Міністерство освіти і науки України
Київський національний університет імені Тараса Шевченка
Навчально-науковий інститут філології
кафедра української мови та прикладної лінгвістики

АВТОМАТИЧНЕ УКЛАДАННЯ ЧАСТОТНОГО СЛОВНИКА УКРАЇНСЬКОЇ МОВИ ЯК ЗАСОБУ МОВНОГО РОЗВИТКУ ДІТЕЙ ДОШКІЛЬНОГО ВІКУ

Кваліфікаційна робота бакалавра
студентки 4 курсу
освітньої програми
«Прикладна (комп'ютерна) лінгвістика
та англійська мова»
спеціальності – 035.10 Філологія
(Прикладна лінгвістика)
галузі знань – 03 «гуманітарні науки»
Вероніки ГІЛЬТАЙЧУК
Науковий керівник:
д.філол.н., проф. Наталія ДАРЧУК

«Допущено до захисту»
Протокол засідання
кафедри української мови та прикладної лінгвістики
протокол № 15 **від «** 06 **»** 06 **2024 року**

завідувач кафедри _____(підпис)
к.філол.н., доц. Сергій РІЗНИК

КИЇВ — 2024

АНОТАЦІЯ

Кваліфікаційна робота «Автоматичне укладання частотного словника української мови як засобу мовного розвитку дітей дошкільного віку» присвячена створенню частотного словника для дітей дошкільного віку з метою підтримки їх мовного розвитку. Актуальність дослідження обумовлена важливістю раннього мовленнєвого розвитку для успішної соціалізації та навчання дітей, а також недостатністю існуючих лінгвістичних ресурсів, спрямованих на цю вікову групу.

Об'єктом дослідження є українська дитяча література, предметом – лексичний склад текстів. Метою роботи є розробка та автоматизація процесу створення частотного словника.

Основними завданнями дослідження були: аналіз існуючих публікацій та проєктів, присвячених частотним словникам і розвитку дитячого мовлення, створення бази даних текстів дитячої літератури та укладання частотного словника на 1000 слів, що охоплює іменники, прикметники, прислівники та дієслова.

Методологічна основа дослідження поєднує комп'ютерну лінгвістику та психолінгвістику з особливим акцентом на застосуванні інструментів обробки природної мови (NLP) для аналізу текстів. Теоретичне підґрунтя ґрунтується на теоріях когнітивного та лінгвістичного розвитку Жана Піаже та Ноема Хомського. Інноваційність дослідження полягає у використанні сучасних технологій для автоматизації процесу укладання словника та його адаптації до потреб дошкільнят.

Результати дослідження підтверджують ефективність частотного словника як інструменту сприяння мовленнєвому розвитку дітей

дошкільного віку. Укладений словник може стати цінним ресурсом для вихователів, батьків і логопедів, допомагаючи їм у розробці цільових навчальних матеріалів. Крім того, частотний словник сприятиме популяризації української мови, наголошуючи на культурно значущій лексиці.

Ключові слова: частотний словник, мовний розвиток, дошкільна освіта, українська мова, комп'ютерна лінгвістика, дитяча література, NLP, лексичний аналіз, Жан Піаже, Ноем Хомський.

ABSTRACT

The qualification work “Automatic compilation of the frequency dictionary of the Ukrainian language as a means of language development of preschool children” is devoted to the creation of a frequency dictionary for preschool children to support their language development. The relevance of the study is due to the importance of early language development for the successful socialization and education of children, as well as the lack of existing linguistic resources aimed at this age group.

The object of the study is Ukrainian children's literature, the subject is the lexical composition of texts. The purpose of the study is to develop and automate the process of creating a frequency dictionary.

The main objectives of the study were to analyze existing publications and projects on frequency dictionaries and the development of children's speech, create a database of children's literature texts, and compile a 1000-word frequency dictionary covering nouns, adjectives, adverbs, and verbs.

The methodological basis of the study combines computational linguistics and psycholinguistics, with a special emphasis on the use of natural language processing (NLP) tools for text analysis. The theoretical basis is based on the theories of cognitive and linguistic development by Jean Piaget and Noam Chomsky. The innovativeness of the study lies in the use of modern technologies to automate the process of compiling a dictionary and its adaptation to the needs of preschoolers.

The results of the study confirm the effectiveness of the frequency dictionary as a tool for promoting the speech development of preschool children. The compiled dictionary can be a valuable resource for educators, parents, and speech therapists, helping them to develop targeted teaching materials. In addition, the frequency dictionary will contribute to the popularization of the Ukrainian language, emphasizing culturally significant vocabulary.

Keywords: frequency dictionary, language development, preschool education, Ukrainian language, computational linguistics, children's literature, NLP, lexical analysis, Jean Piaget, Noam Chomsky.

ЗМІСТ

ВСТУП

Розвиток мови відіграє надзвичайно важливу роль у ранньому дитинстві, слугуючи засобом спілкування та найважливішим інструментом когнітивного розвитку і соціальної взаємодії. Здатність розуміти та використовувати мову безпосередньо впливає на навчання та адаптацію дитини в суспільстві, коли вона дорослішає.

Розвиток мовних навичок у дітей дошкільного віку є важливою та цікавою темою в дошкільній освіті. Дослідження показують, що перші кілька років життя дитини є критично важливими для засвоєння мови, коли закладається фундаментальний словниковий запас та основні мовні структури. Освітяни та дослідники, які працюють з українськомовним населенням, можуть отримати значну користь від розуміння того, які слова найчастіше вживаються дітьми чи в їхньому оточенні, оскільки це може покращити освітні методики та стратегії.

Пропонований частотний словник, розроблений для дітей дошкільного віку в Україні, узгоджується з реальним використанням мови, що спостерігається у повсякденному житті. Цей ресурс надає вихователям і педагогам уявлення про лексичні пріоритети в житті дитини, гарантуючи, що слова, які вивчаються, не лише відповідають віку, але є актуальними та практичними у повсякденному спілкуванні. Створюючи більш цікавий та ефективний навчальний процес, цей словник сприяє збереженню та використанню мови.

Розробка частотного словника з 1000 слів української мови для дошкільнят - це не просто теоретичне завдання. Він заповнює практичну прогалину в існуючих лінгвістичних ресурсах. Більшість матеріалів для вивчення мови призначені для дорослих або дітей шкільного віку, залишаючи поза увагою унікальні потреби дітей раннього віку. Надаючи матеріали, спеціально розроблені для того, щоб охопити їхні інтереси та

відповідати їхнім раннім стадіям розвитку, цей словник безпосередньо підтримує освітні практики для дошкільнят.

Частотний словник також відіграє важливу роль у збереженні та популяризації української мови. Зосереджуючись на лексиці, яка є культурно та контекстуально релевантною, цей ресурс допомагає ґрунтувати освітні практики на мовній спадщині.

Розробка частотного словника для дітей дошкільного віку в Україні є важливою справою, яка може суттєво вплинути на дошкільну освіту. Надаючи науково обґрунтований словник, який резонує з мовним і культурним середовищем учнів, такі ініціативи є життєво важливими для розвитку міцних мовних навичок, що формують фундамент для подальшого навчання. Раннє набуття мовних навичок є основою для розвитку граматичних і когнітивних здібностей. Тому необхідно ретельно відбирати мовні матеріали, які потрапляють у поле зору і слуху дітей. Цей словник має на меті включити активну та пасивну лексику, щоб закласти фундамент мовних навичок і звичок до моменту вступу дітей до школи. Словник може слугувати універсальним освітнім інструментом, який сприятиме вдосконаленню навчальних програм, що вже існують у дошкільних навчальних закладах. Батьки, які хочуть підтримувати мовний розвиток своїх дітей вдома, можуть використовувати словник для підбору матеріалів для читання або створення історій та оповідань, які сприятимуть розширенню словникового запасу їхніх дітей.

Укладання частотного словника на базі дитячої літератури – ***актуальне*** завдання сучасної лінгвістики.

Метою кваліфікаційної роботи є розробка та автоматичне укладання частотного словника на основі лексичного складу українськомовної дитячої літератури. Цей словник націлений на систематизацію 1000 слів, що найчастіше вживаються в дитячих текстах і сприятимуть розширенню активного та пасивного словникового запасу дітей. Кінцева мета полягає в

тому, аби забезпечити основу для ефективного засвоєння мови, яка сприятиме успішному мовленнєвому розвитку та підготовці дітей до школи.

Мета передбачає виконання таких ***завдань дослідження:***

1) проаналізувати наявні публікації та проєкти, в яких аналізуються частотні словники та, відповідно, мовний розвиток дітей;

2) визначити основні напрямки та рекомендації, які могли б вплинути на вибір текстів для аналізу, з особливою увагою до контексту дошкільної освіти;

3) завантаження та конвертація текстів до формату, придатного для комп'ютерного аналізу, забезпечення їх доступності для подальшої обробки;

4) створення структурованої бази даних для ефективного зберігання та управління текстами, що включають інформацію про авторів, назви та інші ключові метадані;

5) використання сучасних інструментів комп'ютерної лінгвістики для детального аналізу текстів, ідентифікації частотності слів та підготовки аналітичних висновків;

6) компіляція словника, що включатиме 1000 найуживаніших слів (у тому числі іменники, прикметники, прислівники та дієслова) у контексті мовного розвитку дошкільнят, з подальшим оформленням результатів у вигляді доступного довідника.

Об’єктом дослідження цієї роботи є дитяча література українськомовних текстів (включаючи українських та зарубіжних авторів).

Предмет — лексичний склад зібраних текстів, який аналізується з метою укладення частотного словника з 1000 слів для сприяння мовного розвитку дітей дошкільного віку. У частотний словник будуть включені іменники, прикметники, прислівники та дієслова.

Теоретичне значення роботи полягає в науковому розумінні мовного розвитку дітей дошкільного віку на основі систематизації та кількісної характеристики лексики дитячої дошкільної літератури. Дослідження є корисним для академічних студій в галузі прикладної лінгвістики, зокрема методології укладання частотних словників. яка відіграє важливу роль у педагогічних та мовних процесах. Завдяки всебічному аналізу дитячої літератури та мовного середовища в дослідженні визначено слова, які мають важливе значення для мовного розвитку та соціалізації дітей. Таким чином, це створює теоретичну основу для подальших досліджень і практичних застосувань.

Практичне значення цього дослідження не менш важливе, оскільки воно має безпосередній вплив на вдосконалення освітніх практик і розвиток мовних навичок у дітей дошкільного віку. Частотний словник, розроблений у рамках проєкту, слугуватиме цінним ресурсом для педагогів, батьків та логопедів, дозволяючи їм застосовувати цілеспрямований підхід до навчання та підтримки мовленнєвого розвитку дошкільників. Використання цього словника може суттєво підвищити мовну компетенцію дітей, що забезпечить їхню належну підготовку до школи. Крім того, словник можна використовувати для розробки цільових навчальних матеріалів та ігор, які сприятимуть мовному та когнітивному розвитку дітей в ігровій формі, роблячи процес навчання більш ефективним і приємним.

Структура роботи: Робота складається з анотації, змісту, вступу, чотирьох розділів, висновків, списку використаних джерел та додатків.

Список використаних джерел включає 47 наукову працю вітчизняних та зарубіжних авторів, а також електронні ресурси.

РОЗДІЛ 1. ОСНОВИ ПСИХОЛІНГВІСТИЧНОГО АНАЛІЗУ РАННЬОГО МОВЛЕННЄВОГО РОЗВИТКУ ДІТЕЙ: РОЛЬ ЧАСТОТНОГО СЛОВНИКА

1.1. Психолінгвістика дитячого мовлення

1.1.1 Поняття онтогенезу мовлення

Онтогенез мовлення – це складне явище, яке охоплює етапи розвитку мовлення та засвоєння мови від немовляти до дорослого віку. Поняття онтогенезу охоплює біологічну, психологічну, лінгвістичну та педагогічну перспективи, які спрямовані на розуміння складних механізмів, що дозволяють людині набувати та вдосконалювати свої мовленнєві здібності з плином часу [26].

З біологічної точки зору, онтогенез стосується біологічних і неврологічних процесів, які лежать в основі набуття і вдосконалення мовленнєвих здібностей протягом життя людини.

Психологічна перспектива фокусується на когнітивних і перцептивних процесах, які підтримують розпізнавання і продукування мовлення, включаючи розвиток фонетичних модулів і здатність розрізняти звуки різних мов.

Лінгвістична перспектива вивчає етапи, через які дитина засвоює фонологічні, синтаксичні та прагматичні компоненти рідної мови, тоді як освітня перспектива підкреслює важливість втручань і стратегій навчання, які відповідають етапам розвитку мовленнєвих і мовних навичок дітей.

Разом ці перспективи підкреслюють взаємодію біологічних, психологічних і соціальних чинників, які впливають на розвиток мовленнєвих здібностей людини. Таким чином, всебічне розуміння онтогенезу мовлення вимагає мультидисциплінарного підходу, який інтегрує знання з вищезгаданих перспектив.

1.1.2 Теорія мовного розвитку Жана Піаже

Теорія когнітивного розвитку Жана Піаже є фундаментальною працею в психології розвитку дитини, яка описує процес інтелектуального зростання людини від дитинства до дорослого віку. Теорія виділяє чотири основні стадії когнітивного розвитку, кожна з яких характеризується певними когнітивними етапами та набуттям навичок. Ці стадії забезпечують структуровану основу для розуміння інтелектуального розвитку дітей, підкреслюючи складну взаємодію між природою, вихованням і дозріванням у їхньому когнітивному зростанні [27].

Початкова стадія — це **сенсомоторна стадія**, що охоплює період від народження до приблизно двох років. У цей період немовлята навчаються переважно за допомогою своїх органів чуття та рухової взаємодії з найближчим оточенням. Вони досліджують світ, торкаючись, пробуючи на смак та маніпулюючи предметами, повільно розвиваючи важливі когнітивні здібності, такі як постійність об'єктів — розуміння того, що об'єкти продовжують існувати, навіть коли вони зникають із поля зору. Ця стадія знаменує початок цілеспрямованої поведінки та раннього мовного розвитку [36].

Друга стадія – передопераційна стадія, яка охоплює вік від двох до семи років. Діти на цій стадії беруть участь у символічних іграх і починають використовувати слова та малюнки для позначення предметів. Їхнє мислення все ще дуже егоцентричне, що означає, що їм важко бачити речі з інших точок зору, крім власної. Під час цієї фази в дітей також починає розвиватися пам'ять та уява, що дозволяє їм брати участь у складних іграх та розуміти хід розповіді.

Третя стадія, **конкретна операційна стадія**, відбувається приблизно у віці від семи до одинадцяти років. Діти починають логічно мислити, охарактеризовувати конкретні події та розуміти концепцію збереження — кількість речовини залишається незмінною, незважаючи на

зміну її форми або контейнера й ідею оборотності та можуть класифікувати об'єкти за кількома ознаками. Ця стадія має вирішальне значення для розвитку логічного мислення, хоча й обмежується матеріальними та практичними питаннями.

Четвертою стадією, яку виділив Піаже, є **формальна операційна стадія**, що починається приблизно з дванадцяти років і триває до дорослого віку. Ця стадія характеризується здатністю абстрактно мислити, логічно міркувати та формулювати гіпотези. Підлітки й дорослі можуть розглядати різні перспективи й думати про майбутнє, беручи участь у дискусійних і гіпотетичних міркуваннях.

Стадії Піаже були і є впливовими в освітній та психологічній галузях, хоча з роками їх критикували та вдосконалювали. Дехто стверджує, що когнітивний розвиток є більш плавним і безперервним, на нього значною мірою впливає соціальний і культурний контекст, який оригінальна модель Піаже не повністю враховує. Однак, його теорія залишається фундаментальною частиною розуміння когнітивного розвитку і продовжує інформувати як освітні практики, так і батьківські стратегії.

1.1.3 Вікові характеристики мовленнєвого розвитку дитини.

Рушійна сила процесу мовлення криється в самій природі людини, а саме в потребі у спілкуванні. Новонароджена дитина звертається до дорослого за допомогою крику. Таким чином вона виявляє свою потребу в їжі або сповіщає про якийсь дискомфорт [2].

Для узагальнення етапів розвитку дитячого мовлення ми створили таблицю, спираючись на дослідження О. Воротинцевої «Основні етапи розвитку мовлення в дітей» [4] та Л.Мейгеш «Вікові особливості мовленнєвого розвитку дитини» [5; 3] :

Таблиця 1.1 *Особливості мовленнєвого розвитку дітей*

Вік	**Особливості мовленнєвого розвитку дитини**
1 рік	- **1–2 міс.** Дитина починає спілкування з дорослим. Малюк намагається спілкуватися за допомогою міміки та активних рухів. - **3 міс.** Починає виражати емоції за допомогою звуків. - **6 міс.** Відгукується на своє ім'я, впізнає мовлення рідною мовою, може інтонаціями голосу передавати настрій. - **9–12 міс.** У цьому віці діти розуміють найпростіші слова, використовують більшу кількість приголосних звуків та інтонацій.
2 роки	Фрази з двох-трьох слів — є найчастішими висловлюваннями малюка в цьому віці. На цьому етапі фраза є простою та граматично не оформленою.
3 роки	Між 2 та 3 роками активно формується фразове мовлення. Висловлювання дитини стають граматично оформленими. Діти в цьому віці починають засвоювати граматичну будову мовлення. До трьох років у дитини формуються всі основні граматичні категорії. - Відбувається активне зростання словникового запасу. - Засвоює мовлення в діалозі. - Вимовляє більшість приголосних, окрім сонорних [л] і [р].
4 роки	- Словниковий запас мовлення дитини стрімко зростає до 1500 словоформ). - Починають будувати складні речення. - У лексиконі малюка з'явиться нове запитання: «Для чого?», «Чому?», яке вказує вже на достатньо високий рівень розвитку не лише його мови, але й мислення.
5 років	- Запас словоформ у дитини збільшується до 3500 одиниць і більше. - У реченні використовуються всі частини мови.

	- Засвоюються граматичні правила зміни слів і з'єднання їх у речення. - Дитина здатна до фонетичного аналізу та синтезу.
6–7 років	- Словник включає в себе від 4000 до 7000 словоформ. - Діти самостійно складають розповідь, переказують казки, що говорить про опановування одним із найвищих видів мовлення — монологічним мовленням. - Здатні підбирати однокореневі слова, утворювати складні граматичні форми іменників, прикметників, дієслів.
8 — 10 років	- Словник включає в себе від 6000 до 10000 словоформ. - Вони використовують не лише базові слова, але й більш складну та спеціалізовану лексику, пов'язану з різними предметами та інтересами. - Діти краще розуміють часи дієслів, займенники, сполучники та інші аспекти синтаксису. Їхні речення стають складнішими, і вони починають використовувати більш складні граматичні конструкції.

1.1.4 Вплив теорій мовного розвитку на освітні практики

Основа будь-якого освітнього процесу, зокрема мовного навчання, глибоко вкорінена в наукові теорії, які пояснюють, як люди навчаються та розвиваються. Вивчення теорій мовного розвитку, на яких ґрунтується сучасна педагогічна практика, може дати новий погляд на те, як розвивати та вдосконалювати мовні навички у дітей від немовлят до шкільного віку, тим самим безпосередньо впливаючи на освітні підходи.

Теорія когнітивного розвитку Жана Піаже, одна з основоположних у галузі психології, акцентує на ідеї, що діти проходять через послідовні стадії розвитку, кожна з яких характеризується унікальними способами мислення та сприйняття світу. Ця теорія підкреслює необхідність

адаптувати стратегії навчання до конкретного вікового періоду дитини, оскільки кожна стадія передбачає різні когнітивні здібності.

Теорія універсальної граматики Ноема Хомського пропонує відмінний від попередніх підхід, зосереджуючись на вродженому аспекті мовного розвитку. Згідно з Хомським, усі люди народжуються із вродженою здатністю до мови, яка розвивається через взаємодію з мовним середовищем. Ця теорія підкреслює необхідність раннього мовного занурення в мовне середовище та стимуляції індивідуального мовного розвитку, що є основою для створення навчальних матеріалів і програм, спрямованих на розвиток цього природного мовного інстинкту [25].

Комбінація цих теорій може трансформувати освітні практики, надаючи фахівцям засоби для розробки більш ефективних підходів до навчання мови. Частотні словники, засновані на ґрунтовних теоретичних знаннях, можуть допомогти у розробці освітніх ресурсів, які сприятимуть природному і послідовному засвоєнню мови дітьми. Такі словники не лише навчають дітей нових слів, але й допомагають їм зрозуміти, як ці слова використовуються в різних контекстах, що є важливим для глибшого засвоєння мови.

Отже, всебічне розуміння теорій мовного розвитку та їх практичного застосування може значно покращити освітні стратегії, заклавши міцний фундамент для майбутнього академічного успіху та соціальної адаптації дітей. Це підкреслює неминущу цінність інвестицій у вивчення та впровадження цих теорій у нашу освітню практику.

1.2. Поняття частотного словника та його значення для мовного розвитку

1.2.1 Визначення та функції частотного словника

Частотний словник – унікальний інструмент у сфері лінгвістики, який являє собою упорядкований список слів на основі частоти їх появи тексті. На відміну від традиційних словників, які надають визначення

слову, фонетичні особливості та приклади вживання, частотні словники заглиблюються в кількісні дані. Ці дані, такі як абсолютна та відносна частота, пропонують кількісну перспективу використання мови, дозволяючи порівнювати різні тексти чи вибірки тексту.

У сфері лінгвістичних досліджень частотні словники – це більше, ніж просто інструменти. Вони дають науковцям можливість проводити детальний аналіз корпусу, сприяючи глибшому розумінню динаміки мови в різних контекстах і жанрах. Дослідники використовують ці словники, щоб відстежувати еволюцію мови, досліджуючи, як певні слова набувають або втрачають популярність з часом. Це відстеження має вирішальне значення для вивчення мовних змін, допомагаючи лінгвістам задокументувати еволюцію мовних моделей і зміни мовних уподобань.

Для тих, хто вивчає мову, частотні словники є практичним ресурсом, який може значно покращити їхній досвід навчання. Зосереджуючись на високочастотних словах, студенти можуть ефективніше вивчати лексику іноземної мови, яка найчастіше зустрічається в реальному спілкуванні. Такий методичний підхід сприяє розвитку функціональних мовних навичок, роблячи їх вивчення більш ефективним і практичним. Крім того, викладачі покладаються на частотні дані для розробки ефективних навчальних програм, планів уроків, які надають перевагу в основному лексиці, таким чином оптимізуючи результати навчання та гарантуючи, що учні оволодіють найбільш відповідними мовними навичками.

Частотні словники відіграють важливу роль у NLP (англ. *Natural language processing; укр. Обробка природної мови – це галузь, що поєднує комп'ютерні науки, штучний інтелект і лінгвістику. Метою є створення алгоритмів і систем, які дозволяють комп'ютерам розуміти, інтерпретувати та генерувати людську мову у той чи інший спосіб*), безпосередньо впливаючи на розробку більш точних систем розпізнавання мовлення та покращуючи технології передбачення тексту. Інформуючи

алгоритми про ймовірність вживання слів, ці словники допомагають підвищити точність «цифрових помічників» і передбачити характеристики тексту, роблячи технологічну взаємодію інтуїтивно зрозумілішою. Крім того, в аналізі тональності частотні словники допомагають розрізнити, які слова переважно вживаються в позитивному чи негативному контексті, тим самим підвищуючи точність алгоритмів, розроблених для інтерпретації емоцій у текстових даних. Це підкреслює актуальність і результативність роботи аудиторії у сфері NLP.

У лексикографії частотні словники є не просто впливовими, а вирішальними для укладання та оновлення традиційних словників. Лексикографи використовують частотні дані, щоб вирішити, які «нові» слова включити, а які «старі» слова виключити з лексичного фонду. Ці дані дають змогу зрозуміти тенденції та активне використання лексики, яке можна охопити в нових виданнях словників, гарантуючи, що ці ресурси залишатимуться актуальними та корисними. Цей ретельний процес прийняття рішень на основі даних переконує аудиторію в точності та актуальності мовних ресурсів.

Тож частотні словники – це більше, ніж просто списки слів; вони є важливими аналітичними інструментами, які забезпечують розуміння використання мови, підтримують освітні зусилля, сприяють лінгвістичним дослідженням і покращують технологічні інтерфейси в програмах NLP. Відображаючи частоту вживання слів, ці словники пропонують фундаментальний ресурс, який сприяє глибшому розумінню механізму мови та ефективнішому спілкуванню. Незалежно від того, чи йдеться про академічні, освітні чи практичні цілі, важливість частотних словників для розуміння та використання мови неможливо переоцінити.

1.2.2 Приклади використання частотних словників у розробці навчальних матеріалів

Oxford Wordlist [41] (англійська мова) – це лексичний ресурс для батьків і вчителів, які хочуть допомогти дітям розвинути мовні навички в реальному житті. Цей унікальний словник було створено на основі аналізу мовлення австралійських школярів, приділяючи особливу увагу словам, які діти віком від 5 до 9 років вживають найчастіше. Це важливий інструмент для будь-якої освітньої програми, оскільки дозволяє створювати вправи та ігри, які безпосередньо розвивають практичні мовні навички в англійськомовному середовищі.

Для французьких учнів початкових класів було створено інструмент **Manulex [28]**. Цей ресурс містить вичерпний корпус слів, зібраних із підручників, які використовуються в першому-п'ятому класах. Слова згруповані за різними лексичними групами та частотою вживання, що робить його незамінним ресурсом для дослідників і вчителів, які займаються мовним розвитком французьких учнів початкової школи. Manulex також пропонує такі розширення, як Manulex-infra та Manulex-morpho, які досліджують вплив морфології на навички читання та орфографії.

Проєкт **ChildLex [17]** (німецька мова) є результатом співпраці Потсдамського університету та Берлінсько-Бранденбурзької академії наук. Він містить 117952 лем із 500 дитячих книжок, призначених для читачів віком від 6 до 12 років, що робить його неоціненним ресурсом, який допомагає дослідникам, педагогам і терапевтам розробляти цільові освітні та терапевтичні матеріали [37]. ChildLex розвиває почуття спільності та спільної мети, пропонуючи ресурси для створення навчальних матеріалів німецькою мовою, адаптованих для дітей.

CREA Junior [22] (іспанська мова) – словник, який відображає фактичне використання мови серед іспанськомовних дітей. Ця адаптація

відомого словника CREA спеціально зосереджена на мовленні дітей, дозволяючи педагогам розробляти навчальні програми, які відображають фактичне використання мови серед іспанськомовних дітей. Він дає можливість створювати ефективні навчальні матеріали, адаптовані для дітей, допомагаючи розвивати мовні навички та прищеплювати любов до навчання.

ВИСНОВКИ ДО РОЗДІЛУ 1

У цьому розділі подано огляд сучасного стану досліджень у галузі психолінгвістики дитячого мовлення, висвітлено основні напрями вивчення мовленнєвого онтогенезу дітей на підставі сучасної теорії Ж.Піаже. Підкреслюється, що для вивчення мовленнєвого розвитку необхідний комплексний підхід, який поєднує лінгвістичні та психологічні аспекти, що має вирішальне значення для інформування та залучення аудиторії.

Також було подано детальний опис того, як розвиваються мовні навички у дітей від народження до 10 років, визначаючи ключові етапи формування словникового запасу та граматичних структур. Таке всебічне розуміння прогресу мовленнєвого розвитку є важливим як для дослідників, так і для практиків.

Крім того, вводиться поняття частотної лексики та її значення для мовного розвитку, з особливим наголосом на ефективності частотних словників у систематизації навчального процесу. Інтеграція найуживаніших слів допомагає дітям розширювати активний і пасивний словниковий запас, формуючи таким чином основи мовленнєвих навичок.

Отже, отримані дані підкреслюють важливість комплексного підходу до вивчення мовленнєвого розвитку дітей, а також вирішальну роль частотних словників у формуванні основ мовленнєвих навичок до початку формальної мовної освіти.

РОЗДІЛ 2. СТВОРЕННЯ БАЗИ ДАНИХ ТЕКСТІВ ДИТЯЧОЇ ЛІТЕРАТУРИ

2.1. Підготовка текстів

2.1.1. Процедура відбору текстів для частотного словника

Процес відбору текстів для частотного словника проводився ретельно, щоб забезпечити включення репрезентативного набору дитячої літератури, яка точно відображає мовне середовище дітей дошкільного віку. Основними критеріями відбору текстів були доступність джерела, відповідність віку та рекомендована література.

Відповідність віковій категорії була ключовим критерієм у процесі відбору, оскільки вона гарантувала, що відібрані тексти підходять для дітей віком до 7 років. Цей віковий критерій збігається з етапами розвитку раннього засвоєння мови та навичок грамотності у дітей дошкільного віку, що робить його важливим фактором у процесі відбору.

Тексти відбиралися із загальнодоступних ресурсів, щоб забезпечити вільний доступ до літератури для аналізу та її широке розповсюдження без будь-яких обмежень щодо авторських прав. Такий підхід також дозволив включити різноманітний спектр текстів, щоб забезпечити всебічне представлення мови, яка використовується в дитячій літературі.

Велика частина текстів була відібрана зі списків рекомендованої літератури з листа МОН “Про переліки навчальної літератури та навчальних програм, рекомендованих Міністерством освіти і науки України для використання в освітньому процесі закладів освіти у 2023/2024 навчальному році” [12]. Зокрема зі списку навчальних програм для дошкільних закладів [7] було використано “Я у Світі” [9] та “Впевнений Старт” [11].

Також частина текстів була відібрана на основі списку рекомендованої літератури, розміщеного на сайті “Barabooka” [14], укладеного Центром досліджень літератури для дітей та юнацтва [13]. У

цьому списку зібрані книги, спрямовані на розвиток мовлення та навичок читання у дітей молодшого віку, і він включає різноманітні жанри для забезпечення різноманітного знайомства з мовою.

Відібрані тексти пройшли процес перевірки, щоб підтвердити їхню актуальність та відповідність цільовій віковій групі. Цей процес відбору мав на меті створити надійну базу даних, яка б забезпечила міцну основу для аналізу частотності та вживання слів у контексті дитячої літератури, що, зрештою, допоможе у створенні практичного та функціонального частотного словника.

2.1.2. Формат файлів та завантаження текстів

Усі документи зберігалися у примітивному текстовому форматі, без жодного форматування, з розширенням .txt. Ім'я файлу повинно було відповідати певному номенклатурному формату, а саме: прізвище автора - назва твору - рік видання - жанр *(див. Малюнок 2.1)*. Ця номенклатура пізніше була використана для категоризації та індексування текстів, що полегшило ідентифікацію та організацію контенту. Включення такого вичерпного масиву інформації в назву файлу є особливо корисним при завантаженні текстів до бази даних. Програма може автоматично витягувати метадані та класифікувати тексти за різноманітними параметрами.

Малюнок 2.1. Номенклатурний формат

Файли були завантажені двома різними методами, де перший підхід включав веб-скрапінг за допомогою програмного коду, а другий – ручне вилучення текстів з відкритих джерел.

Веб-скрапінг – це комп'ютеризований метод, який використовується для вилучення та збору даних з вебсайтів. Він передбачає автоматизований пошук вебсторінок і вилучення з них значущої інформації, як правило, за допомогою спеціалізованого програмного забезпечення або скриптів. Цей процес дозволяє отримувати дані з різних вебсайтів швидко та ефективно, без необхідності ручного введення даних.

Програмний код ***'scrap.py',*** який знаходиться в репозиторії за посиланням у Додатку 1, мовою програмування Python використовувався для збору текстів оповідань з вебсайту за допомогою веб-скрапінгу.

Спочатку скрипт завантажує HTML сторінку з використанням бібліотеки 'requests' [35] і парсить її за допомогою бібліотеки 'BeautifulSoup' [15] для витягування посилань на інші сторінки. Кожне посилання, що веде до сторінки з текстом, обробляється окремо: скрипт переходить за цими

посиланнями, завантажує відповідні сторінки і витягує з них текст оповідань.

Модуль 'requests' в Python - це широко використовувана і багатогранна HTTP-бібліотека, яка полегшує надсилання різних типів HTTP-запитів до серверів з Python-додатків. Завдяки своєму зручному та ефективному інтерфейсу вона стала улюбленим вибором серед веброзробників та програмістів для веб-скрапінгу, взаємодії з API (*укр. Прикладний програмний інтерфейс*) та загального вебпрограмування.

Beautiful Soup - це бібліотека Python, спеціально розроблена для спрощення процесу синтаксичного аналізу HTML та XML документів в інтуїтивно зрозумілий спосіб. Вона пропонує безліч інструментів, які полегшують навігацію, пошук і модифікацію структур дерева розбору, що робить її ідеальним вибором для завдань веб-скрапінгу, які передбачають вилучення інформації з вебсайтів.

Після витягування тексту оповідання, скрипт створює ім'я файлу, засноване на тексті посилання, обмежуючи його до 50 символів і замінюючи пробіли та інші спеціальні символи. В назву файлу було включено ім’я автора та назву твору, рік та жанр твору додавалися в назву файлу вручну, оскільки ця інформація не подається на вебсторінці.

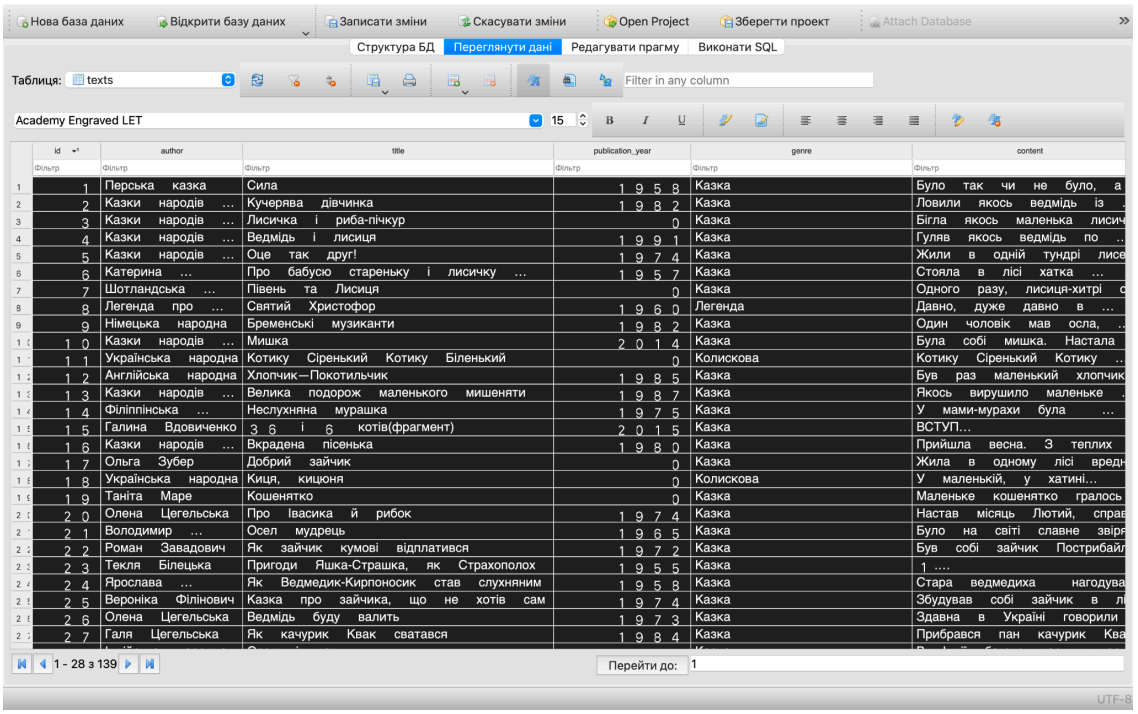
Другий підхід полягає в дослідженні загальнодоступних джерел текстів дитячої літератури. Згодом ці тексти були ретельно вилучені та збережені у визначеному форматі, що відповідає відповідним номенклатурним нормам, а потім збережені на комп'ютері для подальшого аналізу та використання.

2.2. Розробка та аналіз бази даних дитячої літератури

2.2.1. Створення основної таблиці бази даних та екстракція файлів

Скрипт ***'insert_data.py'*** (який знаходиться в репозиторії за посиланням у Додатку 1) – це код, який було розроблено для автоматизації

збору текстових даних і метаданих з TXT-файлів, що зберігаються у зазначеному директорії, та їх збереження у структурованій базі даних 'kidus.db'(див. Додаток 2). Він складається з декількох основних функцій та операцій, які обробляють тексти, витягують важливі метадані та систематично зберігають їх для подальшого аналізу.



Малюнок 2.2. *Вигляд таблиці texts*

На початку код ініціює з'єднання з базою даних за допомогою sqlite3 [39] і створює таблицю *texts (див. Малюнок 2.2)* у структурованій базі даних 'kidus.db', в якій зберігаються основні дані, такі як автор, назва, рік видання, жанр і сам текст. Після цього код переглядає всі файли у вказаному директорії, відкриває кожен файл з розширенням .txt і визначає його кодування за допомогою бібліотеки chardet [16], перш ніж прочитати його вміст.

SQLite3 - це широко використовувана система управління базами даних, яка забезпечує ефективний, простий і надійний спосіб зберігання даних. У Python можна отримати доступ до баз даних SQLite3 через модуль 'sqlite3', який є частиною стандартної бібліотеки Python. Цей модуль

дозволяє розробникам створювати, керувати та отримувати доступ до баз даних SQLite безпосередньо з Python, без необхідності використання окремого сервера.

Модуль sqlite3 у Python пропонує широкий спектр стандартних операцій SQL, а також деякі специфічні функціональні можливості SQLite. За допомогою курсорів, створених з об'єктів підключення, можна виконувати команди SQL, які дозволяють читати, вставляти, оновлювати і видаляти дані в базі даних.

Основна функція 'extract_metadata_from_filename' – аналіз назви файлів для вилучення метаданих, таких як автор, назва, рік видання і жанр. Ця інформація разом з текстовим вмістом файлу додається до бази даних за допомогою функції 'add_text_to_db', що дозволяє систематично і структуровано збирати текстові дані для подальшого аналізу або обробки.

Такий підхід полегшує структуроване зберігання текстових ресурсів і метаданих у базі даних, спрощує доступ до них і забезпечує їх зручність для дослідницьких та аналітичних цілей. Це особливо корисно в академічній науці, яка потребує ретельної організації та швидкого доступу до великої кількості текстових даних.

База даних 'kidus.db' (див. Додаток 2) налічує 198 текстів, що представляють різні жанри та різних авторів. Колекція включає твори як українських, так і зарубіжних авторів, а також казки з усього світу, від українських до американських за походженням. Наразі у базі даних налічується:

- 98 казок, серед яких є також казка-притча, казка-п’єса, казка-оповідання та казкова повість;
- 57 оповідань;
- 13 віршів, з яких 4 збірки з віршами;
- 9 повістей, серед яких є також повість-казка, пригодницька повість, казкова повість;

- 9 збірок оповідань, казок та віршів;
- 1 збірка купальських пісень, 1 збірка жниварських пісень, 1 українська народна пісня, 3 колискові (серед яких одна збірка), 2 збірки колядок та щедрівок;
- 1 збірка скоромовок;
- 1 збірка лічилок;
- 1 п’єса;
- 1 ігрова поезія;
- 1 легенда.

У Додатках 1 та 2 наведено покликання на директорії, де можна детальніше ознайомитися з програмним кодом та базою даних.

2.2.2. Створення другорядних таблиць

Скрипт ***'main_tables.py'*** (див. Додаток 1) слугує зразковою моделлю обробки та аналізу текстових даних за допомогою мови програмування Python, бібліотеки баз даних SQLite3 та механізму обробки мови Stanza [40].

Stanza - це широко використовувана бібліотека Python для обробки природної мови (NLP), яка була розроблена Стенфордською групою NLP. Вона пропонує повний набір інструментів для різних завдань NLP, включаючи токенізацію, модуль багатослівних токенів (MWT [29]), тегування частин мови (POS [32]), лематизацію, синтаксичний розбір залежностей та розпізнавання іменованих об'єктів (NER [30]). Універсальність цієї бібліотеки проявляється в підтримці багатьох мов і наданні попередньо навчених моделей для багатьох лінгвістичних завдань.

Цей скрипт виконує повний набір процедур, які дозволяють аналізувати словоформи, леми та частини мови в текстах, що зберігаються в базі даних, а також надають структурований формат для результатів.

Скрипт починає роботу зі встановлення з'єднання з базою даних SQLite 'kidus.db' і встановлення курсора для виконання SQL-запитів. Потім

він перевіряє наявність таблиць 'words_forms_freq', 'lemmas_freq' і 'part_of_speech_freq' і видаляє їх, щоб запобігти дублюванню даних при повторному запуску скрипта. Після цього створюються нові таблиці для зберігання даних про частоту вживання словоформ, лем і частин мови.

Для обробки текстів скрипт використовує бібліотеку Stanza, яка підтримує українську мову. Скрипт завантажує модель для української мови та ініціалізує обробку з увімкненими механізмами токенізації, морфологічного розбору та лематизації.

Скрипт зчитує текстові дані з таблиці 'texts' бази даних. Для кожного тексту він виконує різні операції з очищення даних, такі як перетворення тексту в нижній регістр і видалення небажаних символів. Потім текст обробляється за допомогою Stanza NLP, і дані про кожну словоформу, її лему та частину мови аналізуються та заносяться до бази даних за допомогою структури 'Counter' для підрахунку частот.

Для кожної словоформи, леми та частини мови код вставляє або оновлює частоту їх вживання в базі даних за допомогою SQL-запитів з умовою 'ON CONFLICT', що дозволяє автоматично виправляти наявні записи в разі повторення.

Після завершення аналізу скрипт зберігає зміни в базі даних і закриває з'єднання. Такий підхід дозволяє систематично обробляти великі обсяги текстових даних для лінгвістичного аналізу, забезпечуючи високу продуктивність і масштабованість обробки. Комплексність, ефективність і корисність скрипта роблять його цінним інструментом в академічних дослідженнях, зокрема в галузі лінгвістики.

В результаті ми отримуємо три таблиці в базі даних: ***'words_forms_freq', 'lemmas_freq' і 'part_of_speech_freq'***. Кожна таблиця має певне призначення для зберігання оброблених лінгвістичних даних.

1. ***'words_forms_freq'***: ця таблиця зберігає частоту кожної словоформи , що зустрічається в текстах (Малюнок 2.3). Таблиця містить такі дані:

– 'word_form' представляє конкретну форму слова, як воно з'являється в тексті. Це поле є первинним ключем, що гарантує унікальність кожної словоформи в таблиці;

– 'gen_freq' представляє загальну частоту кожної словоформи в усіх оброблених текстах цілим числом.

Загалом таблиця ***'words_forms_freq'*** містить 101 235 записів.

	word_form	gen_freq
	Фільтр	Фільтр
1	і	29700
2	не	22163
3	на	18380
4	що	15641
5	а	13278
6	з	12298
7	в	11846
8	я	11208
9	у	11198
10	він	10957
11	й	10022
12	до	9237
13	як	7862
14	та	6997
15	його	5985
16	це	5980
17	за	5559
18	але	5220
19	так	5083
20	вона	4687
21	ти	4589
22	ж	4351
23	вони	3915
24	то	3809
25	було	3751

Малюнок 2.3. Таблиця *'words_forms_freq'*

2. ***'lemmas_freq'***: ця таблиця зберігає леми слів, відповідні частини мови, а також частоту появи кожної леми (Малюнок 2.4). Таблиця містить такі дані:

– 'lemma' представляє словникову форму слова;

– 'part_of_speech' представляє граматичну категорію леми, наприклад, іменник, дієслово, прикметник тощо;

- 'gen_freq' підраховує входження кожної леми у всіх текстах

Первинним ключем є поєднання леми та частини мови що , гарантує, що кожен запис є унікальним для даної леми та її частини мови.

Загалом таблиця ***'lemmas_freq'*** містить 51353 записів.

Таблиця: lemmas_freq

	lemma	part_of_speech	gen_freq
	Фільтр	Фільтр	Фільтр
1	бути	VERB	27588
2	так	ADV	5044
3	чи	CCONJ	1959
4	не	PART	22160
5	а	CCONJ	13188
6	жити	VERB	1076
7	на	ADP	18379
8	світ	NOUN	861
9	горобець	NOUN	106
10	якось	ADV	440
11	узимку	ADV	9
12	у	ADP	111192
13	великий	ADJ	1676
14	мороз	NOUN	123
15	вилетіти	VERB	63
16	із	ADP	2313
17	свій	DET	4479
18	гніздечка	NOUN	26

Малюнок 2.4. Таблиця *'lemmas_freq'*

3. ***'part_of_speech_freq'***: ця таблиця відстежує частоту вживання кожної частини мови в текстах (Малюнок 2.5). Таблиця містить такі дані:

– 'part_of_speech' представляє граматичну категорію, таку як іменник, дієслово, прикметник тощо;

- 'gen_freq' представляє загальну кількість входжень кожної частини мови в усіх текстах цілим числом.

Первинним ключем є частина мови, що гарантує, що кожна граматична категорія представлена один раз.

Загалом таблиця ***'part_of_speech_freq'*** містить 17 записів, де:

- VERB – дієслово;

- ADV – прислівник;
- CCONJ – сполучник;
- PART – частка;
- PUNCT – пунктуаційний знак;
- ADP – прийменник;
- NOUN – іменник ;
- ADJ – прикметник;
- DET – сюди потрапили вказівні займенники та порядкові числівники, оскільки в англійській мові вони є визначниками (*англ. determiner*) ;
- PRON – займенник ;
- SCONJ – сполучники підрядності;
- AUX – допоміжне дієслово (в англійській мові слова be – бути, have – мати);
- PROPN – іменник(власна назва) ;
- INTJ – вигук ;
- NUM – числівник;
- X – інше, таку позначку отримали слова іншими мовами, або які не відповідають сучасному правопису;
- SUM – символи і знаки.

Таблиця: part_of_speech_freq

	part_of_speech	gen_freq
	Фільтр	Фільтр
1	VERB	191280
2	ADV	104436
3	CCONJ	67942
4	PART	53084
5	PUNCT	8986
6	ADP	93718
7	NOUN	251642
8	ADJ	71557
9	DET	49275
1	PRON	99813
1	SCONJ	28025
1	AUX	12524
1	PROPN	4041
1	INTJ	1391
1	NUM	4633
1	X	1692
1	SYM	63

Малюнок 2.5. Таблиця *'part_of_speech_freq'*

2.2.3. Створення таблиці з абсолютною частотою леми у кожному тексті

Наступний скрипт ***'fr_lemmas_in_text.py'*** *(див. Додаток 1)* на Python був розроблений для аналізу лем, які присутні в текстах, що зберігаються в базі даних SQLite. Скрипт оснащено кількома чітко визначеними функціями, які призначені для створення таблиці в базі даних (Малюнок 2.6), всебічного аналізу текстів і, нарешті, зберігання результатів аналізу.

lemma	text_id	frequency ▲1
Фільтр	Фільтр	Фільтр
з	13	293
так	10	287
ніщо	10	286
в	11	286
я	97	281
на	10	280
Пітер	11	280
і	10	279
би	10	278
на	13	277
за	10	276
що	12	276
у	13	276

Малюнок 2.6. Таблиця *'lemma_text_frequency'*

Спочатку у скрипті визначається шлях до файлу бази даних SQLite. Потім за допомогою функції 'create_lemma_text_frequency_table' створюється таблиця 'lemma_text_frequency' у базі даних, яка містить поля для леми, ідентифікатора тексту ('text_id') та частоти лем у цьому тексті. Ця таблиця має первинний ключ, що складається з комбінації леми та 'text_id'. Це гарантує, що записи для кожної леми у кожному тексті є унікальними.

Функція 'populate_lemma_frequencies' є основною функцією, яка відповідає за аналіз текстів, що зберігаються в базі даних. Ця функція використовує для обробки тексту модуль 'stanza.Pipeline', який налаштовано для української мови і який містить процесори для токенізації, визначення частин мови та лематизації. Під час проходження тексту скрипт розбиває його на речення та слова, збирає частоти вживання лем і зберігає цю інформацію в базі даних. Для кожної леми в тексті виконується операція SQL-вставки. Якщо виникає конфлікт ключів, то оновлюється існуючий запис.

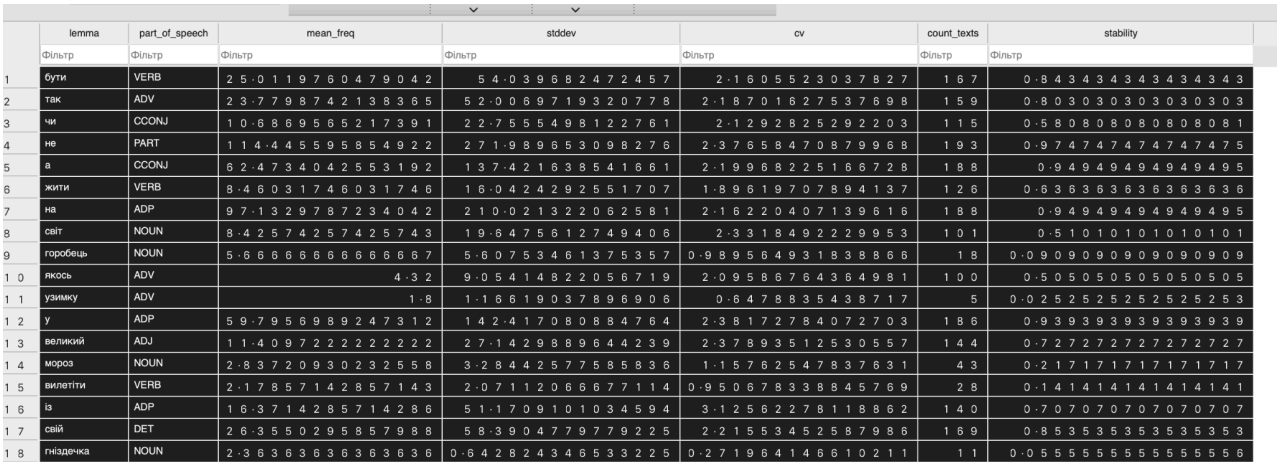
Основна функція 'main' викликає вищевказані функції для створення таблиці та заповнення її даними. Після завершення цих дій з'єднання з базою даних розривається.

Цей скрипт використовує можливості NLP і баз даних для зберігання та аналізу лінгвістичної інформації, надаючи механізм для детального аналізу вживання слів у текстах. Це може бути надзвичайно корисно для лінгвістичних досліджень, освітніх цілей або для підтримки інструментів обробки природної мови.

2.2.4. Статистичні параметри

Код ***'statistics.py'*** (див. Додаток 1) розроблений для інтеграції різних інструментів і технологій, включаючи бібліотеку Stanza NLP, базу даних SQLite і NumPy [31] для статистичних обчислень, для аналізу частоти лем (основних форм слів) у кількох текстах, що зберігаються в базі даних.

Скрипт обробляє текстові дані для генерування та зберігання статистичних вимірювань (Малюнок 2.7).



	lemma	part_of_speech	mean_freq	stddev	cv	count_texts	stability
	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр	Фільтр
1	бути	VERB	25.011976047904 2	54.039682472457	2.160552303782 7	167	0.843434343434343
2	так	ADV	23.779874213836 5	52.006971932077 8	2.187016275376 98	159	0.803030303030303
3	чи	CCONJ	10.686956521739 1	22.755549812276 1	2.129282529220 3	115	0.580808080808081
4	не	PART	114.44559585492 2	271.98965309827 6	2.376584708799 68	193	0.974747474747475
5	а	CCONJ	62.473404255319 2	137.42163854166 1	2.199682251667 28	188	0.949494949494949
6	жити	VERB	8.4603174603174 6	16.042429255170 7	1.896197078941 37	126	0.636363636363636
7	на	ADP	97.132978723404 2	210.02132206258 1	2.162204071396 16	188	0.949494949494949
8	світ	NOUN	8.4257425742574 3	19.647561274940 6	2.331849222995 3	101	0.510101010101010
9	горобець	NOUN	5.666666666666 7	5.607534613753 57	0.989564931838 866	18	0.090909090909090
10	якось	ADV	4.32	9.054148220567 19	2.095867643649 81	100	0.505050505050505
11	узимку	ADV	1.8	1.166190378969 06	0.647883543871 7	5	0.025252525252525
12	у	ADP	59.795698924731 2	142.41708088476 4	2.381727840727 03	186	0.939393939393939
13	великий	ADJ	11.409722222222 2	27.142988964423 9	2.378935125305 57	144	0.727272727272727
14	мороз	NOUN	2.837209302325 58	3.284425775858 36	1.157625478376 31	43	0.217171717171717
15	вилетіти	VERB	2.178571428571 43	2.071120666771 14	0.950678338845 769	28	0.141414141414141
16	із	ADP	16.371428571428 6	51.170910103459 4	3.125622781188 62	140	0.707070707070707
17	свій	DET	26.355029585798 8	58.390477977922 5	2.215534525879 86	169	0.853535353535353
18	гніздечка	NOUN	2.363636363636 36	0.642824346533 225	0.271964146610 211	11	0.055555555555556

Малюнок 2.7. Таблиця *'lemma_stat'*

Код складається з кількох ключових компонентів, кожен із яких відіграє вирішальну роль в успішному виконанні запланованої мети. По-перше, код ініціалізує бібліотеку Stanza для української мови ('uk'), який виконує токенізацію, розгортання багатослівних токенів (MWT), тегування частини мови (POS) і лематизацію. Цей крок необхідний для токенізації та нормалізації вхідних текстових даних, що робить їх придатними для подальшого аналізу.

Далі програма підключається до бази даних SQLite, розташованої за вказаним шляхом, і створює таблицю з назвою 'lemma_stat', якщо вона не існує. Ця таблиця призначена для зберігання кількох статистичних даних для кожної леми, включаючи її середню частоту, стандартне відхилення, коефіцієнт варіації (CV), кількість текстів, у яких вона з'являється, і показник стабільності. З'єднання з базою даних і налаштування таблиці мають вирішальне значення для забезпечення ефективного керування даними та безперебійного пошуку даних.

Перед наповненням новими даними, код очищає таблицю 'lemma_stat', щоб уникнути дублювання та забезпечити збереження лише

останніх результатів аналізу. Цей етап підготовки даних забезпечує точність і актуальність результатів аналізу.

Після цього скрипт зчитує текстові дані з іншої таблиці ('texts') у базі даних і використовує вкладену структуру defaultdict для збереження підрахунків частоти кожної леми в різних текстах, що полегшує ефективне агрегування даних. Функція 'is_valid_token' фільтрує токени шляхом перевірки на регулярний вираз, який перевіряє дійсні українські символи, забезпечуючи включення лише дійсних слів. Цей крок гарантує, що аналіз виконується на дійсних і релевантних текстових даних.

Згодом для кожної леми код обчислює декілька статистичних даних, включаючи середню частоту, стандартне відхилення, коефіцієнт варіації (CV), кількість текстів, у яких з'являється лема, і показник стабільності. Ці статистичні дані дають повний огляд лексичного багатства та різноманітності корпусу, сприяючи створенню адаптованого навчального контенту та розробці навчальних програм.

На фінальному етапі код вставляє обчислену статистику в таблицю 'lemma_stat' за допомогою команди SQL 'INSERT з пунктом 'ON CONFLICT, щоб оновити існуючі записи, якщо вони існують, гарантуючи, що дані залишаються актуальними.

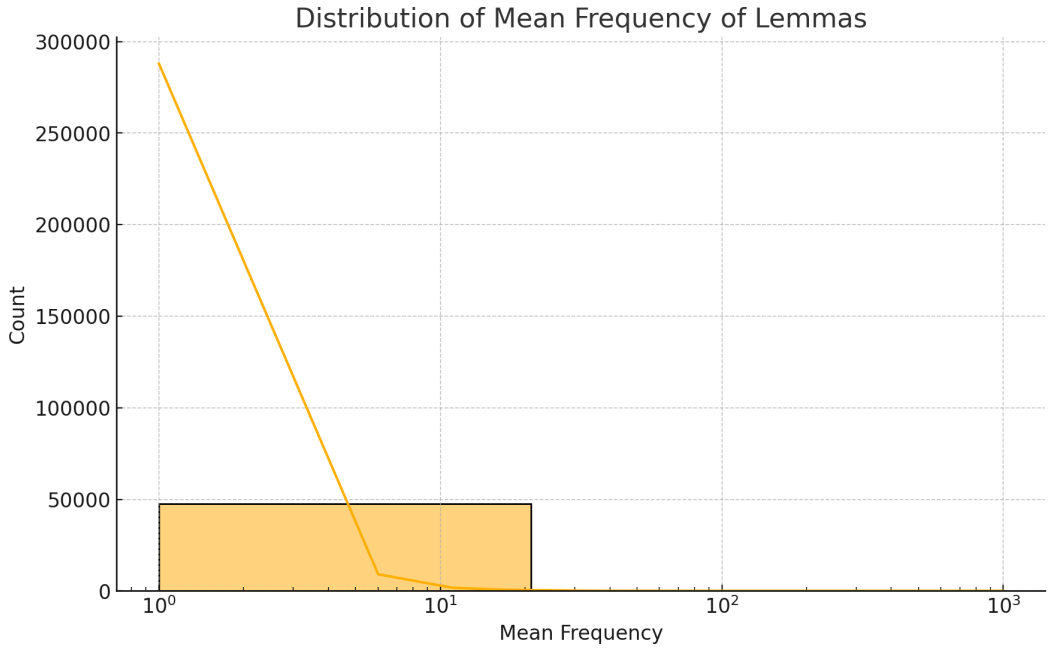
Загалом, цей код демонструє ефективний метод обробки та аналізу великих обсягів текстових даних в академічному чи освітньому середовищі. Він надає значну користь для лінгвістичних досліджень і мовної освіти, дозволяючи дослідникам і педагогам отримати уявлення про частоту та поширення лем в українській дитячій літературі та оцінити стабільність і спільність певних слів чи понять, що є цінним як для вивчення мови, так і для лінгвістики.

2.2.5. Візуалізація та аналіз статистичних параметрів

На основі таблиці 'lemma_stat' було створено візуалізацію даних та їх інтерпретація, отже:

1. Розподіл середньої частоти лем

Логарифмічна гістограма середніх частот показує дуже асиметричний розподіл частот лем, з «довгим хвостом», де невелика кількість лем використовується часто, тоді як переважна більшість використовується нечасто (Малюнок 2.8).



Малюнок 2.8. Гістограма середніх частот

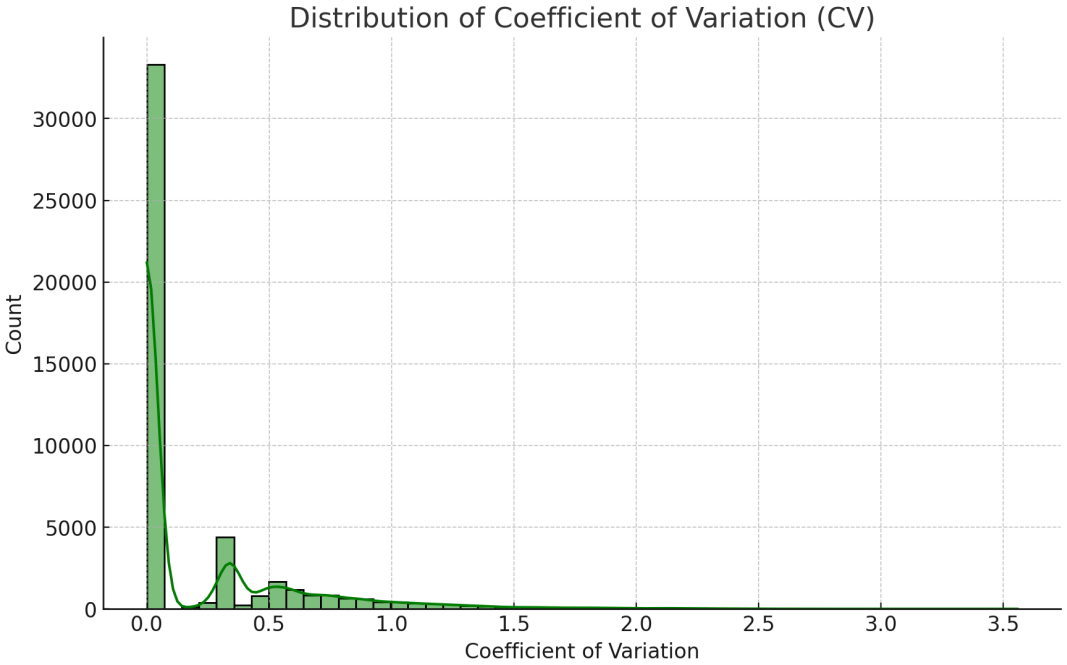
Як бачимо, більшість лем з'являються нечасто, що може свідчити про спеціалізовану лексику або контекстно-залежне використання. Невелика група лем має високу середню частоту, що вказує на їхню важливість у базовій структурі речення та комунікації, наприклад, сполучники, прийменники або загальні дієслова та іменники.

2. Розподіл коефіцієнта варіації (CV)

Коефіцієнт варіації вимірює відношення стандартного відхилення до середнього значення, за формулою:

$CV = \frac{\sigma}{\mu}$, де σ – стандартне відхилення, μ – середня частота вживання леми.

Низький коефіцієнт варіації свідчить про те, що вигляд леми в різних текстах є відносно послідовним, тоді як вищий коефіцієнт свідчить про більшу варіативність (Малюнок 2.9).



Малюнок 2.9. Гістограма коефіцієнта варіації

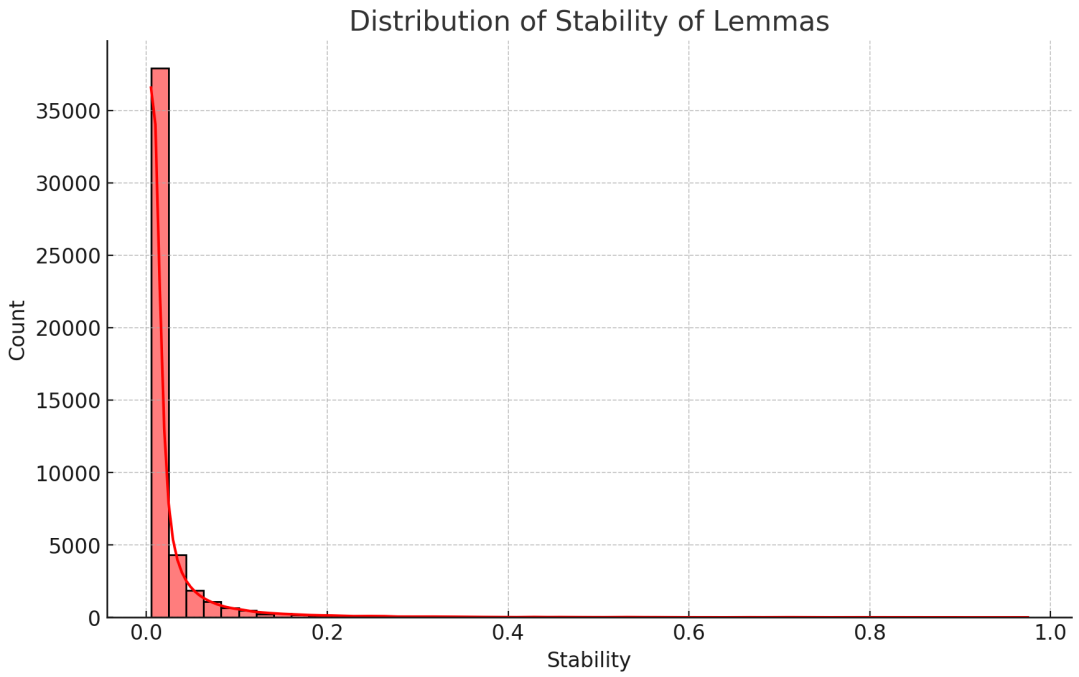
Концентрація лем з низькими значеннями CV вказує на те, що коли лема з'являється в корпусі, її частота в різних текстах має тенденцію до стабільності. Вище значення CV вказує на те, що деякі леми використовуються більш хаотично. Такі леми, ймовірно, є контекстно-залежними або спеціалізованими термінами, які з'являються лише за певних обставин.

3. Розподіл стабільності лем (D)

Стабільність вимірює частку текстів, в яких з'являється лема, за формулою:

$$D = \frac{\text{кількість текстів у яких зустрічається лема}}{\text{загальна кількість текстів}}.$$

Ця метрика допомагає визначити, наскільки поширеною є лема в корпусі (Малюнок 2.10).



Малюнок 2.10. Гістограма коефіцієнта варіації

На графіку показано, що багато лем з'являються в невеликому відсотку текстів, що свідчить про різноманітність лексики з обмеженою кількістю повторюваних вживань у корпусі.

Дуже мало лем демонструють високу стабільність, що свідчить про те, що лише невеликий набір слів часто вживається в текстах. Такі слова, найімовірніше, є функціональними частинами мови, такими як займенники, допоміжні дієслова або загальні іменники та прикметники.

Велика кількість лем з низькою стабільністю може також свідчити про різноманітність тем у корпусі, що призводить до різноманітного використання лексики.

Аналіз показує, що словниковий склад корпусу є різноманітним, більшість слів вживається нечасто, і лише невелика підгрупа слів є спільною для кількох текстів.

ВИСНОВКИ ДО РОЗДІЛУ 2

У Розділі 2 основна увага зосереджена на процесі створення бази даних текстів дитячої літератури, що має на меті формування частотного словника, який точно відображає мовне середовище дошкільної вікової

групи. Підготовка текстів передбачала ретельний відбір матеріалів відповідно до критеріїв доступності, вікової відповідності та репрезентативності з подальшою їх адаптацією в електронний формат для подальшого опрацювання. Важливу роль у цьому процесі відіграло використання автоматизованих технологій збору текстів за допомогою веб-скрапінгу, а також ручного збору з відкритих джерел.

Створена база даних слугує основою для подальшого структурування та аналізу текстів, що забезпечує належне вилучення та категоризацію метаданих, тим самим забезпечуючи високу якість лінгвістичних досліджень. Такий підхід не лише оптимізує процеси зберігання даних, але й забезпечує відповідність даних встановленим науковим та освітнім стандартам.

Таким чином, у Розділі 2 представлено комплексний підхід до створення валідної та репрезентативної бази даних дитячої літератури, яка є основою для розробки частотного словника.

РОЗДІЛ 3. УКЛАДЕННЯ ЧАСТОТНОГО СЛОВНИКА

3.1. Складова програмного забезпечення для компіляції словника

3.1.1. Укладення спільного словника

Скрипт ***'vocab.py'*** *(див. Додаток 1)* на Python слугує ефективною і потужною демонстрацією того, як можна використовувати базу даних SQLite для обробки та агрегування лінгвістичних даних. Скрипт фокусується на вилученні інформації про частоту певних частин мови з однієї таблиці та перенесенні цієї інформації в іншу таблицю, спрощуючи доступ до конкретних даних для подальшого аналізу або візуалізації.

Скрипт починає роботу зі встановлення з'єднання з вказаним файлом бази даних SQLite, 'kidus.db' і встановлення курсора, який використовується для виконання SQL-запитів. Потім скрипт виконує два важливі кроки:

1. Спочатку скрипт перевіряє наявність таблиці 'vocab' і видаляє її, якщо вона вже існує.

2. Після цього скрипт переходить до створення нової таблиці 'vocab', якщо вона ще не існує. Таблиця 'vocab' складається з двох стовпців: 'lemma', яке представляє лексичну основу слова, що слугує первинним ключем, і 'gen_freq', яке позначає загальну частоту вживання леми.

Основною функцією скрипта є вибірка та агрегування даних з таблиці 'lemmas_freq'. Запит відбирає тільки леми іменників, прикметників прислівників і дієслів, після чого обчислює суму їхніх частот ('gen_freq'), групуючи результати за лемами.

Скрипт заповнює даними таблицю 'vocab'. Якщо лема вже існує в таблиці (що теоретично не повинно відбуватися в цьому контексті через попереднє видалення таблиці), частота оновлюється за допомогою 'ON CONFLICT'.

Після виконання всіх запитів скрипт фіксує зміни в базі даних за допомогою функції 'commit', а потім закриває з'єднання з базою даних.

У процесі створення даної таблиці ми приділили особливу увагу не лише кількісному аналізу використаних лем, але і якісній оцінці отриманих даних. У результаті ми вирішили не обмежувати набір даних традиційними 1000 найуживанішими лемами. Замість цього ми ретельно відібрали і проаналізували більш широкий набір лем, що дозволило нам виявити і виправити помилки, які могли виникнути в результаті автоматизованої обробки даних. Наше рішення здійснити ручне очищення даних гарантує вищий ступінь точності та релевантності словника, що, своєю чергою, сприяє глибшому та точнішому розумінню лінгвістичних особливостей текстів. Такий підхід має важливе значення для забезпечення якісної та надійної бази даних, яка слугує основою для подальших досліджень та аналізу.

3.1.2. Укладення словників за частинами мови

Код ***'by_parts_of_speech.py'*** (див. Додаток 1) призначений для взаємодії з базою даних SQLite, призначеною для управління лінгвістичними даними. Спочатку він встановлює з'єднання з вказаним файлом бази даних, а потім створює чотири окремі таблиці, призначені для зберігання лексичних записів, класифікованих за відповідними частинами мови: іменники, прикметники, дієслова та прислівники. Ці таблиці структуровані таким чином, що лема кожного слова є первинним ключем, а відповідна частота - цілим числом.

Після налаштування таблиць програма заповнює їх даними, витягнутими з існуючої таблиці бази даних, яка містить леми, відповідні частини мови та загальні частоти. Записи групуються за лемами, а частоти агрегуються. Цей процес гарантує, що, якщо лема з'являється серед даних кілька разів, її частоти підсумовуються для забезпечення повного підрахунку.

Крім того, кожна частина мови обробляється окремо за допомогою спеціальних SQL-запитів для вибору та вставки даних у відповідні таблиці.

Код також включає додатковий крок для об'єднання дієслів та інфінітивів в одну категорію, таким чином спрощуючи та структуруючи процес пошуку даних для подальшої комп'ютерної обробки або лінгвістичного аналізу.

3.2. Практичне застосування словника

Останніми роками системний аналіз текстів і створення словника найчастотніших слів у дитячій літературі стали важливими напрямами досліджень. Частково це пов'язано з практичним і науковим значенням такого словника в кількох сферах.

Однією з ключових переваг такого словника є його цінність для лінгвістичних і лексикографічних досліджень. Він може слугувати цінним ресурсом для лінгвістів і лексикографів, які вивчають лексичний склад і розвиток мови в дитячій літературі. Завдяки визначенню найуживаніших слів, тематичних категорій і понять, які часто зустрічаються в дитячих текстах, ці дані можуть бути використані для підготовки навчальних матеріалів, словників або довідників, орієнтованих на юних читачів. Ці ресурси можуть забезпечити глибше розуміння мови та збагатити словниковий запас учнів.

Окрім цих переваг, словник має значний потенціал в освітніх цілях. Він може слугувати основою для розробки навчальних програм, підручників і методичних матеріалів, спрямованих на підвищення літературної грамотності дітей. Вчителі можуть використовувати словник для підготовки уроків, які розкривають значення та вживання ключових слів у контексті, сприяючи таким чином глибшому розумінню мови та збагаченню словникового запасу учнів. Крім того, включення найуживаніших слів у навчальні матеріали може сприяти природному та послідовному розвитку мовних навичок.

Словник також має потенціал для вивчення особливостей мови, що використовується в дитячій літературі. Він може стати основою для створення матеріалів для читання, які відповідають віковим і мовним

потребам дітей. Це особливо важливо, оскільки включення найуживаніших слів у навчальні матеріали може сприяти природному і послідовному розвитку мовних навичок.

У наукових дослідженнях словник можна використовувати для автоматизованого контент-аналізу дитячої літератури. Він дозволяє дослідникам визначити не лише частоту вживання слова, а і його роль та значення в контексті різних тем і жанрів. Це дає можливість проаналізувати, які теми та поняття є пріоритетними в дитячій літературі та як вони подаються юним читачам та слухачам. Такий аналіз може допомогти у створенні більш ефективних навчальних матеріалів та у розробці більш релевантних матеріалів для читання для дітей.

Отже, створення та використання словника найчастотніших слів у дитячій літературі має важливе практичне та наукове значення в кількох сферах. Фахівці з дитячої літератури, освіти та лінгвістики можуть використовувати цей інструмент, щоб отримати уявлення про читацькі уподобання дітей та мовні тенденції, а також для розробки більш ефективних навчальних матеріалів та матеріалів для читання для дітей.

ВИСНОВКИ ДО РОЗДІЛУ 3

Розділі 3 присвячено основним аспектам побудови та використання частотного словника в дитячій літературі. Дослідження пропонує надійний підхід до аналізу лексичного матеріалу, який полегшує як кількісне, так і якісне вивчення текстів. Використовуючи скрипти для обробки лінгвістичних даних, ми змогли виокремити та систематизувати інформацію, що стосується частотності лем у дитячій літературі, тим самим проклавши шлях до нових шляхів дослідження лінгвістичних особливостей.

Частотний словник, укладений у результаті нашого дослідження, виявився безцінним ресурсом для різноманітних сфер застосування, включаючи лінгвістичні дослідження та педагогічну практику. Він може

слугувати незамінним інструментом для укладання навчальних програм, підручників і методичних матеріалів, спрямованих на підвищення грамотності дітей. Крім того, словник може бути використаний для проведення автоматизованого контент-аналізу, що сприятиме більш нюансованому розумінню тематичних категорій і понять, які найчастіше зустрічаються в дитячих текстах.

РОЗДІЛ 4. АНАЛІЗ ЧАСТОТНОГО СЛОВНИКА ДИТЯЧОЇ ЛІТЕРАТУРИ

4.1. Обсяг отриманого словника

Таблиця 'lemmas_freq' у базі даних містить детальний розподіл різних частин мови серед 51 353 лем. А саме (див. Малюнок 4.1):

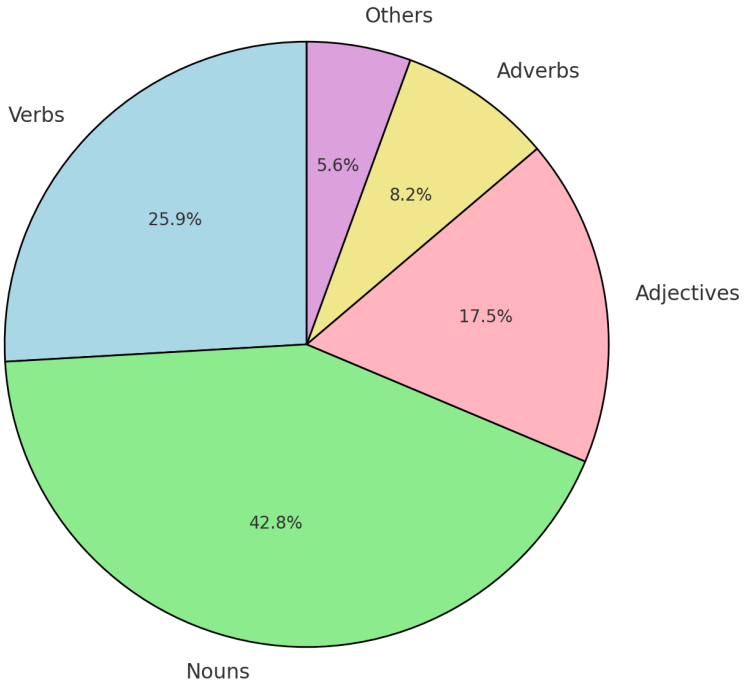
- Іменники: Це найчисленніша категорія в таблиці, яка налічує 21 962 (42,8%) леми. Це свідчить про широке розмаїття суб'єктів, об'єктів та ідей, представлених у текстах.

- Дієслова: Друга за чисельністю група з 13 307 (25,9%) записами, що відображає дії та стани, описані у текстах.

- Прикметники: Вони описують або модифікують іменники і складають 8992 (17,5%) леми, пропонуючи різноманітний набір описових термінів.

- Прислівники: 4 229 (5,6%) прислівників, які модифікують дієслова, прикметники або інші прислівники, надаючи інформацію про спосіб, ступінь, частоту та інші аспекти.

- Інші: Ця категорія, яка включає 2 863 леми, включає частки, вигуки, сполучники, прийменники та інші частини мови, які не були віднесені до основних груп.

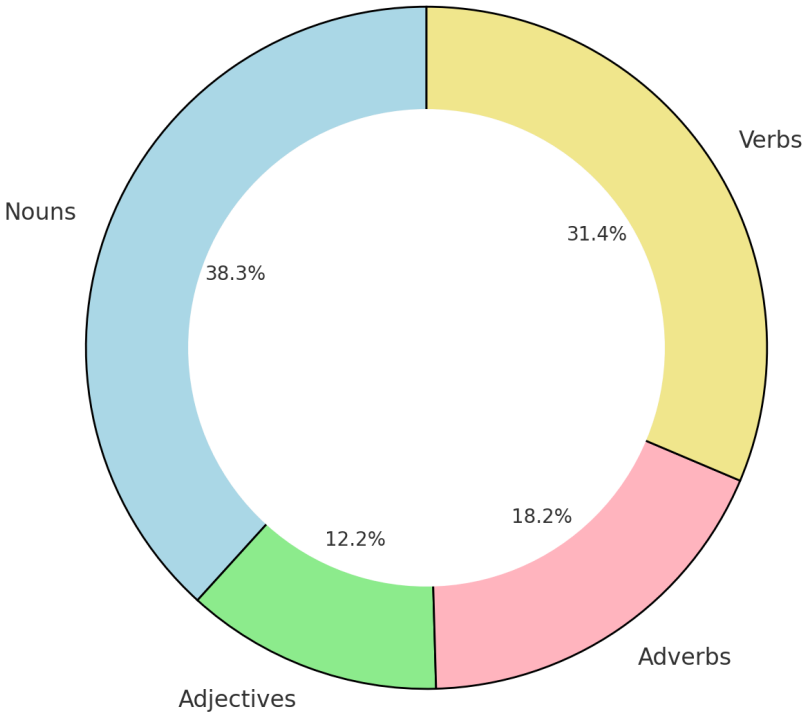


Малюнок 4.1 Розподілення частин мови у відсотковому співвідношенні в таблиці 'lemmas_freq'

Після заповнення таблиці 'vocab' лемами з таблиці 'lemmas_freq' , класифікованими строго як іменники, прикметники, дієслова та прислівники, таблиця налічує загалом 47 327 записів. Для того, щоб полегшити цілеспрямований лінгвістичний аналіз, було прийнято методологічне рішення про дистиляцію цього великого набору даних. В результаті для більш детального вивчення було відібрано першу тисячу лем, які демонструють найвищу частотність. Цей стратегічний відбір має на меті визначити пріоритетність лем, які є найбільш поширеними в корпусі, таким чином надаючи цінну інформацію про основні лексичні компоненти проаналізованих текстів.

На малюнку 4.2 зображений розподіл частин мови серед сформованого словника з 1000 слів. Діаграма ілюструє розподіл 1 000 вибраних слів за частинами мови, демонструючи значне переважання іменників та дієслів, які становлять 38,3% та 31,4% відповідно. Це

підкреслює акцент на предметах і діях у текстах, відображаючи їхню важливу роль у побудові речень і передачі змісту. Прикметники та прислівники, що становлять 12,2% та 18,2% відповідно, вказують на багатий описовий шар, хоча й менш різноманітний, ніж іменники та дієслова..



Малюнок 4.2 Розподілення частин мови у відсотковому співвідношенні в укладеному словнику

4.2. Аналіз частоти лексики

У таблиці 'vocab' відображено розподіл лем залежно від їх частоти, який варіюється від мінімальної частоти 1 до максимальної 12592. Ці дані поділяються на три категорії залежно від частотності використання лем:

- Високочастотні леми. Ця категорія налічує 2504 леми з частотою від 40 до 12592. Ці слова є найбільш часто використовуваними у

корпусі текстів, що вказує на їх велику лінгвістичну та комунікативну значущість.

- Середньочастотні леми. У цю групу входять 5895 лем з частотою від 10 до 39. Ці слова зустрічаються досить часто, але не є настільки розповсюдженими, як високочастотні леми.
- Низькочастотні леми. Найбільша група, яка містить 38928 лем з частотою від 1 до 9. Ці слова з'являються рідко, що може свідчити про їх специфічне використання в текстах або обмежене поширення.

У роботі з лінгвістичними базами даних частотність використання лем є ключовим показником для розуміння їх ролі та впливу на мову. Високочастотні леми, такі як "бути", "так", "сказати", "коли", і "ще" мають частоту використання від 3601 до 12592 разів. Ці слова є невід'ємною частиною повсякденного мовлення та текстових структур, підкреслюючи їхню фундаментальну роль у комунікативних процесах і структурній цілісності мови.

Середньочастотні леми, з частотою від 10 до 39 разів, такі як "багатство", "безперечно", "качан", "вата", і "вдруге", зустрічаються менш регулярно. Вони не є так часто вживаними як високочастотні леми, але все ще відіграють важливу роль у формуванні більш складних та специфічних мовних конструкцій. Ця категорія словникового запасу часто включає слова, які можуть нести спеціалізоване значення або бути характерними для певних стилів мовлення.

Низькочастотні леми, які включають слова з частотою від 1 до 9 разів, такі як "автомашина", "агент", "аеропорт", "родимка", і "акулячий", використовуються рідко.

В укладеному словнику, що складається з 1000 лексичних одиниць, діапазон частот вживання слів коливається від мінімальної 92 до максимальної 12 591. Такий розподіл вказує на значну варіативність

частоти вживання термінів у корпусі, що відображає різну лексичну значущість у наборі даних.

4.3. Лексичні групи слів

Лексичні групи у лінгвістиці — це категорії, до яких класифікуються слова на основі їхніх семантичних, синтаксичних, морфологічних, або інших ознак. Це один зі способів систематизації лексичного матеріалу мови, що дозволяє глибше аналізувати та розуміти мовні структури та їх використання.

У лінгвістиці іменники можуть бути класифіковані за різними критеріями, зокрема за природою об'єктів, які вони позначають. Одними з базових категорій серед іменників є абстрактні, конкретні та власні іменники.

1. Конкретні іменники позначають фізичні об'єкти, які можна сприйняти за допомогою п'яти чуттів. Ці іменники описують реальні предмети у світі, такі як "стіл", "кіт", "дерево".

2. Абстрактні іменники належать до понять, станів або ідеалів, які не мають фізичної форми і не можуть бути сприйняті через чуття. Вони виражають ідеї, якості або стани, такі як "любов", "свобода", "радість".

3. Власні іменники — це назви осіб, місць, організацій, іноді подій або унікальних об'єктів, які вживаються для конкретизації і відрізняються від загальних назв. Вони пишуться з великої літери та включають імена людей (наприклад, "Марія"), географічні назви ("Київ", "Амазонка"), назви компаній ("Google") та інші унікальні об'єкти.

Щоб підвищити чистоту і точність частотного аналізу, словник був ретельно очищений від власних іменників. З первинного набору даних були виключені власні назви, які включають імена конкретних осіб, місць, організацій, а іноді й унікальні події. Таке виключення має вирішальне значення, оскільки власні імена мають унікальну референційну функцію в мові, суттєво відрізняючись від загальних іменників у використанні та

частоті вживання. Вилучення власних назв гарантує, що статистичний аналіз частотності слів точніше відображає закономірності вживання загальних абстрактних і конкретних іменників, уникаючи таким чином викривлення, спричиненого специфікою власних назв.

Прикметники поділяють на два окремі типи: якісні та відносні. Якісні прикметники описують невід'ємні або внутрішні якості іменника, які можна сприйняти або виміряти об'єктивно. Ці прикметники необхідні для опису сенсорних атрибутів, фізичних властивостей або психологічних станів. Наприклад, 'великий', 'маленький', 'старий', 'білий'.

Відносні прикметники, навпаки, вказують на зв'язок або асоціацію між іменником, що змінюється, та іншим предметом, поняттям або характеристикою. Ці прикметники передусім стосуються не внутрішніх якостей, а таких аспектів, як походження, тип, призначення або зв'язок з іншими іменниками. Наприклад, 'сусідній', 'морський', 'дитячий'.

Серед прислівників є якісно-означальні *(напр. швидко)*, кількісно-означальні *(напр. дуже)*, прислівники способу дії *(напр. верхи)* , обставинні *(напр. нарешті),* прислівники часу *(напр. вранці),* місця *(напр. додому)* , причини *(напр. дуже),* власне модальні *(напр. напевно),* предикативні модальні *(напр. треба),* вказівні займенникові *(напр. тут),* означальні займенникові *(напр. інакше),* питальні займенникові *(напр. куди?),* відносні займенникові *(напр. звідки),* неозначені займенникові *(напр. якось),* заперечні займенникові *(напр. ніде)*.

Серед дієслів наявні як і предикати дії, так і предикати стану. Предикати дії – це дієслова, які характеризують предмет чи особу. Тоді як, предикати стану – це дієслова, які також виражають ознаки предмета, але не завжди характеризуються значенням особи [2].

ВИСНОВКИ ДО РОЗДІЛУ 4

У Розділі 4 проводиться поглиблений аналіз частотної лексики, присутньої в дитячій літературі, що проливає світло на значущі мовні

тенденції та характеристики. Завдяки аналізу частотності лексичних одиниць, розподілу за частинами мови та лексичними групами досягається всебічне розуміння структурних і семантичних аспектів мови в дитячих текстах.

Дані, витягнуті з таблиці 'lemmas_freq', розкривають 51 353 леми, 42,8% з яких належать до категорії іменників, що демонструє значне розмаїття суб'єктів, об'єктів і понять, представлених у дитячій літературі. Дієслова та прикметники, що становлять 25,9% та 17,5% відповідно, відображають динамічність та описовість текстів. Прислівники, хоча й менш поширені (5,6%), роблять значний внесок у непряму характеристику дій та явищ.

Аналіз також включав класифікацію та відбір першої тисячі лем з найвищою частотністю, що дозволило виявити найуживаніші слова, які відіграють вирішальну роль у комунікації та забезпеченні структурної цілісності мови. Розподіл частин мови в цій вибірці підтверджує домінування іменників та дієслів, що вказує на зосередженість текстів на об'єктах та діях.

Загалом, результати частотного аналізу лексики в дитячій літературі дають цінну інформацію про фундаментальні мовні аспекти, що слугує основою для подальших досліджень у галузі дитячої літератури, лінгвістики та освіти. Ці знання допоможуть у розробці навчальних матеріалів і стратегій викладання, спрямованих на підтримку розвитку мовних навичок дітей.

ВИСНОВКИ

Ця наукова робота є комплексним дослідженням, спрямованим на розробку частотного словника української мови з метою сприяння мовленнєвому розвитку дітей дошкільного віку. Дослідження охоплює кілька важливих аспектів, зокрема теоретичне обґрунтування важливості частотного словника, аналіз сучасного стану досліджень у галузі психолінгвістики дитячого мовлення та практичне застосування розробленого словника в освітньому процесі.

Перший розділ дослідження присвячений психолінгвістичним аспектам розвитку мовлення дітей, в якому розглядаються основні теорії та дослідження, що формують сучасні уявлення про засвоєння мови. Обговорення теорій Жана Піаже та інших видатних вчених підкреслює складність і багатовимірність мовного розвитку, підтверджуючи необхідність індивідуального підходу в освіті.

Другий розділ присвячено технічним аспектам створення бази даних текстів дитячої літератури, використаних для укладання словника. Описано процеси відбору текстів, вимоги до їхнього змісту, вікові особливості цільової аудиторії, методи комп'ютерної обробки даних. У цьому розділі підкреслено роль сучасних технологій у лінгвістичних дослідженнях, наголошено на їхньому внеску в забезпечення точності та ефективності аналізу.

Третій розділ присвячено укладанню самого частотного словника та його практичному використанню. У ньому описано, як словник може бути використаний у педагогічній практиці, зокрема в контексті дошкільної освіти, де він відіграє життєво важливу роль у розвитку мовних навичок дітей та підготовці їх до школи. Словник сприяє підвищенню літературної грамотності, розвитку мовленнєвих навичок, а також є незамінним помічником для педагогів, батьків та логопедів.

Четвертий розділ присвячено всебічному аналізу частотної лексики, що зустрічається в дитячій літературі, який висвітлює важливі нюанси вживання мови в цих текстах. Словник охоплює загалом 51 353 леми, більшість з яких - іменники (42,8%), що відображають широкий спектр суб'єктів та об'єктів, зображених у книжках. Дієслова (25,9%) та прикметники (17,5%) також займають значне місце в лексиці, демонструючи живий та описовий характер текстів. Крім того, прислівники та інші частини мови збагачують словник, надаючи додаткову інформацію про дії та характеристики. Аналіз виходить за рамки простого підрахунку частотності словесних одиниць, заглиблюючись у їхній розподіл у текстах і пропонуючи цінну інформацію про структурну та семантичну взаємодію, притаманну дитячій літературі.

Це дослідження є вагомим внеском у лінгвістичну науку і практику, яке відкриває нові можливості для подальших наукових розвідок і застосування в освітньому процесі. Інтеграція теоретичних знань і практичних навичок у створенні частотного словника дозволяє дослідникам і педагогам поглибити своє розуміння мовного розвитку дітей. Крім того, це створює міцний фундамент для їхньої успішної академічної та соціальної адаптації, що має вирішальне значення для їхніх майбутніх перспектив.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. АКТУАЛЬНІ ПИТАННЯ ГУМАНІТАРНИХ НАУК:. // Міжвузівський збірник наукових праць молодих вчених Дрогобицького державного педагогічного університету імені Івана Франк / , 2023. – С. 165–172. – Режим доступу до ресурсу: http://www.aphn-journal.in.ua/archive/63_2023/part_2/63-2_2023.pdf#page=165

2. Алексієнко Л. МОРФОЛОГІЯ СУЧАСНОЇ УКРАЇНСЬКОЇ МОВИ / Л. Алексієнко, О. Зубань, І. Козленко. – Київ, 2013.

3. Вікові особливості мовленнєвого розвитку дитини [Електронний ресурс] – Режим доступу до ресурсу: https://zhashkiv4.dytsadok.org.ua/logoped-20-35-43-20-12-2018/.

4. Воротинцева О. «Основні етапи розвитку мовлення в дітей» // Батькам про дітей. Поради логопеда No3. – URL: http://zakinppo.org.ua/images/2020/docs/05/Батькам_про_дітей_3.pdf (дата звернення 07.06.2021).

5. Гільтайчук В. ДОСЛІДЖЕННЯ УСНОГО ДИТЯЧОГО МОВЛЕННЯ [Електронний ресурс] / В. Гільтайчук, В. Басіста – Режим доступу до ресурсу: https://docs.google.com/document/d/1hPmJOhLcPnMts2nupQxZzx-UvAJLqU__kscZv0SvOIM/edit?usp=sharing.

6. Дарчук Н. П. КОМП'ЮТЕРНА ЛІНГВІСТИКА (автоматичне опрацювання тексту) / Наталя Петрівна Дарчук. – Київ, 2008.

7. Заклади дошкільної освіти. Перелік навчальної літератури [Електронний ресурс] – Режим доступу до ресурсу: https://docs.google.com/spreadsheets/d/1-6Qn3PRPqSpDreBZkBFwaPO_ZVWwhAkO-FGGutGpmC8/edit#gid=1947501369.

8. Карпіловська Є. А. ВСТУП ДО ПРИКЛАДНОЇ ЛІНГВІСТИКИ: КОМП'ЮТЕРНА ЛІНГВІСТИКА / Є. А. Карпіловська. –

Донецьк, 2006.

9. Кононко О. Л. Програма розвитку дитини від народження до шести років / О. Л. Кононко // Я У Світі. / О. Л. Кононко. – Київ, 2019. – С. 464–478.

10. Перебийніс В. І. Статистичні методи для лінгвістів / Валентина Ісидорівна Перебийніс. – Вінниця, 2002.

11. Піроженко Т. О. ОСВІТНЯ ПРОГРАМА / Т. О. Піроженко // ВПЕВНЕНИЙ СТАРТ / Т. О. Піроженко., 2017. – С. 62–64.

12. ПРО ПЕРЕЛІКИ НАВЧАЛЬНОЇ ЛІТЕРАТУРИ ТА НАВЧАЛЬНИХ ПРОГРАМ, РЕКОМЕНДОВАНИХ МІНІСТЕРСТВОМ ОСВІТИ І НАУКИ УКРАЇНИ ДЛЯ ВИКОРИСТАННЯ В ОСВІТНЬОМУ ПРОЦЕСІ ЗАКЛАДІВ ОСВІТИ У 2023/2024 НАВЧАЛЬНОМУ РОЦІ [Електронний ресурс] – Режим доступу до ресурсу: https://mon.gov.ua/ua/npa/pro-pereliki-navchalnoyi-literaturi-ta-navchalnih-program-rekomendovanih-ministerstvom-osviti-i-nauki-ukrayini-dlya-vikoristannya-v-osvitnomu-procesi-zakladiv-osviti-u-20232024-navchalnomu-roci.

13. Центр дослідження літератури для дітей та юнацтва [Електронний ресурс] – Режим доступу до ресурсу: https://www.barabooka.com.ua/tsentr-doslidzhennya-literaturi-dlya-ditej-ta-yunatstva/.

14. Читацькі списки для сучасних школярів: українська література [Електронний ресурс] – Режим доступу до ресурсу: http://www.barabooka.com.ua/chitats-ki-spiski-dlya-suchasnih-shkolyariv-ukrayins-ka-literatura/.

15. Beautiful Soup [Електронний ресурс] – Режим доступу до ресурсу: https://pypi.org/project/beautifulsoup4/.

16. Chardet: The Universal Character Encoding Detector [Електронний ресурс] – Режим доступу до ресурсу: https://pypi.org/project/chardet/.

17. ChildLex (German Children's Book Corpus) [Електронний ресурс] – Режим доступу до ресурсу: https://web.archive.org/web/20170516233242/https://www.mpib-berlin.mpg.de/de/forschung/max-planck-forschungsgruppen/mpfg-read/projekte/childlex.

18. Christian Lehmann. Frequency dictionary [Електронний ресурс] / Christian Lehmann – Режим доступу до ресурсу: https://www.christianlehmann.eu/ling/ling_meth/ling_description/lexicography/frequency_dict.html.

19. Collections — Container datatypes [Електронний ресурс] – Режим доступу до ресурсу: https://docs.python.org/3/library/collections.html.

20. Computer-Assisted Vocabulary Load Analysis [Електронний ресурс]. – 2012. – Режим доступу до ресурсу: https://onlinelibrary.wiley.com/doi/10.1002/9781405198431.wbeal0179.

21. Developing Dovyko I: The Czech Adaptation of the MacArthur-Bates Communicative Development Inventory [Електронний ресурс]. – 2024. – Режим доступу до ресурсу: https://ceskoslovenskapsychologie.cz/index.php/csps/article/view/525.

22. El Corpus de Referencia del Español Actual (CREA) [Електронний ресурс] – Режим доступу до ресурсу: https://www.rae.es/banco-de-datos/crea.

23. Emotion-specific vocabulary and its relation to emotion understanding in children and adolescents [Електронний ресурс]. – 2024. – Режим доступу до ресурсу: https://www.tandfonline.com/doi/full/10.1080/02699931.2024.2346745.

24. Innovation in Language Learning and Teaching: Historical Perspectives [Електронний ресурс]. – 2023. – Режим доступу до ресурсу: https://books.google.com.ua/books?hl=en&lr=&id=KIrGEAAAQBAJ&oi=fnd&pg=PR1&dq=compilation+of+frequency+dictionaries+applied+linguistics&ots=

MbuiylYS__&sig=HbZbRoDl1iM1W3DGJZEkC758wwo&redir_esc=y#v=onepage&q&f=false.

25. JAMES HIGGINBOTHAM. Noam Chomsky\'s Linguistic Theory [Електронний ресурс] / JAMES HIGGINBOTHAM – Режим доступу до ресурсу: https://www.jstor.org/stable/40970856?searchText=Noam%20Chomsky%27s%20theory%20of%20universal%20grammar&searchUri=%2Faction%2FdoBasicSearch%3FQuery%3DNoam%2BChomsky%2527s%2Btheory%2Bof%2Buniversal%2Bgrammar%26so%3Drel&ab_segments=0%2Fbasic_search_gsv2%2Fcontrol&refreqid=fastly-default%3Afb3879630e72511a217f7236e9412100.

26. Janet Werker. The Ontogeny of Speech Perception [Електронний ресурс] / Janet Werker – Режим доступу до ресурсу: https://www.taylorfrancis.com/chapters/edit/10.4324/9781315807942-8/ontogeny-speech-perception-janet-werker.

27. Jean Piaget. The Language and Thought of the Child [Електронний ресурс] / Jean Piaget. – 2002. – Режим доступу до ресурсу: https://books.google.com.ua/books?id=WYoEXQLGRLEC&pg=PA50&hl=uk&source=gbs_toc_r&cad=2#v=onepage&q&f=false.

28. Manulex [Електронний ресурс] – Режим доступу до ресурсу: http://www.manulex.org.

29. Multi-Word Token (MWT) Expansion [Електронний ресурс] – Режим доступу до ресурсу: https://stanfordnlp.github.io/stanza/mwt.html.

30. Named Entity Recognition [Електронний ресурс] – Режим доступу до ресурсу: https://stanfordnlp.github.io/stanza/ner.html.

31. NumPy 1.26.0 released [Електронний ресурс] – Режим доступу до ресурсу: https://numpy.org.

32. Part-of-Speech & Morphological Features [Електронний ресурс] – Режим доступу до ресурсу: https://stanfordnlp.github.io/stanza/pos.html.

33. Putting frequencies in the dictionary [Електронний ресурс]. – 1997. – Режим доступу до ресурсу: https://academic.oup.com/ijl/article-abstract/10/2/135/963654?redirectedFrom=fulltext.

34. Python SQLite [Електронний ресурс]. – 2021. – Режим доступу до ресурсу: https://www.geeksforgeeks.org/python-sqlite/.

35. Requests [Електронний ресурс] – Режим доступу до ресурсу: https://pypi.org/project/requests/.

36. Rick Ansorge, Frances Gatta, Amy Gopal. Piaget Stages of Development [Електронний ресурс] / Rick Ansorge, Frances Gatta, Amy Gopal – Режим доступу до ресурсу: https://www.webmd.com/children/piaget-stages-of-development.

37. Sascha Schroeder. childLex: a lexical database of German read by children [Електронний ресурс] / Sascha Schroeder. – 2014. – Режим доступу до ресурсу: https://www.researchgate.net/publication/266975924_childLex_a_lexical_database_of_German_read_by_children.

38. Should they look it up? The role of dictionaries in language learning [Електронний ресурс]. – 2001. – Режим доступу до ресурсу: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=eb9153aeae76d02bf9e9c7e10fdeb716e8e3469a.

39. SQLite3 [Електронний ресурс] – Режим доступу до ресурсу:https://docs.python.org/uk/3.9/library/sqlite3.html

40. Stanza – A Python NLP Package for Many Human Languages [Електронний ресурс] – Режим доступу до ресурсу: https://stanfordnlp.github.io/stanza/.

41. The Oxford 3000 [Електронний ресурс] – Режим доступу до ресурсу: https://www.oxfordlearnersdictionaries.com/about/oxford3000.

42. The Role of Onomatopoeia in Children\'s Early Language Development [Електронний ресурс]. – 2022. – Режим доступу до ресурсу: https://reshare.ukdataservice.ac.uk/855606/.

43. Web-based frequency dictionaries for medium density languages [Електронний ресурс]. – 2006. – Режим доступу до ресурсу: https://dl.acm.org/doi/10.5555/1628297.1628298.

44. What corpora can offer in language teaching and learning [Електронний ресурс]. – 2011. – Режим доступу до ресурсу: http://ndl.ethernet.edu.et/bitstream/123456789/18149/1/75pdf.pdf#page=383.

45. What is an API (application programming interface)? [Електронний ресурс]. – 2024. – Режим доступу до ресурсу: https://www.ibm.com/topics/api.

46. Word frequency and collocation: Using children's literature in adult learning [Електронний ресурс]. – 2015. – Режим доступу до ресурсу: https://sciendo.com/article/10.1515/icame-2015-0004.

47. Word frequency and the importance of context in vocabulary learning [Електронний ресурс]. – 1977. – Режим доступу до ресурсу: https://journals.sagepub.com/doi/pdf/10.1177/003368827700800202.

ДОДАТОК 1

Доступ до папки з програмним забезпеченням знаходиться за посиланням:
https://drive.google.com/drive/folders/1WM7tm5PNoCiHWuYbKnTPEUkHlkNMirWP?usp=sharing

ДОДАТОК 2

Доступ до папки з базою даних 'kidus.db' та словником знаходиться за посиланням:

https://drive.google.com/drive/folders/1lNUVng4Xb7k__TdCyKIc-iRpKZyQNOEH?usp=sharing