конкретного дослідницького або практичного завдання?

Завдання для самоперевірки

1. Створіть власну міні-таблицю (4–6 параметрів порівняння), де зіставте два ресурси на вибір: а) VESUM; б) MULTEXT-East; в) CELEX; г) UD Lexicons; д) Apertium dictionaries. Порівняйте, наприклад, формат тегів, структуру словоформ, цілі створення, переваги й недоліки.
2. Візьміть одну довільну словоформу (будь-яку українську чи англійську) та:
 - розпишіть її морфологічні характеристики у форматі VESUM;
 - спробуйте зробити те саме у форматі UD;
 - прокоментуйте різницю – які графеми збігаються, а які відсутні/спрощені.

3. Напишіть міні-есе (5–7 речень) на тему: «Який лексикон найбільше відповідає завданням сучасної корпусної лінгвістики і чому?» Використайте те, що вивчили про фокус кожного ресурсу.

4. Візьміть невеликий текст (5–6 речень) та:

- оберіть тегсет (VESUM або UD);
- позначте всі частини мови та основні графеми;
- зробіть висновок: який тегсет був зручнішим і чому.

5. Сформулюйте, як саме можна розширити словник VESUM для кращої інтероперабельності з міжнародними ресурсами (наприклад, UD). Придумайте 2–3 пропозиції – це допоможе перевірити та поглибити розуміння структури ресурсів.

Питання для поглибленого вивчення матеріалу

1. Використання текстових корпусів (історичних, навчальних галузевих) у зарубіжній лексикографічній практиці.

2. Корпуси в термінографії як джерела формування галузевих словників та термінологічних баз даних.

3. Інструменти для розмічення текстових корпусів: TreeTagger, GATE (General Architecture for Text Engineering), UD Annotatrix.

4. Великий словник української мови (VESUM) та Електронний словник мови Тараса Шевченка (з морфемною розміткою).

5. Корпус ГРАК як джерело лінгвістичного матеріалу для укладання комп'ютерних словників.

6. Створення паралельних корпусів (ParaRook) та їх використання у комп'ютерній лексикографії.

7. Паралельні корпуси (ParaRook) як джерело фраземних конструкцій та їх еквівалентів.

9. ШТУЧНИЙ ІНТЕЛЕКТ В ЕЛЕКТРОННОМУ СЛОВНИКАРСТВІ

План заняття:

9.1. Можливості ШІ інструментів для створення словників

9.2. Огляд ШІ інструментів для роботи з мовним матеріалом

9.3. Створення словників і глосаріїв засобами ChatGPT

Упродовж останніх років штучний інтелект став одним із ключових драйверів розвитку електронної лексикографії. Якщо на початку 2000-х інтеграція корпусів і словників була технологічним проривом, то сьогодні аналогічну роль відіграють нейронні моделі, які здатні автоматично аналізувати велику кількість текстів, виокремлювати закономірності, формувати лексичні зв'язки та навіть проєктувати структуру словникових статей. Завдяки цьому словник перетворюється на динамічний сервіс, що здатний вберігати відкритість, оновлюваність і гнучкість у масштабах, недосяжних для традиційних методів.

9.1. Можливості ШІ інструментів для створення словників

Сучасні мовні моделі та спеціалізовані інструменти ШІ (Claude Sonnet, Gemini, Llama, Mistral Nemo, OpenAI GPT та інші) забезпечують широкий спектр функцій, важливих для лексикографічної роботи. До ключових можливостей належать автоматична екстракція лексичних даних, генерація та редагування лексикографічних матеріалів, структурування та семантична розмітка даних, а також підтримка користувача та оптимізація функціональності електронного словника. Розглянемо детальніше кожен із цих аспектів.

Автоматична екстракція лінгвістичних даних. Вилучення мовних одиниць із великих текстових масивів – одне з найтрудомісткіших завдань словникарства. Формування реєстру та добір релевантних прикладів потребує значного часу та ресурсів. Якщо короткі тексти ще дозволяють ручне опрацювання, то збільшення їхнього обсягу неминуче спричиняє втому, помилки, дублювання або пропуск важливих одиниць. У таких випадках автоматизовані ШІ-інструменти стають незамінними. ШІ ефективно здійснює такі операції:

– *збір та опрацювання корпусу:* мовні моделі здатні сканувати великі корпуси (іноді – мільярди слововживань) для виявлення нових слів, професійних

термінів, неологізмів, діалектизмів та семантичних зрушень у вже наявній лексиці;

– *виявлення колокацій та фразем*: алгоритми машинного навчання можуть автоматично ідентифікувати типові стійкі поєднання (наприклад, «*кричущий успіх*», «*непорушна істина*»). Це полегшує укладачеві формування якісних та достовірних прикладів.

На рисунку 9.1 представлено приклад роботи AI Term Extractor – інструмента, що здійснює автоматичне вилучення термінів із технічних інструкцій, зокрема документів про технічне обслуговування та ремонт електричних насосів. Такий підхід забезпечує значну економію часу та підвищує точність опрацювання спеціалізованих текстів, що є особливо важливим для укладання технічних та галузевих словників. Варто зауважити, що інструмент здатний вилучати не лише вузькоспеціальні терміни (*slipring, carbon brush, brush holder*), а й загальнотехнічні одиниці (*painting, tagging, revision*), що розширює його застосовність у мультижанрових текстах. У процесі автоматичної екстракції термінів важливо пам'ятати, що ШІ-моделі працюють за статистичними закономірностями, а не за фаховою інтуїцією. Тому інструмент може пропонувати терміни, які формально часто зустрічаються у тексті, але не мають незалежної термінологічної цінності (наприклад, *equipment, part, service* у технічних документах). Щоб уникнути перенасичення реєстру випадковими одиницями, рекомендується застосовувати двоетапну процедуру відбору: (1) автоматичне вилучення, (2) обов'язкова лексикографічна верифікація.

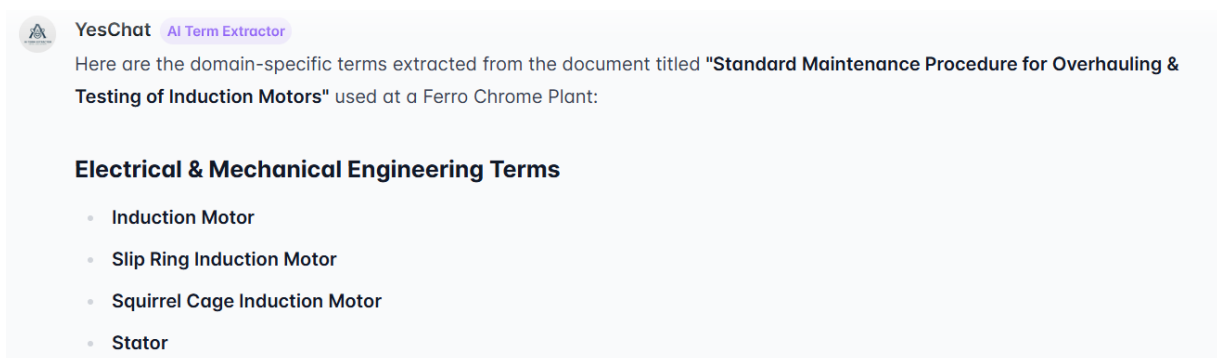


Рис. 9.1 – Вилучення термінів інструментом AI Term extractor

AI Term Extractor нерідко пропускає терміни, які мають низьку частотність або зустрічаються в документах у варіативних формулюваннях (наприклад, *seal*

ring / sealing ring, impeller cover / cover of the impeller). Такі одиниці можуть залишитися поза межами автоматичного аналізу. Для виявлення малочастотної термінології доцільно комбінувати два методи:

- автоматичне вилучення,
- пошук ключових конструкцій, характерних для технічних термінів (наприклад, *N + N, Adj + N, component + of + N*).

Поєднання обох стратегій підвищує повноту реєстру та забезпечує відлов спеціалізованих словосполучень, які не завжди добре розпізнаються моделлю.

Якість результатів безпосередньо залежить від ступеня натренованості моделі, доменної специфіки тренувального корпусу та налаштувань алгоритму. Час опрацювання, у свою чергу, визначається обсягом текстового матеріалу та складністю структури документації. У практичній роботі лексикографа це означає, що повна автоматизація поки неможлива, проте ШІ інструменти суттєво скорочують трудомісткість етапів попереднього аналізу та підготовки даних. Розглянемо тепер використання ШІ на етапах опрацювання лексикографічного матеріалу, зокрема створення та опрацювання лексикографічного матеріалу.

Генерація визначень. Другою за складністю лексикографічною задачею після добору реєстру є укладання **дефініцій** – формулювань, що точно й повно відтворюють зміст слова або терміна. Власне створення якісного визначення традиційно потребує глибокого знання предметної сфери, інтуїції, досвіду та значних часових витрат. Сучасні великі мовні моделі (англ. LLM або укр. ВЛМ) значно полегшують цю частину роботи, оскільки здатні генерувати первинні чернетки словникових статей. Такі чернетки можуть бути швидко адаптовані під різні рівні користувачів: школярів, студентів, фахівців, широкої аудиторії. Лексикограф заощаджує час на створення «нульового варіанту» і зосереджується на ключовому – точності, повноті та відповідності фаховим стандартам.

На рисунку 9.2 наведено приклад визначень, згенерованих моделлю Gemini для слова *мураха* з урахуванням різних цільових груп: учні 7–8 класів, студенти-біологи та пересічні користувачі. Залежно від аудиторії, модель варіює обсяг

інформації, термінологічне навантаження та деталізацію семантичної структури, виокремлюючи ті ознаки поняття, які є найбільш релевантними конкретній групі.

Визначення слова "Мураха"

А) Для школяра (7-8 класи)

Мураха (мурашка) — це маленька комаха, яка живе **великими родами** (колоніями) у спеціальних будинках — **мурашниках**. Вони дуже **працьовиті** та мають **чіткий розподіл обов'язків**: є **робочі мурахи**, які **шукають їжу і будують**, **солдати**, які **охороняють гніздо**, і **матка**, яка **відкладає яйця**. Мурахи **відіграють важливу роль у природі**, наприклад, **розносять насіння рослин і полюють на шкідливих комах**.

Б) Для студента біологічного відділення

Мураха (Formicidae) — це космополітична родина комах ряду **Перетинчастокрилі (Hymenoptera)**, що належить до надродини **Vespoidea** (застарілий погляд) або **Formicoidea** (сучасна філогенія). Характеризуються яскраво вираженою соціальною організацією (евсоціальність) з поліморфізмом каст (матка, робочі особини, самці). Анатомічно відрізняються наявністю одного або двох сегментів, які формують вузол (**петіоль**) між мезосомою (груди) та метасомою (черевце), а також особливою будовою вусиків (колінчасті). Мурахи є важливими компонентами екосистем, виконуючи функції хижаків, детритофагів, мірмекохорів та інженерів екосистеми. Їхня соціальна поведінка є класичною моделлю для вивчення соціальної біології та екології.

В) Для пересічного користувача

Мураха (мурашка) — це поширена, зазвичай невелика, **соціальна комаха**, відома своєю **працьовитістю** та життям у великих групах — **мурашниках**. Мурахи **відіграють важливу роль у навколишньому середовищі**, оскільки харчуються іншими комахами, рештками рослин і допомагають розносити насіння. Вони мають **чітку ієрархію**, де кожна комаха (робоча мураха, солдат, матка) виконує свою роль.

Рис. 9.2 – Генерація визначень моделлю Gemini

Попри високу якість генерації, визначення, створені мовними моделями, потребують обов'язкового опрацювання лексикографом. У процесі редагування варто звертати увагу на такі моменти:

(1) *Перевірка родо-видової структури.* Деякі моделі формують визначення описового типу («мурахи — це маленькі істоти, які...»), пропускаючи родовий термін («комахи»). Редактор має чітко відновити класичну структурну формулу.

(2) *Уникнення зайвої енциклопедичності.* ВЛМ доволі часто додає фонову інформацію (ареал, поведінкові особливості, історичні факти), яка не належить до дефініції. Її потрібно перенести в довідкові розділи або вилучити.

(3) *Коректне добирання термінології.* Моделі інколи пропонують надто складні або, навпаки, надто побутові слова. Завдання редактора — забезпечити рівень термінологічності, відповідний цільовій групі.

(4) *Перевірка семантичної повноти.* Інструмент ШІ інколи опускає важливі диференційні ознаки або подає їх неповно. Редактор має уточнити систему ключових семантичних компонентів.

(5) *Редукція мовних надмірностей*. Моделі схильні до довших і більш розлогих формулювань. Визначення потрібно зробити стислішим без втрати змісту.

Крім цього ВЛМ додає до дефініції оцінні судження або універсальні твердження, які не відповідають лексикографічним нормам. Наприклад: *Мурахи – це неймовірно організовані комахи, які славляться своєю працелюбністю*. Таке формулювання містить оцінні характеристики, які не є частиною нейтральної дефініції. У цій ситуації редактор замінює оцінні елементи на нейтральні дескриптори, наприклад: *Мурахи – це соціальні комахи, що живуть колоніями з розподілом функцій між особинами*. ВЛМ може запропонувати стильний або образний опис, але лексикограф перетворює його на нормативне, структуроване й нейтральне визначення, узгоджене з жанром словника.

Варто зазначити, що ВЛМ можуть подавати хибні дефініції до слів, особливо до специфічних. Розглянемо як приклад визначення до слова *оскопсування*, що запропонувала модель Gemini та порівняємо її з реальною дефініцією:

Gemini

Термін, похідний від назви соціальної мережі «ОК» ... позначає пасивний моніторинг цифрової активності певної особи...

Реальна дефініція зі словника

інновацій (А. Нелюба, Є. Редько)

Опредметнена дія (процес), пов'язана зі Скопусом (наукометрична база).

Цей приклад показує, що замість Scopus вона «винайшла» ОК, O-scoring і дієслово *копати*. Звучить гладко, але повністю відірване від реальності. Типовою галюцинацією є вигадана етимологія. Іншими словами, ВЛМ дуже часто намагаються реконструювати етимологію методом «здогадування» → це майже завжди помилково для неологізмів. Наступний момент – підміна предметної галузі, тобто модель тягне термін у сферу цифрової поведінки («моніторинг активності»), хоча реальний контекст – академічні наукометричні платформи. Вигаданий текст стилістично гладенький і переконливий, що викликає ілюзію правдоподібності. Саме тому потрібна людська верифікація.

Такий матеріал чудово демонструє: генерування дефініції ШІ – це лише чорнова стадія, а остаточна валідація завжди потребує лексикографа. Причиною

генерації хибних визначень є відсутність або недостатня репрезентація терміна у тренувальних даних. ВЛМ тоді не «дістає» інформацію, а конструює її за схожими патернами. Для уникнення хибних дефініцій слід:

- надавати моделі *контекст* або *вихідні дані* перед генерацією;
- вимагати посилань на джерела або пояснення;
- перевіряти всі вузькоспеціальні дефініції вручну;
- використовувати ШІ не для значень «з нуля», а для *редагування* та *уніфікації* вже відомих даних.

Отримавши контекст із критикою «абсолютизації наукометричних показників» і метафорами про науковців, яких «поголовно прагнуть оскопусити», модель згенерувала інтерпретацію:

Gemini

Реальна дефініція зі словника інновацій (А. Нелюба, Є. Редько)

Оскопування – явище в академічній сфері, коли від науковців вимагають масово публікуватися у престижних міжнародних базах на кшталт *Scopus*, зосереджуючись на формальних індексах і створюючи надмірний тиск на викладачів.

Опредметнена дія (процес), пов'язана зі Скопусом (наукометрична база).

Як видно, модель ігнорує словотвірну мотивацію («від *Scopus*» → *оскопування*), дописує соціологічний та етичний компонент, якого немає у словнику, а також створює енциклопедичну довідку, а не дефініцію як жанр, драматизуючи контекст та підсилюючи те, що бачить у підказці. Виходом з цієї ситуації може бути:

(1) Надання моделі словотвірну мотивацію, наприклад: *Термін утворено від **Scopus** + суфікса -ування. Створіть дефініцію, зберігаючи вихідну мотивацію.*

(2) Просити саме дефініцію, а не опис явища: *Дай коротке словникове визначення, без оцінок, без прикладів і без публіцистики.*

(3) Якщо дати моделі приклад дефініції з цього ж словника, вона краще відтворить стиль і структуру.

(4) Не залишати емоційних або дискурсивних натяків.

Виходячи з вищесказаного, при генерації дефініцій ШІ поводитьсь не як «фіксатор значення», а як інтерпретатор світу. Тому лексикографові потрібно керувати ним через чіткі інструкції, жорсткі рамки жанру та надану мотивацію. Інакше модель продукує не словникову інформацію, а публіцистичні або аналітичні тексти.

Однією з найбільш трудомістких частин підготовки словникового матеріалу є **структурування та маркування даних**. У традиційній лексикографії це означає ручне визначення частини мови, граматичних характеристик, значення слова в конкретному контексті, його синтаксичної та семантичної ролі. Інструменти штучного інтелекту значно полегшують ці кроки, адже здатні працювати з великими масивами даних, швидко «прочитувати» морфологічну та семантичну структуру тексту й пропонувати попередню автоматичну розмітку.

Автоматичне розпізнавання частин мови та морфологічних характеристик або **POS-tagging** (part-of-speech tagging) – базова технологія, яка дає змогу автоматично визначити, до якої частини мови належить слово, та встановити його граматичні категорії: рід, число, відмінок, час, вид тощо (залежно від мови). У контексті роботи над електронними словниками сучасні ШІ-системи:

- визначають частину мови в *різних* контекстах (наприклад, *право* як іменник vs *право* як прислівник у фразі *право стояти*);
- пропонують розгорнуту морфологічну характеристику для кожної словоформи;
- пояснюють, чому слово визначено саме так, що корисно для навчання й контролю якості;
- швидко опрацьовують великі корпуси, де вручну розмітка тривала б місяці.

Для лексикографа це означає значне скорочення ручної роботи на початкових етапах. Він має можливість зосередитися не на механічному позначенні форм, а на інтерпретації даних. Алгоритм розмічає текст більш послідовно, ніж людина, яка може втомлюватися або змінювати критерії під час роботи, що сприяє підвищенню однорідності опрацювання. У сучасних ВЛМ (ChatGPT, Gemini,

Claude) морфологічний аналіз уже інтегрований у моделі, тому вони можуть працювати без зовнішніх тегерів, хоча для проєктів високої точності часто застосовують гібридні системи: ВЛМ як інтерпретатора + спеціалізований морфологічний розділ.

Зняття багатозначності (Word Sense Disambiguation – WSD). Багатозначність – це один із ключових викликів лексикографії. Для кожного багатозначного слова необхідно визначити, які саме значення реалізовані в корпусних прикладах, щоб потім коректно упорядкувати словникову статтю. Зняття багатозначності – це автоматичне визначення того, яке значення слова активується в конкретному контексті. Для прикладу:

- *коса* може бути зачіскою («дівчина заплела косу»),
- знаряддям праці («зібрати сіно косою»),
- географічним утворенням («Арабатська коса»),
- навіть частиною імен (*Коса-Дзур*), залежно від системи.

У традиційній лінгвістиці WSD – складна й ресурсомістка задача. Для ШІ вона значно легша, оскільки сучасні моделі оперують контекстом у межах кількох абзаців або навіть сторінок, вміють зіставляти слово з найімовірнішими сценаріями вживання та навчаються на величезних корпусах, де зустрічаються практично всі типові значення. У межах електронного словникарства WSD корисний для:

- автоматичного групування контекстів за значеннями;
- підбору ілюстративних прикладів для кожного значення;
- попереднього структурування статті;
- уникання «суміші значень» у дефініціях.

Нижче подано демонстрацію того, як модель штучного інтелекту виконує розпізнавання значення багатозначного терміна *steam gauge* у двох різних контекстах.

Контекст 1:

Every steam boiler shall have a water-gauge glass and a pressure gauge with a siphon.

Контекст 2:

The 757 is a transitional aircraft in another sense – it was designed in the era after the "steam gauges" of the 727 and early model 747s.

Результат роботи ВЛМ Gemini: надані контексти демонструють полісемію терміна *steam gauge*, а також здатність ШІ правильно інтерпретувати значення залежно від галузі:

1. *Технічне пряме значення (котельне обладнання)*. У першому контексті термін позначає **реальний вимірювальний прилад**, який відображає тиск пари у паровому котлі. Це класичне, буквальне інженерне значення.

2. *Метафоричне авіаційне значення (кабіна літака)*. У другому прикладі *steam gauges* вживається метафорично – не для опису парових приладів, а для позначення старих аналогових стрілочних індикаторів у кабіні літака. Цей вислів підкреслює технологічний перехід від аналогових панелей до сучасних цифрових систем (*glass cockpit*).

Як можна побачити, ШІ коректно виділяє ці значення, оскільки враховує не тільки форму слова, а й галузь, тематику тексту, а також типові терміни, що зустрічаються поруч. Такий аналіз важливий у словникарстві, оскільки дозволяє моделі відрізняти прямі й переносні значення, працювати з галузевими термінами, розуміти метафоризацію у професійних жаргонах (авіація, морська термінологія, ІТ тощо) та формувати коректні словникові статті у разі полісемії.

Попри високу ефективність сучасних моделей, проблема розв'язання полісемічності технічних термінів залишається зоною ризику для ШІ. Найпоширеніша помилка – надмірне покладання на загальномовну статистику, через що модель інколи «притягує» поширеніше значення слова, навіть якщо контекст підказує інше. ШІ може інтерпретувати технічний термін як більш розповсюджене побутове значення, якщо:

- текст містить лише часткові маркери галузі,
- термін використано в метафоричному або спеціалізованому значенні, відмінному від загальномовного,
- сусідні слова не належать до типової технічної лексики,
- у корпусах, на яких тренувалась модель, переважає інше значення.

Отже, ШІ чудово справляється з попереднім сортуванням значень, але остання інтерпретація завжди потребує людського контролю, особливо в технічних,

наукових і галузевих словниках, де помилка може коштувати значно дорожче, ніж у загальномовній лексикографії.

Одним із ключових здобутків сучасних ІІІ-технологій є те, що словник перестає бути «статичним довідником» і дедалі більше перетворюється на інтерактивне середовище, яке підлаштовується під потреби конкретної людини. У цьому напрямі ІІІ виконує дві стратегічно важливі функції: **персоналізацію досвіду користувача** та **розширення можливостей пошуку**. Завдяки аналізу мовного рівня, типу завдань і цифрової поведінки користувача, ІІІ може формувати «адаптивні словникові статті». Це означає, що один і той самий термін може виглядати по-різному:

- для школяра: спрощене визначення, без професіоналізмів, з коротким прикладом;
- для студента: повніша дефініція, базові класифікаційні ознаки, типові контекстні вживання;
- для фахівця: повна структурування стаття з термінотворчими моделями, стилістичними уточненнями, порівняннями, міжмовними відповідниками.

Це дає змогу уникати перевантаження надлишковою інформацією, робити словник доступним для різних аудиторій без дублювання контенту та покращувати навчальну цінність ресурсів. Фактично, системи на базі ІІІ виконують роль «розумного фільтра», який формує подання матеріалу відповідно до компетентності користувача.

Ще одна революційна можливість – перехід від пошуку «за словом» до пошуку «за значенням». Традиційні словники шукають точні збіги: ввів слово – отримав статтю. ІІІ-системи використовують **embedding-моделі** – *числові векторні представлення слів, побудовані на основі контекстів, у яких вони з'являються*. Такі моделі відкривають низку нових функцій:

- Пошук за смисловою близькістю: якщо користувач вводить слово *теплий*, система пропонує *помірний, м'який, лагідний* – слова, які близькі за змістом, але не є прямими синонімами.

– Тематика замість формальної межі: запит *forest industry* може повернути *logging, timber harvesting, sawmills* — хоча ці слова формально не є синонімами.

– Пошук у стилістичних і галузевих зонах: наприклад, знаходження лише технічних або лише розмовних відповідників.

– Виявлення «забутих» або маловживаних відповідників: модель повертає близькі за значенням слова, які користувач міг навіть не знати.

Embedding-пошук фактично перетворює словник на семантичний навігатор, який дозволяє рухатись смисловими полями, а не лише буквальними формами. Підтримка користувача через персоналізацію та семантичний пошук змінює саму філософію словника, тобто він стає адаптивним, контекстно-чутливим і здатним працювати як інтелектуальна система супроводу мовної діяльності.

Сучасні онлайн-словники дедалі частіше поєднують лексикографічні бази з моделями штучного інтелекту. Один із показових прикладів – адаптивний пошук і контекстне пояснення, які реалізовані у системах Reverso Context, DeepL Dictionary та Cambridge Search. Розглянемо умовний приклад, коли користувач вводить слово *pitch* без уточнень. Якщо раніше користувач шукав авіаційні або технічні терміни, система першими покаже відповідники *pitch angle, blade pitch, control pitch* тощо. У цьому випадку система враховує профіль користувача. Якщо користувач не вводить конкретного контексту, система аналізує слова, які він шукав до цього, й формує «ймовірний семантичний сценарій». Наприклад, запит *pitch* після пошуків *airfoil, ailerons, altitude* буде трактований як авіаційний термін. Іншими словами, відбувається **векторний пошук** за сенсом. Такий підхід робить словник не просто «місцем, де зберігається інформація», а інструментом, який активно допомагає користувачеві орієнтуватись у семантичному просторі. ШІ дозволяє формувати потрібне значення автоматично, зменшуючи кількість хибних інтерпретацій і прискорюючи пошук релевантної інформації.

Попри значний потенціал штучного інтелекту в автоматизації лексикографічних процесів, важливо усвідомлювати його принципові обмеження. Алгоритми здатні швидко генерувати визначення, приклади вживання й навіть моделі словникових статей, однак ці результати не завжди ґрунтуються на

перевірених джерелах. Моделі можуть створювати помилкові, неповні або штучно сконструйовані дефініції, які не відповідають фактичному слововживанню. Тому обов'язковою умовою залишається людська експертна верифікація – лексикограф має перевіряти кожен елемент, зіставляти його з корпусними даними та нормами мовної системи.

Крім того, дані, на основі яких навчаються моделі, часто містять культурні, соціальні та ідеологічні упередження. Це може призводити до некоректних семантичних акцентів, нерівномірного представлення лексики або навіть непомітного відтворення дискримінаційних конотацій у словникових статтях. Таким чином, робота з ШІ вимагає критичного підходу та розуміння того, як саме формується результат.

Окремим викликом є ризик професійної деградації: повна автоматизація процесів здатна знецінити роль лексикографа, якщо фахівці почнуть надмірно покладатися на автоматичні рішення. У результаті можуть втрачатися навички глибинного аналізу, інтерпретації та стилістичного відчуття мови – ті компетенції, які принципово не відтворюються машинними моделями.

Отже, ШІ варто розглядати насамперед як інструмент, а не повну заміну професійної лексикографії. Ефективність його застосування залежить від балансу між автоматизацією та людською експертизою, між швидкістю обробки даних і якісною мовною оцінкою. Тільки за такої умови технології можуть суттєво розширити можливості словникарства, не нівелюючи його наукових стандартів.

Питання для самоконтролю

1. Які основні ризики пов'язані з автоматичною генерацією визначень штучним інтелектом?
2. Чому навіть високоточні моделі ШІ потребують людської верифікації в лексикографії?
3. Які типи упереджень можуть переноситися у словникові статті разом із даними, на яких тренується модель?

4. У чому полягає небезпека надмірного покладання на автоматичні системи під час укладання словників?

5. Чим штучний інтелект як інструмент у лексикографії принципово відрізняється від повної автоматизації процесів?

6. Чому збереження фахових компетентностей лексикографа є критично важливим навіть у разі широкого використання ШІ?

7. Які аспекти словникової статті (структурні, семантичні, стилістичні) найуразливіші до помилок моделі?

Завдання для самоперевірки

1. Візьміть будь-яке автоматично згенероване визначення (можете вибрати зі своїх попередніх матеріалів або придумати слово). Проаналізуйте його за критеріями:

- точність змісту;
- відповідність корпусному слововживанню;
- присутність чи відсутність стилістичних зрушень;
- можливі “галюцинації” (додані неіснуючі факти).

Напишіть короткий висновок: де модель «помилилася» або що зробила правильно.

2. Створіть (або попросіть модель створити) визначення слова, пов’язаного з соціальною роллю, етнічною групою, професією, віковою категорією. Проаналізуйте, чи є в дефініції оцінні судження, культурно специфічні конотації чи некоректні узагальнення. Поясніть, як би ви нейтралізували визначення.

3. Візьміть спрощене, помилкове або “сирувате” визначення, автоматично створене ШІ, і відредагуйте його, щоб: уточнити значення, усунути двозначності, додати реальні приклади вживання та навести інтерпретацію у відповідність до норм лексикографічного опису. Опишіть, чому ці зміни були необхідними.

4. Обираємо будь-яке слово (краще полісемічне). Попросіть модель створити: (1) лаконічну дефініцію, (2) розширену дефініцію, (3) приклади вживання. Потім створіть власний опис і порівняйте:

- де модель надто узагальнила;

- де «перебрала» контекстів;
- що пропустила або додала зайвого.

5. Напишіть коротке міні-есе (5–7 речень) про те, де, на вашу думку, людський контроль є безумовно обов'язковим. Наведіть приклади типових помилок ШІ, які не може «впіймати» автоматична система.

9.2. Огляд ШІ інструментів для роботи з мовним матеріалом

На сучасному етапі розвитку електронного словникарства сформувався широкий спектр інструментів штучного інтелекту, які можуть бути безпосередньо залучені до лексикографічного виробництва. Ці інструменти відрізняються за своїм призначенням, архітектурою та рівнем спеціалізації, однак усі вони спрямовані на автоматизацію, прискорення та інтелектуалізацію роботи з мовним матеріалом. Умовно такі інструменти можна поділити на кілька груп, серед яких провідну роль відіграють **великі мовні моделі (ВЛМ)**. До найбільш поширених та універсальних інструментів цього класу належать **ChatGPT (OpenAI)**, **Claude (Anthropic)**, **Gemini (Google)**, **LLaMA (Meta)** та інші сучасні великі мовні моделі. Вони навчаються на масивних корпусах текстів різних жанрів і мов, що дає їм можливість оперувати складними лінгвістичними структурами, семантичними зв'язками та контекстами. У лексикографічній практиці ВЛМ можуть виконувати низку важливих функцій.

По-перше, мовні моделі здатні здійснювати узагальнення великих текстових масивів. Це дозволяє швидко отримувати стислий опис тематичних корпусів, окремих галузей знань або семантичних полів, що є корисним на етапі підготовки реєстру словника чи аналізу предметної області. По-друге, ВЛМ широко застосовуються для генерації дефініцій. Моделі можуть створювати чернеткові визначення слів і термінів, адаптуючи їх до різних типів користувачів (школярі, студенти, фахівці), різних стилів (науковий, популярний, навчальний) та різного рівня деталізації. У цьому контексті ШІ виступає як інструмент підтримки лексикографа, а не як повноцінний замітник авторської роботи. По-третє, мовні моделі здатні до розпізнавання та класифікації значень слова на основі контексту. Це дає змогу використовувати їх у завданнях попереднього розмежування

полісемії, підготовки багатозначних словникових статей та аналізу семантичної структури лексеми. По-четверте, ВЛМ ефективні у створенні перекладних відповідників, зокрема в багатомовних або перекладних словниках. Моделі можуть пропонувати не лише прямі еквіваленти, а й контекстно зумовлені варіанти перекладу, що є важливим для термінологічних та спеціалізованих словників. Нарешті, великі мовні моделі здатні виконувати аналіз семантичних відтінків і стилістичних особливостей лексичних одиниць, виявляти оцінні компоненти, прагматичні значення, конотації та асоціативні зв'язки. Це відкриває нові можливості для опису лексики, яка раніше важко піддавалася формалізованому словниковому представленню (сленг, розмовна мова, інтернет-дискурс).

Водночас слід наголосити, що мовні моделі не мають власної мовної інтуїції та не «розуміють» значення в лінгвістичному сенсі. Їхні результати базуються на статистичних закономірностях, що зумовлює необхідність експертного контролю, критичної оцінки та редакторського опрацювання з боку людини-лексикографа. Варто зазначити, що великі мовні моделі та традиційні інструменти обробки природної мови, попри спільну мету – автоматизоване опрацювання мовного матеріалу – ґрунтуються на різних теоретичних і технологічних підходах та виконують різні функції у лексикографічному процесі. Для чіткого розуміння їхніх можливостей, обмежень і ролі в електронному словникарстві доцільно зіставити ці підходи за ключовими параметрами. Нижче у таблиці 9.1 наведено порівняльну характеристику ВЛМ та NLP-інструментів.

Таблиця 9.1 – Порівняльна характеристику ВЛМ та NLP-інструментів

Критерій	Великі мовні моделі (LLM)	Традиційні NLP-інструменти
Архітектура	Неймережеві трансформерні моделі, навчені на великих корпусах різножанрових текстів	Роздільні або статистичні системи, часто побудовані на правилах і/або класичних ML-алгоритмах
Контекстуальність	Глибоке врахування широкого контексту (речення, абзац, дискурс)	Обмежений контекст (зазвичай речення або фіксоване вікно слів)
Робота з полісемією	Контекстно-залежне розрізнення значень без попередньої розмітки	Потребує спеціально навчених моделей WSD або словникових правил

Генерація тексту	Повноцінна генерація дефініцій, прикладів, пояснень	Генерація відсутня або сильно обмежена
Структурування даних	Може створювати напівструктуровані дані, потребує подальшої формалізації	Чітко структуровані вихідні дані (теги, леми, залежності)
Точність формальних характеристик	Можливі помилки в морфології, тегах, термінах	Висока точність у вузьких завданнях (POS-tagging, лематизація)
Гнучкість і універсальність	Висока: одна модель – багато завдань	Низька або середня: один інструмент – одне завдання
Пояснюваність результатів	Низька (ефект «чорної скриньки»)	Вища: зрозумілі правила та алгоритми
Типові інструменти	ChatGPT, Claude, Gemini, LLaMA	Stanza, UDPipe, spaCy, Apertium
Роль у лексикографії	Інструмент підтримки творчої та аналітичної роботи	Інструмент точного автоматичного аналізу

Отже, великі мовні моделі та традиційні NLP-інструменти не слід розглядати як взаємовиключні підходи. Навпаки, у сучасному електронному словникарстві вони виконують комплементарні функції. ВЛМ забезпечують гнучкість, генеративність і семантичну інтерпретацію, тоді як традиційні ВЛМ-системи гарантують формальну точність, відтворюваність результатів і надійну структуровану розмітку мовних даних. Найефективніші лексикографічні системи поєднують обидва підходи в межах єдиного технологічного процесу.

Розглянемо формулювання дефініцій різними великими лінгвістичними моделями на прикладі слова **оскопусування** відповідно до такого контексту: *«Абсолютизація наукометричних показників і прагнення всіх поголовно “оскопусити”, вимоги від усіх демонструвати свою довжину Гіриша і бути присутніми в наукометричних базах...»*. У результаті моделі пропонують визначення, які істотно різняться за стилем, рівнем узагальнення та глибиною концептуалізації явища, позначуваного аналізованим словом (див. додаток Д.23). Це наочно демонструє як потенціал ШІ у формуванні дефініцій, так і ризики їх некритичного використання в лексикографічній практиці. У таблиці 9.2 подано порівняльну характеристику створених дефініцій.

Таблиця 9.2 – Порівняння дефініцій, згенерованих ВЛМ

Модель	Стиль і тип дефініції	Потенційні ризики для словника
Gemini	Публіцистично-описова, узагальнена, орієнтована на зрозумілість	Спрощення явища, втрата термінологічної чіткості

Claude Sonnet	Процесуальна, інституційно зорієнтована, близька до наукового стилю	Звуження значення до адміністративного аспекту
Llama 4	Абстрактно-метрична, нейтральна, без оцінної експлікації	Розмитість семантики, недостатня експресивність неологізму
DeepSeek	Розгорнута аналітична дефініція з чіткою негативною оцінкою	Надмірна деталізація, енциклопедичність замість словникової лаконічності
GLM	Концептуально-оціночна, високий рівень узагальнення	Втрата конкретних диференційних ознак терміна
ChatGPT	Збалансована науково-описова дефініція з соціальною рамкою	Потреба у скороченні для відповідності словниковому формату

Поданий приклад показує, що великі лінгвістичні моделі здатні адекватно вловлювати семантичне ядро неологізму, проте результати їхньої роботи коливаються між публіцистичним описом, енциклопедичним тлумаченням і власне словниковою дефініцією. Це зумовлює необхідність людського редагування, нормалізації стилю та виокремлення диференційних ознак, без яких дефініція не може вважатися повноцінною лексикографічною одиницею. Побудова «ідеальної» дефініції на основі чорнених варіантів здійснюється у чотири кроки.

Крок 1 – виокремлення спільного семантичного ядра. Так, усі подані варіанти (незалежно від стилю) збігаються в таких ознаках:

- *сфера*: академічне / наукове середовище;
- *джерело тиску*: інституційні, адміністративні, оціночні механізми;
- *ключовий маркер*: наукометричні бази (передусім Scopus);
- *механізм*: абсолютизація формальних кількісних показників;
- *наслідок*: підміна змістовної якості формальною відповідністю.

Це і є інваріант значення, який має бути збережений у словниковій дефініції.

Крок 2 – усунення надлишкового (що не входить у дефініцію). З погляду нормативної лексикографії виключаємо:

- розгорнуті переліки (індекс Гірша, квартилі тощо);
- публіцистичні формули типу «втрата індивідуальності»;
- метафоричні або оціночні пасажі, не формалізовані в дефініційній схемі;
- причинно-наслідкові ланцюги, притаманні радше аналітичному тексту.

Вилучаючи перелічені вище елементи, лексикограф отримує чітку і лаконічну родо-видову структуру.

Крок 3 – вибір дефініційної моделі. Для неологізму типу *оскопсування* оптимальною є модель: «процес / практика + сфера + ключова ознака + наслідок». Вона, по-перше, відповідає академічним словникам, і, по-друге, дає зафіксувати негативну конотацію без публіцистики, і, по-третє, легко масштабується на подібні неологізми (*індексофілія, метрикоцентризм* тощо).

Крок 4. Формування нормативної словникової дефініції. Базовий нормативний варіант дефініції для тлумачного словника може мати такий вигляд: **Оскопсування** – *практика в академічному середовищі, що полягає в абсолютизації вимог до присутності наукових публікацій у наукометричних базах (передусім Scopus), унаслідок чого формальні кількісні показники домінують над змістовною якістю наукових досліджень.*

Отже, «ідеальна» словникова дефініція не є середнім арифметичним між варіантами, згенерованими моделями. Вона є результатом відбору інваріантних семантичних ознак, усунення стилістичних надлишків та приведення формулювання до нормативної дефініційної схеми. У цьому процесі ВЛМ виступають джерелом семантичного матеріалу, але не остаточним суб'єктом лексикографічного рішення. Інші варіанти зредагованих дефініцій подано в додатку Д.24, які можна використовувати для інших типів словників.

Крім побудови дефініцій, великі лінгвістичні моделі можуть виконувати контекстуальний семантичний аналіз, тобто визначати, яке саме поняття позначає слово в конкретному вживанні. Це особливо важливо для полісемічних одиниць, значення яких змінюється залежно від синтаксичного та предметного контексту. У наведених нижче прикладах англійське слово *design* функціонує в різних значеннях, пов'язаних з авіабудуванням, але не зводиться до одного семантичного типу:

*The Boeing 7X7 prototype was **designed** as an all-new twin engine wide body aircraft.*

*The 7N7 prototype was a twin-engine variant of the 727 **designed for** superior range and economics.*

*The 7X7, of course, became the famous Boeing 767, while the 7N7 designed was ultimately scrapped in favor of an all-new narrow body **design**.*

*Since the 757 and 767 were **designed** at the same time, both aircraft were fitted with identical cockpits.*

Результат роботи моделей подано в таблиці 9.3, у якій наведено особливості стилів інтерпретацій, типи виявлених значень, а також потенційні ризики для словників.

Таблиця 9.3 – Результат роботи ВЛМ з семантичного аналізу

Модель	Домінантний стиль інтерпретації	Типи значень <i>design</i> , що розрізняються	Потенційний ризик
Gemini	Прикладно-перекладний	конструкція проєкт процес	обмежена формалізація семантики
Claude	Лексико-семантичний	результат призначення процес іменникове значення	потребує термінологічного узгодження
Llama	Пояснювально-описовий	задум ціль конструкція результат	розмиті межі значень
DeepSeek 3.1	Структурно-полісемічний	концепція функція тип конструкції процес	надлишкова деталізація
GLM 4.5	Концептуально-аналітичний	ідентичність функція архітектура процес	абстрактність
ChatGPT	Системно-лексикографічний	концепція призначення результат тип процес	потребує редакторського контролю

Поданий аналіз демонструє, що великі лінгвістичні моделі здатні виявляти полісемічну структуру слова та групувати контекстні вживання за типами значень, однак спосіб такого групування суттєво залежить від моделі. Для лексикографа це означає, що результати ВЛМ не можуть бути безпосередньо перенесені до словника, але можуть слугувати інструментом попереднього семантичного зондування, після якого необхідна фахова нормалізація значень і встановлення їх ієрархії. Розглянемо процес побудови словникової статті для аналізованого терміна, ґрунтуючись на поданих вище інтерпретаціях:

design [dɪˈzaɪn], *noun, verb*

I. *noun* (іменник)

1. Концепція, загальний задум або архітектурний принцип об'єкта. Походить з інтерпретацій Gemini, GLM, ChatGPT (concept / identity / architecture).

Сукупність основних ідей та рішень, що визначають структуру й ідентичність виробу, системи або проекту.

An all-new narrow body design.

2. Проект, технічний план або завершене конструкторське рішення. Зафіксовано всіма моделями як іменникове значення “result / project”.

Документований або концептуально оформлений результат інженерного проектування, який може бути реалізований або скасований.

The 7N7 design was ultimately scrapped.

3. Тип конструкції або конструктивне компонування. Узагальнення DeepSeek, GLM, Claude.

Категоріальний різновид технічного виконання об'єкта.

narrow-body design; wide-body design.

II. verb (дієслово)

4. Проектувати, розробляти як процес створення об'єкта. Стабільне значення у всіх ВЛМ: process.

Здійснювати інженерне, технічне або концептуальне проектування.

The aircrafts were designed at the same time.

5. Проректувати з певною метою; Виділено Claude, Llama, DeepSeek створювати для виконання визначеної як окреме цільове значення. функції.

Розробляти об'єкт з орієнтацією на задані експлуатаційні або функціональні характеристики.

Designed for superior range and economics.

6. Задумувати, визначати Поєднання значень “concept / початкову концепцію об'єкта. intent”, узгоджене з GLM і ChatGPT.

Формувати ідею або первинний задум майбутнього виробу чи системи.

The prototype was designed as an all-new aircraft.

Наведена словникова стаття демонструє, що результати аналізу ВЛМ дозволяють виявити повний спектр контекстних значень, відокремити процес, результат і концепцію та побудувати ієрархію значень, необхідну для нормативного словника. Водночас остаточне формулювання дефініцій, їх групування та послідовність подання є результатом фахового лексикографічного рішення, а не автоматичної генерації. Розглянутий кейс зі словом *design* дає змогу наочно порівняти два підходи до встановлення значення та багатозначності лексичної одиниці: традиційний корпусний аналіз і аналіз за допомогою великих мовних моделей (ВЛМ).

Корпусний аналіз ґрунтується на опрацюванні великої кількості реальних текстів з подальшим: виявленням типових колокацій (*design of, design for, new design*), частотним аналізом уживань, групуванням контекстів за семантичними ознаками. На основі корпусних даних лексикограф зазвичай виокремлює:

- *design* як процес проєктування;
- *design* як результат / проєкт;
- *design* як тип конструкції або архітектурне рішення;

– рідше *design* як задум або намір.

Проте корпусний аналіз має як переваги, так і обмеження, які потрібно враховувати під час роботи з мовним матеріалом:

Переваги:

- емпірична надійність;
- відтворюваність результатів;
- можливість кількісної верифікації.

Обмеження:

- значні часові витрати;
- потреба у наявності великого та репрезентативного корпусу;
- слабша чутливість до швидких семантичних зсувів;
- складність роботи у випадку з вузькоспеціалізованими технічними контекстами без спеціальних корпусів.

На відміну від корпусних інструментів ВЛМ працюють за іншою схемою, тобто вони здійснюють контекстну інтерпретацію значення в межах конкретного фрагмента, узагальнюють типи значень без попереднього частотного підрахунку та утворюють концептуальні групи значень (процес → результат → концепція → функція). Усі проаналізовані ВЛМ незалежно одна від одної:

- чітко розрізнили дієслівні й іменникові вживання;
- виокремили процесуальне, результативне та цільове значення;
- інтерпретували *design* як концепцію / ідентичність об'єкта, що не завжди одразу фіксується корпусною статистикою.

Так само, як і корпусні системи, ВЛМ мають переваги та обмеження, що мають вплив на якість опрацювання мовного матеріалу:

Переваги:

- швидкість аналізу;
- висока чутливість до контексту;
- здатність працювати з малими фрагментами тексту;
- зручність для здійснення первинного семантичного картування.

Обмеження:

- відсутність кількісної перевірки;
- можливе «розмивання» меж між значеннями;
- вплив стилю конкретної моделі на формулювання дефініцій;
- потреба редакторського контролю.

Нижче у таблиці 9.4. подано порівняльну характеристику корпусного та ВЛМ-аналізу на основі критеріїв, що впливають на ефективність і якість опрацювання даних.

Таблиця 9.4 – Порівняльна характеристика двох типів аналізу

Критерій	Корпусний аналіз	ВЛМ-аналіз
Джерело даних	Масив реальних текстів	Узагальнені мовні патерни
Тип аналізу	Частотний, емпіричний	Контекстно-концептуальний
Швидкість	Низька	Висока
Виявлення полісемії	Поступове, ручне	Швидке, індуктивне
Робота з технічним контекстом	Потребує спеціалізованого корпусу	Висока адаптивність
Ризик суб'єктивності	Низький	Середній
Роль лексикографа	Основна	Ключова (редактор, верифікатор)

На прикладі слова *design* видно, що корпусний і ВЛМ-підходи не є конкурентними, а взаємодоповнювальними:

- корпусний аналіз забезпечує надійність і верифікацію;
- ВЛМ прискорюють виявлення значень і побудову первинної семантичної структури;
- фінальна словникова стаття формується шляхом експертного поєднання обох підходів.

Отже, великі мовні моделі можуть розглядатися як інтелектуальний інструмент попереднього аналізу, тоді як нормативне словникарство залишається сферою відповідальності фахівця, добре обізнаного з предметною галуззю. Показовим є випадок роботи з вузькогалузевою технічною термінологією, коли ВЛМ здатні запропонувати інтерпретацію значення терміна, його узагальнене тлумачення та перекладний еквівалент, однак остаточна редакція, термінологічна верифікація й нормування залишаються за технічним редактором або галузевим спеціалістом. У поданому нижче фрагменті технічного тексту ВЛМ мають проаналізувати зміст термінів (тобто визначити, які саме поняття вони позначають) *main inlet valve*, *turbine inlet pipe*, *power house*, *upstream pipe*, *dismantling joint* та подати їхні перекладні еквіваленти українською мовою:

The Main Inlet Valve (spherical valve) will be installed upstream of the turbine inlet pipe in the power house. The valve body will be bolted to the upstream pipe which

will be welded at lower end of the penstock. Downstream side will be connected to the turbine inlet pipe by a dismantling joint.

Нижче у таблиці 9.5 подано результат роботи різних моделей з аналізу вище перелічених термінів.

Таблиця 9.5 – Результати інтерпретації термінів різними ВЛМ

Модель	main inlet valve	turbine inlet pipe	power house	upstream pipe	dismantling joint
Gemini	головний впускний затвор / головний запірний клапан	вхідний патрубок турбіни	машинний зал / будівля ГЕС	трубопровід верхнього б'єфа	демонтажний (компенсаційний) вузол
Claude	головний вхідний затвор / кульовий затвор	підвідна труба турбіни / вхідний трубопровід турбіни	машинний зал / будівля ГЕС	трубопровід вище за течією	демонтажна вставка / демонтажний компенсатор
Llama	головний впускний клапан (кулястий)	вхідний патрубок турбіни	машинний зал (генераторна зала)	верхня ділянка трубопроводу	розбірне з'єднання
DeepSeek 3.1	головний вхідний клапан (кулястий)	впускний патрубок турбіни	машинний зал (генераторний зал)	верхня ділянка трубопроводу	розбірне з'єднання
GLM 4.5	головний затвор / головний клапан	трубопровід входу турбіни / вхідний патрубок турбіни	машинна зала	верховий трубопровід	демонтажний / розбірний з'єднувач

Аналіз результатів, поданих у таблиці 9.5, засвідчує, що великі мовні моделі загалом коректно ідентифікують предметну віднесеність термінів і здатні відтворити їх базове поняттєве ядро. Водночас між моделями спостерігаються розбіжності у виборі перекладних еквівалентів, ступені термінологічної точності та рівні спеціалізації. Частина моделей (Gemini, Claude) тяжіє до пояснювально-перекладного підходу, пропонуючи кілька варіантів відповідників, що є корисним на етапі первинного аналізу, але потребує подальшого нормування. Інші (Llama, DeepSeek) демонструють схильність до узагальнення, інколи спрощуючи галузеву специфіку. GLM 4.5, своєю чергою, пропонує

концептуально коректні, але інколи стилістично неоднорідні варіанти. Усі наведені приклади підтверджують, що ВЛМ ефективні як інструмент попередньої семантичної та перекладної розвідки, однак остаточне рішення щодо вибору нормативного терміна, його дефініції та словникової фіксації має ухвалювати фахівець з урахуванням галузевих стандартів і контексту вживання.

Таблиця 9.6 демонструє результат фінального етапу лексикографічної роботи – нормативного редагування термінів галузевим фахівцем. Порівняння узагальнених інтерпретацій, запропонованих великими мовними моделями, з фаховими відповідниками засвідчує, що ВЛМ здатні коректно ідентифікувати функціональне призначення технічних об'єктів і окреслити їх поняттєве ядро, однак не завжди відтворюють усталені нормативні назви, прийняті в галузі.

Таблиця 9.6 – Тлумачення ВЛМ та зредаговані відповідники

Термін	Узагальнене тлумачення ВЛМ	Фаховий (нормативний) відповідник
Main Inlet Valve	Головний вхідний / запірний клапан перед турбіною, призначений для перекриття потоку	Основний турбінний затвор (кульовий затвор)
Turbine inlet pipe	Трубопровід або патрубок, що подає воду до турбіни	Вхідний патрубок гідротурбіни
Power house	Будівля, у якій розміщене обладнання ГЕС (турбіни, генератори)	Будівля ГЕС / машинний зал
Upstream pipe	Трубопровід, розташований вище за течією, перед клапаном	Підвідний патрубок (частина напірного трубопроводу)
Dismantling joint	Розбірне з'єднання для демонтажу обладнання	Монтажний патрубок

Зокрема, моделі схильні використовувати описові або універсалізовані формулювання (*запірний клапан, розбірне з'єднання*), тоді як галузева практика закріплює точні терміни, пов'язані з конструктивними, монтажними або експлуатаційними особливостями об'єкта (*основний турбінний затвор, монтажний патрубок*). Це підтверджує, що ВЛМ доцільно застосовувати на етапі первинного аналізу, семантичного узагальнення та варіантного перекладу, тоді як остаточне термінологічне нормування має здійснюватися за участі фахівця та з урахуванням галузевих стандартів.

Проведений огляд прикладів використання великих мовних моделей у лексикографічному аналізі – від формування дефініцій і розрізнення полісемії

до добору перекладних відповідників у вузькогалузових текстах – засвідчує, що різні ВЛМ демонструють відмінні стилі інтерпретації мовного матеріалу та різний ступінь придатності для конкретних лексикографічних завдань. З огляду на це доцільність використання тієї чи іншої моделі визначається не лише її загальними можливостями, а й типом словника та етапом лексикографічного процесу. Узагальнені рекомендації щодо застосування ВЛМ у різних типах словників подано в таблиці 9.7.

Таблиця 9.7 – Типологія використання ВЛМ у словникарстві

Тип словника / завдання	Доцільні моделі	Пояснення
Навчальні словники (шкільні, для вивчення мови)	Gemini, Claude	Моделі добре адаптують визначення до широкої аудиторії, уникають надмірної термінологічності, забезпечують зрозумілу й нейтральну подачу матеріалу
Тлумачні загальномовні словники	Claude, GLM 4.5 (з редакторським контролем)	Claude забезпечує структурованість і нормативність, GLM – виразний семантичний компонент; обидві потребують верифікації
Словники неологізмів, сленгу, публіцистичної лексики	DeepSeek 3.1, GLM 4.5	Ці моделі добре фіксують оцінність, соціальний та ідеологічний контекст, що є критичним для опису нових мовних явищ
Термінологічні та галузеві словники	Gemini, Claude	Забезпечують лаконічні, функціонально точні дефініції, придатні для стандартизованого опису понять
Енциклопедичні словники та довідники	Gemini, Llama 4	Моделі схильні до узагальнення, добре працюють із контекстуалізацією понять у ширших знанєвих структурах
Перекладні словники та багатомовні глосарії	Claude, Gemini, Llama 4	Забезпечують зіставлення значень між мовами, генерацію перекладних відповідників і пояснювальні ремарки
Чернеткова лексикографічна розробка (drafting)	Усі моделі	ВЛМ ефективні на етапі первинного формування дефініцій, які надалі мають бути відредаговані людиною
Автоматичне збагачення словникових баз	Llama 4, DeepSeek 3.1	Придатні для масового опрацювання даних, але потребують фільтрації та нормалізації результатів

Виходячи з вищесказаного, великі мовні моделі не є універсальним інструментом для всіх типів словників. Їх ефективність визначається відповідністю між стилем моделі, типом словника та етапом лексикографічного процесу. Оптимальною є модель «ШІ як асистент лексикографа», де генерація поєднується з науковою експертизою, редакторським контролем і чіткою лексикографічною концепцією.

Питання для самоконтролю

1. У чому полягає принципова відмінність між великими мовними моделями (ВЛМ) та традиційними NLP-інструментами в лексикографічній практиці?
2. Які основні типи лексикографічних завдань можуть виконувати ВЛМ на етапі попереднього аналізу мовного матеріалу?
3. Чому використання ВЛМ у нормативному словникарстві потребує обов'язкового редакторського та експертного контролю?
4. Які стилістичні та методологічні відмінності було виявлено у дефініціях терміна *оскопсування*, згенерованих різними моделями?
5. Як ВЛМ здійснюють розрізнення значень полісемічних слів на прикладі слова *design*?
6. Які потенційні ризики виникають у разі прямого включення згенерованих ВЛМ дефініцій до словникової статті без доопрацювання?
7. У чому полягає різниця між корпусним аналізом і аналізом, здійсненим ВЛМ, у контексті формування словникової статті?
8. Яку роль відіграє галузевий фахівець у фінальному етапі укладання термінологічного словника за участю ВЛМ?
9. Чому різні типи словників потребують різних моделей або комбінацій ВЛМ?
10. Як концепція «ШІ як асистент лексикографа» змінює сучасну практику електронного словникарства?

Завдання для самоперевірки

1. Прочитайте поданий фрагмент наукового або публіцистичного тексту та:
 - Визначте одне полісемічне слово.
 - Запропонуйте щонайменше два можливі значення цього слова у різних контекстах.
 - Опишіть, які з цих значень може розрізнити ВЛМ, а які потребуватимуть втручання лексикографа.
2. Зіставте два підходи до аналізу значення терміна: (а) корпусний аналіз; (б) аналіз за допомогою великої мовної моделі. Оформіть відповідь у вигляді

короткої таблиці або структурованого переліку, зазначивши переваги та обмеження кожного підходу.

3. На основі кількох згенерованих ВЛМ визначень одного слова:

- Виокремте спільні семантичні компоненти.
- Усуньте надмірну оцінність або стилістичну неоднорідність.
- Сформулюйте власну нормативну словникову дефініцію.

4. Проаналізуйте поданий технічний або спеціалізований текст та:

- Визначте 3–5 термінів.
- Подайте узагальнене тлумачення, яке могла б запропонувати ВЛМ.
- Запропонуйте нормативний відповідник з позиції галузевого фахівця.

5. Сформулюйте коротке есе (5–7 речень) на тему: **«Чи може велика мовна модель замінити лексикографа?»** Обґрунтуйте свою позицію, спираючись на матеріал підрозділу.

9.3. Створення словників і глосаріїв засобами ChatGPT

ChatGPT є одним з інструментів, що найшвидше інтегрувався у практику сучасних лексикографів. Він дає змогу виконувати як дрібні операції, так і комплексні завдання, пов'язані з аналізом, інтерпретацією та первинним структуруванням мовного матеріалу. На відміну від спеціалізованих корпусних систем або традиційних лексикографічних програмних середовищ, ChatGPT функціонує як універсальний текстоорієнтований інтерфейс, здатний реагувати на гнучкі інструкції, уточнення та поетапні запити дослідника.

У контексті словникарства ChatGPT може розглядатися передусім як інструмент чернеткового (draft) етапу: генерації попередніх дефініцій, виявлення потенційної багатозначності, формування списків семантичних компонентів, пропозиції перекладних відповідників і варіантів термінологічного найменування. Його сильна сторона полягає у швидкому узагальненні великих масивів мовного досвіду та відтворенні типових моделей словникового опису, однак ці результати не можуть розглядатися як нормативні без подальшої фахової верифікації.

Важливо підкреслити, що ChatGPT не замінює класичних лексикографічних процедур – корпусного аналізу, експертного відбору джерел, термінологічної

нормалізації, – а радше доповнює їх, виконуючи роль інтелектуального асистента. У практиці укладання словників і глосаріїв це означає перерозподіл зусиль: рутинні аналітичні операції частково автоматизуються, тоді як відповідальність за концепцію словника, тип дефініції, стиль викладу та нормативний статус одиниці залишається за людиною.

У цьому підрозділі буде розглянуто можливості використання ChatGPT для створення різних типів словникових продуктів – від навчальних глосаріїв і перекладних списків термінів до чернеткових тлумачних і термінологічних словників. Особливу увагу буде приділено формулюванню запитів (prompting), типології лексикографічних завдань, які доцільно делегувати моделі, а також обмеженням і типовим помилкам, що виникають у разі некритичного використання результатів генерації.

У процесі створення словників і глосаріїв ChatGPT доцільно використовувати не як автономний «словникогенератор», а як інструмент виконання окремих лексикографічних операцій. Йдеться передусім про ті етапи роботи, які мають аналітичний або допоміжний характер і не потребують остаточного нормативного рішення. Нижче подано основні типи таких операцій.

Первинна інтерпретація значення слова у контексті. ChatGPT ефективно визначає, яке саме поняття актуалізується в конкретному текстовому фрагменті: розрізняє предметні, процесуальні, результативні та абстрактні значення, а також виявляє можливі оцінні або метафоричні нашарування. Лексикографічна цінність цього етапу полягає у швидкому попередньому встановленні значення слова та підготовці матеріалу до подальшого розмежування полісемії.

Водночас без достатнього галузевого контексту модель може не завжди чітко відрізняти близькі термінологічні значення. Наведемо приклад, у якому системі подано мінімальний контекст *Gear down!* і поставлено завдання визначити значення слова *gear* та підібрати українські відповідники. Результат інтерпретації показано на рис. 9.3. Зауважимо, що в цьому випадку йдеться про стандартну авіаційну команду, пов'язану з випусканням шасі.

Значення:

Gear down! = «Зменшити швидкість!», «Перейти на нижчу передачу!», «Гальмувати / сповільнюватися!»

Пояснення:

- **gear (down)** походить від іменника *gear* «передача, механізм»;
- у формі **to gear down** означає перемикається на нижчу передачу, а ширше — знижувати темп або інтенсивність дії;
- вигук **Gear down!** — це наказ / попередження, типове для:
 - керування транспортом (авіація, ↓.омобіль),

Типові українські відповідники (залежно від ситуації):

- «На нижчу передачу!»
- «Зменш швидкість!»
- «Пригальмуй!»
- «Сповільнися!»

Рис. 9.3 – Результат аналізу значення слова у вузькому контексті

Після введення ширшого контексту – *When the runway is in sight and the aircraft is on the visual profile and a safe landing is assured, select gear down and flaps 15 and reduce speed* – модель коректно ідентифікує термінологічне значення слова *gear* як *landing gear* («посадкове шасі») та узгоджує його з іншими елементами інструкції (рис. 9.4). Отже, додатковий контекст дозволяє усунути потенційну двозначність і перейти від загальної або командної інтерпретації до точного галузевого значення.

У цьому реченні **gear**:

- не означає «механізм» загалом;
- не є метафорою («сповільнитись»);
- виступає як технічний термін авіаційної сфери.

У поєднанні з дієсловом **select** і командою **down**:

- *gear* означає **landing gear** — посадкове шасі;
- **gear down** — стандартна авіаційна команда під час заходу на посадку.
- ♦ Паралель із другим елементом:
 - **flaps 15** — конкретний кут випуску за ↓ лків;
 - отже, **gear** тут теж позначає конкретний елемент конструкції літака, а не

Рис. 9.4 – Результат аналізу значення слова у ширшому контексті

Наведений нижче приклад показує покроковий алгоритм роботи моделі з аналізу значення вузькоспеціального слова (таблиця 9.8). Це демонструє, що її можна ефективно використовувати як інструмент допоміжного лексикографічного аналізу, за умови контрольованого введення контексту.

Таблиця 9.8 – Алгоритм лексикографічної операції ChatGPT

Етап	Дія	Результат
1. Подання контексту	Введення мовного фрагмента (Gear down!)	Визначено, що одиниця функціонує у командній формі; активується дієслівно-фразовий шаблон
2. Первинна інтерпретація	Аналіз граматичної форми, синтаксису, прагматики	gear не трактується як абстрактний іменник; висуваються ситуативні значення («зменшити швидкість», «перейти на нижчу передачу»)
3. Гіпотеза значення	Формування кількох можливих інтерпретацій	Значення залишається багатоваріантним; фіксується потенційна полісемія
4. Розширення контексту	Додавання професійного тексту (авіаційна інструкція)	Активується авіаційний домен; відсіюються побутові та метафоричні тлумачення
5. Уточнення значення	Співвіднесення gear з іншими термінами (flaps 15)	gear інтерпретується як landing gear
6. Семантична фіксація	Визначення функції одиниці в системі	gear down ідентифікується як стандартна авіаційна команда
7. Формування дефініції	Генерація нормативного словникового тлумачення	«Команда на випускання посадкового шасі під час заходу на посадку»
8. Підготовка до словника	Вибір способу подання (підзначення / фразова одиниця)	Матеріал готовий до включення у словникову статтю gear

Працюючи з вузькогалузевими термінами, крім інструменту ШІ слід також послуговуватися спеціальними корпусами, мова про які йшла по темі 8. Вони дають людині та системі обирати потрібні контексти, групувати їх за значенням та аналізувати контекстні оточення. Нижче у таблиці 9.9 подано алгоритм аналізу значення слова *gear*, використовуючи корпусний підхід.

Таблиця 9.9 – Корпусний підхід до розв’язання завдання

Етап	Корпусна дія	Результат
1. Формування запиту	Пошук <i>gear down</i> у спеціалізованих корпусах	Отримання набору контекстів різної галузевої приналежності
2. Вибір релевантних прикладів	Фільтрація авіаційних текстів	Відібрані інструкції, чек-листи, транскрипти
3. Контекстуальний аналіз	Аналіз оточення (runway, flaps, landing)	Виявлено стійке співвіднесення з посадкою
4. Частотний аналіз	Порівняння вживань у різних доменах	Авіаційне значення домінує в технічних текстах
5. Семантичне узагальнення	Групування прикладів за значенням	Формується окреме значення «випуск шасі»
6. Терміносистемна перевірка	Зіставлення з авіаційними стандартами	Підтвердження нормативності команди
7. Формування дефініції	Ручне узагальнення лексикографом	Створення словникового тлумачення

8. Редакторський контроль	Перевірка формулювання	Остаточна словникова стаття

Ефективність обох підходів визначається низкою критеріїв, поданих у таблиці 9.10. Водночас їх комбіноване використання дає змогу значною мірою нівелювати притаманні кожному з них обмеження.

Таблиця 9.10 – Зіставлення корпусного та ІІІ-підходів

Критерій	ChatGPT	Корпусний підхід
Швидкість первинної інтерпретації	Дуже висока	Низька–середня
Потреба у великому масиві даних	Ні	Так
Чутливість до ширшого контексту	Висока	Залежить від вибірки
Ризик хибної інтерпретації	Є за браку контексту	Менший, але не нульовий
Формування дефініції	Автоматизоване	Ручне
Роль людини	Контроль і верифікація	Аналіз і узагальнення

У наведеному кейсі ChatGPT відтворює скорочений лексикографічний цикл, який у корпусному підході потребує кількох окремих процедур. Водночас корпусний аналіз забезпечує емпіричну перевірку та частотну обґрунтованість значення. Отже, оптимальним є комбінований підхід: ChatGPT – для первинної інтерпретації та формування гіпотез; корпус – для підтвердження, уточнення й нормативної фіксації.

Генерація чернеткових дефініцій. ChatGPT здатний пропонувати кілька варіантів дефініцій одного слова – пояснювальні, описові, контекстуальні, перекладні – що дає змогу швидко оцінити можливі стилі словникового опису. Це прискорює початковий етап створення статті й дає можливість зіставляти дефініції різного типу. Проте вони потребують редагування щодо нормативності, термінологічної точності та стислості.

Так, на основі художнього контексту з оповідання І. Франка «*Борислав сміється*» було здійснено експеримент із автоматизованою генерацією тлумачень діалектної та професійної лексики. Метою експерименту було перевірити здатність ChatGPT розпізнавати значення слова в насиченому художньому контексті, пропонувати різні типи дефініцій та слугувати інструментом чернеткового лексикографічного опису. Результати експерименту з генерацією чернеткових дефініцій подано в таблиці 9.11. Як показує експеримент, ChatGPT:

- не обмежується одним типом словникового опису, а продукує дефініції різного формату;
- коректно виводить значення з контексту, навіть у разі архаїчної або діалектної лексики;
- фактично виконує функцію лексикографічного «генератора варіантів».

Таблиця 9.11 – Генерація типів дефініцій у ChatGPT

Лексема	Тип дефініції	Приклад згенерованого тлумачення	Лексикографічна цінність
дерев'яна сука	пояснювально-контекстуальна	Діалектна назва механічного сигнального пристрою (часто дерев'яного), що видає різкий звук для оповіщення робітників	Добре фіксує функцію та соціально-побутовий контекст
брила	дефініційна	Великий уламок каменю; масивний кам'яний блок	Нормативно коректна, придатна для тлумачного словника
оскарб	пояснювальна	Каменярський інструмент для обробки каменю; різновид молота або зубила	Вдало поєднує родовидову модель і контекст
оскарб	контекстуальна	Інструмент, яким муляри б'ють по піщанику, вибиваючи іскри	Корисна для ілюстративної частини статті
брила	перекладна	кам'яний блок; stone block	Придатна для перекладних і навчальних словників

Водночас згенеровані дефініції: (1) не завжди оптимальні за нормативною стислістю, (2) можуть потребувати термінологічного уточнення та (3) можуть потребувати термінологічного уточнення. Отже, генерація чернеткових дефініцій за допомогою ChatGPT є ефективною на підготовчому етапі створення словникової статті. Модель дає можливість: швидко окреслити семантичне поле слова, зіставити різні стилі тлумачення та підготувати матеріал для подальшої нормативної обробки. Тому, ChatGPT доцільно розглядати як інструмент первинної семантичної розвідки, тоді як остаточна структура та формулювання словникової статті залишаються сферою відповідальності фахового лексикографа (див. таблицю 9.12).

Таблиця 9.12 – Зіставлення дефініцій

Лексема	Дефініція Chat GPT	Дефініція лінгвіста / лексикографа
дерев'яна сука	Діалектна назва механічного сигнального пристрою (часто дерев'яного), що видає різкий звук для оповіщення робітників	Дерев'яний гучник, яким подають знак початку чи кінця роботи

брила	Великий уламок каменю; масивний кам'яний блок	Великий безформний шматок каменя, землі, глини, льоду і т. ін.
оскарб	Каменярський інструмент для обробки каменю; різновид молота або зубила	Молоток для насікання млинових каменів

Зіставлення дефініцій, згенерованих ChatGPT, і дефініцій, сформульованих фахівцем-лексикографом, виявляє системні відмінності між машинним і нормативним підходами до тлумачення. Розглянемо приклад з лексемою *дерев'яна сука*. Модель подає функціонально коректне, але узагальнене технічно-описове визначення, не фіксуючи конкретної сфери вживання та традиційної номінації. Лексикографічна дефініція є коротшою, конкретнішою й жанрово типовою, зосередженою на соціально-побутовій функції предмета. У випадку з лексемою *брила* ВЛМ обмежує значення кам'яним матеріалом і подає звужене тлумачення, тоді як нормативна дефініція відображає семантичну широту лексеми та її здатність уживатися щодо різних матеріалів (камінь, земля, лід тощо). Це приклад типової помилки моделі – недооцінка полісемійної потенції слова. Машинна дефініція до слова *оскарб* є родово правильною, але надто абстрактною: вона не відбиває спеціалізованого призначення інструмента. Лексикограф уточнює галузеву функцію та сферу застосування, що є критично важливим для термінологічно точного опису.

Поданий матеріал демонструє, що ChatGPT добре відтворює базову родову схему дефініції. До того ж, він є ефективним на етапі чернеткового формування значень, однак схильним до узагальнення, семантичного звуження або втрати спеціалізації. Натомість лексикографічна дефініція орієнтована на традицію словникового опису, точніше фіксує обсяг значення та його межі та враховує нормативність, типову сферу вживання. Отже, оптимальною є модель «ВЛМ → первинна генерація → фахове редагування», у якій ChatGPT виступає інструментом прискорення, але не заміни лексикографічної експертизи. Типові помилки ВЛМ у лексикографічному тлумаченні наведено в додатку Д.25.

Виявлення потенційної багатозначності. На основі набору контекстів ChatGPT може запропонувати перелік можливих значень лексеми або терміна та згрупувати їх за семантичними ознаками. У такий спосіб лексикограф отримує

змогу автоматично сформувати первинну модель полісемії й підготувати структуру словникової статті для багатозначного слова. Водночас слід враховувати, що модель інколи схильна або до надмірного дроблення значень, або, навпаки, до їхнього узагальнення, що потребує подальшого редакторського контролю. Так, модель за запитом розрізнила значення (семантичні стани) дієслова *комунікувати*, спираючись на дані з корпусу ГРАК (див. таблицю 9.13).

Таблиця 9.13 – Семантична структура слова *комунікувати*

Контекст	Значення	Семантичні ознаки	Сучасний відповідник
...наскільки добре комунікують між собою Мінсоцполітики, Мінфін та МОН.	здійснювати інституційну взаємодію, обмін інформацією та рішеннями між органами.	офіційна сфера; двосторонній / багатосторонній процес; не просто «говорити», а координувати дії.	взаємодіяти, координуватися, узгоджувати позиції
...залізниць, котрі б краще комунікували шахти поміж собою... (1919)	забезпечувати транспортне або технічне сполучення між об'єктами (<i>іст.</i>).	матеріальна інфраструктура; калька з лат. <i>communicare</i> / нім. <i>kommunizieren</i> ; архаїчне для сучасної мови.	сполучати, з'єднувати, забезпечувати зв'язок
Нам комунікують: 3 днем 9. цвітня с. р. увійшов... (1917)	передавати офіційне повідомлення, інформувати адресатів (<i>канц.-оф., заст.</i>).	безособова конструкція; адміністративно-правовий стиль; пасивний адресат.	повідомляють, доводять до відома, інформують
...по Інтернету комунікувати з колегами...	здійснювати міжособистісне спілкування з використанням каналів зв'язку.	двосторонність; акцент на процесі обміну; сучасне, нейтральне.	спілкуватися, контактувати
...буде, дійсно, комуніковано всі функції...	бути переданим / інтегрованим / узгодженим у межах системи управління (<i>книжн. абстр.</i>).	пасивна форма; управлінський дискурс; дещо канцеляризована мова.	буде передано, узгоджено, реалізовано
...музей активно комунікував з американським суспільством...	здійснювати цілеспрямовану культурну або інформаційну взаємодію з публікою.	стратегічна комунікація; публічний простір; маркетингово-культурний контекст.	взаємодіяти з аудиторією, вести комунікацію

Як показує таблиця, аналізоване слово є полісемантичним, стилістично рухливим (від канцеляризму до нейтральної розмовної мови) та прикладом сучасного розширення в українській мові (таблиця 9.14).

Таблиця 9.14 – Узагальнювальна схема значень

№	Значення	Сфера
1	Інституційна взаємодія	адміністрування
2	Технічне / транспортне сполучення	історична
3	Офіційне повідомлення	канцелярська
4	Міжособистісне спілкування	сучасна
5	Системне передання функцій	управлінська
6	Публічна / стратегічна комунікація	соціокультурна

Поданий приклад демонструє, що ChatGPT здатний ефективно виявляти потенційну полісемію слова, спираючись на різночасові та різностильові контексти, а також пропонувати первинне групування значень за сферою вживання. Водночас остаточне розмежування значень, визначення їх ієрархії, стилістичного статусу та доцільності включення до нормативного словника залишається за лексикографом. Таким чином, III виконує функцію інструмента попереднього семантичного моделювання, тоді як наукова інтерпретація й нормалізація значень потребують фахової експертизи.

Але слід мати на увазі, що не кожне контекстуально виокремлене значення автоматично відповідає окремій словниковій семемі. Частина з них може бути стилістичними або дискурсивними реалізаціями одного значення. У таблиці 9.13 ChatGPT, спираючись на корпусні контексти, виокремив шість семантичних станів дієслова *комунікувати*. Проте з лексикографічного погляду не всі з них доцільно фіксувати як окремі словникові значення. Частина виявлених інтерпретацій є стилістичними, дискурсивними або жанровими реалізаціями базового значення. Так, лексикограф об'єднує близькі реалізації, залишаючи лише ті, що відображають стійку семантичну відмінність (див. таблицю 9.15).

Таблиця 9.15 – Об'єднання семантичних реалізацій

Базова семема	Реалізації	Пояснення
1. Спілкуватися, здійснювати обмін інформацією	1, 4, 6	Різняться сферою застосування, але мають спільну семантичну основу
2. Повідомляти, передавати інформацію (офіц.)	3, 5	Канцелярські й управлінські вживання без зміни ядра значення
3. З'єднувати, забезпечувати сполучення (іст.)	2	Окреме історичне значення, неактуальне для сучасної норми

У підсумку замість шести «значень», запропонованих ВЛМ, нормативна словникова стаття фіксує три семеми, з можливим стилістичним маркуванням:

1. **Комунікувати** – *спілкуватися, здійснювати обмін інформацією* (нейтр., публіц.).

2. **Комунікувати** – *повідомляти, доводити до відома* (офіц.-канц.).

3. **Комунікувати** – *з'єднувати, забезпечувати сполучення* (іст.).

Отже, ВЛМ ефективно виявляють потенційну полісемію, але схильні до надмірної сегментації. З огляду на це, завдання лексикографа – звести аналітичні інтерпретації до нормативно обґрунтованої кількості семем, зберігаючи баланс між точністю опису та економією словникової структури.

Добір перекладних відповідників. Для перекладних словників і глосаріїв ChatGPT може пропонувати кілька варіантів перекладу з короткими семантичними коментарями або вказівками на сферу вживання. Модель забезпечує швидке зіставлення міжмовних відповідників та надає попередній відбір термінів для подальшої нормалізації, але може змішувати загальномовні та фахові відповідники. Розглянемо роботу ВЛМ з підбору еквівалентів до слова *design* у наведених контекстах (таблиця 9.16).

Таблиця 9.16 – Перекладні еквіваленти до слова *design*

Контекст	Значення	Еквіваленти	Коментар
The Boeing 7X7 prototype was designed as an all-new twin-engine wide body aircraft...	бути спроектованим / розробленим як певний тип технічного об'єкта.	спроектований розроблений (рідше) сконструйований	Йдеться не про зовнішній вигляд, а про інженерну розробку концепції літака.
The 7N7 prototype was a twin-engine variant of the 727 designed for superior range and economics	створений (у проєкті) з певною функціональною метою.	розроблений для спроектований з метою орієнтований на	Тут design включає <i>цільові характеристики</i> (дальність, економічність).
...the 7N7 designed was ultimately scrapped in favor of an all-new narrow-body design.	конструктивна схема / проєктне рішення	проєкт конструкція проєктна схема тип конструкції	Це вже результат проєктування, а не процес.
Since the 757 and 767 were designed at the same time, both aircraft were fitted with identical cockpits	перебували в процесі паралельного проєктування.	проєктувалися розроблялися	

Як видно з наведених контекстів, *design* є терміном-гіперонімом, що охоплює проектування, конструкцію, архітектуру, та інженерну концепцію. Найчастіший коректний відповідник українською – проєкт / проєктування / конструкція / розробка, залежно від контексту (таблиця 9.17).

Таблиця 9.17 – Узагальнені результати підбору еквівалентів

Англійське <i>design</i>	Контекстуальне значення	Український еквівалент
designed as	інженерне проєктування	спроєктований, розроблений
designed for	цільова функція	розроблений для
design (noun)	результат проєктування	проєкт, конструкція
narrow-body design	тип конструкції	конструктивна схема
aircraft were designed	процес	проєктувалися
designed in the era	історико-техн. контекст	спроєктований у період

У поданому прикладі ChatGPT коректно ідентифікує багатозначність англійського *design* та пропонує релевантні перекладні відповідники залежно від контексту. Водночас модель демонструє схильність до варіативного перекладу без чіткого розмежування між процесуальним, результативним і типологічним значеннями, що потребує подальшої лексикографічної нормалізації. Типовою помилкою ВЛМ є злиття рівнів перекладу, як це показано в таблиці 9.18.

Таблиця 9.18 – Злиття рівнів перекладу

Рівень	Приклад	Коментар
Процес	<i>aircraft were designed</i> → проєктувалися	Коректно
Результат	<i>narrow-body design</i> → конструкція / проєкт	Коректно
Тип	<i>new design</i> → дизайн	Потенційна помилка
Стиль	<i>design</i> → дизайн (універсально)	Ненормативне узагальнення

ChatGPT схильний використовувати калькований відповідник *дизайн* як універсальний, навіть у суто інженерних або технічних контекстах, де в українській нормативній традиції закріплені інші терміни (*проєкт, конструкція, проєктування*). У перекладних словниках англійське *design* доцільно розглядати як термін-гіперонім, що в українській мові реалізується через низку контекстуально зумовлених еквівалентів (*проєкт, проєктування, конструкція, розробка*). Використання калькованого відповідника *дизайн* у технічних контекстах має бути обмежене або марковане стилістично.

Отже, модель ChatGPT ефективно генерує первинний набір перекладних відповідників до *design* та правильно ідентифікує контекстуальні значення. Водночас остаточне формування нормативної словникової статті можливе лише

за умови редукції варіантів, усунення міжмовної інтерференції та чіткого розмежування семантичних рівнів.

Семантичне групування і класифікація термінів. ChatGPT добре працює з класифікаційними завданнями: групує терміни за функцією, тематикою, родовидовими відношеннями, типами процесів або об'єктів. Систему можна використовувати для формування концептуальних полів та підготовки словників-тезаурусів. Під час розв'язання цього завдання результати класифікації мають рекомендаційний характер, оскільки їхня побудова відбувається без повноцінного онтологічного аналізу та формальної валідації відношень.

Показовим є приклад семантичного групування терміна *design*, який у технічному дискурсі функціонує як багаторівневий гіперонім. На основі корпусних контекстів і машинної інтерпретації ChatGPT виділяє кілька стабільних семантичних груп, що формують концептуальне поле проєктування (таблиця 9.19).

Таблиця 9.19 – Семантичне групування терміна *design*

Семантична група	Тип значення	Опис	Українські відповідники
Процес	процесуальне	Дія зі створення технічного об'єкта або системи	проєктування, розробка
Результат	предметне	Матеріалізований або концептуальний результат проєктування	проєкт, конструкція
Тип / клас	класифікаційне	Усталений різновид технічного рішення	тип конструкції, схема
Цільова орієнтація	функціональне	Проєктування з наперед визначеною метою	розроблений для
Естетичний аспект	оцінне	Зовнішній вигляд або стилістичне оформлення	дизайн

Концептуальну схему для аналізованого терміна можна подати у такому вигляді:

design

інженерне проєктування

результат *проєкт / конструкція*

тип *конструктивна схема*

функція *розроблений для*

естетичне оформлення (периферійна зона)

Використовуючи ChatGPT для семантичного групування, лексикограф отримує швидку первинну модель концептуального поля, однак має враховувати, що модель схильна:

- поєднувати різні рівні абстракції (процес ↔ результат ↔ тип);
- виділяти периферійні значення (напр., естетичне *design*) нарівні з ядреними;
- не завжди чітко розрізняти лексичне значення і функціональну роль у контексті.

Тому машинно згенеровані класифікації доцільно використовувати як інструмент попереднього структурування матеріалу, що надалі уточнюється шляхом корпусної перевірки та експертної лексикографічної інтерпретації. Так, для терміна *design* модель може запропонувати таку структуру:

design – процес створення технічного об'єкта; проєкт; конструкція; тип літака; схема.

У цьому переліку поєднано: процес (*проєктування*); результат процесу (*проєкт, конструкція*); класифікаційний тип (*тип літака*); репрезентацію результату (*схема*). З погляду лексикографії така структура є помилковою, оскільки порушує принцип єдиного критерію членування (див. таблицю 9.20).

Таблиця 9.20 – Лексикографічний аналіз помилки

Елемент	Статус	Проблема
design = процес	процесуальне значення	коректне
design = проєкт	результативне значення	змішано з процесом
design = конструкція	предметне значення	інший онтологічний рівень
design = тип літака	класифікаційне	не семема, а термінологічна роль
design = схема	репрезентація	не самостійне значення

Модель не фіксує, що проєкт і конструкція не є різними значеннями *design* у строгому сенсі, а лише реалізаціями результативної семени. З погляду лексикографа коректною є ієрархічна модель:

design

1. **Проєктування** – процес створення технічного об'єкта.
2. **Результат проєктування** – проєкт, конструкція, схема (контекстуальні реалізації).

У цьому випадку процес і результат розведено, внутрішня варіативність результату не дробиться на окремі значення і, до того ж, зберігається економність словникової структури. Змішування процесу і результату зумовлене тим, що ВЛМ орієнтується на поверхневу дистрибуцію в корпусах, не має вбудованої онтології подій і об'єктів та схильна перетворювати контекстні ролі на «окремі значення». Виходячи з вищесказаного, ChatGPT ефективно генерує первинні класифікаційні моделі, але не здатний самостійно гарантувати онтологічну коректність. Тому на етапі семантичного групування його слід розглядати як інструмент попереднього аналізу, а остаточне рішення щодо структури значень залишається за лексикографом.

Аналіз стилістичних і оцінних компонентів. Модель здатна виявляти не лише денотативне значення лексеми, а й її стилістичне та прагматичне навантаження: оцінність, експресивність, іронічність, а також належність до певного функціонального стилю (публіцистичного, розмовного, есеїстичного тощо). Це особливо важливо для укладання словників неологізмів, сленгу та соціально маркованої лексики, де значна частина смислу закладена не у «що означає слово», а як і з якою інтенцією воно вживається.

Показовим у цьому плані є неологізм *айфонізація*, ужитий у публіцистичному контексті наприкінці ХХ – на початку ХХІ століття. Формально слово побудоване за продуктивною моделлю з абстрактним суфіксом *-ізація*, характерним для нейтральної науково-офіційної лексики. Однак у реальному мовленні воно набуває іронічного та критично-жартівливого забарвлення, оскільки поєднує псевдонаукову форму з побутовим, брендовим компонентом (*iPhone*). Модель коректно ідентифікує:

- **стиль:** публіцистичний / есеїстичний;
- **забарвлення:** іронічне;
- **ступінь експресії:** помірний;
- **оцінність:** критично-жартівлива;
- **мовний прийом:** іронічна деривація.

Таким чином, *айфонізація* постає не як нейтральний термін, а як риторичний інструмент, що дистанціює автора від позначуваного явища й водночас оцінює його. У контексті 1990-х років така лексема підкреслювала новизну й демонстративність західних технологічних трендів; у сучасному ж українському дискурсі подібні утворення (*комп'ютеризація, інтернетизація, стільниковізація, айфонізація*) можуть втрачати ефект новизни й сприйматися вже радше як стилістично маркований перелік, ніж як реальна оцінка технологічного рівня суспільства.

Отже, модель може не лише фіксувати стилістичні характеристики лексеми, а й попереджати про можливі стилістичні помилки в дефініціях, зокрема у випадках, коли іронічно або публіцистично забарвлене слово помилково подається як нейтральний термін. Водночас ефективність такого аналізу безпосередньо залежить від чітко заданого контексту, оскільки саме контекст актуалізує прагматичні та оцінні компоненти значення.

Формування заготовок словникових статей. ChatGPT може створювати структурні шаблони словникових статей, що охоплюють основні зони лексикографічного опису: заголовне слово, граматичну характеристику, стилістичні та функціональні ремарки, дефініцію, ілюстративний матеріал. Такі заготовки зручно використовувати як робочий матеріал на початковому етапі укладання словника. Використання модельно згенерованих шаблонів сприяє уніфікації структури словника, зменшує ризик пропуску важливих параметрів опису та суттєво економить час на технічному оформленні словникових статей. Особливо ефективним цей підхід є у роботі з неологізмами, сленгом і стилістично маркованою лексикою, де необхідно фіксувати не лише значення, а й прагматичні характеристики слова.

Водночас такі заготовки не є остаточним продуктом: фінальна редакція структури й наповнення словникової статті залежить від концепції конкретного словника, його типу (тлумачний, неологічний, термінологічний, навчальний тощо) та цільової аудиторії. На рисунку 9.5 подано приклад словникової статті для слова *айфонізація*, що ілюструє можливість поєднання граматичної

інформації, стилістичних ремарок (*розм., публіц., ірон.*), чітко сформульованої дефініції та контекстного прикладу в межах єдиної уніфікованої структури.

айфонізація, -ї, ж., неодн.

розм., публіц., ірон.

1. Процес масового поширення смартфонів (переважно iPhone) у повсякденному житті суспільства та зростання їхньої ролі в комунікації й побуті.

◆ У сучасній Україні рівень комп'ютеризації, інтернетизації, стільниковизації, айфонізації вже давно перевершив західний.

Рис. 9.5 – Словникова стаття, згенерована ChatGPT

Таким чином, ChatGPT доцільно використовувати на етапах аналізу, генерації варіантів і попереднього структурування, тоді як нормативне тлумачення, відбір єдиного варіанта дефініції та встановлення статусу словникової одиниці залишаються сферою відповідальності лексикографа. Найефективнішою є модель співпраці, у якій ChatGPT виконує роль інтелектуального асистента, а фахівець – роль концептуального й редакторського центру.

Питання для самоконтролю

1. Які типи лексикографічних операцій доцільно делегувати ChatGPT на початковому етапі укладання словника або глосарія?
2. У чому полягає різниця між контекстною інтерпретацією значення та словниковою семемою?
3. Чому контекстуально виокремлене значення не завжди слід фіксувати як окреме словникове значення?
4. Які ознаки свідчать про надмірну сегментацію значень у результатах роботи ВЛМ?
5. У чому полягає типова помилка змішування **процесу** і **результату** під час семантичної класифікації?
6. Як приклад зі словом *design* демонструє полісемію терміна-гіпероніма?
7. Які ризики виникають під час автоматичного добору перекладних відповідників?

8. У чому полягає принципова відмінність між корпусним аналізом і аналізом, виконаним ВЛМ?

9. Які критерії використовує лексикограф для об'єднання або розведення значень?

10. Яку роль відіграє людина-фахівець у фінальній нормалізації словникової статті, створеної за допомогою ChatGPT?

Завдання для самоперевірки

1. Подайте три різні контексти для багатозначного слова (наприклад, *модель, канал, платформа*) і:

- визначте значення за допомогою ChatGPT;
- зіставте їх із нормативним словниковим описом;
- вкажіть, де модель схильна до узагальнення або надмірного розщеплення.

2. Візьміть перелік значень, згенерованих ВЛМ, і:

- визначте, які з них є стилістичними або дискурсивними варіантами;
- запропонуйте скориговану семантичну структуру словникової статті.

3. Проаналізуйте одне дієслово або іменник, у якого можливе змішування процесу і результату (*реалізація, управління, дизайн, переклад*), і:

- розмежуйте онтологічні рівні;
- побудуйте нормативну схему значень.

4. Для терміна з фахової галузі:

- згенеруйте перекладні відповідники за допомогою ChatGPT;
- позначте загальнономовні та термінологічні еквіваленти;
- укладіть міні-статтю перекладного словника.

5. Виконайте зіставлення:

- корпусного аналізу;
- аналізу ВЛМ

на одному й тому ж лексичному матеріалі, описавши сильні та слабкі сторони кожного підходу.

Питання для поглибленого вивчення матеріалу

1. Полісемія та омонімія в автоматично згенерованих словникових структурах.
2. Роль жанрового та дискурсивного контексту в інтерпретації значення.
3. Межі застосування ВЛМ у термінологічній лексикографії.
4. Стилїстичне маркування в словниках, створених із залученням ШІ.
5. Використання ChatGPT для побудови тезаурусів і онтологічних схем.
6. Нормативність і відповідальність лексикографа в умовах ШІ-асистованої роботи.
7. Перспективи комбінованих моделей: корпус + ВЛМ + експерт.
8. Етичні та методологічні аспекти використання ШІ в мовознавчих дослідженнях.

ПІСЛЯМОВА

Комп'ютерна лексикографія пройшла довгий шлях, перетворившись із допоміжної галузі на повноцінний напрям сучасної лінгвістики з власною теорією методами та науковою проблематикою. На сьогоднішній етап розвитку дисципліни – це постійний пошук нових рішень, оптимізація процесів та інтеграція цифрових технологій у найрізноманітніші типи словникових ресурсів. Зберігаючи найкращі традиції паперового словникарства та поєднуючи методи суміжних гуманітарних і комп'ютерних наук, комп'ютерна лексикографія сформувала власні підходи до створення, використання та інтеграції словників у єдиному інформаційно-цифровому середовищі.

Теоретико-методичний апарат комп'ютерної лексикографії збагачено не лише традиційними лінгвістичними напрямками – структурною лінгвістикою, теорією мови, перекладознавством, лінгводидактикою, – а й сучасними фреймворками, зокрема теорією лексикографічних систем. Ці підходи забезпечують створення концептуальних моделей словників, внутрішніх лексикографічних баз даних та зовнішніх інтерфейсів, що формують повноцінну екосистему цифрових лексикографічних ресурсів.

Сучасні проблеми комп'ютерної лексикографії пов'язані з динамікою мови (неологізми, сленг), опрацюванням великих обсягів даних (Big Data), створенням якісних електронних словників для ШІ (розуміння контексту, семантики), інтеграцією словників у мультимедійні та крос-платформні середовища, а також із масштабною автоматизацією процесів за допомогою машинного навчання. Виклики тут різноманітні: від підвищення точності та інтерактивності електронних словників до забезпечення їхньої доступності й адаптивності.

Для опрацювання величезних обсягів текстів (інтернет, соціальні мережі) активно розробляються алгоритми автоматичного вилучення лексичної інформації, факт-чекінгу та верифікації даних. Вирішення проблеми контекстуальної семантики включає створення моделей, здатних коректно відображати багатозначність, нові контекстні відтінки значень, а також явища комп'ютерно-опосередкованої комунікації, як-от емодзі та скорочення. Водночас

проводяться активні розробки систем, що інтегруються з навчальними платформами, пошуковими сервісами та мобільними додатками, забезпечуючи динамічне оновлення та персоналізацію контенту. Особливу роль відіграє створення багатих лексикографічних баз даних, призначених для навчання моделей штучного інтелекту – сфера, що вже отримала назву «нейромережева лексикографія».

З огляду на вищесказане, комп'ютерна лексикографія рухається від простого оцифрування паперових словників до створення «живих», інтелектуальних, адаптивних систем, здатних працювати з динамічною, багатовимірною сучасною мовою. Це шлях від статичного опису до моделювання картини мовного світу в реальному часі.

Наостанок хочемо висловити сподівання, що цей посібник стане надійним фундаментом у вашому подальшому розвитку як філолога, лінгвоінженера чи лексикографа. Попереду – широчезне поле досліджень, експериментів і творчих відкриттів у сфері, яка стрімко змінює те, як людство зберігає, організовує та переосмислює мовне знання. Нехай здобуті знання стануть для вас не лише професійним інструментом, а й джерелом натхнення, що відкриватиме нові горизонти в розумінні мови та її цифрового майбутнього.

З повагою,

Автор.

ГЛОСАРІЙ

АПАРАТ

Семантичний апарат – сукупність об'єктів предметної галузі, об'єднаних між собою різними типами смислових відношень, такі як рід-вид, синонімія, антонімія тощо. Семантичний апарат використовують інтелектуальні системи (онтології) для інтерпретації слів у конкретному контексті.

АРХАЇЗМ

Слово, вираз або граматична форма, які застаріли й вийшли з активного вжитку, але мають сучасні синоніми. Наприклад: *лексикон* (застаріле слово на позначення словника) – *словник*.

АУДІОСЛОВНИК

Електронний словник, у якому мовний матеріал репрезентовано у вигляді звукозаписів, що демонструють нормативну вимову, інтонаційні особливості, діалектні варіанти.

БАЗА ДАНИХ

Організована та структурована сукупність даність про систему мови або її певний рівень (фонетичний, морфологічний, лексичний, синтаксичний). База даних є сховищем лінгвістичної інформації для **електронних словників**.

ВЕРСІЯ

Комп'ютерна версія словника – цифрове відтворення друкованого словника, що зберігає його основну структуру без суттєвих змін, але надає можливість інтерактивного пошуку та гіпертекстових переходів.

ВИБІРКА

Частина генеральної сукупності даних, відібрана відповідно до запиту користувача з **бази даних**. Наприклад, користувач може отримати сукупність словникових статей, що містять заголовкові слова іншомовного походження.

ВИЗНАЧЕННЯ

Формулювання або вислів, що розкриває значення слова або зміст поняття. У лексикографії основними способами подання визначення є **тлумачення** й **дефініція**.

ВІДЕОСЛОВНИК

Електронний словник, у якому значення слів і термінів подається за допомогою відеороликів або анімації, що демонструють предмети, поняття, процеси та дії, важкі або неможливі для повноцінного опису виключно словесними засобами. Відеословники особливо ефективні для фіксації та опису засобів невербальної комунікації, зокрема мови жестів.

ВХІД ДО СЛОВНИКА

Функціональний компонент словникової системи, призначений для забезпечення доступу до лінгвістичного матеріалу та навігації в межах словника. Реалізується на рівні **макроструктури** (алфавітний, тематичний чи інший реєстр), **мікроструктури** (елементи словникової статті), а також за допомогою пошукових механізмів і сервісних інструментів (**індекси**, історія перегляду тощо).

ГІПЕРТЕКСТ

Нелінійний спосіб подання й організації інформації в електронних документах, заснований на системі гіперпосилань, що забезпечують швидкий перехід між пов'язаними фрагментами тексту. У лексикографії гіпертекст використовується для організації словникового матеріалу, навігації між розділами й статтями та візуалізації системних зв'язків між мовними одиницями.

ГЛОСАРІЙ

Спеціалізований словник або список термінів, понять, аббревіатур та скорочень з їхніми визначеннями та поясненнями в межах певної галузі знань. Може створюватися наприкінці книги або окремо для забезпечення єдності термінології, уникнення непорозумінь і допомоги читачам, перекладачам чи фахівцям у розумінні специфічних слів.

ГНІЗДО

Об'єднання слів або словосполучень, пов'язаних спільним коренем або спільним семантичним центром, у межах якого виділяють словотвірні та лексико-семантичні гнізда.

Словотвірне гніздо – група слів, утворених від одного кореня; центральним елементом такого гнізда є просте непохідне слово, від якого мотивуються інші спільнокореневі слова з різним ступенем словотвірної мотивації (пор.: *гніздо* – *гніздечко*, *гніздище*).

Лексико-семантичне гніздо – група словосполучень, об'єднаних навколо одного слова, що позначає родові поняття і поєднує назви видових понять (пор.: *гніздо* – *осине гніздо*, *пташине гніздо*).

ГРАМАТИКА

Підсистема мови, що містить сукупність правил для породження цілісних об'єктів із мовних одиниць, враховуючи інформацію про їхні лексико-семантичні властивості зі словника.

ДЕНОТАТ

Об'єкт дійсності або уяви, позначуваний мовним знаком.

ДЕФІНІЦІЯ

Стисле, чітке й логічно побудоване визначення, яке відображає найістотніші ознаки визначуваного поняття; зазвичай використовується для **термінів** і є основним засобом фіксації їхнього наукового значення. *Див. також* тлумачення.

Відсилкова дефініція – дефініція, що містить посилання на іншу словникову статтю, у якій подано повне тлумачення поняття; застосовується переважно для дублетних термінів або абсолютних синонімів.

Контекстуальна дефініція – дефініція, що розкриває значення слова або поняття через контекст його вживання, зокрема через приклади, ситуації або типові мовленнєві оточення.

Пояснювальна дефініція – дефініція, метою якої є пояснення функцій та способу вживання мовної одиниці, зокрема через відповіді на запитання: «*що виражає?*», «*для чого використовується?*», «*яким чином використовується?*». Застосовується переважно для службових частин мови (прийменників, сполучників, часток).

Перелічувальна дефініція – дефініція, що розкриває значення слова шляхом переліку предметів, понять або ознак, які позначаються цим словом.

Родо-видова дефініція – дефініція, у якій значення слова встановлюється шляхом указівки на родові поняття (гіперонім) та виокремлення видових ознак, що відрізняють це поняття від інших у межах того самого роду.

Дефініція через синонім – дефініція, у якій значення слова пояснюється шляхом заміни його тотожним або семантично близьким словом, без розгорнутого опису поняття.

ДІЄСЛОВО

Самостійна частина мови, що означає дію або стан як процес і має такі основні граматичні категорії: часу, виду, способу, стану, особи та числа.

ЕФЕКТ

Лексикографічний ефект – поняття з теорії лексикографічних систем, що репрезентує певний мовний феномен, який має власну форму подання, зміст і механізм утворення у словниковому середовищі.

ЗАЙМЕННИК

Самостійна частина мови, що не називає предметів, ознак, кількостей чи обставин, а вказує на них або заступає їх у мовленні.

ЗВ'ЯЗКИ

Дублетні зв'язки – зв'язки між двома рівнозначними лексемами або між двома термінами, що позначають один і той самий **денотат** (пор: *палатальний – м'який, номінація – найменування*).

Партитивні зв'язки – лексичні відношення між назвами цілого предмета та його складових частин (пор.: *дерево – гілка, стовбур, коріння*).

Родо-видові зв'язки – зв'язки між лексемами, за яких одна з них позначає родові поняття, а інші – відповідні видові поняття.

Синонімічні зв'язки – зв'язки між лексемами, що мають близьке, але не тотожне лексичне значення. (пор. *когнітивний – пізнавальний*).

ЗНАЧЕННЯ

Відображення у слові його змісту або функції. Залежно від характеру цього відображення розрізняють лексичне значення (пов'язане з предметом, поняттям, процесом, явищем) і граматичне значення (пов'язане з роллю слова в системі мови).

Денотативне значення – значення мовної одиниці, що відображає клас предметів, явищ або об'єктів дійсності, на які вона безпосередньо вказує. (пор.: *словник – книжка; комп'ютерна програма; онлайн-ресурс; окремий файл у системах NLP*).

Конотативне значення – додаткові емоційно-експресивні та оцінні відтінки значення, що накладаються на основне (денотативне) значення слова (пор.: *зелений* – у денотативному значенні: колір, в конотативному: молоду недосвідчену людину).

Сигніфікативне значення – значення мовної одиниці, що відображає зміст поняття та сукупність його інваріантних ознак, спільних для всіх предметів або явищ одного класу.

Узуальне значення – загальноприйняте значення мовної одиниці, що склалося історично в мовній практиці певного суспільства та фіксується у словниках.

ІМЕННИК

Самостійна частина мови, яка позначає предметність у широкому розумінні: особи, речі, явища, абстрактні поняття та стани.

ІНДЕКС

Пошуковий індекс – спеціальна структура даних, призначена для прискорення пошуку, сортування та фільтрації інформації. У **базі даних** електронного словника пошуковими індексами можуть бути заголовкове слово, а також його граматичні, лексичні й стилістичні характеристики, що дають змогу формувати **вибірки** словникових записів.

ІНДЕКСУВАННЯ

Індексування даних – процес створення спеціальних структур даних (**індексів**) для стовпців, що дозволяє значно прискорити пошук, сортування та фільтрацію записів **бази даних** словника.

ІНТЕГРАЦІЯ

Семантична інтеграція – узгодження структури даних між різними словниковими базами, наприклад: від лема в тлумачному словнику до відповідної статті в граматичному словнику.

ІСТОРИЗМ

Слово або словосполучення, що називає предмет, явище чи реалію минулого, які зникли з сучасного життя й не мають активних сучасних відповідників (пор.: *чумак, капітан*).

КАРТИНА

Картина мовного світу – ментально опрацьована й закріплена мовою система уявлень про навколишній світ, сформована в результаті мовно-розумової та комунікативної діяльності людини, спільноти або нації.

КОЛОКАЦІЯ

Комбінація двох або більше слів, які формують стійке словосполучення, що відіграє важливе значення для природного звучання мови.

КОМПОЗИЦІЯ

Композиція словника – внутрішня будова та організація словника, що визначає спосіб розміщення, структурування й взаємозв'язок його компонентів (словникових статей, їхніх розділів і підрозділів) та забезпечує цілісність і зручність користування словником.

КОМПОНЕНТ

Граматичний компонент семантичного стану – компонент **семантичного стану** мовної одиниці, що відображає реалізацію її морфологічних і синтаксичних характеристик у конкретних контекстах уживання.

Лексичний компонент семантичного стану – компонент **семантичного стану** мовної одиниці, що відображає реалізацію її лексичної семантики або функціонального призначення в конкретних контекстах уживання.

КОНКОРДАНС

Список усіх уживань даного слова в лівосторонньому та правосторонньому контексті.

КОНЦЕПЦІЯ

Концепція словника – загальна ідея словника, що визначає його призначення, обсяг і принципи організації.

КОПІЯ

Комп'ютерна копія словника – цифрове відтворення друкованого видання без суттєвих змін його структури. Зазвичай такі словники зберігаються у форматі видавничих систем PDF або у форматі DJVU, створеному для стиснення зображень і забезпечення малого розміру файлу.

КОРПУС ТЕКСТІВ

Спеціально сформоване, структуроване та комп'ютерно оброблене зібрання писемних або усних текстів певної мови, що відображає її сучасний стан або історичний розвиток.

Галузевий корпус текстів – корпус, що охоплює тексти певної галузі знань, професійної сфери або комунікативної ситуації.

Загальномовний корпус текстів – корпус, що відображає сучасну мову в усіх її функціональних різновидах – художньому, публіцистичному, науковому, розмовному тощо.

Паралельний корпус текстів – корпус, що містить тексти двома або кількома мовами, вирівняні за реченнями або фрагментами для зручності зіставлення.

Спеціалізований корпус текстів *див.* **галузевий корпус текстів**.

ЛАБОРАТОРІЯ

Віртуальна лексикографічна лабораторія – цифрова інфраструктура, що забезпечує дистанційне редагування, оновлення та використання лексикографічних ресурсів. *Див. також* середовище розроблення словників.

ЛЕКСИКОГРАФІЯ

Науковий і практичний розділ мовознавства, що охоплює питання створення й використання словників.

Комп'ютерна лексикографія – розділ лексикографії, що займається теорією та практикою створення словників в електронному форматі, використовуючи цифрові технології для зберігання, обробки, пошуку та подання лексичної інформації.

Корпусна лексикографія – напрям лексикографії, що ґрунтується на використанні мовних корпусів як основного джерела емпіричних даних для опису лексичних одиниць, їх значень, сполучуваності та функціонування в реальному мовленні.

Паперова лексикографія *див.* традиційна лексикографія.

Традиційна лексикографія – розділ мовознавства, що займається теорією та практикою створення словників, ґрунтуючись на класичних методах опису словникового складу мови та реалізується переважно в паперовому форматі.

ЛЕМА

Початкова, словникова форма слова, що використовується як основна одиниця під час укладання словника та організації словникових статей. *Див. також* реєстрове слово.

ЛЕМАТИЗАЦІЯ

Процес приведення різних форм слова до його початкової, словникової форми. *Див. також* лема.

ЛІНГВІСТИКА

Наука про мову, її будову, функціонування, розвиток та використання в суспільстві.

Комп'ютерна лінгвістика – розділ лінгвістики, присвячений теоретичним і практичним питанням автоматичного опрацювання природної мови, зокрема створенню мовних технологій, електронних словників, корпусів, систем ІІІ і NLP.

Корпусна лінгвістика – розділ лінгвістики, що вивчає принципи створення, розмітки та використання текстових корпусів як великих електронних зібрань мовних даних для дослідження граматичних, лексичних, синтаксичних, частотних і прагматичних особливостей мови.

Математична лінгвістика – розділ лінгвістики, що застосовує математичні методи, формальні моделі та статистичні підходи для опису, аналізу й моделювання мовних структур і процесів.

Структурна лінгвістика – розділ лінгвістики, що досліджує мову як систему взаємопов'язаних елементів і вивчає її рівневу організацію (фонологічну, морфологічну, синтаксичну, лексичну).

МАКРОРІВЕНЬ

Макрорівень концептуальної моделі – рівень загальної організації словника, що охоплює принципи упорядкування, тип користувацького доступу, пошукову систему та формат подання даних.

МАКРОСТРУКТУРА

Організація словника в цілому, що визначає принципи розміщення словникових статей, логіку побудови словника та способи доступу користувача до інформації (алфавітну, тематичну, гніздову, ієрархічну тощо).

МЕДІОРІВЕНЬ

Медіорівень концептуальної моделі – рівень групування словникових статей за тематичними, частиномовними або функціональними ознаками

МЕРЕЖА

Всесвітня лексикографічна мережа – сукупність електронних словників та інших лексикографічних онлайн-ресурсів, які пов'язані між собою та можуть обмінюватися даними й оновленнями.

МЕТАМОВА

Метамова словника – система мовних і немовних засобів, за допомогою яких у словнику описують мовні одиниці. До неї належать скорочення, **ремарки**, типи **дефініцій**, а також цифри, шрифтові виділення та графічні позначки.

МІКРОРІВЕНЬ

Мікрорівень концептуальної моделі – рівень словникової статті, де представлено опис окремих одиниць (лем, дефініцій, прикладів, граматичних і стилістичних характеристик). *Див. також мікроструктура.*

МІКРОСТРУКТУРА

Внутрішня організація окремої **словникової статті**, що визначає склад елементів інформації про слово та порядок їх подання.

МОДЕЛЬ

Внутрішня модель – рівень архітектури електронного словника, що описує, як дані фізично зберігаються в базі даних. *Див. також:* концептуальна модель, зовнішня модель, модель даних.

Модель даних словника – абстрактна формальна репрезентація елементів лінгвістичної інформації та відношень між ними, що визначає спосіб структурування й опрацювання даних у словнику. *Див. також:* концептуальна модель, зовнішня модель, внутрішня модель.

Зовнішня модель – вигляд словника для користувача: інтерфейс, пошук, меню та способи доступу до інформації, що міститься в базі даних словника. *Див. також:* концептуальна модель, внутрішня модель.

Концептуальна модель – формальна модель, що на абстрактному рівні відображає його структуру і взаємозв'язки між основними елементами словника. *Див. також:* зовнішня модель, внутрішня модель, формальна модель.

– **Ієрархічна модель** – тип концептуальної моделі, що відображає структуру словника як дерево, де вищі рівні (наприклад, макрорівень) підпорядковують нижчі (мікро- та медіорівень).

– **Мережева модель** *див. сіткова модель.*

– **Онтологічна модель** – тип концептуальної моделі, що описує лексичні одиниці та їхні відношення як об'єкти і класи у спеціальній онтологічній схемі.

– **Сіткова модель** – тип концептуальної моделі, що описує взаємозв'язки між лексичними одиницями у вигляді графа, де вузли відповідають словам або поняттям, а ребра – типам відношень (синонімія, антонімія, гіпонімія тощо).

– **Фреймова модель** – тип концептуальної моделі, орієнтований на репрезентацію знань у вигляді сценаріїв або фреймів, що описують типові ситуації, пов'язані з певним концептом.

Формальна модель – спрощене подання реального об'єкта, системи або явища, яке виражене за допомогою спеціалізованої, строго визначеної мови (математичної, логічної, програмної, графічної).

МОДЕЛЬ РЕПРЕЗЕНТАЦІЇ МОВИ

Спосіб подання, структурування та упорядкування лінгвістичної інформації в словнику.

Специфічна модель репрезентації мови – модель, орієнтована на докладний опис конкретної мови з урахуванням її внутрішніх закономірностей і культурного контексту.

Універсальна модель репрезентації мови – модель, що передбачає єдину систему опису для двох або більше мов.

НЕОЛОГІЗМ

Новостворене слово, словосполучення або фраза, що перебуває в процесі входження в загальне використання і ще не включена до загальновживаної мови.

ОНТОЛОГІЯ

формалізована модель певної предметної галузі, у якій задаються поняття (класи), відношення між ними, властивості та правила інтерпретації.

ОПИС

Морфосинтаксичний опис (MSD) – система умовних кодів, що визначають лексичні та формальні властивості слів у контексті.

ПАРАМЕТР

Лексикографічний параметр – одиниця лінгвістичної інформації (частина даних про слово), що може мати самостійну цінність для користувача, але зазвичай функціонує в поєднанні з іншими такими одиницями для формування повного уявлення про слово у словнику.

Асоціативний параметр – параметр, що показує зв'язки лема з іншими мовними одиницями, розкривають семантичне оточення слова.

Атрибутивний параметр – параметр, що відображає належність лема до певного стилю, галузі, територіального чи індивідуального різновиду.

Етимологічний параметр – параметр, що подає інформацію про походження слова, його джерельну мову, етимон (початкову форму) та первісне значення.

Ілюстративний параметр – параметр, що показує, як саме слово функціонує у реальному мовленні.

Інтерпретаційний параметр – параметр, що має на меті пояснити значення реєстрового слова, тобто його семантичний зміст.

Прагматичний параметр – параметр, що описує особливості вживання слова, залежно від часу, частотності, стилістичної доречності та контексту вживання.

Формальний параметр – параметр, що репрезентує план вираження заголовкової одиниці, тобто її зовнішню оболонку в мові.

ПАРСИНГ

Автоматизований процес перетворення тексту словника в комп'ютерну лексикографічну базу даних із чітко визначеною структурою.

ПОЛІСЕМІЯ

Наявність у одного слова чи іншої мовної одиниці двох або більше взаємопов'язаних значень, які мають спільне походження чи асоціативні зв'язки

ПОШУК

Векторний пошук – пошук, який враховує зміст слів і текстів, а не тільки їх точний збіг, завдяки перетворенню даних у числові вектори.

ПРИЙМЕННИК

Службова частина мови, яка виражає залежність іменника, займенника чи числівника від інших слів у реченні, показуючи відношення між ними (місце, час, причину, мету, напрямок тощо).

ПРИКЛАД

Ілюстративний приклад – конкретний, наочний зразок (речення, фраза, словосполучення, текстовий фрагмент, зображення тощо), що використовується для розкриття, пояснення або підтвердження значення реєстрового слова у словниковій статті.

ПРИКМЕТНИК

Самостійна частина мови, що називає ознаку предмета (якість, колір, розмір, належність).

ПРИНЦИП

Принцип розмічення – принцип укладання словників, орієнтований на зовнішню форму подання інформації, зокрема використання кольору, шрифтового накреслення, графічних виділень та інших візуальних засобів.

Структурний принцип – принцип укладання словників, орієнтований на внутрішню організацію лінгвістичних даних, зокрема на їх формальну модель, структуру та взаємозв'язки між елементами.

Комбінований принцип – принцип укладання словників, що поєднує елементи принципу розмічення та структурного принципу, забезпечуючи узгодженість зовнішнього представлення словника з його внутрішньою моделлю даних.

ПРИСЛІВНИК

Самостійна частина мови, що виражає ознаку дії, стан предмета або ознаку якості.

РЕЄСТР

Список слів (реєстрових одиниць), що є об'єктом опису в словнику та утворюють його **макроструктуру**. *Див. також* лема.

Абетковий реєстр – список слів, упорядкованих за алфавітом.

Інверсійний реєстр – список слів, упорядкованих у зворотному алфавітному порядку (від кінця слова до початку).

Тематичний реєстр – список слів, упорядкованих за тематичними або поняттєвими групами.

Частотний реєстр – список слів, упорядкованих за частотою їх уживання в певному текстовому корпусі або масиві текстів.

РЕМАРКА

Елемент метамови, спеціальна позначка у словнику біля заголовкового слова, що містить додаткову інформацію про особливості його вживання, зокрема стилістичні, граматичні, часові або сферу функціонування.

Ареальна ремарка – ремарка, що вказує на географічний ареал уживання слова або його окремого значення.

Ідіолектна ремарка – ремарка, що вказує на належність слова до індивідуального авторського різновиду мови.

Метаідіолектна ремарка – ремарка, що вказує на належність слова або значення до певної наукової школи, напрямку чи традиції.

Тематична ремарка – ремарка, що вказує на вживаність слова або його значення в певній галузі знань чи сфері діяльності.

Функціонально-стильова ремарка – ремарка, що вказує на належність слова або його значення до певного функціонального стилю мовлення.

РОЗГАЛУЖЕННЯ

Асоціативні розгалуження значень – перехід від основного значення слова до додаткових значень, що з'являються в різних контекстах.

РОЗМІТКА

Розмітка текстів – приписування текстам або їхнім компонентам спеціальних міток, які позначають певні види лінгвістичної інформації чи метаданих.

Екстралінгвістична розмітка – система спеціальних міток, що описують обставини створення тексту і не належать до самого змісту.

Лінгвістична розмітка – система спеціальних міток, що описують лінгвістичні характеристики одиниць, що складають текст.

– **Морфологічна розмітка** – тип лінгвістичної розмітки, що додає до кожного слова інформацію про його частину мови, граматичні категорії (рід, число, відмінок, час, аспект тощо) та базову форму (лему).

– **Семантична розмітка** – тип лінгвістичної розмітки, що позначає значення слів або фраз.

– **Синтаксична розмітка** – тип лінгвістичної розмітки, що визначає граматичні зв'язки між словами (підмет, присудок, додаток тощо) та структуру речень.

Структурна розмітка – система спеціальних міток, що відображають структурні одиниці тексту: абзаци, речення, заголовки тощо.

СТАН

Семантичний стан – сукупність граматичних і лексичних значень, які реалізує мовна одиниця у різних мовних контекстах. *Див. також* граматичний компонент семантичного стану, лексичний компонент семантичного стану.

СЕРЕДОВИЩЕ

Лексикографічне середовище – простір, у межах якого відбувається практична реалізація функцій системи.

Динамічне середовище – лексикографічне середовище або краудсорсингова платформа (напр., Urban Dictionary, Вікісловник), у якій наповнення та оновлення словника здійснюється спільнотою користувачів.

Друковане середовище *див.* традиційне середовище.

Електронне середовище – лексикографічне середовище, представлене CD/DVD-словниками, комп'ютерними програмами для інсталяції та локального використання.

Інтегроване середовище – лексикографічне середовище, що являє собою поєднання електронного словника, корпусу, тезауруса, граматики в єдиній платформі – багатофункціональній системі, орієнтованій на комплексний аналіз мови.

Корпусне середовище – лексикографічне середовище, що загальноюмовні та спеціалізовані корпуси текстів.

Офлайнове середовище див. електронне середовище.

Онлайнове середовище – лексикографічне середовище, яке репрезентують сайти-словники, гіпертекстові системи та електронні версії друкованих словників.

Соціально-мережеве середовище див. динамічне середовище.

Традиційне середовище – лексикографічне середовище, що охоплює тлумачні, орфографічні, етимологічні, перекладні словники, енциклопедії, атласи.

Середовище розроблення словників – спеціалізований програмний комплекс, призначений для компіляції, редагування та публікації словників та лексичних ресурсів різного типу.

СИСТЕМА

Лексикографічна система – інформаційно-лінгвістична система, що у формалізованому вигляді репрезентує будову й функціонування мови, а також механізми утворення мовних феноменів (словотвір, словозміна, синтаксис тощо).

Система екстракції знань (knowledge mining) – система, призначена для автоматичного вилучення релевантної інформації з великих масивів текстів.

Система мови – абстрактний механізм, що породжує лінгвістичні об'єкти будь-якого ступеня складності, маючи в основі три важливих компоненти: структуру, суб'єкт та субстанцію.

СКОРОЧЕННЯ

Елемент метамови, що виконує службову функцію в тексті словникової статті, зокрема привертає увагу до інформації або вводить її. Див. також ремарка.

Дефісне скорочення – тип скорочення, що використовується для зменшення довжини слів шляхом опущення їх частини та поєднання компонентів дефісом (напр.: *літ-ра*, *ун-т*).

Крапкове скорочення – тип скорочення, що використовується у словнику для позначення джерел, граматичних характеристик або стилістичних особливостей (напр.: *див.*, *у знач.*).

Комбіноване скорочення – тип скорочення, що поєднує дефісний і крапковий способи скорочення (напр.: *змениш.-пестл.*, *півд.-зах.*).

СЛОВНИК

Упорядкований і систематизований мовно-інформаційний ресурс, здатний інтегруватися та взаємодіяти з іншими ресурсами з метою задоволення різноманітних потреб користувачів – від наукових до практичних.

Автоматизований словник – словник, згенерований або частково сформований автоматично з використанням методів корпусної лінгвістики та обчислювальної обробки мови.

Багатомовний словник – словник, що містить переклади слів, фраз і термінів кількома мовами, подаючи тлумачення, синоніми та інші мовні дані.

Гібридний словник – словник, що поєднує базу даних із можливістю синхронізації через Інтернет. *Див. також* локальний словник, онлайн-словник.

Граматичний словник – словник, що описує морфологічний та синтаксичний рівні мови. *Див. також* словотвірний словник, словозмінний словник.

Двомовний словник – словник, що здійснює зіставний опис двох мов, відображаючи ступінь їхньої лексичної та семантичної еквівалентності. *Див. також* одномовний словник, багатомовний словник.

Експертно укладений словник – словник, створений лексикографами вручну на основі наукового аналізу мовного матеріалу.

Електронний словник – динамічна та інтерактивна лексикографічна система, яка у цифровому середовищі забезпечує оперативний доступ до упорядкованого лінгвістичного матеріалу фактично необмеженого обсягу та надає можливість інтеграції з іншими інформаційними ресурсами. *Див. також* комп'ютерний словник, комп'ютерна версія словника, комп'ютерна копія словника.

Енциклопедичний словник – словник, що містить фактографічну інформацію про осіб, предмети, явища або події, пов'язані з певною галуззю знань або сферою діяльності.

Етимологічний словник – словник, що містить відомості про походження слів однієї мови або групи чи сім'ї споріднених мов. *Див. також* історичний словник.

Ідеографічний словник – словник, у якому слова упорядковано за тематичними або понятійними групами на основі спільності понять; слугує основою для створення **тезаурисів**.

Історичний словник – словник, у якому подається лексика давніх писемних пам'яток із поясненням і документацією її вживання. *Див. також* етимологічний словник.

Комбінований словник – словник, що об'єднує кілька типів словників різного призначення, наприклад граматичний (словозмінний), фразеологічний, словник синонімів і антонімів.

Комп'ютерний словник – словник, призначений для використання на персональному комп'ютері, зазвичай у вигляді програмного забезпечення або локальної бази даних. *Див. також* локальний словник, мобільний словник.

Краудсорсинговий словник – словник, що укладається, редагується та оновлюється спільнотою користувачів з різним рівнем мовної й лексикографічної компетентності. *Див. також* експертно укладений словник.

Локальний словник – словник, установлений на локальному носії, що працює офлайн, тобто без постійного доступу до Інтернету. *Див. також* комп'ютерний словник, мобільний словник, онлайн-словник, локальний словник.

Мобільний словник – словник, установлений на мобільному пристрої (смартфоні, планшеті), що забезпечує доступ до лексикографічної інформації в мобільному середовищі. *Див. також* комп'ютерний словник.

Навчальний словник – словник, що допомагає користувачам опановувати мову або конкретну галузь знань, містить слова разом із поясненнями,

прикладами вживання та практичні завдання на засвоєння лексики або термінології.

Одномовний словник – словник, що описує слова тією ж мовою, якою вони подані. *Див. також* орфографічний словник, орфоепічний словник, тлумачний словник, словник синонімів, словотвірний словник, словозмінний словник, ідеографічний словник.

Онлайновий словник – електронний словник, розміщений в Інтернеті та доступний через веб-браузер або онлайн-сервіси. *Див. також* локальний словник, гібридний словник.

Орфографічний словник – словник, у якому подано нормативне написання слів. *Див. також* орфоепічний словник.

Орфоепічний словник – словник, що подає норми правильної вимови слів і наголосу. *Див. також* орфографічний словник.

Перекладний словник – словник, у якому слова однієї мови подано з перекладом іншою мовою або іншими мовами. *Див. також* багатомовний словник, двомовний словник.

Словозмінний словник – словник, що містить усі граматичні форми реєстрових слів. *Див. також* граматичний словник, словотвірний словник.

Словотвірний словник – словник, який показує способи та моделі утворення слів. *Див. також* граматичний словник, словозмінний словник.

Словник синонімів – словник, у якому слова з близьким значенням об'єднано в синонімічні ряди.

Тезаурусний словник – словник, у якому лексика мови або певної галузі знань подається за семантичними зв'язками (синонімічними, антонімічними, тематичними). *Див. також* ідеографічний словник.

Текстовий словник – тип словника, в якому інформація подається у вербальній формі та спрямована на опис слів, понять і явищ. *Див. також* аудіословник, відеословник.

Термінологічний словник – словник, що описує спеціальну лексику (терміни) певної галузі знань або діяльності.

Тлумачний словник – словник, що пояснює значення слів, подає їхні граматичні характеристики, відтінки значень, а також ілюстративні приклади вживання, фразеологізми та іншу довідкову інформацію.

Фразеологічний словник – словник, що пояснює значення, походження та вживання стійких словосполучень – фразеологізмів (ідіом, крилатих висловів, прислів'їв, приказок).

Частотний словник – словник, що містить список слів (лексем) та/або їхніх форм у певній мові чи текстовому корпусі, упорядкованих за частотою вживання.

АІ-збагачений словник – словник, що використовує нейромережеві моделі, методи машинного навчання, семантичний аналіз та обробку природної мови (NLP) для автоматичного формування, оновлення й персоналізації контенту.

СЛОВО

Повнозначне слово – самостійна частина мови, що має власне лексичне значення, називає предмети, ознаки, дії, кількість або обставини та може виступати самостійним членом речення. *Див. також* іменник, прикметник, дієслово, числівник, займенник, прислівник.

Реєстрове слово – слово, включене до реєстру словника й подане як об'єкт лексикографічного опису в **словниковій статті**. *Див. також* лема.

Службове слово – слово, що не має самостійного лексичного значення та виконує в мові допоміжну, граматичну або синтаксичну функцію.

СЛОВОФОРМА

Конкретна граматична форма одного й того ж слова, яка утворюється під час відмінювання (зміни за відмінками, числами, родами) або дієвідмінювання (зміни за часами, особами).

СПОЛУЧНИК

Службова частина мови, яка не має самостійного значення, не відповідає на питання і служить для зв'язування однорідних членів речення або частин складного речення. *Див. також* прийменник.

СТАТТЯ

Словникова стаття – основний елемент будь-якого словника, що містить вичерпну інформацію про **реєстрове слово** (лему).

СТРУКТУРА

Елемент системної тріади, який в абстрактному вигляді задає будову породжуваного лінгвістичного об'єкта, наприклад: речення, словозмінної парадигми, словоформи тощо.

СУБ'ЄКТ

Елемент системної тріади, здатний ідентифікувати, інтерпретувати структурно-субстанційні властивості й елементи системи та здійснювати з ними операції за певними правилами та алгоритмами. Суб'єкт може бути як людина (або спільнота загалом), так і комп'ютерна система.

СУБСТАНЦІЯ

Елемент системної тріади, що містить набір мовних елементів, необхідних для побудови лінгвістичного об'єкта (наприклад, граматичної форми, речення).

ТАБЛИЦЯ

Таблиця каскадних стилів (CSS – Cascading Style Sheets) – мова опису оформлення вебсторінок, що визначає їхній зовнішній вигляд (кольори, шрифти, розмітку, розташування елементів) і використовується разом із мовою HTML.

ТЕГ

Мітка або ключове слово, що використовується для позначення, опису чи класифікації елемента даних (зокрема текстового) та полегшує його ідентифікацію, пошук і опрацювання.

ТЕЗАУРУС

Див. тезаурусний словник.

ТЕОРІЯ

Теорія мови – комплекс гіпотез та наукових концепцій (лінгвістичних моделей) у межах загального мовознавства, які прагнуть пояснити, як мова виникла, розвивалася та функціонує, її фундаментальні властивості, універсальні ознаки та відмінності між конкретними мовами, слугуючи основою для

розуміння мови як системи знаків, засобу спілкування й мислення. *Див. також* лінгвістика.

ТИПОЛОГІЯ

Структурна типологія словників – класифікація електронних словників за способом організації лексикографічного матеріалу, побудовою макро- та мікроструктури, а також принципами його упорядкування.

Технологічна типологія словників – класифікація електронних словників з погляду використовуваних технічних засобів, платформи, архітектури, способів доступу до даних та рівня інтерактивності.

Функціональна типологія словників – класифікація словників за їхнім призначенням та виконуваними функціями.

ТЛУМАЧЕННЯ

Розкриття або визначення змісту слова чи мовної одиниці. *Див також* визначення, дефініція.

ТРІАДА

Системна тріада – концепт інформаційно-лінгвістичної теорії, згідно з яким мовна система постає як єдність трьох складників: **структури, суб'єкта та субстанції**.

УСТАНОВКА

Авторська установка словника – сукупність концептуальних рішень автора щодо мети й призначення словника, принципів відбору, систематизації та тлумачення лексики, що забезпечують реалізацію авторської ідеї.

Методична установка укладача – сукупність рішень укладача щодо процесу укладання словника, зокрема принципів відбору лексики, способів опрацювання та подання мовного матеріалу.

ФІЛІАЦІЯ

Процес розчленування семантичної структури слова на окремі значення під час лексикографічного опису. *Див також* полісемія.

ФОРМУЛА

Формула тлумачення – елемент метамови, типові синтаксичні конструкції, що використовують для опису лексичного значення або функціональних характеристик леми. *Див. також* примітка, скорочення.

ЧАСТИНА

Ліва частина словникової статті – зона словникової статті, у якій подається формальна характеристика заголовкового слова (леми): граматична, фонетична, структурна, а в окремих словниках – також етимологічна.

Права частина словникової статті – зона словникової статті, що містить змістовий опис леми: її семантичні, стилістичні, узуальні та інші характеристики, зокрема тлумачення та ілюстративні приклади.

ЧАСТКА

Службова частина мови, що надає реченню або словам додаткових смислових (питальні, заперечні, підсилювальні) або граматичних (умовний спосіб дієслів: *б, би*) відтінків, а також служить для утворення нових слів.

ANSI/SPARC

Трирівнева архітектура для систем керування базами даних (СУБД), запропонована підкомітетом SPARC. *Див. також* концептуальна модель, внутрішня модель, зовнішня модель.

AWK

Скриптова мова та утиліта командного рядка в UNIX/Linux, призначена для обробки текстових даних та файлів шляхом сканування і виконання дій над рядками, що відповідають заданим шаблонам.

Concordance

Інструмент **Sketch Engine**, що показує приклади використання слова чи фрази (конкорди) в реальному тексті корпусу, допомагаючи зрозуміти їхнє значення, контекст та граматичні сполучення (колокації).

CQL

Corpus query language – корпусна мова запитів, що використовується в системі **Sketch Engine** для пошуку складних лексичних і граматичних моделей

у корпусах текстів, а також для задання критеріїв пошуку, недоступних у стандартному інтерфейсі користувача.

CSS

Див. таблиця каскадних стилів.

CSV

Comma-Separated Values – простий текстовий формат файлу для зберігання табличних даних, де кожен рядок являє собою запис, а значення (поля) у записі розділені комами (або іншим символом, як-то крапкою з комою)

DWS

Див. середовище розроблення словників.

EAGLES

Expert Advisory Group on Language Engineering Standards – європейська експертна ініціатива 1990-х років, спрямована на розроблення рекомендацій і стандартів для створення та опису лінгвістичних ресурсів (корпусів, лексиконів, анотацій) з особливим акцентом на морфологічні та синтаксичні рівні в багатомовному середовищі.

EMBEDDING-модель

Числові векторні представлення слів, побудовані на основі контекстів, у яких вони з'являються.

GDEX

Good dictionary examples – інструмент **Sketch Engine** для автоматичного добору з текстового корпусу найкращих, тобто найпростіших та найпоказовіших ілюстративних прикладів.

GDEX-score

Числовий коефіцієнт, що автоматично присвоюється інструментом **GDEX** кожному ілюстративному прикладу та відображає ступінь його зрозумілості, простоти й стислості для користувача словника.

GraphQL API

Мова запитів і маніпуляції даними з відкритим кодом для API і середовище виконання для обслуговування запитів з наявних даних. *Див. також* REST API.

HTML

Мова розмітки, за допомогою якої створюють і впорядковують вміст вебсторінок. *Див. також* тег, CSS.

ICON

Мова програмування, яка широко використовується для обробки мовних даних і реалізує генеративну модель обчислень, де вирази можуть породжувати множину значень. *Див. також* CQL, корпусна лінгвістика, векторний пошук.

JavaScript

Мова програмування, що використовується для створення інтерактивних і динамічних вебсторінок, забезпечуючи обробку подій, зміну вмісту сторінки та взаємодію користувача з веб інтерфейсом. *Див. також* HTML, CSS, зовнішня модель.

JSON

Формат запису даних у вигляді тексту, який застосовують для обміну інформацією між програмами та вебсайтами. *Див. також* JavaScript, зовнішня модель, модель даних.

JSON-LD

Формат і метод подання зв'язаних даних, що використовує синтаксис JSON для опису семантичних зв'язків між об'єктами та інтеграції даних у семантичне веб-середовище. *Див. також* JavaScript, зовнішня модель, модель даних.

Keywords

Інструмент Sketch Engine для автоматичного вилучення термінів із текстів під час роботи зі спеціалізованими (галузевими) корпусами.

KWIC

Key Word In Context – спосіб подання результатів корпусного пошуку, за якого ключове слово відображається разом із безпосереднім лівим і правим контекстом. *Див. також* Sketch Engine, NoSketch Engine.

LMF

Lexical Markup Framework – міжнародний стандарт ISO 24613 для структурованого представлення лексиконів (словників) для опрацювання

природної мови (NLP) та машиночитаних словників (MRD). *Див. також* TEI, TEI-XML.

MySQL

Вільна система керування реляційними базами даних (СУРБД) з відкритим кодом, яка дозволяє ефективно зберігати, керувати та отримувати доступ до структурованих даних за допомогою мови запитів SQL. *Див. також* NoSQL, SQLite

Node.js

Кросплатформне середовище виконання **JavaScript** з відкритим кодом, що дає змогу запускати JavaScript-код поза браузером, зокрема на сервері, для створення мережових застосунків і сервісів.

NoSketch Engine

Безкоштовна, локально встановлювана версія **Sketch Engine**, орієнтована на навчальні та дослідницькі потреби.

NoSQL

Нереляційні бази даних, що не використовують таблиці SQL і добре працюють з великими та неструктурованими даними. *Див. також* MySQL, SQLite.

OAI-PMH

Протокол, що застосовують для масового збору метаданих і використовують переважно у наукових, архівних і корпусних проєктах.

OntoLex-Lemon Model

Стандарт для опису словників і лексичних даних у форматі RDF, який поєднує слова з їхніми значеннями та поняттями в графах знань.

OWL

Формальна мова опису онтологій для семантичної павутини, призначена для моделювання класів, властивостей, індивідів та відношень між ними, а також для задання логічних обмежень і правил інтерпретації знань. *Див. також* RDF.

PDF

Універсальний формат файлів, який дозволяє зберігати та переглядати документи з фіксованим макетом. *Див. також* HTML, XML, CSS.

PHP

Популярна серверна скриптова мова програмування з відкритим кодом, що використовується для створення динамічних веб-сторінок та веб-додатків. *Див. також* JavaScript.

POS-tagging

Процес присвоєння кожному слову в реченні мітки частини мови, наприклад, іменника, дієслова, прикметника тощо.

Python

Інтерпретована об'єктно-орієнтована мова програмування високого рівня із суворою динамічною типізацією.

RDF

Resource Description Framework – модель даних, яка описує інформацію у вигляді простих зв'язків між об'єктами (суб'єкт – властивість – значення), щоб комп'ютери могли розуміти зміст даних.

REST API

Representational State Transfer Application Programming Interface – спосіб, за допомогою якого програми обмінюються даними через Інтернет, використовуючи звичайні HTTP-запити. *Див. також* GraphQL API, SOAP API.

Sketch Engine

Комерційна онлайн-платформа, призначена для створення, обробки та аналізу великих текстових корпусів. *Див. також* NoSketch Engine.

SQLite

SQLite – полегшена реляційна система керування базами даних (СУБД), реалізована у вигляді вбудованої бібліотеки, яка не потребує окремого сервера та зберігає всю базу даних у одному файлі. *Див. також* MySQL, NoSQL.

TEI

Text Encoding Initiative – міжнародний стандарт розмітки текстових та лексикографічних даних, який використовує XML для структурованого опису текстів (корпусів, словників, рукописів, літературних та історичних джерел). *Див. також* XML, HTML, RDF, OntoLex-Lemon Model.

TEI/XML

Див. TEI

Thesaurus

Інструмент Sketch Engine для побудови переліку синонімів або групи слів, що належать до спільної (семантичної) категорії.

TreeTagger

програма, яка автоматично визначає частини мови та початкові форми слів у тексті. *Див. також* POS-tagging, лематизація, корпусна лінгвістика, Sketch Engine.

Trends

Інструмент Sketch Engine для аналізу слів (неологізмів або архаїзмів), частота вживання яких змінюється з часом (діахронічний аналіз).

UDPipe

Інструмент автоматичної лінгвістичної обробки текстів, призначений для токенизації, **морфологічної розмітки (POS-tagging)**, лематизації та **синтаксичної розмітки** відповідно до стандарту Universal Dependencies.

Webhooks

Автоматизовані повідомлення (HTTP-запити) від одного веб сервісу до іншого, які миттєво сповіщають про певну подію в реальному часі.

Word Sketch

Інструмент Sketch Engine для автоматичного створення словникових статей на основі частотних колокацій.

Word Sketch Difference

Інструмент Sketch Engine для порівняння лексичних профілів двох слів (спільні та диференційні риси).

Wordlist

Інструмент Sketch Engine для генерації словникового реєстру на основі частот уживання в текстах.

WordNet

Велика лексична база даних англійської мови, що групує слова в набори синонімів (синсети) та показує їхні семантичні зв'язки, створюючи електронний словник-тезаурус для комп'ютерної обробки та ІІІ.

XLSX

Розширення файлу для електронних таблиць Microsoft Excel, що використовується починаючи з версії Microsoft Office 2007.

XML

Розширювана мова розмітки, яка використовується для структурування, зберігання та передачі даних у текстовому форматі, зрозумілому як для людей, такю і для програм. *Див. також* HTML, XSLT.

XSLT

Мова програмування на основі XML, що використовується для перетворення XML-документів з одного формату в інший. *Див. також* XML.

СПИСОК ДЖЕРЕЛ ІНФОРМАЦІЇ

1. Бобкова Т. В. Корпус текстів: основні аспекти визначення. Науковий вісник кафедри Юнеско КНЛУ Сер.: Філологія, Педагогіка, Психологія. Вип. 29. 2014. С. URL: http://www.mova.info/corpus_papers/bobkova-corpus.pdf.
2. Бойко М. І. Українська комп'ютерна лексикографія: суспільні запити, проблеми та перспективи. Науковий вісник Міжнародного гуманітарного університету. Сер.: Філологія. 2022. № 56. С. 12–15.
3. Вступ до корпусної лінгвістики: навчальний посібник / В. В. Жуковська. Житомир: Вид-во ЖДУ ім. І. Франка, 2013. 142 с. URL: https://eprints.zu.edu.ua/18909/1/korpusna_lingv.pdf.
4. Демська О. Вступ до лексикографії. К.: Вид. дім «Києво-Могилянська академія», 2010. 266 с. URL: <https://archive.org/details/demaska2010/page/n3/mode/2up>
5. Демська-Кульчицька О. М. Лексикографія: між ремеслом і наукою. Наукові записки НаУКМА. 2008. Т. 85: Філологічні науки. С. 3–9.
6. Дубічинський В. В. Лексикографія : навч.-метод. посібник. Харків: НТУ «ХП», 2012. 69 с.
7. Зубань О. Комп'ютерне лексикографічне моделювання морфемної системи української мови: монографія. К.: Видавничо-поліграфічний центр «Київський університет», 2020. 259 с.
8. І.В. Мейзерська. Лексикографічна параметризація: структурний та функціональний аспекти (на прикладі великого тлумачного словника сучасної української мови). Вісник СумДУ. Сер.: філологія. 2007, № 1. Т.1. С. 151–155. URL: <https://essuir.sumdu.edu.ua/server/api/core/bitstreams/9967b9d5-7783-4244-a419-66f1b7c5e7b8/content>
9. Капранов Я. В., Тронь Т. В. Основи лексикографічної теорії і практики: підручник для закладів вищої освіти, які здійснюють професійну підготовку майбутніх здобувачів першого (бакалаврського) рівня вищої освіти зі спеціальності 035 Філологія. К.: Вид. центр КНЛУ, 2023. 308 с.

10. Комп'ютерна лексикографія: [монографія] / В. А. Широков. Київ : Наукова думка, 2011. 351 с.
11. Купріянов Є. В. Лексикографічна система іспанської мови. Феноменологія інтегрального опису / Є. В. Купріянов. К. : УМІФ НАНУ, 2018. 254 с.
12. Купріянов Є. Developing software for compiling electronic inflectional dictionary of the Spanish language. Біоніка інтелекту. 2023. № 99. С. 83-87.
13. Купріянов Є. Theory of L-systems as a framework for creating inflectional e-dictionaries. Біоніка інтелекту. 2022. № 98. С. 89-97.
14. Купріянов Є. В. Інтегральний опис системи іспанської мови в парадигмі теорії семантичних станів і лексикографічних систем. Вчені записки таврійського національного університету імені В. І. Вернадського. 2021. Т. 32. № 5. С. 28–34.
15. Купріянов Є. В., Остапова І. В., Яблочков М. М. Технологічні аспекти створення віртуальної лексикографічної лабораторії на базі тлумачного словника (на прикладі ВЛЛ DLE 23). 2024. № 2. С. 17–22.
16. Лінгвістично-інформаційні студії : Праці Українського мовно-інформаційного фонду НАН України: у 5 т. Т. 3 : Тлумачна лексикографія. Кн. 2 : Системна семантика тлумачних словників / В. А. Широков та ін. Київ : Український мовно-інформаційний фонд НАН України, 2018. 250 с.
17. Лінгвістично-інформаційні студії : праці Українського мовно-інформаційного фонду НАН України: у 5 т. / В. А. Широков та ін. Т. 3 : Тлумачна лексикографія. Кн. 1 : Словник української мови у двадцяти томах / В. А. Широков та ін. Київ : Український мовно-інформаційний фонд НАН України, 2018. 276 с.
18. Лінгвістично-інформаційні студії : праці Українського мовно-інформаційного фонду НАН України: у 5 т. / В. А. Широков та ін. Т. 2 : Граматичні системи. Київ. Український мовно-інформаційний фонд НАН України. 2018. 300 с.

19. Лінгвістично-інформаційні студії: праці Українського мовно-інформаційного фонду НАН України : у 5 т. / В. А. Широков та ін. Т. 5 : Віртуалізація лінгвістичних технологій. Київ. Український мовно-інформаційний фонд НАН України. 2018. 239 с.
20. Лінгвістично-інформаційні студії: праці Українського мовно-інформаційного фонду НАН України : у 5 т. / В. А. Широков та ін. Т. 4 : Корпусна та когнітивна лінгвістика. Київ. Український мовно-інформаційний фонд НАН України. 2018. 246 с.
21. Надутенко М. В. , Надутенко М. В., Яблочков М. М. Лексикографічні інструменти нового покоління: від словникової спадщини до інноваційної дидактичної практики. Слобожанський науковий вісник. Серія: Філологія. 2025. Вип. 10. С. 79-86. URL: <https://journals.spu.sumy.ua/index.php/philology/article/download/708/663/1356>.
22. Сивак О., Шадура В. Типологія електронних словників. Південний архів (філологічні науки), 2020. № 84. С. 124–131. URL: <https://pa.journal.kspu.edu/index.php/pa/article/view/715/713>.
23. Сікалюк А., Литвин С., Шевченко Ю. Передумови появи комп'ютерної лексикографії. Вісник науки та освіти. 2024. № 1(19). С. 80–89.
24. Широков В. А., Олександрук І. В. Віртуальна лексикографічна лабораторія «словник української мови» як інструмент семантичних досліджень. Мовознавство, 2019. № 5. С. 3–20. URL: <https://movoznavstvo.org.ua/component/attachments/download/1370.html>
25. Abduvahobov, G. About the concept of computer lexicography. ISJ Theoretical & Applied Science. 2021. 6(98). P. 664-668. URL: <https://www.t-science.org/arxivDOI/2021/06-98/PDF/06-98-89.pdf>.
26. Béjoint H. Modern lexicography: an introduction. Oxford; New York: Oxford University Press, 2000. 276 p.
27. Blynova, I. A., & Davydenko, V. I. (2021). Contemporary Ukrainian lexicography: status, typology, and development trends. Scientific Journal of Mykhailo Dragomanov State University of Ukraine. Series 9. Current Trends in Language

<https://sjnpu.com.ua/index.php/journal/article/view/302/262>.

28. Chemerys Y., Nakhlik O., Rysin A., Shvedova M. Normalization of a Historic Western Ukrainian orthographic system Zhelekhivka in the Ukrainian Language Reference Corpus (GRAC). Proc. of the IEEE 18th International Conference on Computer Sciences and Information Technologies (CSIT) (20 Oct. 2023. LVIV), Lviv, 2023. DOI: 10.1109/CSIT61576.2023.10324255.

29. Dorozhynska A. Formal Model of Explanatory Trilingual Terminology Dictionary. CEUR. Kharkiv, 2021. Vol. 2870. P. 475–485. URL: [https://ceur-
ws.org/Vol-2870/paper35.pdf](https://ceur-ws.org/Vol-2870/paper35.pdf).

30. Frankenberg-Garcia A. Contrasting a new AI-powered dictionary designed for on-screen reading with electronic dictionaries that have evolved from print editions. *Electronic lexicography in the 21st century (eLex 2025): Intelligent lexicography: Proceedings of the eLex 2025 conference* (Bled, 18–20 November 2025). Bled, 2025. P. 391–413.

31. Fuertes-Olivera Pedro A., Tarp S. Theory and Practice of Specialised Online Dictionaries: Lexicography versus Terminography. Berlin/Boston: De Gruyter, 2014. 271 p. URL: https://www.researchgate.net/publication/259745049_Theory_and_Practice_of_Specialised_Online_Dictionaries_Lexicography_versus_Terminography.

32. Fuertes-Olivera P.A. Making Lexicography Sustainable: Using ChatGPT and Reusing Data for Lexicographic Purposes. *Lexikos*. 2024. No 24. P. 123–140. URL: <https://www.semanticscholar.org/reader/5646f5b128d0cd544cc5e3bab7e86f3564b80e59>.

33. Fuertes-Olivera, P., Tarp, S. Innovative lexicography: new products and methods in the generative AI age. *Revista Internacional De Linguas Extranjeras / International Journal of Foreign Languages*. 2025. 24. P. 71–95. URL: <https://www.revistes.urv.cat/index.php/rile/article/view/4164/4905>.

34. Gilles-Maurice de Schryver. Lexicographers' dreams in the electronic dictionary age. URL: <https://tshwanedje.com/publications/dreams.pdf>.

35. Jakubíček M., Rundell M. The end of lexicography? Can ChatGPT outperform current tools for post-editing lexicography? Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference* (Brno, 27-29 June 2023). Brno, 2023. P. 518–533. URL: <https://www.sketchengine.eu/wp-content/uploads/102.pdf>.
36. Jumalieva G., Bektenova R., Hašková A., Gadušová Z., Kabaeva G., Kyrgyzbayev A. Peculiarities of Creating Computer Lexicography. *Journal of Education Culture and Society*. 2024. Vol. 15 No. 2. P. 403–418. URL: <https://jecs.pl/index.php/jecs/article/view/1704>.
37. Kupriianov Y. Computer-assisted Linguistic Researches on the Basis of Spanish Dictionary Text. CEUR. Lviv, 2022. Vol. 3171. P. 27–37. URL: <https://ceur-ws.org/Vol-3171/paper5.pdf>.
38. Kupriianov Y., Ostapova I., Shyrov V., Yablochkov M. Virtual lexicographic laboratory as a linguist assistant in conducting dictionary-based researches: case of VLL DLE 23. Book of Abstracts of the XX EURALEX International Congress (12-16 July 2022, Mannheim, Germany). Mannheim, 2022. P. 3–15.
39. Kupriianov Y., Ostapova I., Yablochkov M. Design of the User's Interface of Virtual Lexicographic Laboratory for Explanatory Dictionary of the Spanish Language. CEUR. Lviv, 2020. Vol. 2604. P. 121–139. URL: <https://ceur-ws.org/Vol-2604/paper10.pdf>.
40. Kuprijanov E. V. Electronic dictionary classification as problem of modern computer lexicography. Вісн. Харківського нац. ун-ту ім. В. Н. Каразіна. Сер.: Філологія, 2014. № 1127. С. 29-33.
41. Landau S. The Art and Craft of Lexicography. New York: Scribner, 1984. 370. URL: <https://archive.org/details/dictionaries00sidn/mode/2up>.
42. Language. Information. System / V. Shirokov. Ukrainian Lingua-Information Fund of the NAS of Ukraine. Kyiv: Akademperiodyka, 2021. 160 p. URL: https://ulif.mon.gov.ua/system/files/blok-shirokov_book.pdf.

43. Lew R. Dictionaries and lexicography in the AI era. *Humanities and Social Communications*. 11, 2024. DOI: <https://doi.org/10.1057/s41599-024-02889-7>.
44. Lytvyn V., Kupriianov Y., Ostapova I., Yablochkov M. Parsing technique in computer lexicography (case of Spanish dictionary DLE 23). *CEUR*. Kharkiv, 2025. Vol. 3976. P. 3–15. URL: <https://ceur-ws.org/Vol-3976/paper7.pdf>.
45. Makhachashvili R., Semenist I., Klochkov, V. AI-Enhanced Multilingual Lexicography for Digital Communication. *Proc. of the 16th International Multi-Conference on Complexity, Informatics and Cybernetics: IMCIC 2025*, 1. P. 247–253. URL: <https://elibrary.kubg.edu.ua/id/eprint/51804/1/DR701VS.pdf>.
46. Merx R., Vylomova E., Kurniawan K. Generating bilingual example sentences with large language models as lexicography assistants. *Proceedings of the 22nd Annual Workshop of the Australasian Language Technology Association (December 2–4, 2024) Canberra, 2024*. P. 64–74. URL: <https://aclanthology.org/2024.alta-1.5.pdf>.
47. Müller-Spitzer C. Textual structures in electronic dictionaries compared with printed dictionaries: a short general survey. URL: <https://ids-pub.bsz-bw.de/frontdoor/index/index/year/2014/docId/1709>.
48. Ostapova I., Šyrovkov V., Kupriianov Y., Yablochkov M. Etymological dictionary in digital environment. *Book of Abstracts of the XX EURALEX International Congress, 12-16 (July 2022, Mannheim, Germany)*. Mannheim, 2022. P. 584–593.
49. Pastor V., Amparo A. Search Techniques in Electronic Dictionaries: A Classification for Translators. URL: https://www.researchgate.net/publication/249239354_Search_Techniques_in_Electronic_Dictionaries_A_Classification_for_Translators.
50. Pleskach, V., Tumasonis, R., Yesypchuk, N., Kalmykov, O., & Akimova, A. Investigating on Creating a Software Application for an On-Screen Explanatory Dictionary. *Psycholinguistics*, 2025. 37(1), P. 224–262. DOI: <https://doi.org/10.31470/2309-1797-2025-37-1-224-262>.

51. Rundell, M. Automating the Creation of Dictionaries: Are We Nearly There? Proc. of the 16th International Conference of the Asian Association for Lexicography: Lexicography, Artificial Intelligence, and Dictionary Users (Seoul, 22-24 June 2023), 2023. P. 1–9.
52. Serdiukova O. Basic concepts of computer lexicography / Olha Serdiukova, Olga Lazareva, Inna Cherniavska. Актуальні питання гуманітарних наук : міжвуз. зб. наук. пр. / ред. кол.: М. П. Пантюк [та ін.]. Дрогобич: Гельветика, 2023. Вип. 63, т. 2. С. 165–172.
53. Shvedova M., Lukashevskyi A. Creating Parallel Corpora for Ukrainian: A German-Ukrainian Parallel Corpus (ParaRook || DE–UK) // Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024. Eds. M. Romanyshyn, N. Romanyshyn, A. Hlybovets, O. Ignatenko. Torino, Italia: ELRA and ICCL, 2024. P. 14–22. URL: <https://aclanthology.org/2024.unlp-1.3/>
54. Shvedova M., Prydvorova N., Skibina I. Normalization of Early Modern Ukrainian in GRAC: the Case of Lesia Ukrainka's Works. CEUR. Vol. 3171. P. 71–80. URL: <http://ceur-ws.org/Vol-3171/paper9.pdf>.
55. Shvedova M., Rysin A., Starko V. Handling of Nonstandard Spelling in GRAC. 2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT). 22-25 Sept. 2021. LVIV, Ukraine. Vol. 2. P. 105–108. DOI: 10.1109/CSIT52700.2021.9648834 URL: <https://ieeexplore.ieee.org/document/9648834>.
56. Shvedova M., Starko V., Rysin A. Going Beyond Standard Ukrainian: How a Corpus Informs an E-Dictionary. Medved' M. & Měchura, M. & Tiberius C., & Kosem I. & Kallas J. & Jakubíček M. & Krek S. (eds.) (2023). Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Book of abstracts. Brno, 27-29 June 2023. Brno: Lexical Computing CZ s.r.o. P. 121. URL: <https://elex.link/elex2023/book-of-abstracts/>
57. Shyrovkov V., Ostapova I., Kupriianov Y., Dorozhynska A., Yablochkov M., Verbynenko I. Terminology Dictionary Digitalization. CEUR. Lviv, 2022. Vol. 3171. P. 3–15. URL: <https://ceur-ws.org/Vol-3171/paper3.pdf>.

58. Starko V., Rysin A. Creating a POS Gold Standard Corpus of Modern Ukrainian. Proc. 2nd Ukrainian Natural Language Processing Workshop (UNLP). Association for Computational Linguistics, 2023, P. 91–95. URL: <https://aclanthology.org/2023.unlp-1.11.pdf>.
59. Starko V., Rysin A. VESUM: A Large Morphological Dictionary of Ukrainian as a Dynamic Tool. Proc. 6th Int. Conf. COLINS 2022. Vol. 3171, 71-80. URL: <https://ceur-ws.org/Vol-3171/paper8.pdf>.
60. Tarp S., Nomdedeu-Rull A. Who has the last word? Lessons from using ChatGPT to develop an AI-based Spanish writing assistant. *Círculo de Lingüística Aplicada a la Comunicación*. 2024. 97. P. 309-321. URL: <https://revistas.ucm.es/index.php/CLAC/article/download/91985/4564456568614/4564456701906>.
61. Vakaliuk T., Chernysh O., Figen K., Ferrer-Conill R. Examining the Context of E-Dictionary Use in Language Studies. URL: <https://www.atlantispress.com/proceedings/icntllsc-21/125957104>.
62. Widmann T. A Pipeline for Automated Dictionary Creation with Optional Human Intervention. *Electronic lexicography in the 21st century (eLex 2025): Intelligent lexicography: Proceedings of the eLex 2025 conference* (Bled, 18-20 November 2025). Bled, 2025. P. 33–43.
63. Yablochkov M., Kupriianov Y., Shyrovkov V., Ostapova I. Digital Toolkit to Develop Research Potential of Explanatory Dictionary (Case of Spanish Language Dictionary). CEUR. Kharkiv, 2021. Vol. 2870. P. 388–401. URL: <https://ceur-ws.org/Vol-2870/paper28.pdf>.
64. Zgusta L. Manual of lexicography. Praha: Academia, 1971. 360 p. URL: <https://archive.org/details/manualoflexicogr0000zgus/mode/2up>.

КОРИСНІ ПОСИЛАННЯ

Онлайн-словники

1. Англійсько-український юридичний словник. URL: <https://1546.slovaronline.com/>
2. Англійсько-українські словники. URL: <https://e2u.org.ua/>.
3. Великий електронний словник української мови (ВЕСУМ). URL: <https://vesum.nlp.net.ua/>.
4. Великий тлумачний словник української мови. URL: <https://1531.slovaronline.com/>.
5. Відкритий словник. URL: <http://www.mova.info/wordlist.aspx?l1=179>.
6. Горох – українські словники. URL: <https://goroh.pp.ua/>.
7. Граматичний словник української мови (MOVA.info). URL: <http://www.mova.info/grmasl.aspx>.
8. Електронний словник мови Тараса Шевченка. URL: <http://www.mova.info/cfqsh.aspx>.
9. Енциклопедія сучасної України. URL: <https://esu.com.ua/>.
10. Енциклопедія українознавства. URL: <http://litopys.org.ua/encycl/euii.htm>.
11. Інтегрована лексикографічна система «Словники України» онлайн. URL: <https://lcorp.ulif.org.ua/dictua/>.
12. Короткий українсько-сербський словник сполучуваності слів. URL: <http://www.mova.info/Page3.aspx?l1=65>.
13. Лінгвістична енциклопедія. URL: <http://www.mova.info/Page3.aspx?l1=65>.
14. Орфографічний словник української мови. URL: <https://559.slovaronline.com/>.
15. Семантичний словник української мови. URL: <http://www.mova.info/semvoc.aspx?l1=193>.
16. Словник жестової мови. URL: <https://spreadthesign.com/uk.ua/search/>.
17. Словник іншомовних слів. URL: <https://126.slovaronline.com/>.

18. Словник порівнянь Юрія Андруховича та Оксани Забужко. URL: <http://www.mova.info/Slovnyk.aspx>.
19. Словник синонімів української мови. URL: <https://1675.slovaronline.com/>
20. Словник українських прийменникових колокацій. URL: http://www.mova.info/sl_spol.aspx.
21. Словник української мови в 20-ти томах. URL: <https://sum20ua.com/?wordid=0&page=0>.
22. Словник.UA: портал української мови та культури. URL: <https://slovnyk.ua/>.
23. Словники й енциклопедії. URL: <https://slovnyk.me/>.
24. Список словників української мови. URL: <https://litcentr.in.ua/index/0-55>.
25. Сучасний лінгвістичний словник / А. Загнітко. Вінниця : ТВОРИ, 2020. 920 с. URL: https://r.donnu.edu.ua/bitstream/123456789/469/1/Словник_RD-Суч_лінгв_словник%20%281%29.pdf.
26. Сучасний українсько-англійський юридичний словник. URL: <https://18.slovaronline.com/>.
27. Тезаурус з комп'ютерної лексикографії. URL: <http://www.mova.info/Page3.aspx?l1=188&volid=1>.
28. Тезаурус з лінгвістичної термінології. URL: http://www.mova.info/mov_thes.aspx?l1=68.
29. Універсальний словник-енциклопедія. URL: <https://558.slovaronline.com/>.
30. Фразеологічний словник української мови. URL: <https://123.slovaronline.com/>.
31. Cambridge Dictionary. URL: <https://dictionary.cambridge.org/>.
32. Collins Online Dictionary. URL: <https://www.collinsdictionary.com/>
33. Encyclopædia Britannica. URL: <https://www.britannica.com/topic/Encyclopaedia-Britannica-English-language-reference-work>.
34. English Picture Dictionary. URL: <https://dictionary.langeek.co/>

35. Longman Dictionary of Contemporary English Online. URL: <https://www.ldoceonline.com/>.
36. Merriam-Webster's Dictionary. URL: <https://www.merriam-webster.com/>.
37. Multitran. URL: <https://www.multitran.com/>.
38. Online Etymology Dictionary. URL: <http://etymonline.com/ua>.
39. Online Oxford Collocation Dictionary. URL: <https://m.freecollocation.com/>.
40. Oxford English Dictionary. URL: <https://www.oed.com/?tl=true>.
41. Oxford Learner's Dictionaries. URL: <https://www.oxfordlearnersdictionaries.com/>.
42. Reverso Dictionary. URL: <https://dictionary.reverso.net/>.
43. The American Heritage Dictionary of the English Language. URL: <https://ahdictionary.com/>.
44. The Britannica Dictionary. URL: <https://www.britannica.com/dictionary>.
45. The Thesaurus for the Real World. URL: <https://www.thesaurus.com/>.
46. Visual Dictionary Online. URL: <https://www.visualdictionaryonline.com/>.
47. WordNet. URL: <https://hyperdic.net/en/doc/wordnet>.
48. Wordnik. URL: <https://www.wordnik.com/>.
49. World History Encyclopedia. URL: <https://www.worldhistory.org/index/>.
50. The World Atlas of National Online Encyclopedias. URL: <https://wanoe.encyclopedia.kyiv.ua/>

Текстові корпуси

51. Корпус ГРАК та інші текстові корпуси української та слов'янської мов. URL: <https://uacorpus.org/inshi-korpusi-ukrayinskoyi-movi-ta-slovyanskih-mov>.
52. Bank of English (BoE). URL: <https://www.bankofengland.co.uk/>.
53. British Academic Written English Corpus (BAWE). URL: <https://www.sketchengine.eu/british-academic-written-english-corpus/>.
54. British Law Report Corpus. URL: <https://www.sketchengine.eu/blarc-british-law-reference-corpus/>.
55. British National Corpus (BNC). URL: <https://www.english-corpora.org/bnc/>.

56. Brown University Standard Corpus of Present-Day American English. URL: <https://www.sketchengine.eu/brown-corpus/>.
57. Corpus of Contemporary American English (COCA). URL: <https://www.english-corpora.org/coca/>.
58. EuroParl Parallel Corpus. URL: <https://www.statmt.org/europarl/>.
59. JRC-Acquis Multilingual Parallel Corpus. URL: https://joint-research-centre.ec.europa.eu/language-technology-resources/jrc-acquis_en.
60. Medical English Corpus. URL: <https://www.sketchengine.eu/english-medical-corpus/>.
61. Open American National Corpus (ANC). URL: <https://anc.org/>.
62. OPUS (Open Parallel Corpus). URL: <https://opus.nlpl.eu/>.
63. Oxford English Corpus. URL: <https://www.sketchengine.eu/oxford-english-corpus/>.
64. ParaCrawl Corpus. URL: <https://paracrawl.eu/>.
65. ParaRook Corpora. URL: <https://uacorporus.org/poshuk-u-graku/paralelni-korpusi-pararook>.
66. ParlaMint Corpora. URL: <https://www.clarin.eu/parlamint>.

Додаткові Інтернет-ресурси

67. Український мовно-інформаційний фонд НАН України: український лінгвістичний портал. URL: <https://www.ulif.org.ua/>.
68. Elex Conferences: electronic lexicography in the 21st century. URL: <https://elex.link/>.
69. Euralex Conferences. URL: <https://euralex.org/conferences/>.
70. Internet Journal on Lexicography. Oxford University Press. URL: <https://academic.oup.com/ijl/issue/38/4>.
71. Journal of ASIALEX. University of Toronto Press. URL: <https://utppublishing.com/journal/LEXI>.
72. List of lexicography publications. URL: <https://euralex.org/list-of-lexicography-publications/>.

ДОДАТКИ

Д.1. Лексикографічні параметри словникової статті

Тип параметра	Що відображає	Приклади подання
Формальні	Графічна й фонетична форма, наголос, граматичні особливості, словозміна, варіанти	<i>СУМ'ЯТНИЙ</i> , <i>a</i> , <i>e</i> , <i>розм.</i> ; <i>acknowledge</i> [ək'nɒlɪʒ]; <i>water</i> /'wɔ:tə/
Етимологічні	Походження слова, етимон, мова-посередник	<i>dieta</i> (грец. <i>diata</i> – <i>спосіб життя</i>); <i>water</i> ← <i>Old English wæter</i> ← <i>Proto-Indo-European</i>
Атрибутивні	Стиль, галузь, ареал, ідіолект, школа	<i>АДСОРБЕНТ</i> , <i>фіз.</i> , <i>хім.</i> ; <i>thou2 informal</i> ; <i>ЄВГЕНИКА</i> (термін <i>Ф. Гальтона</i>)
Асоціативні	Семантичне оточення, родо-видові й партитивні відношення	<i>microfauna</i> → <i>macrofauna</i> ; <i>ремарка</i> <i>див. примітка, помітка</i>
Прагматичні	Хронологія, уживаність, комунікативні обмеження	<i>АЕРОПЛАН</i> <i>заст.</i> ; <i>КОБИЛА</i> , <i>перен.</i> , <i>вульг.</i> ; <i>cock</i> <i>archaic</i>
Інтерпретаційні	Значення слова: сигніфікативне, конотативне, денотативне	<i>run</i> : « <i>to move swiftly on foot</i> »; <i>конот.</i> – « <i>керувати, очолювати</i> »
Ілюстративні	Приклади вживання у контексті	<i>acercar... Los dos países se han acercado políticamente</i> ; <i>to break the ice</i>

Д.2. Метамова у різних типах словників

Тип словника	Формули для опису значення	Ремарки	Пояснення	Скорочення	Символи
Тлумачний (СУМ, OED)	«Предмет, що...», «Процес, при якому...», «Той, хто...»	<i>заст.</i> , <i>поет.</i> , <i>розм.</i>	уточнення: «Уживається тільки в минулому часі»	<i>імен.</i> , <i>чол.</i> , <i>жін.</i>	нумерація значень (1., 2.), ♦ для прикладів
Перекладний (англ.-укр., укр.-фр.)	передача значення через еквіваленти: <i>house</i> – <i>дім</i> , <i>будинок</i>	<i>Ат.</i> , <i>Вр.</i>	коментарі про відтінки значення	<i>n.</i> , <i>adj.</i> , <i>v.</i>	стрілки (→) для вказівки на близькі слова
Термінологічний (мед., фіз.)	«Прилад для...», «Явище, що...»	<i>біол.</i> , <i>фіз.</i> , <i>мат.</i>	дефініції з уточненнями	<i>pl.</i> , <i>sg.</i> , <i>abbr.</i>	грецькі/лат. символи, формули
Фразеологічний	«У переносному значенні...», «Вживається для позначення...»	<i>ірон.</i> , <i>жарт.</i> , <i>книжн.</i>	тлумачення стійкого виразу	<i>фразеол.</i> , <i>присл.</i>	лапки, тире, стрілки до синонімічних виразів

Синонімічний / Антонімічний	перелік варіантів: <i>гарний</i> – <i>красивий</i> , <i>вродливий</i> , <i>симпатичний</i>	<i>поет.</i> , <i>розм.</i> , <i>заст.</i>	уточнення про відтінки значення	<i>syn.</i> , <i>ant.</i>	стрілки (→) для позначення відповідності

Д.3. Порівняльна таблиця лексикографічних середовищ

Середовище	Основні форми	Характеристика	Переваги	Недоліки
Традиційне (друковане)	Тлумачні, орфографічні, перекладні, етимологічні словники	Статичний друкований формат, фіксована структура	Авторитетність, надійність, культурна цінність	Важко оновлювати, обмежений обсяг, громіздкість
Електронне (офлайнове)	CD/DVD- словники, програмні продукти	Цифровий формат, пошук за ключовими словами, мультимедіа	Швидкість пошуку, компактність, зручність	Потребує носія, відсутність оперативного оновлення
Онлайнове	Веб-сайти, гіпертекстові бази, онлайн- версії друкованих словників	Доступність через Інтернет, інтерактивність, оновлюваність	Широка аудиторія, багатомовність, швидке редагування	Залежність від Інтернету, нерівна якість
Корпусне	Національні та спеціалізовані корпуси текстів	База анотованих текстів, статистика слововживання, контексти	Емпірична достовірність, реальний ужиток, аналіз граматичних зв'язків	Складність для новачків, потреба спецінтерфейсу
Інтегроване	Поєднання словника, корпусу, тезауруса, граматики	Багатофункціональна система, комплексний опис мови	Повнота, універсальність, аналітичні можливості	Складність реалізації, великі ресурси
Динамічне (соціально- мережеве)	Краудсорсингові словники, інтерактивні сервіси (Urban Dictionary, Вікісловник)	Контент створює спільнота, постійне оновлення	Оперативність, відображення живої мови	Відсутність академічності, можливі помилки

Д.4. Стандартизовані формати репрезентації даних.

Формат	Призначення	Переваги	Обмеження
XML TEI (Text Encoding Initiative)	Репрезентація словникових статей та обмін структурованими лінгвістичними даними	Стандартизовані машиночитані структури; висока інтероперабельність;	Складний у розробці; громіздкі структури; незручний для

		гнучкість; можливість глибокого аналізу	неієрархічних даних
TEI Lex-0	Спрощений профіль TEI для лексикографії	Уніфікує різні підходи TEI; забезпечує кращу сумісність між інструментами; чіткі та мінімалістичні правила	Порівняно новий стандарт; тяжіє до мінімалізму (менше свободи у розмітці)
LMF (Lexical Markup Framework, ISO 24613)	Модель опису лексичних ресурсів, стандартизація структур словникових баз	Міжнародний ISO-стандарт; ідеальний для складних лексичних БД; розділність	Вимагає складного моделювання; XML-реалізації іноді надто технічні
OntoLex-Lemon (W3C)	Представлення лексичних даних у форматах семантичного вебу; пов'язання словника з онтологіями	Добре працює в RDF/Linked Data; ідеальний для інтеграції словників; підтримка BFO, WordNet, Wikidata	Складний поріг входу; потребує знання семантичних технологій
RDF (Resource Description Framework)	Представлення лексичних і семантичних даних у вигляді зв'язаних триплетів	Гнучкість; ідеальний для зв'язаних даних; легко інтегрується з онтологіями	Потребує спеціальних інструментів; крива навчання
JSON-LD	Легковагова репрезентація лексичних та семантичних структур у вебсередовищі	Легко інтегрується у вебдодатки; більш компактний, ніж XML; сумісний із Linked Data	Менш формальний для складних лексичних структур; потребує узгоджених схем
CSV / TSV	Прості таблиці для експорту/імпорту даних, частот, списків лем	Простота; універсальність; підтримка всіма інструментами; ідеальний для швидкої обробки	Немає ієрархії; немає семантичних зв'язків; непридатний для складних словникових статей

Д.5. Характеристика мережевих протоколів та API

Протокол / API	Призначення	Переваги	Обмеження
REST API	Стандартна модель обміну даними між словниками, корпусами та вебсервісами	Простота; сумісність з HTTP; легка інтеграція; підтримує JSON/XML	Обмежена гнучкість у складних транзакціях; немає вбудованого типування
GraphQL	Запити до лексикографічних та корпусних даних з кастомною структурою відповіді	Отримуєш рівно ті поля, що потрібні; ефективність; ідеально для складних словників (з вкладеннями)	Потребує складнішої серверної реалізації; менш широко підтримується
SPARQL	Доступ до онтологій та лексичних баз у форматі RDF/Linked	Ідеальний для семантичних словників;	Крута крива навчання;

	Data (напр., OntoLex-Lemon)	складні запити; інтеграція зі знаннями Вікіданих	потребує RDF-схем
OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting)	Обмін метаданими між словниками, корпусами, лінгвістичними репозитаріями	Стандарт у цифровій гуманітаристиці; автоматизує синхронізацію великих ресурсів	Старіший протокол; обмежений типами даних; не для інтерактивних запитів
Lexical Markup Framework Services (LMF-based APIs)	Обмін лексичними даними відповідно до ISO 24613	Стандартизований обмін між лексикографічними платформами; структурованість	Складний; потребує спеціальних моделей даних
W3C Web Annotation API	Обмін анотаціями корпусів і словників (коментарі, позначки, лінгвістичні теги)	Добре підходить для глосаріїв і корпусів; інтероперабельність	Вимагає RDF-сумісних інструментів
OpenSearch API	Пошук у лексикографічних та корпусних базах через стандартизовані запити	Простота; інтеграція з браузерами; використовується в великих словникових порталах	Обмеженість формату; слабка підтримка складних лінгвістичних запитів
Microdata / Schema.org (Dictionary, DefinedTerm)	Презентація словникових даних у вебсторінках з машинною доступністю для Google/Bing	Підвищує видимість словника; оптимізація для пошуку; легка реалізація	Підтримує лише базову структуру словникової статті
Webhook-інтерфейси	Автоматичне оновлення словників/корпусів при зміні даних	Синхронність; зручні для динамічних ресурсів	Потребує сервера, що постійно доступний

Д.6. Внесок окремих дисциплін у комп'ютерну лексикографію

Дисципліна	Методи / інструменти	Що саме надає для лексикографії
Етимологія	Порівняльно-історичний метод, реконструкція праформ, аналіз запозичень	Відтворення походження слова, історії його значення та форм; допомагає укласти етимологічні словники й коментарі до статей
Діалектологія	Польові дослідження, анкетування, картографування мовних явищ	Виявлення територіальних варіантів лексики; створення діалектних словників; збереження локальних мовних одиниць
Стилістика	Функціонально-стилістичний аналіз, зіставлення регістрів, опис експресивності	Класифікація слів за стилістичним маркуванням (нейтральні, розмовні, книжні); створення словників синонімів, антонімів, фразеологізмів
Семантика	Компонентний аналіз, поля значень, семантичні карти	Виявлення структури значень слова; опис полісемії, омонімії; побудова тезаурусів і тематичних словників

Фонологія	Визначення фонемного складу, фонематичний аналіз, мінімальні пари	Точна передача вимови, орфографічних і орфоепічних характеристик у словниках
Морфологія	Морфемний аналіз, парадигматика словозміни, моделювання словотвірних типів	Подання словозміни у словниках; відображення продуктивних та непродуктивних словотвірних моделей
Синтаксис	Синтаксичний аналіз, моделювання словосполучень і конструкцій	Ілюстрація сполучуваності слів; створення словників керування та граматичних словників
Психолінгвістика	Експериментальні методи (асоціативні тести, час реакції), корпусні дослідження	Дані про частотність і асоціативні зв'язки; побудова частотних та асоціативних словників
Соціолінгвістика	Соціальні опитування, аналіз варіантів мови за групами	Включення соціально маркованої лексики; опис жаргонів, молодіжних слів, професіоналізмів
Структурна лінгвістика	Аналіз парадигматичних та синтагматичних відношень, моделювання систем	Побудова ієрархій і мережевих структур у словниках; перехід від описових словників до системних; розвиток тезаурусів

Д.7. Внесок лінгвістичних дисциплін у створенні словників

Дисципліна	Методи / інструменти	Що надає для лексикографії	Приклади словників
Етимологія	Порівняльно-історичний метод, реконструкція праформ, аналіз запозичень	Походження слова, історія значень, зв'язки між мовами	<i>Етимологічний словник української мови</i> (О. Мельничук та ін.), словники Фасмера, Черниха
Діалектологія	Польові дослідження, анкетування, діалектні карти	Виявлення регіональних варіантів, збереження локальної лексики	<i>Атлас української мови</i> , <i>Словник бойківських говірок</i> , <i>Поліський словник</i>
Стилістика	Функціонально-стилістичний аналіз, зіставлення реєстрів	Маркування стилістичних шарів (нейтр., розм., книжн., поет.)	<i>Словник синонімів</i> , <i>Словник антонімів</i> , <i>Словник омонімів</i> , <i>Словник паронімів</i>
Семантика	Компонентний аналіз, семантичні поля, карти значень	Систематизація значень, розмежування полісемії й омонімії	<i>Тлумачний словник української мови</i> (СУМ), <i>Thesaurus Linguae Latinae</i> , тематичні словники
Фонологія	Фонематичний аналіз, мінімальні пари	Точна вимова, орфоепічні норми	<i>Орфоепічний словник української мови</i>
Морфологія	Морфемний аналіз, парадигматика словозміни	Подання словозміни, словотвірних моделей	<i>Граматичний словник української мови</i>

			<i>мови, словники афіксів</i>
Синтаксис	Аналіз конструкцій, словосполучень, керування	Ілюстрація сполучуваності слів	<i>Словник керування української мови (Кочерга), Collins COBUILD Grammar Patterns</i>
Психолінгвістика	Експериментальні методи (асоціативні тести, частотність)	Асоціативні поля, частотність, психолінгвістичні дані	<i>Український асоціативний словник (П. Селігей та ін.), частотні словники (Нац. корпус мови)</i>
Соціолінгвістика	Опитування, аналіз соціально маркованої лексики	Жаргони, молодіжна мова, професіоналізми	<i>Словник молодіжного сленгу, Словник українського жаргону (Тараненко)</i>
Структурна лінгвістика	Аналіз парадигматичних і синтагматичних відношень	Побудова ієрархій, системних описів	<i>Тезаурус української мови (Широков), WordNet, EuroWordNet</i>

Д.8. Реалізація принципу розмічення і структурного принципу

Приклад словника / проекту	Принцип реалізації	Ключові особливості
Turbolex (Словник гідротурбінних термінів, укл. Є. В. Купріянов)	Принцип розмічення (HTML + CSS)	– Словникові статті подані як HTML-сторінки з блоковою версткою; – Використання гіпертексту для структурування статей; – Реєстри (алфавітний, тематичний) реалізовані як окремі сторінки; – Тезаурусна організація матеріалу з візуалізацією термінологічних зв'язків.
ВЛЛ СУМ-20 (Український мовно-інформаційний фонд НАН України)	Структурний принцип (ієрархічна модель даних)	– Словникова стаття зберігається як набір елементів у структурі бази даних; – Можливість редагування окремих елементів; – Пошук здійснюється не лише за реєстром, а й за іншими компонентами (значення, граматики, приклади); – Дозволяє створювати нові елементи в межах заданої структури.
WordNet (США)	Структурний принцип (семантична мережа)	– Лексичні одиниці представлені як «синсети» (множини синонімів); – Організація за семантичними відношеннями (гіпонімія, антонімія, меронімія тощо);

		– Дас змогу виконувати пошук не лише за словом, а й за семантичною категорією чи зв'язками.
EuroWordNet (Європейський багатомовний проєкт)	Структурний принцип (інтегрована база даних)	– Об'єднання національних WordNet'ів різних мов; – Єдина міжмовна інтерлінгва для інтеграції; – Підтримка багатомовного пошуку та крослінгвістичних досліджень; – Широке застосування в комп'ютерній лінгвістиці та NLP.

Д.9. Зразок аналізу електронного словника

Назва словника	Входи (алфавітний, тематичний, історія перегляду тощо)	Пошукові індекси (частина мови, галузь, етимологія, інші)	Принцип побудови (розмічення, структурний, комбінований)	Переваги	Недоліки
Merriam-Webster Online	- Алфавітний реєстр - Пошук через рядок запиту - Список переглянутих слів (History)	- Частина мови - Етимологія - Приклади вживання - Синонімія/антонімія	Комбінований: словникові статті структуровані, з розміткою HTML та індексами для пошуку	Швидкий доступ, розширена система індексів, інтеграція з тезаурусом, аудіо-вимова	Обмежений доступ до повних історичних статей, частина функцій доступна лише в платній версії

Д.10. Об'єднана система принципів створення електронного словника

Рівень	Принцип	Зміст
Лінгвістичний	Системності	Словник розглядається як цілісна лексикографічна система, де кожна одиниця пов'язана з іншими за типами значень, відношеннями та категоріями.
	Концептуального моделювання	Побудова концептуальної моделі, що формалізує словник як систему понять та зв'язків (лексема – значення – граматична інформація – контекст).
	Семантичної узгодженості	Відображення узгодженості граматичної і лексичної інформації, представленої у різних форматах (текст, таблиця, БД, теґована розмітка).
	Полісистемності	Опис лексем у різних функціональних системах мови: літературній, діалектній, професійній, сленговій тощо.
Інформаційний	Формалізації	Представлення даних у чітко визначених структурах (БД, XML, JSON, RDF).

	Інтероперабельності	Забезпечення сумісності словникових даних з іншими лексикографічними та корпусними ресурсами.
	Розділності	Розділення даних за рівнями: лексичним, граматичним, семантичним, прагматичним.
Технологічний	Автоматизації	Використання алгоритмів автоматичного збирання, нормалізації та оновлення словникових статей.
	Динамічності	Можливість постійного оновлення словника за рахунок моніторингу нових текстів і джерел.
	Інтерактивності	Реалізація користувацького інтерфейсу, що дозволяє шукати, фільтрувати, зіставляти й експортувати дані.
	Візуалізації	Подання лексичних зв'язків у вигляді графів, карт понять, мереж значень.

Д.11. Порівняльна характеристика форматів PDF і HTML

Параметр	PDF	HTML
Призначення	Збереження фіксованого макета (репліка друкованої сторінки)	Веб-представлення, гіпертекстова навігація, інтерактивність
Структура	Візуально-сторінкова; структура прихована за макетом	Семантична/логічна (теги), легко виражається в DOM
Відтворення макета	Фіксоване (точне відтворення шрифтів, колонок, нумерації)	Адаптивне; макет керується стилями (CSS) і залежить від пристрою
Пошук і індексація	Пошук по тексту можливий, але часто обмежений; складніше виділяти структурні елементи	Пошук за елементами/тегами; добре індексується пошуковими системами
Машиночитаність	Низька (якщо немає додаткової семантики або XML-розмітки)	Висока (за наявності структурних тегів та семантичних атрибутів)
Оновлення контенту	Потребує регенерації файлу; менш зручне для частих змін	Динамічне оновлення; зміни вмісту без перевидавання
Інтерактивність	Обмежена (статичний документ; можлива лише базова навігація)	Повна (гіперпосилання, фільтри, скрипти, мультимедіа, інтерактивні вибірки)
Стилізація	Статична (вбудоване форматування); обмежені можливості адаптації	Гнучка (CSS), підтримка адаптивного дизайну та різних тем відображення
Доступність (accessibility)	Можлива, але залежить від якості тегування; складніше підтримувати адаптацію для екранів читання	Краща можливість забезпечення доступності (ARIA, семантичні теги)
Метадані / семантична розмітка	Обмежені; метадані можуть бути вкладені, але не стандартизовано як для словників	Легко інтегруються (microdata, RDFa, schema.org), можна прив'язати до TEI/LMF
Мультимедіа	Можливе (вбудовані зображення), але інтерактивність обмежена	Натуральна підтримка аудіо, відео, інтерактивних елементів

Типові застосування у лексикографії	Архівні копії, наукові репліки, завантажувані довідки, друк-підготовка	Онлайн-словники, інтегровані платформи, віртуальні лексикографічні лабораторії
Переваги (кратко)	Точне відтворення макета; зручність для друку та цитування	Гнучкість, інтерактивність, машиночитаність, легкість оновлення
Обмеження (кратко)	Погана машинна обробка; складне оновлення; обмежена інтерактивність	Якщо без додаткової семантики — слабка лексикографічна розмітка; потребує веб-інфраструктури

Д.12. Порівняльна характеристика стандартів TEI та LMF

Параметр	TEI (Text Encoding Initiative)	LMF (Lexical Markup Framework, ISO 24613)
Призначення	Загальний стандарт для розмітки текстів гуманітарного спрямування, включно з лексикографічними ресурсами.	Міжнародний стандарт ISO для моделювання лексичних ресурсів і забезпечення сумісності між ними.
Тип ресурсів	Лексикографічні, корпусні, архівні, літературні та інші текстові ресурси.	Лексикони, термінологічні бази, корпусні лексичні ресурси, машинні словники.
Структурна одиниця	<entry> — словникова стаття, що містить формальну, граматичну, семантичну й ілюстративну частини.	<LexicalEntry> — базовий об'єкт, що включає форму, значення, приклади, варіанти.
Ключові елементи	<form>, <orth>, <sense>, <def>, <cit>, <gramGrp>, <etym>, <usg> тощо.	<Form>, <Sense>, <Example>, <Equivalent>, <Context>, <Synset> тощо.
Рівень формалізації	Високий, але допускає варіативність (гнучкість у виборі тегів та атрибутів).	Дуже високий, із чітко визначеними структурами та обов'язковими зв'язками між об'єктами.
Фокус опису	Людиночитаний формат, орієнтований на гуманітарні дослідження та архівування текстів.	Машиночитаний формат, орієнтований на автоматизовану обробку, інтеграцію й обмін даними.
Сумісність з іншими системами	Добре інтегрується з HTML, RDF, JSON-LD через розділі TEI Publisher.	Використовується у проєктах NLP, машинного перекладу, корпусних дослідженнях.
Приклади проєктів	СУМ-20 (Україна), Oxford English Dictionary (TEI XML), Slovniky.cz, Wiktionary TEI XML.	Global WordNet, LLOD (Linguistic Linked Open Data), PanLex, FrameNet.
Переваги	Гнучкість, відкритість, можливість поєднання з іншими гуманітарними форматами.	Стандартизованість, придатність до автоматичного аналізу та обміну лексичними ресурсами.
Недоліки	Висока варіативність може призвести до відсутності уніфікації між словниками.	Менша гнучкість для гуманітарних описів; потребує точного дотримання моделі ISO.

Д.13. Порівняльна характеристика форматів зберігання

лексикографічної інформації

Формат	Тип призначення	Основні переваги	Основні недоліки	Приклади використання в лексикографії
PDF	Статичний формат представлення друкованих або видавничих матеріалів	Збереження автентичного вигляду друкованого словника; можливість офлайн-доступу; простота поширення	Не підтримує машинну обробку; відсутня структура даних; складність інтеграції у бази даних	Електронні копії друкованих словників (наприклад, «Словник української мови» у 20 т.)
HTML	Формат гіпертекстової розмітки для веб-представлення словникових даних	Інтерактивність, підтримка гіперпосилань, зручність для користувачів; легко поєднується з CSS та JavaScript	Відсутність чіткої лексикографічної структури без додаткових схем; залежність від браузерів	Cambridge Dictionary Online, Словник української мови онлайн, Glosbe
XML	Формат для структурування, обміну та зберігання даних	Машиночитаність; чітка структуризація елементів словникової статті; можливість конвертації в інші формати; сумісність зі схемами TEI, LMF	Потребує високого рівня технічної підготовки; громіздкий для ручного редагування	TEI-based проекти, Словник української біологічної термінології, проекти УМІФ
Бази даних (SQL / NoSQL)	Формати організації, збереження та взаємодії структурованих даних	Ефективний пошук; можливість оновлення; масштабованість; інтеграція з вебінтерфейсами	Вимагає складної архітектури й технічного супроводу; менш наочний для кінцевого користувача	СУМ-20 онлайн, Словники України on-line, Lingvo Live, Turbolex

Д.14. Рівні лексикографічного моделювання

Рівень	Характеристика	Основні операції / принципи	Приклад	Тип словника / модель
1. Макрорівень	Структура всього словника, його типологія, обсяг, порядок подання одиниць	Визначення типу (тлумачний, перекладний, термінологічний тощо); алфавітне чи тематичне впорядкування	Визначення, що словник буде двомовним українсько-англійським технічним	Проект макроструктури словника
2. Медіорівень	Організація словникових статей за різними принципами	Групування за темами, частинами мови, семантичними відношеннями	Усі іменники зі спільним коренем <i>рух</i> - (рух, рухати, відношеннями)	Мережа пов'язаних статей або словотвірне гніздо

		(омонімія, полісемія, синонімія, антонімія, гіпонімія, словотвір)	рухомий, рухливість)	
3. Мікрорівень	Внутрішня структура окремої словникової статті	Формалізація значень, граматичних, стилістичних, дериваційних та перекладних характеристик	У статті: <i>рух</i> — 1) <i>переміщення</i> ; 2) <i>суспільне явище</i>	Формат словникової статті з мітками, позначками, перекладом
4. Метарівень	Принципи метамови та системи позначень, якими описується словникова інформація	Визначення лейблів, символів, скорочень, семантичних тегів	Використання <POS=poun> або [політ.] для позначення частини мови чи сфери вживання	Формальна модель словникової бази даних
5. Технологічний рівень	Технічна реалізація словникової бази	Кодування в XML, JSON, база даних; створення API	Модель XML-словника з тегами , ,	Електронний словник або корпусна база

Д.15. Порівняльна таблиця типів комп'ютерних словників

Критерій	«Словники України online»	Wiktionary (Вікісловник)	ABBYY Lingvo
1. Функціональний тип	Комбінований: граматичний, фразеологічний, синонімів, антонімів	Універсальний, енциклопедично-лексикографічний (включає значення, етимологію, приклади, переклади)	Перекладний, термінологічний, навчальний
2. Структурний тип	Абетковий	Абетковий + гіперпосилання між статтями	Абетковий, із вбудованими тематичними групами
3. Технологічна реалізація	Вебінтерфейс, база даних на сервері УЛПФ НАН України	Вебінтерфейс на платформі MediaWiki, хмарне збереження даних	Локальний (desktop) та онлайнний застосунок із серверною синхронізацією
4. Користувацьке призначення	Загальнодоступний ресурс для української мови	Масовий, краудсорсний, мультимовний ресурс	Професійний лексикографічний інструмент і навчальний продукт
5. Мовне охоплення	Одномовний (українська)	Мультимовний (усі основні мови світу)	Багатомовний (українська, англійська, німецька, французька тощо)

6. Спосіб створення	Експертно укладений, частково автоматизований	Краудсорсинг, відкрите редагування спільнотою	Експертне укладання, закриті авторські бази
7. Ступінь інтерактивності	Низька: лише пошук	Висока: редагування, коментарі, обговорення	Середня: користувацькі налаштування, інтеграція з іншими додатками
8. Доступність / Ліцензія	Безкоштовний, відкритий	Повністю безкоштовний, відкритий (Creative Commons)	Комерційний продукт
9. Актуалізація даних	Періодичне оновлення фахівцями	Безперервна спільнотна актуалізація	Регулярні версійні оновлення
10. Тип користувацького інтерфейсу	Простий пошуковий інтерфейс	Гіпертекстовий, з навігацією між мовами	Інтерактивний, мультимедійний (аудіо, приклади, вправи)

Д.16. Порівняльна характеристика типологій комп'ютерних словників

Типологія	Критерій класифікації	Основні типи словників	Ключові особливості	Приклади
Функціональна	Призначення словника, ціль використання	• Тлумачні • Перекладні • Термінологічні • Енциклопедичні • Фразеологічні, частотні, етимологічні тощо	Визначає, яку функцію виконує словник – пояснення значення, переклад, фахову інтерпретацію або довідкову інформацію.	<i>CYM-20 online, Oxford English Dictionary, e2u.org.ua, COSMOS, Wikipedia</i>
Структурна	Організація словникового матеріалу	• Однорівневі (простий перелік) • Багаторівневі (з мікро-, медіо-, макрорівнями) • Алфавітні, тематичні, ономасіологічні • Тезауруси, частотні, онтологічні бази	Описує внутрішню будову словника, принцип упорядкування та зв'язки між одиницями. Визначає модель словникової статті та структуру бази даних.	<i>Словники України online, WordNet, EuroWordNet, FrameNet</i>
Технологічна	Технічна реалізація та платформа	• Локальні (офлайн) • Онлайн • Гібридні • Мобільні додатки • Краудсорсингові • AI-збагачені	Визначає спосіб зберігання, доступу та інтеракції: де розміщено дані, як забезпечено пошук, оновлення, інтерактивність і мультимедійність.	<i>ABBYY Lingvo, Lingvo Live, Wiktionary, Reverso Context, DeepL Glossary</i>

Д.17. Класифікація середовищ укладання електронних словників

Тип середовища	Характеристика	Приклади
1. Професійні лексикографічні системи (Digital Writing Systems, DWS)	Призначені для створення багатотомних і багатоаспектних словників. Забезпечують роздільність, контроль версій, командну роботу, інтеграцію з корпусами.	<i>TLex Suite, IDM DPS, Lexique Pro, TshwaneLex</i>
2. Віртуальні лексикографічні лабораторії (ВЛЛ)	Онлайн-платформи для колективної роботи над словниковими проектами; підтримують спільне редагування, коментування, експортування даних у різні формати.	<i>ELEXIS Platform, Sketch Engine Lexonomy, WebLicht</i>
3. Академічні та навчальні середовища	Використовуються у ВНЗ для навчання лексикографії, проєктної діяльності студентів; мають спрощений інтерфейс і базові функції для створення невеликих словників.	<i>Lexonomy (навчальні акаунти), Lexique Pro (освітні версії)</i>
4. Онлайн-конструктори та відкриті інструменти	Безкоштовні або частково безкоштовні сервіси, що дозволяють створювати власні словники, додавати аудіо, відео, зображення.	<i>WordPress + Lingua plugin, GlossaryTech, DictionaryMaker, Airtable</i>
5. Програмні розділі інтегрованих видавничих систем	Входять до складу редакційних платформ великих видавництв; забезпечують автоматичне верстання та лінгвістичну верифікацію.	<i>InDesign Dictionary Module, IDM DPS LexiCat</i>

Д.18. Порівняльна характеристика різних систем розроблення словників (DWS / CPC)

Назва системи	Розробник / країна	Основні можливості	Тип ліцензії	Приклади використання
TLex (TshwaneLex)	TshwaneDJe HLT, ПАР	Професійне середовище для створення, редагування та публікації словників; підтримка XML, автоматичне форматування, багатомовність, розділ статистики	Комерційна	<i>Oxford Bilingual Dictionaries, Macmillan Dictionary Projects</i>
IDM DPS (Dictionary Production System)	IDM, Франція	Комплексна платформа для великих видавництв; інтеграція з базами даних, підтримка командної роботи, керування версіями	Комерційна	<i>Collins, Le Robert, Larousse</i>
Lexique Pro	SIL International, США	Орієнтований на академічні та польові лінгвістичні проєкти;	Безкоштовна	Словники малих і корінних мов

		підтримує імпорт з FieldWorks, гнучкий пошук і створення онлайн-словників		
TLex Online / TLex Cloud	TshwaneDJe HLT, ПАР	Хмарна версія TLex для командної роботи онлайн; автоматичні оновлення і синхронізація	Комерційна	Вебверсії академічних і двомовних словників
Lexonomy	Lexical Computing CZ, Чехія	Відкрите вебсередовище для створення та редагування словників; підтримує гіпертекст, багатомовність, API	Безкоштовна / з відкритим кодом	Академічні та навчальні проєкти в Європі
Sketch Engine for Lexicography (SkELL)	Lexical Computing CZ	Поєднує корпусний аналіз із лексикографічним редагуванням; автоматично витягує приклади вживання	Комерційна	Національні корпусні словники
FieldWorks Language Explorer (FLEX)	SIL International	Інструмент для польових лінгвістів; збір лексем, морфологічний аналіз, експорт у Lexique Pro	Безкоштовна	Польові лексикографічні дослідження

Д.19. Характеристика загальномовних корпусів

Назва	Платформа	Тексти	Розмітка	Доступ	Додаткове ПЗ
Український національний лінгвістичний корпус (УНЛК)	Microsoft. NET Remoting	Художні та публіцистичні (газетні статті, інтернет-ресурси, рекламні матеріали)	Морфологічна, синтаксична, семантична	Лише для зареєстрованих користувачів	Потребує
British National Corpus (BNC)	IMS Corpus Workbench / BNCweb	Художні, науково-популярні, газетні, усні транскрипції, академічні тексти	Морфологічна, частиномовна, синтаксична, семантична	Вільний доступ через BNCweb (реєстрація бажана)	Не потребує
DWDS Core Corpus	DWDS Search Platform (Berlin-Brandenburgische Akademie der	Сучасні німецькі тексти: художні, наукові, публіцистичні, усні, технічні	Морфологічна, частиномовна, частково семантична, лематизація	Відкритий доступ (онлайн-пошук DWDS.de)	Не потребує

	Wissenschaft en)				
ELEXIS French Web 2020	Sketch Engine	Веб-корпус французької мови (блоги, новини, форуми, соцмережі, онлайн-медіа)	Автоматична морфологічна, частиномовна, лематизація	Доступ через підписку Sketch Engine	Працює у Sketch Engine
Corpus del Español del Siglo XXI (CORPES XXI)	Real Academia Española (CORPES XXI Interface)	Літературні, наукові, газетні, усні, інтернет-тексти з 2001 р.	Морфологічна, частиномовна, лематизація, регіональна маркованість	Безкоштовно після реєстрації (corp.es.rae.es)	Не потребує

Д.20. Характеристика спеціалізованих корпусів

Назва	Платформа	Тексти	Розмітка	Доступ	Додатков е ПЗ
The JRC-Acquis Multilingual Parallel Corpus	Joint Research Centre (JRC)	Юридичні тексти законодавства ЄС 231 мовної пари	Морфологічна, частиномовна, синтаксична, багатомовне вирівнювання	Відкритий (на сайті JRC)	Не потребує
Parallel English-Croatian Corpus in the Domain of Air Traffic	Faculty of Humanities and Social Sciences, University of Zagreb	Законодавчі тексти з авіаційної термінології (англійською та хорватською)	Морфологічна, частиномовна, термінологічна, паралельне вирівнювання	Вільний для академічного використання	Не потребує
MedCorp (Medical Corpus)	University of Gothenburg (Språkbanken)	Медичні тексти, наукові статті, клінічні протоколи	Морфологічна, частиномовна, термінологічна, NER-розмітка	Вільний після реєстрації	Не потребує
EcoLexicon English Corpus (EEC)	University of Granada	Екологічні тексти: наукові статті, глосарії, навчальні матеріали	Морфологічна, частиномовна, семантична, тематична	Відкритий (ecolexicon.ugr.es)	Не потребує
EUR-Lex Corpus	European Commission	Тексти законодавчих актів ЄС різними мовами	Морфологічна, юридична, багатомовна (паралельна)	Вільний на eur-lex.europa.eu	Не потребує

Д.21. Характеристика паралельних корпусів

Назва	Платформа	Тексти	Розмітка	Доступ	Додаткове ПЗ
OPUS (Open Parallel Corpus)	University of Helsinki	Тексти з фільмів, технічних документів, вебресурсів (технічні, художні, субтитри)	Морфологічна, частиномовна, паралельне вирівнювання	Відкритий (opus.nlpl.eu)	Не потребує
Europarl Corpus	European Parliament	Стенограми засідань Європарламенту (21 мова)	Морфологічна, паралельна (на рівні речень)	Відкритий	Не потребує
UN Parallel Corpus	United Nations	Офіційні документи та промови ООН (6 мов)	Морфологічна, семантична, паралельне вирівнювання	Відкритий	Не потребує
ParaCrawl	European Language Resources Association (ELRA)	Вебтексти різних жанрів, автоматично вирівняні	Морфологічна, частиномовна, паралельна	Відкритий (paracrawl.eu)	Не потребує
The JRC-Acquis Multilingual Parallel Corpus	Joint Research Centre (JRC)	Законодавчі акти ЄС, 231 мовна пара	Морфологічна, частиномовна, паралельна	Відкритий	Не потребує
UkrParCor (Ukrainian Parallel Corpus)	Національний університет «Києво-Могилянська академія»	Переклади художніх, технічних і наукових текстів (укр.-англ.)	Морфологічна, паралельна, частиномовна	Обмежений доступ для дослідників	Не потребує
Tatoeba Parallel Corpus	Tatoeba Project	Короткі приклади речень, перекладені сотнями мов	Морфологічна (частково), паралельна	Відкритий	Не потребує

Д.22. Приклади MSD тегів у MULTEXT-East

Категорія	Літера	Значення
<i>Частини мови</i>		
N	noun	іменник
V	verb	дієслово
A	adjective	прикметник
P	pronoun	займенник
R	adverb	прислівник
C	conjunction	сполучник
T	particle	частка

I	interjection	вигук
M	numeral	числівник
X	residual	інше / некатегоризоване
Граматичні категорії (фрагменти тегів)		
m / f / n	masculine / feminine / neuter	рід
s / p	singular / plural	число
i / p / m	indicative / participle / imperative	спосіб / форма
1 / 2 / 3	—	особа
n / g / d / a / l / i	—	відмінки (називний, родовий тощо, залежно від мови)
p / c / s	positive / comparative / superlative	ступені порівняння
f / s	full / short form (для прикметників у слов'янських мовах)	

Д.23. Приклади дефініцій, створених різними ВЛМ

Gemini: Оскопування – явище в академічній сфері, коли від науковців вимагають масово публікуватися у престижних міжнародних базах на кшталт Scopus, зосереджуючись на формальних індексах і створюючи надмірний тиск на викладачів.

Claude: Оскопування – процес примусового включення науковців / публікацій до наукометричних баз даних (зокрема Scopus), часто штучний чи формальний, заради відповідності формальним вимогам та підвищення показників, без урахування реальної наукової цінності роботи.

Llama 4 Maverik Turbo: Оскопування – це процес або дія, що полягає у приведенні чогось або когось до певного стандарту чи форми, що відповідає вимогам певних метрик чи баз даних, зокрема наукометричних. Це може включати в себе уніфікацію, стандартизацію чи адаптацію під певні критерії оцінки продуктивності чи якості.

Deepseek 3.1: Оскопування (іменник) — неологізм, який описує негативний явище в академічному середовищі, що полягає у:

(1) механічній стандартизації наукової діяльності, зведенні її до набору формальних вимог (наприклад, обов'язкова присутність у певних наукометричних базах даних, публікація у визнаних журналах тощо).

(2) абсолютизації кількісних показників (індекс Хірша, кількість цитувань), які стають головним і часто єдиним критерієм оцінки успішності дослідника чи установи.

(3) вимушеній конформізації та втраті індивідуальності дослідників, які змушені «підганяти» свою роботу під ці критерії, часто шкодячи якісній, інноваційній чи довготривалій науковій праці.

GLM 4.5: «Оскопусування» — це негативно забарвлений термін, що описує перетворення науки на механістичний процес задля досягнення формальних, кількісних показників, де індивідуальний підхід і якість жертвуються заради уніфікації та звітності.

Chat GPT 5: Оскопусування — *соціально й інституційно зумовлений процес у сучасній академічній сфері*, що полягає в абсолютизації вимог до присутності науковців і результатів їхньої діяльності в наукометричних базах (передусім Scopus), унаслідок чого формальні кількісні показники (індекси, квартилі, цитованість) починають домінувати над якістю, змістовною новизною та суспільною значущістю наукових досліджень.

Д.24. Варіанти зредагованих дефініцій

Варіант А – базовий нормативний (універсальний словник)

Оскопусування, *іменник, процес*. Практика в академічному середовищі, що полягає в абсолютизації вимог до присутності наукових публікацій у наукометричних базах (передусім Scopus), унаслідок чого формальні кількісні показники домінують над змістовною якістю наукових досліджень.

Варіант Б – для словника неологізмів / медіамови

Оскопусування – соціально зумовлений процес у сучасній науці, за якого орієнтація на показники наукометричних баз стає визначальним критерієм оцінки наукової діяльності, витісняючи якісні та змістові параметри досліджень.

Варіант В — термінологічно стриманий (для академічного словника)

Оскопусування – інституційна практика оцінювання наукової діяльності, що ґрунтується на пріоритеті формальних наукометричних індикаторів, зокрема індексації у Scopus.

Д.25. Типові помилки великих мовних моделей (ВЛМ)

у лексикографічному тлумаченні

Тип помилки	Суть помилки	Приклад з кейсу	Коментар лексикографа
Надмірне узагальнення	Дефініція описує предмет занадто широко, без фіксації традиційної номінації	<i>дерев'яна сука</i> → «механічний сигнальний пристрій»	Утрачається побутово-історична конкретика; словникова дефініція має бути короткою і типовою
Семантичне звуження	Модель обмежує значення лише одним матеріалом або контекстом	<i>брила</i> → лише «кам'яний блок»	Ігнорується ширший обсяг значення (земля, глина, лід тощо)
Втрачена спеціалізація	Не фіксується вузьке функціональне призначення терміна	<i>оскарб</i> → «інструмент для обробки каменю»	Для термінів принциповою є вказівка на конкретну операцію
Контекстуальна гіперкомпенсація	Контекст художнього тексту помилково сприймається як основне значення	<i>дерев'яна сука</i> як «оповіщення робітників»	Це контекстуальна реалізація, а не ядро словникового значення
Заміна дефініції описом	Дефініція перетворюється на розгорнуте пояснення	«що видає різкий звук для оповіщення...»	Нормативна дефініція має бути стислою і формульною
Ігнорування лексикографічної традиції	Не враховуються усталені словникові формули	Усі три приклади	ВЛМ не “знає”, яка дефініція є типовою для словника
Змішування рівнів значення	Посвідчення родового значення з функціональною інтерпретацією	<i>оскарб</i> = інструмент + спосіб використання	Потрібне розмежування дефініції та енциклопедичного коментаря

Навчальне видання

КУПРІЯНОВ Євген Валерійович

КОМП'ЮТЕРНА ЛЕКСИКОГРАФІЯ

Навчальний посібник
для бакалаврів напряму підготовки
035 «Філологія»
денної та заочної форм навчання

Відповідальна за випуск
Роботу до видання рекомендував

доц. Бабкова Н. В.
проф. Кіпенський А. В.

В авторській редакції

План 2026 р., поз. 33

Гарнітура Times New Roman. Ум. друк. арк. 15,8

Видавничий центр НТУ «ХП».
Свідоцтво про державну реєстрацію ДК № 5478 від 21.08.2017 р.
61002, Харків, вул. Кирпичова, 2 _____

Електронне видання