

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Київський національний університет імені Тараса Шевченка
Навчально-науковий інститут філології
Кафедра української мови та прикладної лінгвістики

**ГУЦУЛІЗМИ У ХУДОЖНІХ ТЕКСТАХ МАРІЇ МАТІОС ТА
ТАРАСА ПРОХАСЬКА
(ЛІНГВОСТАТИСТИЧНЕ ДОСЛІДЖЕННЯ НА МАТЕРІАЛІ
АВТОМАТИЧНО УКЛАДЕНОГО ЧАСТОТНОГО СЛОВНИКА)**

Кваліфікаційна робота
освітнього ступеня «бакалавр»
студентки 4 курсу бакалаврату
ОПП «Прикладна (комп'ютерна)
лінгвістика та англійська мова»,
спеціальності 035 «Філологія»,
спеціалізації 035.10 «Прикладна
лінгвістика»,
галузі знань 03 «Гуманітарні науки»
Владислава РАТУШНОГО

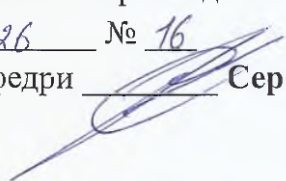
Науковий керівник:
Кандидат філологічних наук, доц.
кафедри української мови та
прикладної лінгвістики
Оксана ЗУБАНЬ

«Допущено до захисту»

Протокол засідання кафедри

української мови та прикладної лінгвістики

від 05.06.2026 № 16

Завідувач кафедри  **Сергій РІЗНИК**

КИЇВ – 2026

ЗМІСТ

АНОТАЦІЯ	4
ВСТУП	8
РОЗДІЛ 1. ЛІНГВІСТИЧНІ ОСОБЛИВОСТІ ГУЦУЛЬСЬКОГО ДІАЛЕКТУ: СИСТЕМНО–МОВНИЙ ТА ФУНКЦІЙНИЙ АСПЕКТИ	10
1.1 Основні риси гуцульського діалекту	10
1.2 Функції гуцулізмів у художніх текстах української літератури	13
Висновки до першого розділу	18
РОЗДІЛ 2. АВТОМАТИЗОВАНЕ УКЛАДАННЯ ЧАСТОТНОГО СЛОВНИКА ГУЦУЛЬСЬКИХ ДІАЛЕКТИЗМІВ ЗА ХУДОЖНІМИ ТЕКСТАМИ МАРІЇ МАТІОС ТА ТАРАСА ПРОХАСЬКА	19
2.1 Частотні словники у Корпусі української мови	19
2.2 Характеристика текстового матеріалу дослідження за даними Корпусу української мови	22
2.3 Автоматичне укладання електронного частотного словника гуцулізмів	25
Висновки до другого розділу	32
РОЗДІЛ 3. ЛІНГВОСТАТИСТИЧНИЙ АНАЛІЗ ГУЦУЛЬСЬКИХ ДІАЛЕКТИЗМІВ У ХУДОЖНІХ ТЕКСТАХ МАРІЇ МАТІОС І ТАРАСА ПРОХАСЬКА	33
3.1 Частотна характеристика гуцулізмів у текстах Марії Матіос і Тараса Прохаська	33
3.2 Лексико–семантична класифікація виявлених гуцульських діалектизмів	38
3.3 Типи гуцулізмів у досліджуваному матеріалі: лексичні, етнографічні, фонетичні, семантичні	43
3.4 Порівняльний аналіз і функційне навантаження гуцулізмів у текстах Марії Матіос і Тараса Прохаська	45
Висновки до третього розділу	50
ВИСНОВКИ	51
ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ	54
СПИСОК ДЖЕРЕЛ ІЛЮСТРАТИВНОГО МАТЕРІАЛУ	57
ДОДАТКИ	58

АНОТАЦІЯ

Бакалаврська робота складається зі вступу, трьох розділів, висновків, списку використаних джерел і додатків. Загальний обсяг роботи становить 49 сторінок. Робота містить 15 таблиць, 7 рисунків, 3 додатків. Список використаних джерел містить 46 найменувань.

У роботі досліджено гуцульські діалектизми в художніх текстах Марії Матіос і Тараса Прохаська. Об'єктом дослідження є гуцульські діалектизми в художніх текстах сучасної української літератури. Предметом дослідження є частотні, лексико–семантичні, типологічні та функційні особливості гуцульських діалектизмів у текстах Марії Матіос і Тараса Прохаська.

Мета роботи полягає в автоматизованому укладанні електронного частотного словника гуцульських діалектизмів за художніми текстами Марії Матіос і Тараса Прохаська та здійсненні лінгвостатистичного аналізу виявлених одиниць. Матеріалом дослідження стали авторські частотні словники Корпусу української мови, сформовані за текстами Марії Матіос і Тараса Прохаська, а також реєстр гуцульських діалектизмів, підготовлений на основі лексикографічних джерел.

У роботі використано описовий, корпусний, лексико–семантичний, типологічний, зіставний і лінгвостатистичний методи, а також елементи автоматизованого опрацювання мовних даних. Описано основні риси гуцульського діалекту, функції гуцулізмів у художньому тексті, структуру вхідних частотних словників і процедуру автоматизованого зіставлення лексем із реєстром гуцульської діалектної лексики.

У результаті дослідження розроблено вебзастосунок для укладання електронного частотного словника гуцулізмів, здійснено очищення й нормалізацію частотних словників, сформовано авторські та спільну порівняльну таблиці, таблицю ручної перевірки й статистичні зведення. У текстах Марії Матіос виявлено 51 гуцулізм, у текстах Тараса Прохаська – 40 гуцулізмів. Практичне значення результатів полягає в можливості їх використання для подальших досліджень діалектної лексики, комп'ютерної лексикографії та корпусного аналізу художніх текстів.

Ключові слова: гуцулізм, гуцульський діалект, діалектизм, частотний словник, корпус української мови, Марія Матіос, Тарас Прохасько, лінгвостатистика, комп'ютерна лексикографія.

ABSTRACT

The bachelor's thesis consists of an introduction, three chapters, conclusions, a list of references and appendices. The total volume of the thesis is 49 pages. The paper contains 15 tables, 7 figures and 3 appendices. The list of references includes 46 sources.

The thesis studies Hutsul dialectisms in the literary texts of Mariia Matios and Taras Prokhasko. The object of the research is Hutsul dialectisms in the literary texts of contemporary Ukrainian literature. The subject of the research is the frequency, lexical–semantic, typological and functional features of Hutsul dialectisms in the texts of Mariia Matios and Taras Prokhasko.

The aim of the thesis is to automatically compile an electronic frequency dictionary of Hutsul dialectisms based on the literary texts of Mariia Matios and Taras Prokhasko and to carry out a linguostatistical analysis of the identified units. The research material includes authorial frequency dictionaries from the Corpus of the Ukrainian Language, compiled on the basis of the texts by Mariia Matios and Taras Prokhasko, as well as a register of Hutsul dialectisms prepared on the basis of lexicographic sources.

The study uses descriptive, corpus–based, lexical–semantic, typological, comparative and linguostatistical methods, as well as elements of automated language data processing. The thesis describes the main features of the Hutsul dialect, the functions of Hutsul dialectisms in literary texts, the structure of the input frequency dictionaries and the procedure of automated comparison of lexemes with the register of Hutsul dialect vocabulary.

As a result of the research, a web application for compiling an electronic frequency dictionary of Hutsul dialectisms was developed. The frequency dictionaries were cleaned and normalized, authorial tables, a joint comparative table, a manual review table and statistical summaries were generated. In the texts of Mariia Matios, 51 Hutsul dialectisms were identified, while in the texts of Taras Prokhasko, 40 Hutsul dialectisms were found. The practical value of the results lies in their possible use for further studies of dialect vocabulary, computer lexicography and corpus analysis of literary texts.

Keywords: hutsulism, hutsul dialect, dialectism, frequency dictionary, corpus of the ukrainian language, Mariia Matios, Taras Prokhasko, linguostatistics, computer lexicography.

ВСТУП

Актуальність теми дослідження зумовлена потребою поєднання традиційного мовознавчого аналізу з методами корпусної та прикладної лінгвістики. У працях Н. Дарчук корпус української мови й частотні словники розглядаються як важливі електронні ресурси для автоматизованого опрацювання мовного матеріалу [Дарчук 2008; Дарчук 2010]. У дослідженнях Т. Мак–Енері, Е. Гарді, П. Бейкера та М. Стаббса обґрунтовано значення корпусного й лінгвостатистичного підходів для кількісного аналізу мовних явищ [McEneaney, Hardie 2012; Baker 2006; Stubbs 2001]. Саме тому використання електронних частотних словників і автоматизованого зіставлення лексичних одиниць є доцільним для дослідження гуцульської діалектної лексики в художніх текстах.

Гуцульський діалект є виразним територіальним різновидом української мови. У художніх текстах гуцулізми виконують номінативну, локалізаційну, етнокультурну, характерологічну та стилістичну функції. Особливий інтерес становить їх функціонування у прозі Марії Матіос і Тараса Прохаська, оскільки творчість обох авторів пов'язана з карпатським культурним простором [Бевзенко 1980; Герман 2010; Грещук 2009; Гуцульські 1997].

Об'єктом дослідження є гуцульські діалектизми в художніх текстах сучасної української літератури.

Предметом дослідження є частотні, лексико–семантичні, типологічні та функційні особливості гуцульських діалектизмів у художніх текстах Марії Матіос і Тараса Прохаська.

Метою роботи є автоматизоване укладання електронного частотного словника гуцульських діалектизмів за художніми текстами Марії Матіос і Тараса Прохаська та здійснення лінгвостатистичного аналізу виявлених одиниць.

Для досягнення мети поставлено такі завдання:

1. Охарактеризувати основні риси гуцульського діалекту та функції гуцулізмів у художньому тексті.
2. Розглянути частотні словники Корпусу української мови як джерело матеріалу для дослідження.
3. Охарактеризувати текстовий матеріал за авторськими частотними словниками Марії Матіос і Тараса Прохаська.
4. Розробити методику до автоматизованого укладання електронного частотного словника гуцулізмів.
5. Здійснити очищення, нормалізацію та зіставлення частотних словників із реєстром гуцульської діалектної лексики.
6. Проаналізувати кількість, частотність, семантичні групи, типи та функції виявлених гуцулізмів.
7. Порівняти особливості авторського вживання гуцулізмів у текстах Марії Матіос і Тараса Прохаська.

Матеріалом дослідження є авторські частотні словники, сформовані за художніми текстами Марії Матіос і Тараса Прохаська в Корпусі української мови, а також реєстр гуцульських діалектизмів, укладений на основі лексикографічних джерел. Після очищення частотний словник за текстами Марії Матіос охоплює 440662 слововживання і 38996 лексем реєстру частотного словника, а частотний словник за текстами Тараса Прохаська – 127350 слововживань і 16759 лексем реєстру частотного словника. У результаті зіставлення цих словників із реєстром гуцульської діалектної лексики виявлено 51 лексему реєстру частотного словника гуцулізмів у матеріалі Марії Матіос і 40 лексем у матеріалі Тараса Прохаська.

У роботі використано описовий **метод** для характеристики гуцульського діалекту та функцій гуцулізмів у художньому тексті; лексико–семантичний метод для класифікації виявлених одиниць за тематичними групами; порівняльний метод для зіставлення матеріалу Марії Матіос і Тараса Прохаська; корпусний метод для роботи з авторськими частотними словниками; лінгвостатистичний метод для визначення абсолютної частоти,

відносної частоти, покриття тексту гуцулізмами та частки гуцулізмів у реєстрі частотного словника. Для реалізації прикладної частини застосовано мову програмування Python, локальний вебінтерфейс Streamlit, бібліотеки pandas, openpyxl, re та collections, а також табличні формати TXT і XLSX для збереження та експорту результатів [Python б. д.; Streamlit б. д.; pandas б. д.; openpyxl б. д.].

Наукова новизна роботи полягає в поєднанні описового аналізу гуцулізмів із автоматизованим частотним опрацюванням авторських словників. У роботі гуцульську діалектну лексику описано не лише якісно, а й кількісно: визначено абсолютну та відносну частоту виявлених одиниць, покриття тексту гуцулізмами, частку гуцулізмів у реєстрі частотного словника, а також спільні й унікальні одиниці в авторських корпусах.

Практичне значення роботи полягає в можливості використання розробленої методики для подальших досліджень діалектної лексики, укладання електронних частотних словників, аналізу художнього тексту та вивчення авторського ідіостилу. Отримані результати можуть бути використані в курсах української діалектології, лексикології, стилістики художнього тексту, корпусної лінгвістики та прикладної лексикографії.

Кваліфікаційна робота складається зі вступу, трьох розділів, висновків, переліку використаних джерел і додатків. У першому розділі розглянуто основні риси гуцульського діалекту та функції гуцулізмів у художньому тексті. У другому розділі обґрунтовано методику автоматизованого укладання електронного частотного словника гуцулізмів. У третьому розділі здійснено лінгвостатистичний аналіз виявлених одиниць у текстах Марії Матіос і Тараса Прохаська. Перелік використаних джерел становить 46 позицій, серед них 19 джерел англійською мовою.

РОЗДІЛ 1. ЛІНГВІСТИЧНІ ОСОБЛИВОСТІ ГУЦУЛЬСЬКОГО ДІАЛЕКТУ: СИСТЕМНО–МОВНИЙ ТА ФУНКЦІЙНИЙ АСПЕКТИ

1.1 Основні риси гуцульського діалекту

Гуцульський діалект належить до південно–західного наріччя української мови й становить один із виразних територіальних різновидів українського діалектного простору. У працях з української діалектології гуцульські говірки розглядаються як складова карпатського ареалу, що має специфічні фонетичні, морфологічні, словотвірні, лексичні та семантичні ознаки [Бевзенко 1980; Жилко 1966; Аркушин 2000; Гриценко 2004]. Їхня своєрідність зумовлена історичним розвитком українських карпатських говорів, географічними умовами функціонування, а також тісним зв'язком із традиційною матеріальною і духовною культурою Гуцульщини.

Територіальне поширення гуцульського діалекту пов'язане з етнографічним регіоном Гуцульщини. Гуцульські говірки функціонують переважно в гірських районах Івано–Франківської, Чернівецької та Закарпатської областей. Географічна локалізація, відносна історична відокремленість окремих поселень, особливості традиційного господарства, побуту й культури сприяли формуванню виразної мовної системи гуцульського говору [Аркушин 2000; Жилко 1966]. Саме тому гуцульська діалектна лексика тісно пов'язана з назвами природного середовища, гірського ландшафту, полонинського господарства, родинного життя, ремесел, одягу, їжі та обрядовості [Грещук 2009; Грещук, Грещук 2020; Гуцульські 1997].

Однією з помітних рис гуцульського діалекту є фонетична своєрідність. У гуцульських говірках простежуються особливості вимови голосних і приголосних, які можуть передаватися в художньому тексті графічно. Наприклад, у словникових джерелах фіксуються форми на зразок **вжинитисі**, **дири́вйиний**, **зимльи**, що засвідчують локальні фонетичні реалізації [Грещук 2009; Гуцульська 2020]. У художній літературі такі форми можуть виконувати

роль засобу стилізації мовлення персонажів і створення регіонального мовного колориту.

Морфологічні особливості гуцульського діалекту виявляються у специфічних формах словозміни, варіантах закінчень, формах займенників, часток, дієслів і службових слів. У художньому тексті такі одиниці можуть функціонувати як маркери усного мовлення. Водночас вони не завжди легко піддаються автоматизованому виявленню, оскільки часто проявляються не на рівні окремої словникової одиниці, а в межах словоформи або синтаксичного контексту [Бевзенко 1980; Жилко 1966; Гриценко 1987].

На словотвірному рівні гуцульські говірки характеризуються наявністю похідних, пестливих, згрубілих, оцінних і локально маркованих форм. Такі одиниці важливі для аналізу художніх текстів, оскільки письменники можуть використовувати не тільки базові діалектні лексеми, а й похідні від них форми. Для автоматизованого аналізу це створює додаткову складність: точний збіг із реєстром словника не завжди дає змогу виявити всі варіантні або похідні форми [Грещук 2009; Гуцульська 2020; Герман 2010].

На основі праць із української діалектології та досліджень гуцульських говірок узагальнено основні мовні ознаки гуцульського діалекту, які подано в таблиці 1.1 [Бевзенко 1980; Жилко 1966; Аркушин 2000; Гриценко 2004; Гуцульські 1997].

Таблиця 1.1

Основні мовні ознаки гуцульського діалекту

Мовний рівень	Основні ознаки	Приклади / прояви	Значення для дослідження
Фонетичний	Відмінності у вимові голосних і приголосних, графічна передача діалектного звучання	форми типу <i>вжсинитисі</i> , <i>диривійиний</i> , <i>зимльи</i>	Допомагає виявляти фонетичні гуцулізми та пояснювати стилізацію мовлення персонажів

Продовження таблиці 1.1

Мовний рівень	Основні ознаки	Приклади / прояви	Значення для дослідження
Морфологічний	Специфічні словоформи, варіанти закінчень, діалектні форми службових слів	локальні форми займенників, дієслів, часток	Потребує ручної перевірки, оскільки не всі форми виявляються автоматично
Словотвірний	Похідні, пестливі, згрубілі, оцінні форми	похідні від базових діалектизмів, наприклад форми навколо <i>тазда</i>	Важливий для пошуку варіантів, які можуть бути відсутні в основному словниковому реєстрі
Лексичний	Самобутні діалектні слова з літературними відповідниками або описовими тлумаченнями	назви осіб, предметів побуту, природних явищ, господарських процесів	Є основним матеріалом для укладання електронного частотного словника
Семантичний	Відмінність значення діалектної одиниці від літературного відповідника	слова із локально специфічним значенням	Дає змогу класифікувати гуцулізми за типами й семантичними групами
Етнокультурний	Назви локальних реалій традиційної культури, побуту, обрядів	лексика одягу, їжі, житла, полонинського господарства	Пояснює культурно–маркувальну функцію гуцулізмів у художньому тексті

Подані в таблиці 1.1 ознаки засвідчують, що гуцульський діалект має комплексний характер і виявляється на різних мовних рівнях. Для цього дослідження особливе значення має саме лексичний рівень, оскільки гуцулізми найчастіше репрезентують назви предметів побуту, осіб, явищ природи, елементів традиційної культури, господарства, одягу, їжі та простору. Саме лексичні одиниці найкраще фіксують зв'язок мови з локальною культурою, тому вони є придатним матеріалом для частотного й лексико–семантичного аналізу.

Найбільш виразною рисою гуцульського діалекту є його лексична самобутність. Гуцульська діалектна лексика охоплює назви осіб, предметів побуту, одягу, їжі, житла, господарських процесів, природних явищ, обрядів, емоційних станів, соціальних ролей і локальних культурних реалій [Грещук 2009; Грещук, Грещук 2020; Гуцульські 1997]. Значна частина таких одиниць має літературні відповідники, однак частина потребує описового тлумачення, оскільки позначає специфічні поняття традиційної гуцульської культури.

За характером співвідношення з літературною мовою гуцульські діалектизми можна поділяти на лексичні, етнографічні, семантичні та фонетичні. Лексичні діалектизми мають відповідники в літературній мові, але відрізняються від них формою або сферою вживання. Етнографічні діалектизми позначають реалії матеріальної та духовної культури певного регіону. Семантичні діалектизми збігаються з літературними словами за формою, але мають інше або додаткове значення в межах говірки. Фонетичні діалектизми відрізняються передусім звуковим оформленням [Бевзенко 1980; Гриценко 2004; Зубань 2018].

Особливе місце в гуцульській діалектній системі посідає етнографічна лексика. Вона безпосередньо пов'язана з традиційним способом життя гуцулів і часто не має точних однослівних відповідників у літературній мові. Такі слова називають предмети одягу, елементи житла, знаряддя праці, страви, обрядові дії, елементи полонинського господарства та родинно-соціальні ролі. У художньому тексті етнографічні гуцулізми виконують не лише номінативну, а й культурно-маркувальну функцію [Грещук, Грещук 2020; Гуцульська 2020; Лесюк 2014].

Важливими джерелами для вивчення гуцульської лексики є словник «Гуцульська діалектна лексика та фраземіка в українській художній мові» [Гуцульська 2020] та словник «Гуцульські говірки. Короткий словник» [Гуцульські 1997]. Ці лексикографічні праці дають змогу встановити словникову фіксацію гуцульських діалектизмів, уточнити їхнє значення, визначити належність до певної семантичної сфери та використати відповідні

одиниці як основу для автоматизованого зіставлення з авторськими частотними словниками.

У межах цієї роботи словниково зафіксовані гуцулізми використано для формування реєстру гуцульської діалектної лексики, з яким зіставлялися частотні словники художніх текстів Марії Матіос і Тараса Прохаська. Такий підхід дає змогу перейти від загального опису гуцульської лексики до її кількісного аналізу в конкретних авторських корпусах.

Отже, гуцульський діалект є важливим об'єктом лінгвістичного дослідження, оскільки поєднує власне мовні, культурні та регіональні ознаки. Для аналізу художніх текстів особливу цінність становить гуцульська діалектна лексика, адже вона фіксує локальні реалії, передає етнокультурну специфіку Гуцульщини й бере участь у створенні мовного образу регіону. Саме тому гуцулізми є релевантним матеріалом для автоматизованого укладання частотного словника та подальшого лінгвостатистичного аналізу.

1.2 Функції гуцулізмів у художніх текстах української літератури

Діалектна лексика в художньому тексті є не лише засобом передавання регіонального мовлення, а й важливим елементом художньої організації твору. У лінгвістичних працях діалектизми розглядаються як одиниці, здатні моделювати мовне середовище, соціокультурний простір, індивідуальне мовлення персонажів і локальну картину світу [Кочерган 2006; Лесюк 2014; Грещук 2009]. Їх використання залежить від авторського задуму, тематики твору, типу оповіді та ступеня занурення художніх подій у конкретний регіональний простір.

У дослідженнях В. Грещука та В. Грещук гуцульська діалектна лексика розглядається як важливий засіб репрезентації регіонального мовного колориту в українській художній мові [Грещук, Грещук 2020]. Дослідники наголошують на тому, що такі одиниці не лише називають локальні реалії, а й беруть участь у створенні етнокультурного образу Гуцульщини. Отже, гуцулізми в художньому

тексті не можна трактувати тільки як декоративні мовні елементи: вони виконують смислову, стилістичну та культурно—маркувальну роль.

Гуцулізми як різновид діалектної лексики мають особливе функційне навантаження, оскільки вони пов'язані не лише з територіальною мовною специфікою, а й з культурним образом Гуцульщини. У художніх текстах такі одиниці можуть позначати предмети побуту, родинні й соціальні ролі, явища природи, особливості господарювання, обрядові дії, елементи традиційного одягу, їжі, житла й духовної культури. Через це гуцулізми не тільки називають певні реалії, а й уводять у текст локальний культурний код [Гуцульська 2020; Гуцульські 1997; Лесюк 2014].

Однією з провідних функцій гуцулізмів є номінативна функція. Вона полягає в називанні предметів, явищ і понять, характерних для гуцульського середовища. Значна частина діалектизмів позначає локальні культурно—побутові реалії, які не завжди мають точні відповідники в літературній мові. У художньому тексті такі слова дають змогу не пояснювати локальний світ зовнішньо, а показувати його через природну для персонажів лексику [Грещук 2009; Гуцульська 2020].

Локалізаційна функція полягає в тому, що гуцулізми закріплюють події твору за конкретним етнографічним і мовним простором. Вони створюють ефект упізнаваного регіону, у якому мова персонажів, побутові деталі, назви предметів і способи мовлення пов'язані з Гуцульщиною. У цьому сенсі діалектизм є не лише мовною одиницею, а й знаком простору [Грещук, Грещук 2020; Лесюк 2014].

Етнокультурна функція гуцулізмів пов'язана з передачею традиційної системи цінностей, звичаїв, моделей поведінки, побутових практик і світоглядних уявлень носіїв говірки. Особливо це стосується етнографічних діалектизмів, які називають предмети та явища, тісно пов'язані з регіональною культурою. Такі слова не можна повністю замінити літературними відповідниками без втрати культурного відтінку [Гуцульська 2020; Гуцульські 1997; Грещук, Грещук 2020].

Характерологічна функція виявляється тоді, коли гуцулізми допомагають розкрити походження персонажа, його соціальну належність, емоційний стан, близькість до традиційного середовища або дистанціювання від нього. У діалогах діалектні слова, фонетичні форми, частки й фразеологізми створюють ефект живого усного мовлення та формують мовний образ персонажа [Зубань 2018; Карпенко 2006; Кочерган 2010].

Стилістична й експресивна функції пов'язані з увиразненням художньої мови. Гуцульська лексика може функціонувати як засіб стилізації, створювати регіональний колорит, передавати емоційне напруження, оцінку, іронію, фамільярність або драматизм ситуації. У таких випадках важливо враховувати не лише словникове значення діалектизму, а й контекст його вживання [Кочерган 2006; Кочерган 2010; Зубань 2018].

У прозових текстах гуцулізми можуть мати й текстотвірну функцію. Повторюваність певних діалектних одиниць може свідчити про їхню семантичну або композиційну значущість. Саме тому частотний аналіз гуцулізмів дає змогу виявити не тільки кількість діалектних одиниць у тексті, а й ті слова, які стають лексичними домінантами авторського мовного світу [McEneaney, Hardie 2012; Baker 2006; Stubbs 2001].

З урахуванням праць, присвячених діалектній лексиці в художньому тексті, можна виокремити кілька основних функцій гуцулізмів: номінативну, локалізаційну, етнокультурну, характерологічну, експресивну, стилістичну та текстотвірну. Узагальнення цих функцій подано в таблиці 1.2 [Грещук 2009; Грещук, Грещук 2020; Лесюк 2014; Зубань 2018].

Таблиця 1.2

Функції гуцулізмів у художньому тексті

Функція	Зміст функції	Прояв у художньому тексті	Значення для дослідження
Номінативна	Називання локальних предметів, явищ, дій і понять	Позначення побутових, природних, господарських, обрядових реалій	Дозволяє виокремити гуцулізми як одиниці словникового реєстру

Продовження таблиці 1.2

Функція	Зміст функції	Прояв у художньому тексті	Значення для дослідження
Локалізаційна	Закріплення подій за конкретним регіональним простором	Створення мовного образу Гуцульщини	Пояснює зв'язок діалектизмів із просторовою організацією тексту
Етнокультурна	Передача традицій, побуту, звичаїв і локальної картини світу	Використання назв одягу, їжі, житла, обрядів, народних уявлень	Дає змогу визначати етнографічні гуцулізми
Характерологічна	Мовна характеристика персонажа	Діалектні слова в репліках, діалогах, внутрішньому мовленні	Допомагає аналізувати зв'язок мови персонажа з його походженням і соціальним середовищем
Стилістична	Створення мовного колориту й авторської манери	Стилізація усного мовлення, увиразнення оповіді	Пояснює роль гуцулізмів у художній організації тексту
Експресивна	Передача емоцій, оцінки, ставлення мовця	Пестливі, згрубілі, лайливі, оцінні, емоційно забарвлені одиниці	Важлива для якісного контекстуального аналізу
Текстотвірна	Участь у формуванні цілісної атмосфери й повторюваних мотивів	Повторення ключових діалектизмів, створення семантичних домінант	Пов'язує функційний аналіз із частотним аналізом

Дані таблиці 1.2 показують, що гуцулізми в художньому тексті не обмежуються роллю декоративного засобу. Вони можуть одночасно називати локальні реалії, маркувати простір, характеризувати персонажа, передавати етнокультурну інформацію та формувати стильову своєрідність тексту.

У цьому дослідженні гуцулізми в художніх текстах Марії Матіос і Тараса Прохаська розглядаються як одиниці, придатні для корпусного й лінгвостатистичного аналізу. Частотний словник дає змогу формалізувати мовний матеріал: виділити повторювані лексеми, визначити їхню абсолютну частоту, частиномовну належність і семантичне поле. Після зіставлення такого словника з реєстром гуцульської діалектної лексики можна отримати

електронний словник гуцулізмів, який відображає конкретний кількісний матеріал [Дарчук 2008; Дарчук 2010; Частотні б. д.].

Водночас автоматизований підхід не усуває потреби в лінгвістичній інтерпретації. Частотність сама по собі не пояснює функцію слова в художньому тексті. Одне й те саме слово може бути нейтральною назвою побутового предмета, засобом характеристики персонажа, емоційним маркером або елементом авторської стилізації [Sinclair 1991; Stubbs 2001; Gries 2009].

Зв'язок системно–мовних ознак гуцульського діалекту з функціями гуцулізмів у художньому тексті узагальнено на рисунку 1.1.



Рисунк 1.1 – Системно–мовні та функційні аспекти гуцульського діалекту в художньому тексті

Як показано на рисунку 1.1, гуцульський діалект у межах дослідження розглядається у двох взаємопов'язаних площинах. З одного боку, він постає як система фонетичних, морфологічних, словотвірних, лексичних і семантичних

ознак. З іншого боку, у художньому тексті ці ознаки набувають функційного навантаження, оскільки беруть участь у створенні регіонального простору, мовної характеристики персонажів, етнокультурної атмосфери й текстових домінант.

Отже, гуцулізми в художніх текстах української літератури виконують комплекс функцій: номінативну, локалізаційну, етнокультурну, характерологічну, стилістичну, експресивну та текстотвірну. Для цієї роботи особливо важливо, що функційне навантаження гуцулізмів може бути досліджене не лише описово, а й за допомогою частотного, корпусного та лінгвостатистичного аналізу.

Висновки до першого розділу

Гуцульський діалект є важливим об'єктом лінгвістичного дослідження, оскільки поєднує територіально–мовні, культурні та етнографічні ознаки. Його специфіка виявляється на фонетичному, морфологічному, словотвірному, лексичному та семантичному рівнях, однак для цієї роботи особливе значення має саме лексичний рівень, пов'язаний із назвами локальних реалій, побуту, природи, родинно–соціальних відносин і традиційної культури.

Функційна значущість гуцулізмів у художньому тексті полягає в тому, що вони не лише називають предмети, явища й поняття гуцульського середовища, а й беруть участь у створенні регіонального мовного колориту, характеристиці персонажів, моделюванні художнього простору та репрезентації етнокультурної інформації. Тому гуцулізми доцільно розглядати не як випадкові діалектні вкраплення, а як важливий елемент художньої організації тексту.

Отже, теоретичний розгляд гуцульського діалекту й функцій гуцулізмів створює основу для подальшого автоматизованого укладання електронного частотного словника, лінгвостатистичного опису виявлених одиниць і зіставлення особливостей їх функціонування в текстах Марії Матіос і Тараса Прохаська.

РОЗДІЛ 2. АВТОМАТИЗОВАНЕ УКЛАДАННЯ ЧАСТОТНОГО СЛОВНИКА ГУЦУЛЬСЬКИХ ДІАЛЕКТИЗМІВ ЗА ХУДОЖНІМИ ТЕКСТАМИ МАРІЇ МАТІОС ТА ТАРАСА ПРОХАСЬКА

2.1 Частотні словники у Корпусі української мови

Для дослідження гуцулізмів у художніх текстах Марії Матіос і Тараса Прохаська використано частотні словники, сформовані за матеріалами Корпусу української мови. Частотний словник у межах цієї роботи розглядається як електронний лексикографічний ресурс, що містить перелік лексем, зафіксованих у відповідній текстовій вибірці, а також додаткові кількісні й лінгвістичні характеристики. Використання таких ресурсів дає змогу перейти від суто описового аналізу діалектної лексики до її кількісного опрацювання [Корпус б. д.; Частотні б. д.; Дарчук 2008; Дарчук 2010].

У сучасній корпусній лінгвістиці частотні словники є важливим інструментом для встановлення складу лексики, частотності окремих одиниць і співвідношення різних груп слів у текстовому матеріалі. За допомогою частотного словника можна виявити не лише наявність певної лексеми в тексті, а й ступінь її повторюваності. Це особливо важливо для аналізу гуцулізмів, оскільки діалектна одиниця може бути представлена в тексті як поодинокі стилістичне вкраплення або як регулярний елемент авторської мовної системи [McEneaney, Hardie 2012; Baker 2006; Stubbs 2001].

У межах роботи використано два авторські частотні словники: частотний словник за текстами Марії Матіос і частотний словник за текстами Тараса Прохаська. Кожен із них містить перелік лексем, зафіксованих у відповідній текстовій вибірці, а також додаткові характеристики: код частини мови, абсолютну частоту й семантичний опис. Вхідні частотні словники, електронний реєстр гуцульських діалектизмів, програмний код застосунку та результати автоматизованого опрацювання подано в електронному додатку й розміщено у [відкритому репозиторії GitHub](#). Основний файл запуску застосунку – [app.py](#),

список залежностей – [requirements.txt](#), файл запуску для Windows – [run.bat](#). Вхідні дані подано в папці [data](#), результати автоматизованого опрацювання – у папці [output](#), а допоміжні модулі програмного коду – у папці [src](#). Скріншоти роботи застосунку наведено в додатку А, фрагменти програмного коду – у додатку Б, а списки гуцулізмів, виявлених у частотних словниках Марії Матіос і Тараса Прохаська, – у додатку В.

Частотний словник за текстами Марії Матіос подано у файлі [matios_frequency.txt](#). Після очищення в ньому зафіксовано 440662 слововживання та 38996 лексем реєстру частотного словника. Частотний словник за текстами Тараса Прохаська подано у файлі [prokhasko_frequency.txt](#). Після очищення він містить 127350 слововживань і 16759 лексем реєстру частотного словника. Отже, авторські корпуси істотно відрізняються за обсягом, що необхідно враховувати під час інтерпретації кількості виявлених гуцулізмів, їхньої сумарної абсолютної частоти та покриття тексту гуцулізмами.

Кількісну характеристику текстового матеріалу подано в таблиці 2.1.

Таблиця 2.1

Кількісна характеристика авторських частотних словників після
очищення

Авторський матеріал	Вхідний файл	Кількість слововживань після очищення	Кількість лексем реєстру частотного словника після очищення
Тексти Марії Матіос	matios_frequency.txt	440662	38996
Тексти Тараса Прохаська	prokhasko_frequency.txt	127350	16759

Дані таблиці 2.1 свідчать, що частотний словник за текстами Марії Матіос є значно більшим за обсягом, ніж частотний словник за текстами Тараса Прохаська. Саме тому в подальшому аналізі використовуються не лише

абсолютні показники, а й відносні: відносна частота, покриття тексту гуцулізмами та частка гуцулізмів у реєстрі частотного словника.

Перед автоматизованим виявленням гуцулізмів частотні словники потребували очищення та нормалізації. Це зумовлено тим, що в електронних частотних словниках можуть міститися порожні значення, пунктуаційні послідовності, числові записи, іншомовні елементи, технічні токени й залишки автоматичного опрацювання. Такі одиниці не становлять матеріалу лінгвістичного аналізу, тому вони були вилучені з подальшого опрацювання.

Очищення частотних словників передбачало вилучення порожніх значень, пунктуаційних послідовностей, чисел, технічних залишків і токенів, що не містять українських літер. Нормалізація лексем полягала у приведенні слів до єдиного графічного вигляду: усуненні зайвих пробілів, переведенні до нижнього регістру, уніфікації апострофів і вилученні зайвих знаків на межах слова. Водночас для гуцульського матеріалу важливо не здійснювати механічної заміни літери *г* на *г*, оскільки така заміна може спотворити діалектну форму.

Очищені авторські частотні словники, програмний код очищення та результати обробки подано в електронному додатку. Це забезпечує можливість перевірити вхідні дані й відтворити основні етапи автоматизованого опрацювання.

2.2 Характеристика текстового матеріалу дослідження за даними Корпусу української мови

Для автоматизованого виявлення гуцулізмів у текстах Марії Матіос і Тараса Прохаська, крім авторських частотних словників, використано реєстр гуцульських діалектизмів. Реєстр укладено на основі лексикографічних джерел, у яких зафіксовано гуцульську діалектну лексику, її значення, тип діалектизму та джерело словникової фіксації [Гуцульська 2020; Гуцульські 1997; Закревська 1976; Грещук 2009].

Електронний реєстр гуцульських діалектизмів подано в електронному додатку у форматі XLSX. Він містить діалектну форму, тлумачення, тип діалектизму, джерело словникової фіксації та додаткові коментарі. Саме цей реєстр використовується як база для автоматизованого зіставлення з лексемами авторських частотних словників.

Структуру вхідних даних, використаних для автоматизованого зіставлення, подано в таблиці 2.2.

Таблиця 2.2

Структура вхідних даних для автоматизованого виявлення гуцулізмів

Вхідні дані	Основні поля	Зміст	Роль у дослідженні
matioi_frequency.txt	lexeme, class, frequency, semantics	Частотний словник за текстами Марії Матіос	Джерело авторських лексем для пошуку гуцулізмів
prokhasko_frequency.txt	lexeme, class, frequency, semantics	Частотний словник за текстами Тараса Прохаська	Джерело авторських лексем для порівняльного аналізу
Реєстр гуцульських діалектизмів, XLSX	dialectism, meaning, type, source, comment	Електронний реєстр гуцульських діалектизмів із тлумаченнями	База для автоматизованого зіставлення й визначення типу діалектизму

Як видно з таблиці 2.2, у дослідженні поєднано два типи вхідних даних: авторські частотні словники та словниковий реєстр гуцульських діалектизмів. Частотні словники дають матеріал для пошуку діалектних одиниць у текстах конкретних авторів, а реєстр гуцулізмів забезпечує контрольовану основу для їх автоматизованого виявлення. Для забезпечення відтворюваності дослідження авторські частотні словники, реєстр гуцульських діалектизмів, очищені таблиці, результати автоматизованого зіставлення та програмний код застосунку подано в електронному додатку й у [відкритому репозиторії GitHub](#).

2.3 Автоматизоване укладання електронного частотного словника гуцулізмів

Для автоматизованого укладання електронного частотного словника гуцульських діалектизмів розроблено локальний вебзастосунок, який виконує допоміжну інструментальну функцію в межах лінгвостатистичного дослідження. Застосунок не є самостійним лінгвістичним ресурсом, а використовується як прикладний інструмент для обробки, зіставлення та експорту даних.

Застосунок реалізовано мовою програмування Python із використанням локального вебінтерфейсу Streamlit. Для опрацювання табличних даних застосовано бібліотеку pandas, для експорту результатів у формат XLSX – openpyxl, для очищення й нормалізації текстових значень – регулярні вирази модуля re, а для підрахунку частот і групування даних – collections [Python б. д.; Streamlit б. д.; pandas б. д.; openpyxl б. д.].

Інтерфейс застосунку побудовано відповідно до послідовності прикладного лінгвістичного дослідження. Користувач може завантажити вхідні файли або використати приклади з папки даних, виконати очищення й нормалізацію частотних словників, переглянути результати зіставлення, статистичні показники та експортувати сформовані таблиці.

Головну сторінку застосунку подано на рисунку 2.1.

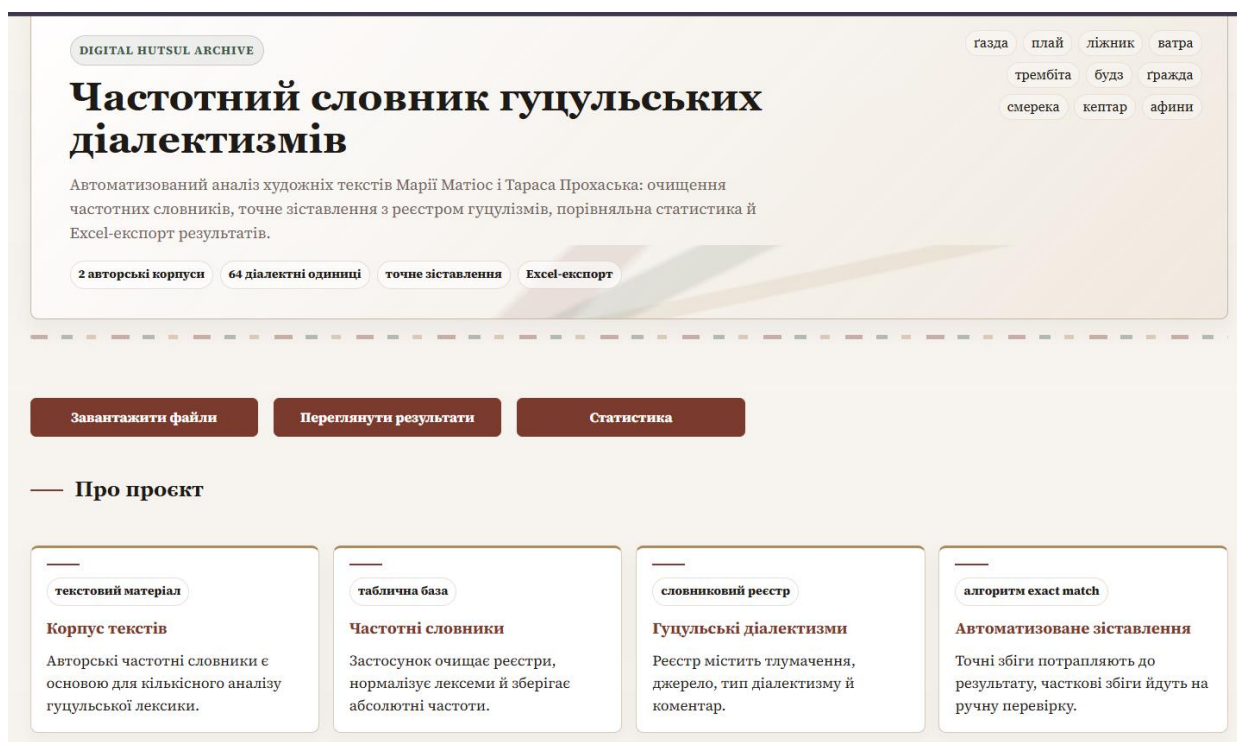


Рисунок 2.1 – Головна сторінка застосунку для укладання частотного словника гуцулізмів

Як видно з рисунка 2.1, інтерфейс застосунку містить основні блоки, потрібні для роботи з частотними словниками: завантаження даних, очищення й нормалізацію, автоматизоване зіставлення, перегляд результатів і експорт сформованих таблиць. Повний програмний код застосунку подано у [відкритому репозиторії GitHub](#), а його основні фрагменти наведено в додатку Б. Скріншоти, що демонструють завантаження вхідних файлів, підсумок автоматизованої обробки, перегляд словника гуцулізмів та аналітичну статистику, подано в додатку А.

Послідовність автоматизованого укладання електронного частотного словника гуцулізмів охоплює кілька етапів: завантаження авторських частотних словників, завантаження реєстру гуцульських діалектизмів, уніфікацію структури таблиць, нормалізацію лексем, очищення частотних словників, автоматизоване зіставлення, пошук потенційних збігів, формування авторських словників, створення спільної таблиці, розрахунок статистичних показників і експорт результатів.

Етапи автоматизованого укладання електронного словника гуцулізмів подано в таблиці 2.3.

Таблиця 2.3

Етапи автоматизованого укладання електронного частотного словника гуцулізмів

Етап	Дія	Вхідні дані	Результат	Програмна реалізація / додаток
1	Завантаження частотних словників	matio_frequency.txt, prokhasko_frequency.txt	Отримання авторських реєстрів лексем	Файли в папці data ; код завантаження у файлі app.py
2	Завантаження реєстру гуцулізмів	Реєстр гуцульських діалектизмів, XLSX	Отримання словникової бази для зіставлення	Реєстр гуцулізмів у папці data
3	Уніфікація структури таблиць	Усі вхідні файли	Перевірка й доповнення відсутніх колонок	Програмна реалізація у папці src

Продовження таблиці 2.3

Етап	Дія	Вхідні дані	Результат	Програмна реалізація / додаток
4	Нормалізація лексем	Поле lexeme і поле dialectism	Створення уніфікованих форм для зіставлення	Програмна реалізація у папці src
5	Очищення частотних словників	Авторські частотні словники	Вилучення службових tokenів, чисел і пунктуації	Програмна реалізація у папці src
6	Автоматизоване зіставлення за нормалізованою формою	Нормалізовані лексеми й реєстр гуцулізмів	Виявлення лексем, що збігаються з реєстром гуцулізмів	Програмна реалізація у папці src
7	Пошук потенційних збігів	Очищені авторські словники	Формування таблиці для ручної перевірки	Файл manual_review.xlsx у папці output
8	Формування авторських словників гуцулізмів	Підтверджені збіги	Таблиці гуцулізмів для Марії Матіос і Тараса	Файли matios_hutsulisms.xlsx і prokhasko_hutsulisms.xlsx у папці output

			Прохаська	
9	Формування спільної таблиці	Авторські таблиці результатів	Порівняльна таблиця гуцулізмів двох авторів	Файл combined_hutsulisms.xlsx у папці output
10	Розрахунок статистики	Підсумкові таблиці	Кількісні показники, групування й типи	Файл statistics_summary.xlsx у папці output
11	Експорт результатів	Усі сформовані таблиці	Excel-файли та ZIP-архів	Файл all_results.zip у папці output

Дані таблиці 2.3 показують, що укладання електронного словника гуцулізмів здійснюється як послідовність формалізованих операцій. Спочатку програма працює з вхідними таблицями, далі очищує й нормалізує матеріал, після чого виконує зіставлення з діалектним реєстром. Завершальним етапом є формування авторських і спільних таблиць, статистичних зведень та файлів експорту. Усі програмні операції підтверджуються програмним кодом, поданим в електронному додатку.

Алгоритм автоматизованого укладання електронного частотного словника гуцулізмів подано на рисунку 2.2.

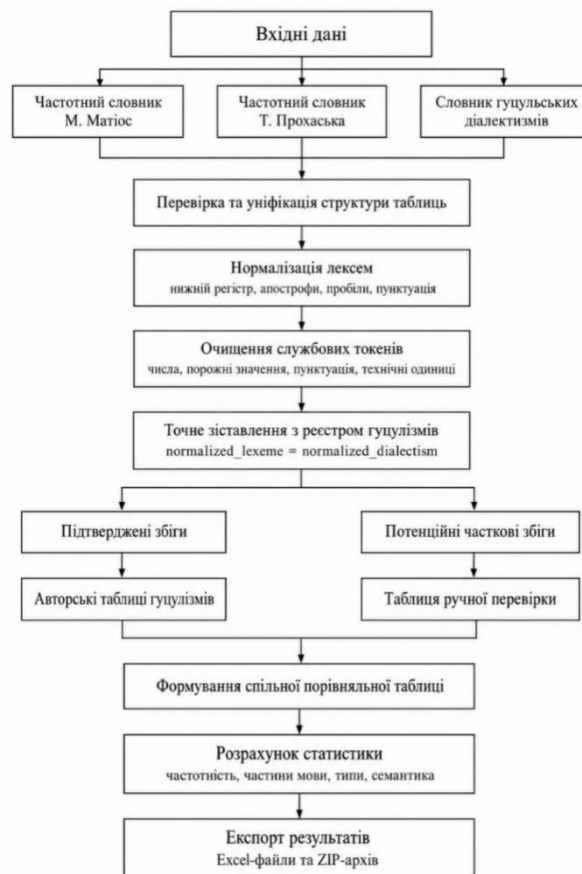


Рисунок 2.2 – Алгоритм автоматизованого укладання електронного частотного словника гуцулізмів

Як показано на рисунку 2.2, алгоритм побудовано за принципом переходу від необроблених частотних словників до структурованого словника гуцульських діалектизмів. Автоматичне зіставлення не є єдиним етапом аналізу: після нього передбачено ручну перевірку потенційних збігів, оскільки частина діалектних одиниць може бути представлена в текстах у похідних, відмінкових або графічно варіантних формах.

На етапі нормалізації лексеми приводяться до єдиного технічного вигляду: переводяться до нижнього регістру, очищуються від зайвих пробілів, уніфікуються за апострофами та пунктуаційними знаками. Водночас літера г не замінюється на ґ, оскільки для гуцульського діалектного матеріалу така відмінність може бути змістовно важливою.

Ручна перевірка застосовувалася до одиниць, які могли бути пов'язані з гуцульською діалектною лексикою, але не завжди збігалися з реєстром у точній графічній формі. До таких випадків належали похідні форми, відмінкові варіанти, можливі фонетичні варіанти, а також слова, які потребували уточнення за словниковим тлумаченням і контекстом. Результати ручної перевірки подано у файлі `manual_review.xlsx` в електронному додатку. У додатку цей файл має містити список потенційних збігів, рішення про підтвердження або відхилення одиниці та короткий коментар до перевірки.

У результаті роботи застосунок формує кілька вихідних файлів, призначених для подальшого аналізу. Їх характеристику подано в таблиці 2.4.

Таблиця 2.4

Вихідні файли застосунку та їх призначення

Вихідний файл	Зміст	Призначення в дослідженні
<code>matios_hutsulisms.xlsx</code>	Таблиця гуцулізмів, виявлених у частотному словнику Марії Матіос	Аналіз складу й частотності гуцулізмів у текстах Марії Матіос
<code>prokhasko_hutsulisms.xlsx</code>	Таблиця гуцулізмів, виявлених у частотному словнику Тараса Прохаська	Аналіз складу й частотності гуцулізмів у текстах Тараса Прохаська
<code>combined_hutsulisms.xlsx</code>	Спільна порівняльна таблиця гуцулізмів двох авторів	Порівняльний аналіз спільних і унікальних одиниць
<code>manual_review.xlsx</code>	Таблиця потенційних часткових або похідних збігів	Ручна лінгвістична перевірка сумнівних одиниць
<code>statistics_summary.xlsx</code>	Зведена статистика за авторами, типами, частинами мови й семантичними групами	Побудова таблиць, рисунків і висновків у третьому розділі
<code>all_results.zip</code>	Архів із підсумковими файлами	Зручне збереження всіх результатів обробки

Сформовані застосунком файли `matios_hutsulisms.xlsx`, `prokhasko_hutsulisms.xlsx`, `combined_hutsulisms.xlsx`, `manual_review.xlsx`, `statistics_summary.xlsx` та архів `all_results.zip` подано в папці [output](#) відкритого репозиторію GitHub. Це забезпечує можливість перевірити результати автоматизованого зіставлення, ручної перевірки та статистичних розрахунків.

У фінальній версії застосунку після автоматизованого опрацювання за текстами Марії Матіос виявлено 51 лексему реєстру частотного словника гуцулізмів із сумарною абсолютною частотою 831 слововживання, а за текстами Тараса Прохаська – 40 лексем реєстру частотного словника гуцулізмів із сумарною абсолютною частотою 163 слововживання. У спільній порівняльній таблиці зафіксовано 64 діалектні одиниці: 23 одиниці наявні в матеріалі обох авторів, 26 одиниць – лише в текстах Марії Матіос, 15 одиниць – лише в текстах Тараса Прохаська. Крім того, сформовано таблицю ручної перевірки, яка містить 39 потенційних часткових або похідних збігів.

Повні списки виявлених гуцулізмів, спільних і унікальних одиниць, а також результати ручної перевірки подано в електронному додатку. У тексті третього розділу використовуються узагальнені статистичні показники, а повні таблиці доступні в додаткових електронних файлах.

Отже, методика автоматизованого укладання електронного частотного словника гуцулізмів ґрунтується на поєднанні корпусного матеріалу, словникового реєстру гуцульських діалектизмів і програмної обробки табличних даних. Такий підхід дає змогу формалізувати процес виявлення діалектних одиниць, отримати авторські й порівняльні таблиці, розрахувати статистичні показники та підготувати матеріал для подальшого лінгвостатистичного й функційного аналізу.

Висновки до другого розділу

У другому розділі обґрунтовано методику автоматизованого укладання електронного частотного словника гуцулізмів на основі авторських частотних словників і реєстру гуцульської діалектної лексики. Значущість цієї методики полягає в тому, що вона дає змогу перейти від ручного пошуку окремих діалектизмів до відтворюваного й кількісно контрольованого опрацювання мовного матеріалу.

Вхідними даними для дослідження стали частотні словники за текстами Марії Матіос і Тараса Прохаська, а також електронний реєстр гуцульських діалектизмів. Очищення й нормалізація цих даних забезпечили можливість коректного автоматизованого зіставлення лексем із діалектним реєстром. Водночас ручна перевірка залишилася необхідним етапом, оскільки частина гуцулізмів може функціонувати в похідних, відмінкових або графічно варіантних формах.

Розроблений застосунок виконує допоміжну інструментальну функцію: завантажує вхідні файли, очищує й нормалізує дані, зіставляє авторські частотні словники з реєстром гуцулізмів, формує авторські та спільні таблиці, розраховує статистичні показники й експортує результати. Отримані файли створюють основу для подальшого лінгвостатистичного аналізу гуцульської діалектної лексики в текстах Марії Матіос і Тараса Прохаська.

РОЗДІЛ 3. ЛІНГВОСТАТИСТИЧНИЙ АНАЛІЗ ГУЦУЛЬСЬКИХ ДІАЛЕКТИЗМІВ У ХУДОЖНІХ ТЕКСТАХ МАРІЇ МАТІОС І ТАРАСА ПРОХАСЬКА

3.1 Частотна характеристика гуцулізмів у текстах Марії Матіос і Тараса Прохаська

Лінгвостатистичний аналіз гуцульських діалектизмів у художніх текстах Марії Матіос і Тараса Прохаська здійснено на основі результатів автоматизованого зіставлення авторських частотних словників із реєстром гуцульської діалектної лексики. Такий підхід дав змогу виокремити одиниці, які мають точний збіг із підготовленим словниковим реєстром, визначити їхню абсолютну частоту в кожному авторському корпусі, розрахувати відносні показники та встановити частку гуцулізмів у загальному обсязі опрацьованого матеріалу.

У межах дослідження частотність розглядається як кількісний показник уживаності діалектної одиниці в корпусі текстів певного автора. Вона дає змогу встановити, які гуцулізми є поодинокими, а які регулярно функціонують в авторському мовному матеріалі. Водночас частотність не ототожнюється автоматично з художньою значущістю слова, оскільки навіть одиничне вживання діалектизму може виконувати важливу стилістичну або етнокультурну функцію [McEnery, Hardie 2012; Baker 2006; Stubbs 2001].

Для зіставлення авторських корпусів різного обсягу в роботі використано не лише абсолютну, а й відносну частоту. Відносну частоту гуцулізму розраховано за формулою:

$$F_{\text{відн}} = \frac{F_{\text{абс}}}{N} \times 1000 \quad (3.1)$$

де $F_{\text{відн}}$ – відносна частота гуцулізму на 1000 слововживань;

$F_{\text{абс}}$ – абсолютна частота відповідного гуцулізму;

N – загальна кількість слововживань в авторському частотному словнику після очищення.

Крім того, для визначення покриття тексту гуцулізмами розраховано частку всіх уживань виявлених гуцулізмів у загальній кількості слововживань авторського корпусу за формулою:

$$P = \frac{\sum F_{\text{абс гуц}}}{N} \times 100\% \quad (3.2)$$

де P – покриття тексту гуцулізмами;

$\sum F_{\text{абс гуц}}$ – сумарна абсолютна частота всіх виявлених гуцулізмів;

N – загальна кількість слововживань у відповідному авторському корпусі.

За результатами автоматизованої обробки в текстах Марії Матіос виявлено 51 лексему реєстру частотного словника гуцулізмів із сумарною абсолютною частотою 831 слововживання. У текстах Тараса Прохаська зафіксовано 40 лексем реєстру частотного словника гуцулізмів із сумарною абсолютною частотою 163 слововживання. Водночас необхідно враховувати різний обсяг авторських корпусів: після очищення частотний словник Марії Матіос охоплює 440662 слововживання та 38996 лексем реєстру частотного словника, тоді як частотний словник Тараса Прохаська – 127350 слововживань і 16759 лексем реєстру частотного словника.

Загальні кількісні показники автоматизованого аналізу подано в табл. 3.1.

Таблиця 3.1

Загальні кількісні показники виявлення гуцулізмів

Показник	Марія Матіос	Тарас Прохасько
Загальна кількість слововживань після очищення	440662	127350

Продовження таблиці 3.1

Показник	Марія Матіос	Тарас Прохасько
Кількість лексем реєстру частотного словника гуцулізмів	51	40
Частка лексем від загального реєстру частотного словника, %	0,131	0,239
Сумарна абсолютна частота гуцулізмів	831	163
Покриття тексту гуцулізмами, %	0,189%	0,128%

Дані таблиці 3.1 свідчать, що в текстах Марії Матіос виявлено більшу абсолютну кількість лексем реєстру частотного словника гуцулізмів, ніж у текстах Тараса Прохаська: 51 одиницю проти 40. Проте корпуси авторів істотно відрізняються за обсягом. Частотний словник Марії Матіос після очищення охоплює 440662 слововживання, тоді як частотний словник Тараса Прохаська – 127350 слововживань. Тому безпосереднє зіставлення лише абсолютної кількості виявлених одиниць не є достатнім.

Частка лексем гуцулізмів від загального реєстру частотного словника становить 0,131 % у матеріалі Марії Матіос і 0,239 % у матеріалі Тараса Прохаська. Це означає, що в лексичному реєстрі частотного словника Тараса Прохаська гуцулізми займають відносно більшу частку, попри меншу абсолютну кількість виявлених одиниць. Натомість сумарна абсолютна частота гуцулізмів у текстах Марії Матіос є вищою і становить 831 слововживання проти 163 слововживань у текстах Тараса Прохаська.

Покриття тексту гуцулізмами становить 0,189 % у корпусі Марії Матіос і 0,128 % у корпусі Тараса Прохаська. Отже, у текстах Марії Матіос гуцулізми мають вищу сумарну частотність і дещо більше покриття тексту, тоді як у текстах Тараса Прохаська вони становлять більшу частку саме на рівні лексичного реєстру. Це свідчить про різні способи функціонування гуцульської діалектної лексики в авторських корпусах: у Марії Матіос вона частіше

повторюється в тексті, а в Тараса Прохаська представлена компактнішим, але відносно помітним набором лексем.

Відносні показники виявлення гуцулізмів у текстах Марії Матіос і Тараса Прохаська подано на рисунку 3.1.

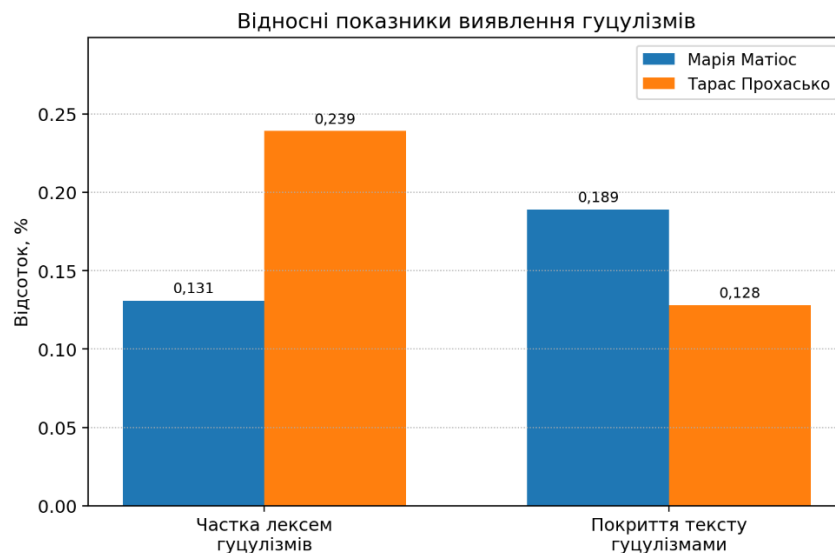


Рисунок 3.1 – Відносні показники виявлення гуцулізмів у текстах Марії Матіос і Тараса Прохаська, %

Як показано на рисунку 3.1, за часткою лексем гуцулізмів від загального реєстру частотного словника вищий показник має матеріал Тараса Прохаська – 0,239 % проти 0,131 % у Марії Матіос. Натомість за покриттям тексту гуцулізмами вищий показник має матеріал Марії Матіос – 0,189 % проти 0,128 % у Тараса Прохаська. Отже, у текстах Марії Матіос гуцулізми частіше повторюються, а в текстах Тараса Прохаська вони займають відносно більшу частку в лексичному реєстрі.

Окремого аналізу потребують найчастотніші гуцулізми в кожному авторському корпусі. У текстах Марії Матіос найвищу абсолютну частоту мають лексеми *газда*, *баба*, *дримба*, *молодиця*, *дідо*, *вуйко*, *трембіта*, *смерека*, *газдиня*, *пивниця*. Ці одиниці переважно пов'язані з родинно-соціальним

середовищем, традиційним побутом, локальною культурою та простором Гуцульщини.

Таблиця 3.2

Найчастотніші гуцулізми в текстах Марії Матіос

Діалектизм	Абсолютна частота	Частка серед уживань гуцулізмів, %	Покриття тексту, %
газда	101	12,15	0,0229
баба	73	8,78	0,0166
дримба	68	8,18	0,0154
молодиця	62	7,46	0,0141
дідо	50	6,02	0,0113
вуйко	48	5,78	0,0109
трембіта	30	3,61	0,0068
смерека	29	3,49	0,0066
газдиня	29	3,49	0,0066
пивниця	21	2,53	0,0048

Дані таблиці 3.2 засвідчують, що серед найчастотніших гуцулізмів у текстах Марії Матіос домінують іменникові назви осіб, родинно–соціальних ролей і культурно–побутових реалій. Найвищу частку серед усіх уживань гуцулізмів має лексема *газда* – 12,15 % від зафіксованого гуцульського слововживання цього авторського корпусу. Далі йдуть *баба* (8,78 %), *дримба* (8,18 %), *молодиця* (7,46 %), *дідо* (6,02 %) і *вуйко* (5,78 %). Отже, у матеріалі Марії Матіос виразно проступає соціально–родинна та етнокультурна семантика гуцульського світу. Повний список гуцулізмів, виявлених у частотному словнику за текстами Марії Матіос, подано в додатку В.

У текстах Тараса Прохаська найчастотнішими є лексеми *дідо*, *ліжник*, *пивниця*, *будз*, *опришок*, *смерека*, *черес*, *коц*, *фіранка*, *ватра*. У межах зафіксованого гуцульського слововживання найбільшу частку має *дідо* (18,40 %), помітними також є *ліжник* (11,04 %), *пивниця* (6,75 %), *будз* (5,52 %) й *опришок* (4,91 %). Порівняно з матеріалом Марії Матіос, тут відносно сильніше представлені назви предметів побуту, традиційної культури, їжі та історико–культурних реалій.

Таблиця 3.3

Найчастотніші гуцулізми в текстах тараса прохаська

Діалектизм	Абсолютна частота	Частка серед уживань гуцулізмів, %	Покриття тексту, %
дідо	30	18,40	0,0236
ліжник	18	11,04	0,0141
пивниця	11	6,75	0,0086
будз	9	5,52	0,0071
опришок	8	4,91	0,0063
смерека	7	4,29	0,0055
черес	7	4,29	0,0055
коц	7	4,29	0,0055
фіранка	7	4,29	0,0055
ватра	6	3,68	0,0047

Як видно з таблиці 3.3, у текстах Тараса Прохаська найчастотніші гуцулізми переважно стосуються предметної, побутової, природної та історико–культурної сфер. Лексема *дідо* має найвищу абсолютну частоту й може бути пов’язана з мотивами пам’яті, спадковості, родової тяглості та локального досвіду. Лексеми *ліжник*, *будз*, *черес*, *коц*, *ватра* репрезентують матеріальну культуру й побут. Повний список гуцулізмів, виявлених у частотному словнику за текстами Тараса Прохаська, подано в додатку В.

Зіставлення таблиць 3.2 і 3.3 показує, що частина гуцулізмів є спільною для двох авторських корпусів, але їхня відносна вага різна. Наприклад, лексема *дідо* становить 6,02 % усіх уживань гуцулізмів у Марії Матіос і 18,40 % у Тараса Прохаська; *пивниця* – відповідно 2,53 % і 6,75 %. Натомість *газда* має дуже високу частку в Марії Матіос (12,15 %) і лише 1,23 % у Тараса Прохаська. Отже, порівняльно значущими є не лише підсумкові абсолютні частоти, а й відносна частка одиниці в межах кожного авторського корпусу.

Для узагальнення спільних найчастотніших одиниць використано спільну таблицю *combined_hutsulisms.xlsx*, у якій зафіксовано 64 діалектні одиниці. Найчастотніші гуцулізми у спільній таблиці подано в таблиці 3.4.

Таблиця 3.4

Найчастотніші гуцулізми у спільній таблиці

Діалектизм	Абс. частота Марії Матіос	Частка серед уживань гуцулізмів Марії Матіос, %	Абс. частота Тараса Прохаська	Частка серед уживань гуцулізмів Тараса Прохаська, %	Загальна абсолютна частота
газда	101	12,15	2	1,23	103
баба	73	8,78	1	0,61	74
дримба	68	8,18	1	0,61	69
молодиця	62	7,46	0	0,00	62
дідо	50	6,02	30	18,40	80
вуйко	48	5,78	3	1,84	51
трембіта	30	3,61	1	0,61	31
смерека	29	3,49	7	4,29	36
газдиня	29	3,49	0	0,00	29
пивниця	21	2,53	11	6,75	32

Дані таблиці 3.4 дають змогу побачити не лише сумарну абсолютну частоту гуцулізмів, а й нерівномірність їх розподілу між двома авторськими корпусами. Лексеми *газда*, *баба*, *дримба*, *молодиця*, *вуйко*, *трембіта* значно активніше представлені в текстах Марії Матіос. Натомість лексеми *дідо* й *пивниця* мають більшу відносну вагу в текстах Тараса Прохаська.

Отже, частотна характеристика засвідчує, що гуцулізми в текстах обох авторів не є випадковими одиничними елементами. У текстах Марії Матіос діалектна лексика представлена більшою кількістю слововживань і сильніше пов'язана з родинно-побутовою та соціальною сферою. У текстах Тараса Прохаська гуцулізми мають нижчу сумарну абсолютну частоту, однак виразно репрезентують карпатський предметний і культурний простір.

3.2 Лексико–семантична класифікація виявлених гуцульських діалектизмів

Після визначення загальної кількості, абсолютної частоти та відносних показників уживання гуцулізмів доцільно розглянути їхній лексико–граматичний і лексико–семантичний склад. Такий аналіз дає змогу встановити, які типи мовних одиниць найактивніше представлені в авторських корпусах, а також які тематичні сфери гуцульської діалектної лексики переважають у текстах Марії Матіос і Тараса Прохаська.

Попереднім етапом лексико–семантичного аналізу є врахування частиномовної належності виявлених гуцулізмів. У частотних словниках ця характеристика подана через поле class, яке фіксує код лексико–граматичного класу одиниці. Такий показник важливий, оскільки дає змогу з'ясувати, які групи слів найчастіше виступають носіями діалектної маркованості в художньому тексті.

Лексико–граматичний розподіл виявлених гуцулізмів подано в таблиці 3.5.

Таблиця 3.5

Лексико–граматичний розподіл гуцулізмів у текстах м. Матіос і т.
Прохаська

Автор	Код	Розшифрування	Кількість	Частка реєстру, %	Частка серед уживань гуцулізмів, %
Марія Матіос	К	іменники	19	37,25	49,10
Марія Матіос	Й	іменники	17	33,33	36,46
Марія Матіос	без коду	без коду	8	15,69	5,90
Марія Матіос	Г	дієслова	2	3,92	2,41
Марія Матіос	Н	прислівники	1	1,96	2,29
Марія Матіос	Л	іменники множинної / збірної семантики	2	3,92	1,93
Марія Матіос	И	іменники pluralia	1	1,96	1,81

		tantum / множинні форми			
Марія Матіос	А	прикметники	1	1,96	0,12
Тарас Прохасько	Й	іменники	17	42,50	59,51
Тарас Прохасько	К	іменники	9	22,50	22,70
Тарас Прохасько	без коду	без коду	12	30,00	15,95
Тарас Прохасько	Л	іменники множинної / збірної семантики	1	2,50	1,23
Тарас Прохасько	Н	прислівники	1	2,50	0,61

Дані таблиці 3.5 показують, що в обох авторських корпусах переважають іменникові одиниці. Коди К та Й у вихідних частотних словниках позначають іменникові класи, тому в інтерпретації їх доцільно об'єднувати як групу іменників. У текстах Марії Матіос ці два коди охоплюють 36 одиниць, тобто 70,58 % реєстру виявлених гуцулізмів. У текстах Тараса Прохаська іменникові коди К та Й охоплюють 26 одиниць, тобто 65,00 % реєстру.

За часткою серед уживань гуцулізмів іменникові одиниці також мають провідне значення. У корпусі Марії Матіос коди К та Й разом становлять 85,56 % усіх уживань гуцулізмів, а в корпусі Тараса Прохаська – 82,21 %. Це підтверджує, що автоматично виявлені гуцулізми переважно називають осіб, предмети, просторові, побутові та етнокультурні реалії. Інші частиномовні групи представлені значно меншою кількістю одиниць.

У текстах Марії Матіос код К охоплює 19 одиниць із часткою реєстру 37,25 %, а код Й – 17 одиниць із часткою 33,33 %. У текстах Тараса Прохаська найбільше одиниць має код Й – 17 одиниць, що становить 42,50 % реєстру, тоді як код К охоплює 9 одиниць, або 22,50 %. Одиниці без частиномовного коду не вилучаються з аналізу автоматично, оскільки вони можуть бути релевантними гуцулізмами, але потребують уважнішої інтерпретації за словниковим тлумаченням і контекстом.

Для наочного представлення лексико–граматичного розподілу гуцулізмів використано відсоткові показники, оскільки саме вони дають змогу зіставити корпуси різного обсягу.

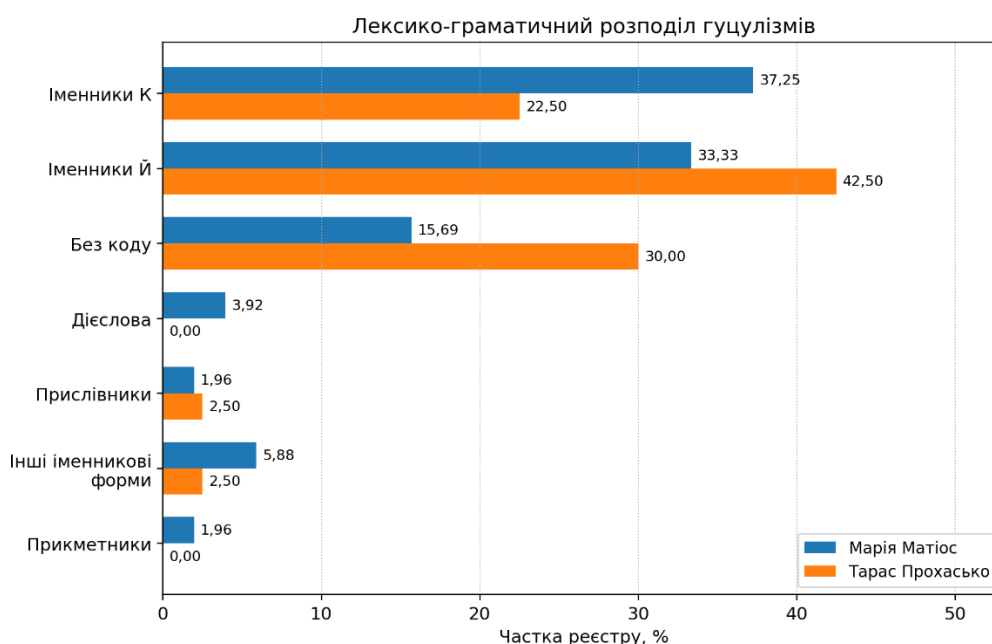


Рисунок 3.2 – Лексико–граматичний розподіл гуцулізмів у текстах Марії Матіос і Тараса Прохаська, %

Як показано на рисунку 3.2, в обох авторських корпусах переважають іменникові одиниці. У текстах Марії Матіос частка іменникових кодів К та Й становить 70,58 %, а в текстах Тараса Прохаська – 65,00 %. Водночас у матеріалі Тараса Прохаська помітнішою є частка одиниць без коду – 30,00 % проти 15,69 % у матеріалі Марії Матіос. Це свідчить про потребу додаткової ручної перевірки таких одиниць, оскільки автоматизований аналіз не завжди дає змогу повністю інтерпретувати їхню лексико–граматичну належність.

Лексико–семантична класифікація виявлених гуцулізмів здійснювалася з урахуванням їх словникового значення, семантичного опису в частотному словнику та додаткових коментарів, поданих у реєстрі гуцульських діалектизмів. У межах дослідження виділено кілька основних тематичних груп: родинні й соціальні назви, побутова та предметна лексика, лексика житла й простору, природна лексика, музична культура, одяг і прикраси, харчова та господарська лексика, історико–культурні й обрядові назви.

Основні семантичні групи виявлених гуцулізмів подано в таблиці 3.6.

Таблиця 3.6

Основні семантичні групи виявлених гуцулізмів

Семантична група	Приклади гуцулізмів	Характеристика
Родинні й соціальні назви	<i>газда, газдиня, баба, дідо, вуйко, молодиця, мамка</i>	Позначають членів родини, соціальні ролі, осіб традиційного гуцульського середовища
Побутова та предметна лексика	<i>ліжник, коц, баняк, фіранка, пивниця</i>	Називають предмети щоденного вжитку, елементи домашнього простору та побуту
Житло й простір	<i>колиба, гражда, хороми, плай, виходок</i>	Пов'язані з організацією житла, гірського простору, поселення та пересування
Природна лексика	<i>смерека, барвінок</i>	Репрезентує рослинний світ і карпатське природне середовище
Музична культура	<i>дримба, трембіта</i>	Позначає традиційні музичні інструменти та елементи звукового образу Гуцульщини
Одяг і прикраси	<i>кептар, кептарик, черес</i>	Називають елементи традиційного одягу й матеріальної культури
Харчова й господарська лексика	<i>будз, газдувати, газдівство</i>	Пов'язана з їжею, господарством, працею та традиційною організацією побуту
Історико-культурні й обрядові назви	<i>опришок, ватра</i>	Актуалізують історичну пам'ять, обрядовість, символічні й культурні смисли

Дані таблиці 3.6 засвідчують, що гуцулізми в досліджуваному матеріалі репрезентують кілька важливих фрагментів гуцульської мовної картини світу. Найпомітнішою є група родинних і соціальних назв, до якої належать *газда, газдиня, баба, дідо, вуйко, молодиця, мамка*. У текстах Марії Матіос ця група має особливо виразне значення, оскільки пов'язана з моделюванням родинного й соціального простору.

Важливою є також побутова та предметна лексика: *ліжник, коц, баняк, фіранка, пивниця*. Такі одиниці називають предмети матеріальної культури, що формують повсякденний простір персонажів. Назви житла й простору, зокрема *колиба, гражда, хороми, плай, виходок*, виконують локалізаційну функцію, оскільки закріплюють художній світ за карпатським регіональним середовищем.

Природна лексика представлена передусім одиницями *смерека* й *барвінок*. Лексема *смерека* є особливо показовою, оскільки вона наявна в обох авторських корпусах і має загальну абсолютну частоту 36. Група музичної культури представлена лексемами *дримба* і *трембіта*, які мають високу частотність у текстах Марії Матіос і посилюють етнокультурне звучання її прози.

Лексика одягу, харчова й господарська лексика, а також історико–культурні назви доповнюють загальну картину гуцульського світу. До них належать *кентар*, *черес*, *будз*, *газдувати*, *газдівство*, *опришок*, *ватра*. Такі одиниці вводять у текст культурно специфічні реалії, пов’язані з матеріальною культурою, працею, обрядовістю та історичною пам’яттю.

Оскільки семантичні групи в таблиці 3.6 подано як якісну класифікацію з прикладами, окремий рисунок для цього матеріалу не використовується, щоб не дублювати зміст таблиці. Для побудови окремої відсоткової діаграми семантичного розподілу потрібно було б кількісно визначити частку кожної семантичної групи в реєстрі кожного автора, однак у межах цього підрозділу семантична класифікація виконує насамперед пояснювальну функцію.

Отже, лексико–семантична класифікація показує, що виявлені гуцулізми репрезентують передусім соціально–родинний, побутовий, предметний, етнокультурний і просторовий компоненти художнього світу. У текстах Марії Матіос сильніше виявлена група назв осіб і соціально–побутових ролей, а в текстах Тараса Прохаська помітнішим є предметно–просторовий і культурно–атмосферний шар гуцульської лексики.

3.3 Типи гуцулізмів у досліджуваному матеріалі: лексичні, етнографічні, фонетичні, семантичні

Після лексико–семантичної класифікації доцільно розглянути виявлені гуцулізми за типами діалектної маркованості. Такий аналіз дає змогу з’ясувати, які саме різновиди гуцульської діалектної лексики переважають у текстах Марії

Матіос і Тараса Прохаська: власне лексичні, етнографічні, семантичні або фонетично марковані одиниці.

Лексичними гуцулізмами вважаються діалектні одиниці, які мають відповідники в літературній мові, але відрізняються від них формою або регіональною сферою вживання. До цієї групи належать, наприклад, *газда*, *баба*, *дідо*, *вуйко*, *молодиця*, *газдиня*. Такі одиниці переважно називають осіб, родинні ролі, соціальні статуси або побутові явища.

Етнографічні гуцулізми позначають специфічні реалії матеріальної та духовної культури Гуцульщини. До них належать назви предметів побуту, елементів одягу, музичних інструментів, їжі, господарських процесів, просторових об'єктів і культурних явищ: *дримба*, *трембіта*, *ліжник*, *кептар*, *черес*, *будз*, *колиба*, *гражда*, *ватра*. Такі слова разом із номінацією передають локальний культурний зміст.

Семантичні гуцулізми пов'язані з відмінністю або уточненням значення слова в межах говіркового вжитку. Фонетичні гуцулізми виявляються через особливості звукового оформлення слова, які можуть передаватися в художньому тексті графічно. У межах автоматизованого зіставлення за точним збігом фонетичні одиниці не виокремилися як окрема статистично представлена група, оскільки їх виявлення потребує аналізу варіантних написань, контексту й ручної лінгвістичної перевірки. Інакше кажучи, автоматизований підхід добре працює з лексемами, але не дуже любить живу мову, бо ж навіщо машині складність, коли можна просто зіставляти рядки.

Розподіл виявлених гуцулізмів за типами подано в таблиці 3.7.

Таблиця 3.7

Розподіл гуцулізмів за типами діалектизмів

Автор	Тип діалектизму	Кількість	Частка реєстру, %	Абсолютна частота	Частка серед уживань гуцулізмів, %
Марія Матіос	Лексичний	32	62,75	617	74,25
Марія Матіос	Етнографічний	18	35,29	207	24,91

Продовження таблиці 3.7

Автор	Тип діалектизму	Кількість	Частка реєстру, %	Абсолютна частота	Частка серед уживань гуцулізмів, %
Марія Матіос	Семантичний	1	1,96	7	0,84
Тарас Прохасько	Лексичний	22	55,00	90	55,21
Тарас Прохасько	Етнографічний	17	42,50	71	43,56
Тарас Прохасько	Семантичний	1	2,50	2	1,23

Дані таблиці 3.7 показують, що в обох авторських корпусах переважають лексичні гуцулізми. У текстах Марії Матіос вони становлять 62,75 % реєстру виявлених одиниць і 74,25 % усіх уживань гуцулізмів. Етнографічні гуцулізми в її матеріалі охоплюють 35,29 % реєстру і 24,91 % уживань. У текстах Тараса Прохаська лексичні гуцулізми становлять 55,00 % реєстру і 55,21 % уживань, а етнографічні – 42,50 % реєстру і 43,56 % уживань.

Отже, у текстах Тараса Прохаська етнографічна лексика має більшу відносну вагу, ніж у текстах Марії Матіос. Це особливо помітно за часткою серед уживань гуцулізмів: 43,56 % у Прохаська проти 24,91 % у Матіос. Натомість у текстах Марії Матіос сильніше виражене повторення лексичних соціально-побутових одиниць, що підтверджує вищу абсолютну частоту лексичних гуцулізмів – 617 слововживань.

Другою за кількістю групою є етнографічні гуцулізми. У текстах Марії Матіос виявлено 18 етнографічних одиниць із сумарною абсолютною частотою 207 слововживань, а в текстах Тараса Прохаська – 17 одиниць із сумарною абсолютною частотою 71 слововживання. Різниця в кількості етнографічних одиниць між авторами незначна, однак абсолютна частота в текстах Марії Матіос майже втричі вища. Водночас у відсотковому співвідношенні етнографічні гуцулізми помітніше представлені саме в матеріалі Тараса Прохаська.

До етнографічних гуцулізмів належать одиниці *дримба, трембіта, ліжник, будз, кептар, кептарик, черес, ватра, колиба, гражда*. У художньому тексті такі одиниці виконують етнокультурну та локалізаційну функції, оскільки вводять у текст реалії, пов'язані саме з гуцульським середовищем. Їхня роль полягає не лише в називанні предметів або явищ, а й у створенні культурного тла, через яке формується образ Гуцульщини.

Семантичні гуцулізми становлять найменшу групу: по одній одиниці в кожному авторському корпусі. Їх невелика кількість пояснюється тим, що автоматизований аналіз спирався передусім на точне зіставлення лексем із реєстром гуцульських діалектизмів, тоді як семантичні відмінності часто потребують ширшого контекстуального аналізу. Отже, кількісно ця група є периферійною, але методологічно важливою, оскільки засвідчує межі автоматизованого підходу.

Для наочного представлення типологічного розподілу гуцулізмів використано відсоткові показники частки реєстру.

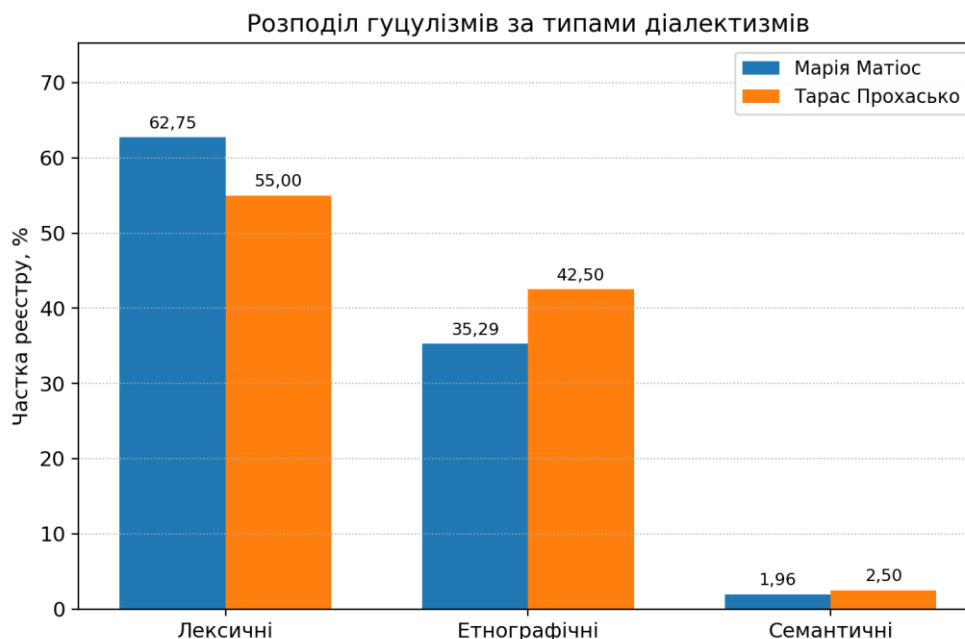


Рисунок 3.3 – Розподіл гуцулізмів за типами діалектизмів, %

Як показано на рисунку 3.3, в обох авторських корпусах переважають лексичні гуцулізми: 62,75 % у текстах Марії Матіос і 55,00 % у текстах Тараса Прохаська. Водночас частка етнографічних гуцулізмів є вищою в текстах Тараса Прохаська – 42,50 % проти 35,29 % у текстах Марії Матіос. Це свідчить про помітнішу відносну роль етнографічної лексики в матеріалі Тараса Прохаська.

Типологічний розподіл підтверджує, що гуцульські діалектизми в досліджуваних текстах виконують не лише функцію регіональної мовної стилізації, а й функцію репрезентації конкретного етнокультурного простору. Лексичні одиниці забезпечують називання персонажів, соціальних ролей, побутових і природних реалій, тоді як етнографічні одиниці увиразнюють матеріальну й духовну культуру Гуцульщини.

Отже, типологічний аналіз підтверджує, що основну частину виявлених гуцулізмів становлять лексичні та етнографічні діалектизми. У текстах Марії Матіос переважає повторюване вживання лексичних соціально–побутових одиниць, тоді як у текстах Тараса Прохаська етнографічні гуцулізми мають вищу відносну частку. Семантичні гуцулізми менш численні, проте вони важливі для якісного контекстуального аналізу та демонструють потребу поєднання автоматизованого пошуку з ручною лінгвістичною інтерпретацією.

3.4 Порівняльний аналіз і функційне навантаження гуцулізмів у текстах М. Матіос і Т. Прохаська

Порівняльний аналіз авторського вживання гуцулізмів дає змогу встановити не лише кількісні відмінності між корпусами текстів Марії Матіос і Тараса Прохаська, а й особливості функціонування діалектної лексики в межах двох різних художніх систем. Обидва автори пов'язані з карпатським культурним простором, однак способи введення гуцульської лексики в текст, її частотність, семантичне наповнення та функційне навантаження в їхній прозі відрізняються.

За результатами автоматизованого аналізу в об'єднаній порівняльній таблиці combined_hutsulisms.xlsx зафіксовано 64 гуцульські діалектні одиниці. Із них 23 одиниці є спільними для текстів обох авторів, 26 одиниць виявлено лише в текстах Марії Матіос, а 15 одиниць – лише в текстах Тараса Прохаська. Для коректної інтерпретації цих даних важливо враховувати не тільки абсолютну кількість одиниць, а й їхню частку в межах спільного реєстру.

Розподіл спільних та унікальних гуцулізмів подано в таблиці 3.8.

Таблиця 3.8

Спільні та унікальні гуцулізми в текстах м. Матіос і т. Прохаська

Група гуцулізмів	Кількість	Частка від спільної таблиці, %	Приклади
Спільні для обох авторів	23	35,94	газда, дідо, баба, дримба, вуйко, смерека, пивниця, ліжник, трембіта, ватра
Лише в текстах Марії Матіос	26	40,62	молодиця, газдиня, мамка, газдувати, хороми, колиба, плай, кептар, постоли
Лише в текстах Тараса Прохаська	15	23,44	черес, жереп, бартка, афини, гогодзи, бриндза, трембітар, бахур

Дані таблиці 3.8 показують, що найбільшу частку становлять гуцулізми, зафіксовані лише в текстах Марії Матіос, – 40,62 % спільного реєстру. Спільні для обох авторів одиниці становлять 35,94 %, а гуцулізми, виявлені лише в текстах Тараса Прохаська, – 23,44 %. Такий розподіл засвідчує наявність спільного ядра гуцульської діалектної лексики, але водночас демонструє індивідуальну авторську вибірковість у доборі діалектних одиниць.

Відсоткове співвідношення спільних та унікальних гуцулізмів подано на рисунку 3.4.

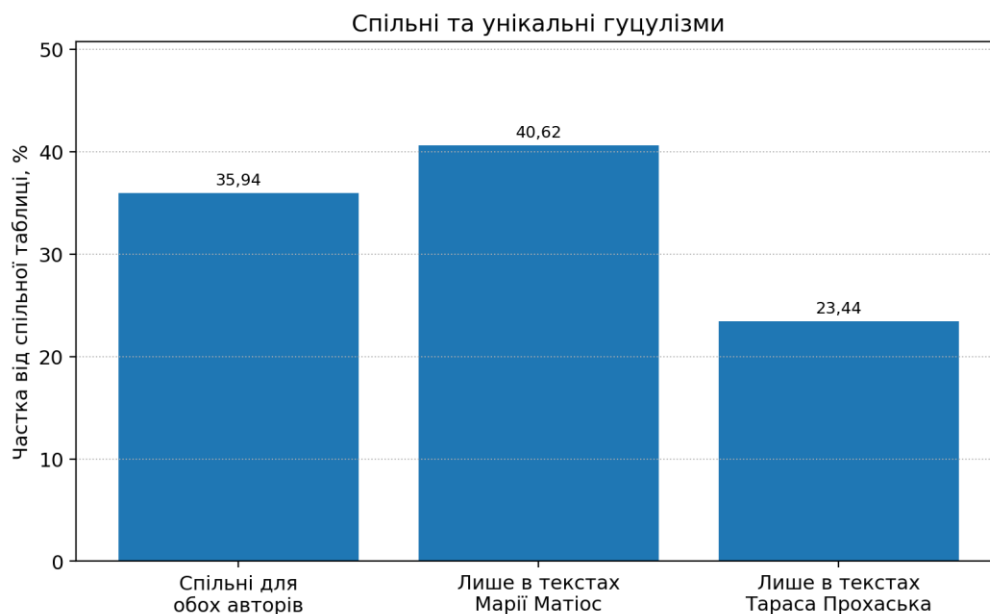


Рисунок 3.4 – Спільні та унікальні гуцулізми в текстах Марії Матіос і Тараса Прохаська, %

Як показано на рисунку 3.4, найбільшу частку становлять гуцулізми, зафіксовані лише в текстах Марії Матіос, – 40,62 %. Спільні для обох авторів одиниці становлять 35,94 %, а гуцулізми, зафіксовані лише в текстах Тараса Прохаська, – 23,44 %. Це свідчить про те, що матеріал Марії Матіос демонструє ширший абсолютний перелік унікальних одиниць, тоді як у текстах Тараса Прохаська гуцулізми представлені компактнішим, але семантично виразним набором.

Водночас частотність спільних одиниць у текстах двох авторів істотно відрізняється, тому її потрібно аналізувати через відносні показники. Наприклад, лексема *газда* становить 12,15 % уживань гуцулізмів у Марії Матіос і 1,23 % у Тараса Прохаська; *дідо* – відповідно 6,02 % і 18,40 %; *півниця* – 2,53 % і 6,75 %. Такі розбіжності засвідчують не лише різну кількість повторів, а й різні тематичні центри авторського мовного моделювання Гуцульщини.

У прозі Марії Матіос гуцульська діалектна лексика має вищу сумарну абсолютну частоту й активніше інтегрується в мовну тканину тексту. Особливо помітними є одиниці, пов'язані з родинно–соціальним середовищем: *газда*,

баба, молодиця, дідо, вуйко, газдиня, мамка. Вони допомагають формувати систему персонажних зв'язків, передають традиційну соціальну ієрархію й підкреслюють належність героїв до локального середовища.

У текстах Тараса Прохаська унікальні гуцулізми мають дещо інший характер. Серед них помітні одиниці, пов'язані з природним середовищем, предметною культурою, локальними назвами їжі, одягу й народних уявлень: *черес, жереп, бартка, афини, тогодзи, бриндза, кєчка, трембітар, гадєрниця*. Такі слова працюють як маркери карпатського простору, культурної пам'яті й предметно–чуттєвого світу.

Порівняльний аналіз показує, що в текстах Марії Матіос гуцулізми частіше пов'язані із соціально–родинною сферою, характеристикою персонажів і побутовим середовищем. Натомість у текстах Тараса Прохаська помітнішою є предметно–просторова, природна та етнокультурна складова. Це не означає повної відсутності цих груп у текстах іншого автора, однак засвідчує різні акценти в авторському використанні гуцульської діалектної лексики.

Функційний аналіз гуцульських діалектизмів дає змогу уточнити, яку роль ці одиниці виконують у художніх текстах. У досліджуваному матеріалі гуцулізми виконують номінативну, локалізаційну, етнокультурну, характерологічну, експресивну, стилістичну та текстотвірну функції. Одна й та сама одиниця може одночасно називати локальну реалію, маркувати простір, характеризувати персонажа й підтримувати загальну атмосферу твору.

Основні функції гуцулізмів у досліджуваному матеріалі подано в таблиці 3.9.

Таблиця 3.9

Функційне навантаження гуцулізмів у художніх текстах

Функція	Зміст функції	Приклади гуцулізмів	Прояв у текстах
Номінативна	Називання локальних предметів, осіб, явищ, дій і понять	газда, дідо, ліжник, будз, черес	Дає змогу позначити реалії, характерні для гуцульського середовища

Продовження таблиці 3.9

Функція	Зміст функції	Приклади гуцулізмів	Прояв у текстах
Локалізаційна	Маркування простору Гуцульщини	смерека, плай, колиба, гражда, пивниця	Закріплює художні події за карпатським регіональним простором
Етнокультурна	Передача традиційної культури, побуту, ремесел, їжі, одягу	дримба, трембіта, кептар, будз, ватра	Вводить у текст культурно специфічні гуцульські реалії
Характерологічна	Мовна характеристика персонажів	газда, баба, молодиця, вуйко, газдиня	Показує соціальну роль, походження, належність персонажа до локального середовища
Експресивна	Передача оцінності, емоційності, фамільярності або напруження	баба, мамка, молодиця, бахур	Уточнює ставлення мовця до особи або ситуації
Стилістична	Створення мовного колориту й авторської манери	дідо, пивниця, коц, фіранка, ватра	Увиразнює мовну тканину тексту й формує регіональний стиль
Текстотвірна	Формування повторюваних лексичних домінант і мотивів	газда, дримба, дідо, смерека, ліжник	Пов'язує окремі фрагменти тексту через повторювані культурно значущі одиниці

Дані таблиці 3.9 показують, що гуцулізми виконують у художньому тексті комплекс взаємопов'язаних функцій. Найбільш очевидною є номінативна функція, оскільки більшість виявлених одиниць називає конкретні предмети, осіб, явища або поняття, пов'язані з гуцульським середовищем. Локалізаційна та етнокультурна функції допомагають закріпити художній світ за карпатським простором і ввести в текст реалії традиційної культури.

У прозі Марії Матіос етнокультурна функція тісно поєднується з характерологічною. Частотні лексеми *газда*, *баба*, *молодиця*, *дідо*, *вуйко*, *газдиня* не тільки називають людей, а й окреслюють соціальні ролі, родинні зв'язки та місце персонажів у традиційній системі координат. Через такі гуцулізми вибудовується соціальний простір твору.

У текстах Тараса Прохаська гуцулізми частіше виконують локалізаційну, етнокультурну та атмосферотворчу функції. Лексеми *ліжник, пивниця, будз, смерека, черес, коц, ватра, опришок* створюють предметно–культурне тло, пов'язане з Карпатами, пам'яттю, побутом і природним середовищем.

Отже, порівняльний і функційний аналіз засвідчує, що гуцулізми в текстах Марії Матіос і Тараса Прохаська мають спільну регіональну основу, але різне авторське функціонування. У текстах Марії Матіос вони утворюють інтенсивний соціально–побутовий і характерологічний шар, тісно пов'язаний із мовленням персонажів і драматичною організацією тексту. У текстах Тараса Прохаська гуцулізми частіше виступають як засоби культурної локалізації, просторової пам'яті й атмосферного моделювання карпатського світу.

Висновки до третього розділу

У третьому розділі здійснено лінгвостатистичний аналіз гуцульських діалектизмів у художніх текстах Марії Матіос і Тараса Прохаська на основі результатів автоматизованого зіставлення авторських частотних словників із реєстром гуцульської діалектної лексики. Аналіз охопив частотні, лексико–граматичні, лексико–семантичні, типологічні та функційні характеристики виявлених одиниць.

У текстах Марії Матіос виявлено 51 лексему реєстру частотного словника гуцулізмів із сумарною абсолютною частотою 831 слововживання та покриттям тексту 0,189 %. У текстах Тараса Прохаська зафіксовано 40 лексем із сумарною абсолютною частотою 163 слововживання та покриттям 0,128 %. Водночас частка лексем гуцулізмів від загального реєстру частотного словника є вищою в матеріалі Тараса Прохаська: 0,239 % проти 0,131 % у матеріалі Марії Матіос. Це засвідчує потребу враховувати не лише абсолютні, а й відносні показники.

Лексико–граматичний аналіз показав переважання іменникових одиниць в обох авторських корпусах: коди К та Й разом охоплюють 70,58 % реєстру гуцулізмів у текстах Марії Матіос і 65,00 % у текстах Тараса Прохаська.

Лексико–семантична класифікація засвідчила, що виявлені гуцулізми репрезентують родинно–соціальну, побутову, предметну, природну, етнокультурну, харчову, господарську, історико–культурну та обрядову сфери.

Типологічний аналіз підтвердив переважання лексичних та етнографічних гуцулізмів. У текстах Марії Матіос лексичні одиниці становлять 62,75 % реєстру, а етнографічні – 35,29 %. У текстах Тараса Прохаська частка лексичних гуцулізмів становить 55,00 %, а етнографічних – 42,50 %. Отже, у матеріалі Прохаська етнографічний компонент має більшу відносну вагу, тоді як у текстах Матіос активніше повторюються соціально–побутові лексеми.

Порівняльний аналіз засвідчив, що із 64 діалектних одиниць 23 є спільними для обох авторів, 26 зафіксовано лише в текстах Марії Матіос, а 15 – лише в текстах Тараса Прохаська. Функційно гуцулізми виконують номінативну, локалізаційну, етнокультурну, характерологічну, експресивну, стилістичну та текстотвірну функції. У текстах Марії Матіос вони активніше формують соціально–побутовий і характерологічний шар, тоді як у текстах Тараса Прохаська частіше виступають засобами культурної локалізації, просторової пам'яті й атмосферного моделювання карпатського світу.

ВИСНОВКИ

Проведене дослідження засвідчує, що гуцулізми в художніх текстах Марії Матіос і Тараса Прохаська становлять не випадкові регіональні вкраплення, а системний шар діалектної лексики, який бере участь у формуванні мовного образу Гуцульщини. Їхня значущість полягає не лише в номінації локальних реалій, а й у створенні етнокультурного простору, характеристиці персонажів, моделюванні побутового середовища та увиразненні авторської художньої картини світу.

Поєднання корпусного, лексикографічного й лінгвостатистичного підходів дало змогу перейти від описового спостереження за діалектною лексикою до її кількісно обґрунтованого аналізу. Автоматизоване зіставлення авторських частотних словників із реєстром гуцульської діалектної лексики виявилось результативним інструментом для встановлення складу, частотності та розподілу гуцулізмів у двох авторських корпусах. Значущість такого підходу полягає в тому, що він робить аналіз діалектної лексики більш відтворюваним і придатним для подальших корпусних досліджень.

Результати частотного аналізу показали, що в текстах Марії Матіос зафіксовано 51 лексему реєстру частотного словника гуцулізмів із сумарною абсолютною частотою 831 слововживання. У текстах Тараса Прохаська виявлено 40 таких лексем із сумарною абсолютною частотою 163 слововживання. Покриття тексту гуцулізмами становить 0,189 % у корпусі Марії Матіос і 0,128 % у корпусі Тараса Прохаська. Це засвідчує більшу інтенсивність повторення гуцульської діалектної лексики в текстах Марії Матіос.

Водночас відносні показники уточнюють інтерпретацію абсолютних даних. Частка лексем гуцулізмів від загального реєстру частотного словника є вищою в матеріалі Тараса Прохаська: 0,239 % проти 0,131 % у матеріалі Марії Матіос. Це означає, що в текстах Тараса Прохаська гуцулізми становлять відносно помітнішу частину лексичного реєстру, хоча загальна кількість їх

уживань є нижчою. Отже, зіставлення авторських корпусів потребує врахування не лише абсолютної частоти, а й відносних показників, оскільки корпуси істотно відрізняються за обсягом.

Лексико–граматичний аналіз засвідчив переважання іменникових одиниць у виявленому матеріалі. У текстах Марії Матіос іменникові коди К та Й разом охоплюють 70,58 % реєстру гуцулізмів, а в текстах Тараса Прохаська – 65,00 %. Це підтверджує, що гуцулізми в досліджуваних текстах насамперед виконують номінативну функцію, називаючи осіб, предмети, елементи побуту, простору, матеріальної культури та соціального середовища. Такий результат узгоджується з природою діалектної лексики, яка часто фіксує конкретні реалії локального світу.

Лексико–семантична класифікація показала, що гуцулізми репрезентують кілька важливих сфер гуцульської мовної картини світу: родинно–соціальну, побутову, предметну, природну, етнокультурну, харчову, господарську, історико–культурну та обрядову. У текстах Марії Матіос особливо виразним є соціально–родинний і характерологічний шар діалектної лексики, представлений одиницями на зразок *газда, баба, молодиця, дідо, вуйко, газдиня*. У текстах Тараса Прохаська помітнішим є предметно–просторовий, природний і культурно–атмосферний компонент, пов'язаний із лексемами *ліжник, пивниця, будз, смерека, черес, коц, ватра, опришок*.

Типологічний аналіз підтвердив, що основну частину матеріалу становлять лексичні та етнографічні гуцулізми. У текстах Марії Матіос лексичні гуцулізми становлять 62,75 % реєстру, етнографічні – 35,29 %. У текстах Тараса Прохаська частка лексичних гуцулізмів становить 55,00 %, а етнографічних – 42,50 %. Отже, етнографічний компонент має вищу відносну вагу в матеріалі Тараса Прохаська, тоді як у текстах Марії Матіос активніше повторюються лексичні соціально–побутові одиниці. Це вказує на різні способи художнього використання гуцульської лексики: у Матіос вона частіше пов'язана із соціальною структурою персонажного світу, а в Прохаська – з культурною пам'яттю, предметністю та атмосферою Карпат.

Порівняння спільних і унікальних одиниць засвідчило наявність спільного ядра гуцульської діалектної лексики та водночас індивідуальну авторську вибірковість. Із 64 діалектних одиниць 23 є спільними для обох авторів, 26 зафіксовано лише в текстах Марії Матіос, а 15 – лише в текстах Тараса Прохаська. Частка спільних одиниць становить 35,94 %, частка одиниць, властивих лише матеріалу Марії Матіос, – 40,62 %, а частка гуцулізмів, зафіксованих лише в текстах Тараса Прохаська, – 23,44 %. Це дає підстави говорити про спільну регіональну основу, але різні авторські принципи добору й актуалізації діалектної лексики.

Функційна значущість гуцулізмів полягає в тому, що вони виконують у текстах не одну, а комплекс взаємопов'язаних ролей: номінативну, локалізаційну, етнокультурну, характерологічну, експресивну, стилістичну та текстотвірну. У прозі Марії Матіос гуцулізми активніше формують соціально–побутовий і характерологічний шар художнього світу, пов'язаний із родинними зв'язками, мовленням персонажів і традиційною системою соціальних ролей. У прозі Тараса Прохаська вони частіше виступають засобами культурної локалізації, просторової пам'яті й атмосферного моделювання карпатського середовища.

Отже, значущість отриманих результатів полягає в тому, що гуцулізми постають як вимірюваний і функційно важливий компонент художнього тексту. Автоматизоване укладання частотного словника дало змогу не лише зафіксувати діалектні одиниці, а й зіставити їхню частотність, семантичні групи, типи та функції у двох авторських корпусах. Запропонований підхід може бути використаний у подальших дослідженнях української діалектної лексики, комп'ютерної лексикографії, корпусного аналізу художнього тексту та вивчення авторського ідіостилу.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Аркушин Г. Л. Діалектні okazіоналізми: як вони виникають, де і скільки живуть. Урок української. 2000. № 5–6. С. 18–21.
2. Аркушин Г. Л. Західнополіська діалектологія. Луцьк : Волинський національний університет імені Лесі Українки, 2012. 257 с.
3. Бевзенко С. П. Українська діалектологія. Київ : Вища школа, 1980. 246 с.
4. Гриценко П. Ю. Діалектологія. Українська мова : енциклопедія. Київ : Українська енциклопедія, 2004. С. 146–148.
5. Гриценко П. Ю. Українська діалектна лексика: реальність і опис. Діалектна лексика: лексикологічний та лексикографічний аспекти. Київ : Наукова думка, 1987. С. 5–23.
6. Гриценко П. Ю. Українська діалектологія: проблеми і перспективи. Мовознавство. 2008. № 2–3. С. 3–14.
7. Грещук В. В. Гуцульська діалектна лексика в українській художній мові. Івано–Франківськ : Місто НВ, 2009. 352 с.
8. Грещук В. В., Грещук В. В. Гуцульська діалектна лексика в українській художній мові: словник. Івано–Франківськ : Місто НВ, 2020. Т. 1. 584 с.
9. Гуцульська діалектна лексика та фраземіка в українській художній мові : словник : у 2 т. / відп. ред. В. Грещук. Івано–Франківськ : Місто НВ, 2020. Т. 2 : Н–Я. 468 с.
10. Герман К. Ф. Лінгвогеографічне дослідження гуцульського діалекту. Вісник Прикарпатського національного університету. Філологія. Вип. XXV–XXVI. Івано–Франківськ, 2010. С. 45–52.
11. Гуцульські говірки. Короткий словник / уклад. Я. Закревська. Львів : Інститут українознавства ім. І. Крип'якевича НАН України, 1997. 232 с.
12. Дзендзелівський Й. О. Українська діалектологія. Ч. 1. Ужгород : Ужгородський державний університет, 1965. 112 с.
13. Жилко Ф. Т. Нариси з діалектології української мови. Київ : Радянська школа, 1966. 307 с.

14. Закревська Я. В. Нариси з діалектного словотвору в ареальному аспекті. Київ : Наукова думка, 1976. 164 с.
15. Закревська Я. В. Українська діалектна лексикографія. Українська мова. 2004. № 2. С. 3–11.
16. Зубань О. М. Корпусна лінгвістика: навчально–методичний посібник. Київ : ВПЦ “Київський університет”, 2018. 132 с.
17. Карпенко Ю. О. Вступ до мовознавства. Київ : Академія, 2006. 336 с.
18. Кобилянський Б. В. Діалект і літературна мова. Київ : Наукова думка, 1960. 276 с.
19. Ковалик І. І. Вчення про словотвір. Львів : Вид-во Львівського університету, 1958. 78 с.
20. Кочерган М. П. Загальне мовознавство. Київ : Академія, 2006. 464 с.
21. Кочерган М. П. Вступ до мовознавства. Київ : Академія, 2010. 368 с.
22. Лесюк М. П. Становлення і розвиток української літературної мови в Галичині. Івано–Франківськ : Місто НВ, 2014. 732 с.
23. Дарчук Н. П. Комп’ютерна лінгвістика: навчальний посібник. Київ : ВПЦ “Київський університет”, 2008. 351 с.
24. Дарчук Н. П. Український національний лінгвістичний корпус: засади і принципи створення. Українське мовознавство. 2010. Вип. 40/1. С. 34–42.
25. Корпус української мови. Лабораторія комп’ютерної лінгвістики Навчально–наукового інституту філології КНУ імені Тараса Шевченка. URL: <http://www.mova.info/corpus.aspx>
26. Частотні словники української мови. Корпус української мови. Лабораторія комп’ютерної лінгвістики КНУ імені Тараса Шевченка. URL: <http://www.mova.info/corpus.aspx>
27. Перебийніс В. С. Статистичні методи для лінгвістів : навчальний посібник. Вінниця : Нова книга, 2002. С. 3–8.
28. McEnery T., Wilson A. Corpus Linguistics: An Introduction. Edinburgh : Edinburgh University Press, 2001. 235 p.

29. McEnery T., Hardie A. Corpus Linguistics: Method, Theory and Practice. Cambridge : Cambridge University Press, 2012. 294 p.
30. Baker P. Using Corpora in Discourse Analysis. London : Continuum, 2006. 208 p.
31. Biber D., Conrad S., Reppen R. Corpus Linguistics: Investigating Language Structure and Use. Cambridge : Cambridge University Press, 1998. 300 p.
32. Sinclair J. Corpus, Concordance, Collocation. Oxford : Oxford University Press, 1991. 179 p.
33. Stubbs M. Words and Phrases: Corpus Studies of Lexical Semantics. Oxford : Blackwell, 2001. 288 p.
34. Kennedy G. An Introduction to Corpus Linguistics. London ; New York : Longman, 1998. 315 p.
35. Meyer C. F. English Corpus Linguistics: An Introduction. Cambridge : Cambridge University Press, 2002. 168 p.
36. Gries S. T. Quantitative Corpus Linguistics with R: A Practical Introduction. New York : Routledge, 2009. 240 p.
37. Manning C. D., Schütze H. Foundations of Statistical Natural Language Processing. Cambridge, MA : MIT Press, 1999. 680 p.
38. Jurafsky D., Martin J. H. Speech and Language Processing. 3rd ed. draft. URL: <https://web.stanford.edu/~jurafsky/slp3/>
39. Python Software Foundation. Python 3 Documentation. URL: <https://docs.python.org/3/>
40. Streamlit Documentation. URL: <https://docs.streamlit.io/>
41. pandas Documentation. URL: <https://pandas.pydata.org/docs/>
42. openpyxl Documentation. URL: <https://openpyxl.readthedocs.io/>
43. Unicode Consortium. The Unicode Standard. URL: <https://www.unicode.org/standard/standard.html>
44. Matplotlib Documentation. URL: <https://matplotlib.org/stable/users/index>
45. W3C. HTML Living Standard. URL: <https://html.spec.whatwg.org/>
46. W3C. CSS Snapshot. URL: <https://www.w3.org/TR/CSS/>

СПИСОК ДЖЕРЕЛ ІЛЮСТРАТИВНОГО МАТЕРІАЛУ

1. ММ – Частотний словник за художніми текстами Марії Матіос. Корпус української мови. Лабораторія комп'ютерної лінгвістики ННІ філології КНУ імені Тараса Шевченка.
2. ТП – Частотний словник за художніми текстами Тараса Прохаська. Корпус української мови. Лабораторія комп'ютерної лінгвістики ННІ філології КНУ імені Тараса Шевченка.
3. СГД – Гуцульська діалектна лексика та фраземіка в українській художній мові : словник : у 2 т. / відп. ред. В. Грещук. Івано–Франківськ : Місто НВ, 2020.

ДОДАТКИ

ДОДАТОК А

Скріншоти роботи застосунку

Рисунок А.1 – Завантаження вхідних файлів

Вхідні файли

Завантажте власні файли або скористайтесь прикладами з папки data.

Частотний словник М. Матіос

↻

Обрати файл

Частотний словник Т. Прохаська

↻

Обрати файл

Реєстр гуцулізмів

↻

Обрати файл

☒ Використати файли з папки data

Буде використано приклади з папки data.

Опрацювати дані

Рисунок А.2 – Підсумок автоматизованої обробки

Підсумок обробки

МАТІОС: ТОКЕНИ 440662 після очищення	МАТІОС: УНІКАЛЬНІ 38996 лексем	МАТІОС: ГУЦУЛІЗМИ 51 сумарна частота 831	ПРОХАСЬКО: ТОКЕНИ 127350 після очищення	ПРОХАСЬКО: УНІКАЛЬНІ 16759 лексем	ПРОХАСЬКО: ГУЦУЛІЗМИ 40 сумарна частота 163	СПІЛЬНІ ОДИНИЦІ 23 в обох авторів	ЛИШЕ МАТІОС 26 із 64 у спільному реєстрі
ЛИШЕ ПРОХАСЬКО 15 із 64 у спільному реєстрі							
РУЧНА ПЕРЕВІРКА 39 потенційні похідні форми				СПІЛЬНИЙ РЕЄСТР 64 для порівняння авторів			

Рисунок А.3 – Перегляд словника гуцулізмів

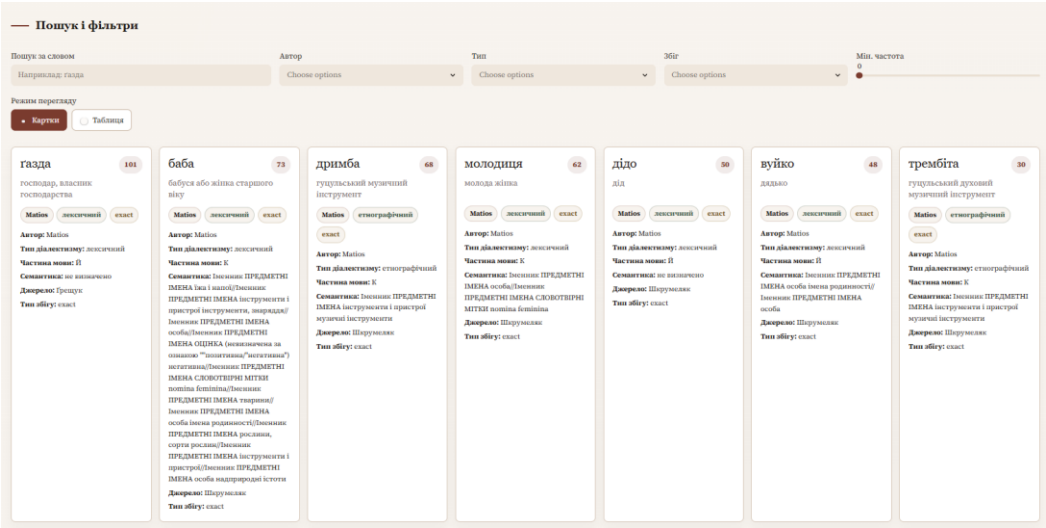
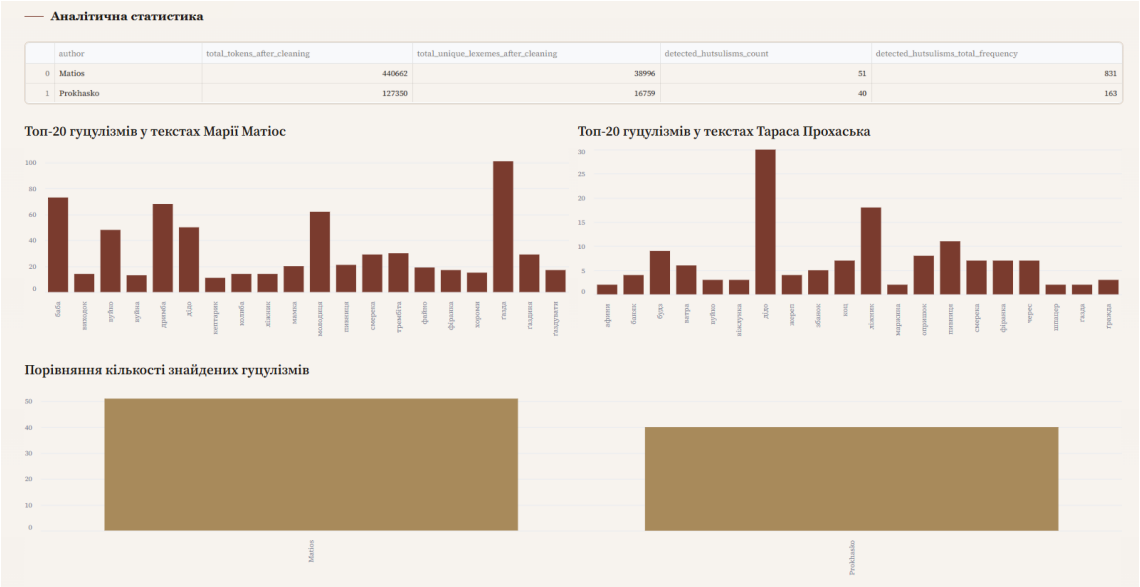


Рисунок А.4 – Аналітична статистика



ДОДАТОК Б

Фрагменти програмного коду застосунку

У додатку Б наведено основні фрагменти програмного коду локального вебзастосунку, за допомогою якого здійснювалося очищення частотних словників, нормалізація лексем, автоматизоване зіставлення з реєстром гуцульських діалектизмів, формування таблиці ручної перевірки, розрахунок статистичних показників та експорт результатів. Повний програмний код застосунку розміщено [у відкритому репозиторії GitHub](#). Основний файл запуску – [app.py](#). Допоміжні модулі обробки даних містяться в папці [src](#), вхідні дані – у папці [data](#), результати обробки – у папці [output](#). Для запуску застосунку потрібно встановити залежності з файла [requirements.txt](#) і виконати команду `streamlit run app.py` або запустити файл [run.bat](#).

Лістинг Б.1 – Завантаження вхідних даних

```
from __future__ import annotations

from pathlib import Path

from typing import BinaryIO

import pandas as pd

FREQUENCY_COLUMNS = ["lexeme", "class", "frequency", "semantics"]

HUTSUL_COLUMNS = ["dialectism", "meaning", "source", "type", "comment"]

def _read_csv_like(file: str | Path | BinaryIO) -> pd.DataFrame:

    try:

        return pd.read_csv(file, sep=";", encoding="utf-8", dtype=str,
                           keep_default_na=False)

    except UnicodeDecodeError:

        if hasattr(file, "seek"):

            file.seek(0)

            return pd.read_csv(file, sep=";", encoding="cp1251", dtype=str,
                               keep_default_na=False)
```

```

def _read_excel(file: str | Path | BinaryIO) -> pd.DataFrame:

    return pd.read_excel(file, dtype=str).fillna("")

def read_table(file: str | Path | BinaryIO, name: str | None = None) ->
pd.DataFrame:

    filename = name or getattr(file, "name", "") or str(file)

    suffix = Path(filename).suffix.lower()

    if suffix in {".xlsx", ".xls"}:

        return _read_excel(file)

    return _read_csv_like(file)

def _add_missing_columns(df: pd.DataFrame, columns: list[str]) ->
tuple[pd.DataFrame, list[str]]:

    missing = [col for col in columns if col not in df.columns]

    result = df.copy()

    for col in missing:

        result[col] = ""

    return result[columns + [col for col in result.columns if col not in
columns]], missing

def load_frequency_dictionary(file: str | Path | BinaryIO, name: str | None =
None) -> tuple[pd.DataFrame, list[str]]:

    df = read_table(file, name)

    df, missing = _add_missing_columns(df, FREQUENCY_COLUMNS)

    return df, missing

def load_hutsul_dictionary(file: str | Path | BinaryIO, name: str | None = None)
-> tuple[pd.DataFrame, list[str]]:

    df = read_table(file, name)

    df, missing = _add_missing_columns(df, HUTSUL_COLUMNS)

    return df, missing

```

Лістинг Б.2 – Очищення та нормалізація лексем

```

from __future__ import annotations

```

```

import re

import pandas as pd

APOSTROPHES = {

    "`": "'",

    "'": "'",

    "‘": "’",

    "’": "’",

    "``": "",

    "``": "",

    "''": "",

    "''": ""

}

def normalize_lexeme(text: object) -> str:

    """Normalize a lexeme for matching without changing r to r."""

    if text is None:

        return ""

    value = str(text).strip().lower()

    for old, new in APOSTROPHES.items():

        value = value.replace(old, new)

    value = value.replace("-", "-").replace("-", "-").replace("-", "-")

    value = re.sub(r"\s+", " ", value)

    value = value.strip(" \t\r\n\"'“”„«»[]{}()<>.,;:!?/\\|")

    return value

def is_service_token(text: object) -> bool:

    value = normalize_lexeme(text)

    if not value:

        return True

    if re.fullmatch(r"[\W_]+", value, flags=re.UNICODE):

        return True

    if re.fullmatch(r"\d+([\.,:/-]\d+)*", value):

```

```

        return True

    if re.fullmatch(r"\d{2,4}(-\d{2,4})?", value):

        return True

    if not re.search(r"[a-щьюяеііѣ]", value):

        return True

    return False

def safe_frequency(value: object) -> int:

    try:

        if value is None or str(value).strip() == "":

            return 0

        return int(float(str(value).replace(",", ".")))

    except ValueError:

        return 0

def clean_frequency_dictionary(df: pd.DataFrame) -> tuple[pd.DataFrame,
pd.DataFrame]:

    result = df.copy()

    if "lexeme" not in result.columns:

        result["lexeme"] = ""

    if "frequency" not in result.columns:

        result["frequency"] = "0"

    result["original_lexeme"] = result["lexeme"].astype(str)

    result["normalized_lexeme"] = result["lexeme"].map(normalize_lexeme)

    result["frequency"] = result["frequency"].map(safe_frequency)

    result["is_service_token"] =
result["normalized_lexeme"].map(is_service_token)

    removed = result[result["is_service_token"]].copy()

    cleaned = result[~result["is_service_token"]].copy()

    cleaned = cleaned[cleaned["frequency"] > 0].copy()

    cleaned = cleaned.sort_values("frequency",
ascending=False).reset_index(drop=True)

```

```
return cleaned, removed
```

Лістинг Б.3 – Зіставлення лексем із реєстром гуцулізмів

```
from __future__ import annotations

import pandas as pd

from .utils import normalize_lexeme

RESULT_COLUMNS = [

    "dialectism",

    "author",

    "original_lexeme",

    "normalized_lexeme",

    "class",

    "frequency",

    "semantics",

    "dialectism_type",

    "meaning",

    "dictionary_source",

    "match_type",

    "comment",

]

def prepare_hutsul_dictionary(df: pd.DataFrame) -> tuple[pd.DataFrame,
pd.DataFrame]:

    result = df.copy()

    if "dialectism" not in result.columns:

        result["dialectism"] = ""

    result["normalized_dialectism"] = result["dialectism"].map(normalize_lexeme)

    result = result[result["normalized_dialectism"] != ""].copy()
```

```

        duplicates = result[result.duplicated("normalized_dialectism",
keep=False)].copy()

        result = result.drop_duplicates("normalized_dialectism",
keep="first").reset_index(drop=True)

        return result, duplicates

def match_exact(author_df: pd.DataFrame, hutsul_df: pd.DataFrame) ->
pd.DataFrame:

    merged = author_df.merge(

        hutsul_df,

        left_on="normalized_lexeme",

        right_on="normalized_dialectism",

        how="inner",

    )

    if merged.empty:

        return pd.DataFrame(columns=RESULT_COLUMNS)

    result = pd.DataFrame(

        {

            "dialectism": merged.get("dialectism", ""),

            "author": merged.get("author", ""),

            "original_lexeme": merged.get("original_lexeme",
merged.get("lexeme", "")),

            "normalized_lexeme": merged.get("normalized_lexeme", ""),

            "class": merged.get("class", ""),

            "frequency": merged.get("frequency", 0),

            "semantics": merged.get("semantics", ""),

            "dialectism_type": merged.get("type", ""),

            "meaning": merged.get("meaning", ""),

            "dictionary_source": merged.get("source", ""),

            "match_type": "exact",

            "comment": merged.get("comment", ""),

        }

```

```

    )

    return result.sort_values("frequency",
                              ascending=False).reset_index(drop=True)

def build_author_result(author_name: str, author_df: pd.DataFrame, hutsul_df:
pd.DataFrame) -> tuple[pd.DataFrame, pd.DataFrame]:

    prepared = author_df.copy()

    prepared["author"] = author_name

    exact = match_exact(prepared, hutsul_df)

    manual = find_manual_candidates(prepared, hutsul_df)

    return exact, manual

```

Лістинг Б.4 – Формування таблиці ручної перевірки

```

from __future__ import annotations

import pandas as pd

def find_manual_candidates(author_df: pd.DataFrame, hutsul_df: pd.DataFrame) ->
pd.DataFrame:

    exact_keys = set(hutsul_df["normalized_dialectism"].dropna())

    exact_author =
set(author_df.loc[author_df["normalized_lexeme"].isin(exact_keys),
"normalized_lexeme"])

    rows: list[dict[str, object]] = []

    dictionary_items = hutsul_df[["dialectism",
"normalized_dialectism"]].dropna().values.tolist()

    for _, row in author_df.iterrows():

        lexeme = row.get("normalized_lexeme", "")

        if not lexeme or lexeme in exact_author:

            continue

        for dialectism, normalized in dictionary_items:

            if len(normalized) < 4:

```

```

        continue

    if lexeme.startswith(normalized) or normalized in lexeme:

        rows.append(

            {

                "author": row.get("author", ""),

                "original_lexeme": row.get("original_lexeme",
row.get("lexeme", "")),

                "normalized_lexeme": lexeme,

                "possible_dictionary_match": dialectism,

                "frequency": row.get("frequency", 0),

                "class": row.get("class", ""),

                "semantics": row.get("semantics", ""),

                "reason": "partial_or_derived_form",

                "decision": "",

                "corrected_dialectism": "",

                "comment": "",

            }

        )

    break

    return pd.DataFrame(rows).sort_values("frequency",
ascending=False).reset_index(drop=True)

```

Лістинг Б.5 – Розрахунок статистичних показників

```

from __future__ import annotations

import pandas as pd

def _unique_lexemes(df: pd.DataFrame) -> int:

    """Distinct normalized lexemes after cleaning (the final unique-lexeme
metric)."""

    if df.empty:

        return 0

```

```

    if "normalized_lexeme" in df.columns:

        return int(df["normalized_lexeme"].nunique())

    return len(df)

def _group_stats(df: pd.DataFrame, group_col: str, author: str) -> pd.DataFrame:

    if df.empty or group_col not in df.columns:

        return pd.DataFrame(columns=["author", group_col, "count",
"total_frequency"])

    result = (

        df.groupby(group_col, dropna=False)

        .agg(count=("dialectism", "count"), total_frequency=("frequency",
"sum"))

        .reset_index()

    )

    result.insert(0, "author", author)

    return result.sort_values("total_frequency",
ascending=False).reset_index(drop=True)

def calculate_statistics(

    matios_result: pd.DataFrame,

    prokhasko_result: pd.DataFrame,

    matios_cleaned: pd.DataFrame,

    prokhasko_cleaned: pd.DataFrame,

    combined_result: pd.DataFrame,

) -> dict[str, pd.DataFrame]:

    general_stats = pd.DataFrame(

        [

            {

                "author": "Matios",

                "total_tokens_after_cleaning":

int(matios_cleaned["frequency"].sum()) if not matios_cleaned.empty else 0,

                "total_unique_lexemes_after_cleaning":

_unique_lexemes(matios_cleaned),

```

```

        "detected_hutsulisms_count": len(matios_result),

        "detected_hutsulisms_total_frequency":
int(matios_result["frequency"].sum()) if not matios_result.empty else 0,

    },

    {

        "author": "Prokhasko",

        "total_tokens_after_cleaning":
int(prokhasko_cleaned["frequency"].sum()) if not prokhasko_cleaned.empty else 0,

        "total_unique_lexemes_after_cleaning":
_unique_lexemes(prokhasko_cleaned),

        "detected_hutsulisms_count": len(prokhasko_result),

        "detected_hutsulisms_total_frequency":
int(prokhasko_result["frequency"].sum()) if not prokhasko_result.empty else 0,

    },

]

)

all_results = pd.concat([matios_result, prokhasko_result],
ignore_index=True)

top_hutsulisms = all_results.sort_values("frequency",
ascending=False).head(40).reset_index(drop=True)

by_part_of_speech = pd.concat(

    [_group_stats(matios_result, "class", "Matios"),
_group_stats(prokhasko_result, "class", "Prokhasko")],

    ignore_index=True,

)

by_dialectism_type = pd.concat(

    [_group_stats(matios_result, "dialectism_type", "Matios"),
_group_stats(prokhasko_result, "dialectism_type", "Prokhasko")],

    ignore_index=True,

)

by_semantics = pd.concat(

    [_group_stats(matios_result, "semantics", "Matios"),
_group_stats(prokhasko_result, "semantics", "Prokhasko")],

```

```

        ignore_index=True,
    )

    shared_rows = []

    for _, row in combined_result.iterrows():

        if row.get("present_in_matos") and row.get("present_in_prokhasko"):

            status = "shared"

        elif row.get("present_in_matos"):

            status = "only_matos"

        else:

            status = "only_prokhasko"

        shared_rows.append(

            {

                "dialectism": row.get("dialectism", ""),

                "frequency_matos": row.get("frequency_matos", 0),

                "frequency_prokhasko": row.get("frequency_prokhasko", 0),

                "status": status,

            }

        )

    shared_and_unique = pd.DataFrame(shared_rows)

    return {

        "general_stats": general_stats,

        "top_hutsulisms": top_hutsulisms,

        "by_part_of_speech": by_part_of_speech,

        "by_dialectism_type": by_dialectism_type,

        "by_semantics": by_semantics,

        "shared_and_unique": shared_and_unique,

    }

```

Лістинг Б.6 – Експорт результатів у Excel і ZIP

```

from __future__ import annotations

from pathlib import Path

from zipfile import ZipFile

import pandas as pd

def ensure_output_dir(path: str | Path = "output") -> Path:

    output = Path(path)

    output.mkdir(parents=True, exist_ok=True)

    return output

def write_excel(path: str | Path, sheets: dict[str, pd.DataFrame]) -> Path:

    output_path = Path(path)

    output_path.parent.mkdir(parents=True, exist_ok=True)

    with pd.ExcelWriter(output_path, engine="openpyxl") as writer:

        for sheet_name, df in sheets.items():

            safe_name = sheet_name[:31] or "Sheet1"

            df.to_excel(writer, sheet_name=safe_name, index=False)

    return output_path

def export_results_to_excel(

    matios_result: pd.DataFrame,

    prokhasko_result: pd.DataFrame,

    combined_result: pd.DataFrame,

    manual_review: pd.DataFrame,

    statistics: dict[str, pd.DataFrame],

    output_dir: str | Path = "output",

) -> dict[str, Path]:

    out = ensure_output_dir(output_dir)

    paths = {

```

```

        "matios": write_excel(out / "matios_hutsulisms.xlsx",
{"matios_hutsulisms": matios_result})),

        "prokhasko": write_excel(out / "prokhasko_hutsulisms.xlsx",
{"prokhasko_hutsulisms": prokhasko_result})),

        "combined": write_excel(out / "combined_hutsulisms.xlsx",
{"combined_hutsulisms": combined_result})),

        "manual": write_excel(out / "manual_review.xlsx", {"manual_review":
manual_review})),

        "statistics": write_excel(out / "statistics_summary.xlsx", statistics),

    }

    zip_path = out / "all_results.zip"

    with ZipFile(zip_path, "w") as archive:

        for path in paths.values():

            archive.write(path, arcname=path.name)

```

ДОДАТОК В

Списки гуцулізмів, виявлених у частотних словниках

В.1 Гуцулізми в текстах Марії Матіос

№	Діалектизм	Абс. частота	Частка в уживаннях, %	Тип	Джерело
1	газда	101	12,15	лексичний	Грещук
2	баба	73	8,78	лексичний	Шкрумеляк
3	дримба	68	8,18	етнографічний	Шкрумеляк
4	молодиця	62	7,46	лексичний	Шкрумеляк
5	дідо	50	6,02	лексичний	Шкрумеляк
6	вуйко	48	5,78	лексичний	Шкрумеляк
7	трембіта	30	3,61	етнографічний	Шкрумеляк
8	смерека	29	3,49	лексичний	Шкрумеляк
9	газдиня	29	3,49	лексичний	Грещук
10	пивниця	21	2,53	лексичний	Шкрумеляк
11	мамка	20	2,41	лексичний	Шкрумеляк
12	файно	19	2,29	лексичний	Шкрумеляк
13	газдувати	17	2,05	лексичний	Грещук

14	фіранка	17	2,05	лексичний	Шкрумеляк
15	хороми	15	1,81	лексичний	Шкрумеляк
16	колиба	14	1,68	етнографічний	Шкрумеляк
17	виходок	14	1,68	лексичний	Шкрумеляк
18	ліжник	14	1,68	етнографічний	Шкрумеляк
19	вуйна	13	1,56	лексичний	Шкрумеляк
20	кептарик	11	1,32	етнографічний	Шкрумеляк
21	кептар	11	1,32	етнографічний	Шкрумеляк
22	челядь	10	1,20	лексичний	Шкрумеляк
23	дідич	10	1,20	лексичний	Шкрумеляк
24	дідич	10	1,20	лексичний	Шкрумеляк
25	баняк	10	1,20	лексичний	Шкрумеляк
26	дараба	10	1,20	етнографічний	Шкрумеляк
27	ватра	9	1,08	етнографічний	Шкрумеляк
28	газдівство	9	1,08	лексичний	Ґрещук
29	дупло	7	0,84	семантичний	Шкрумеляк
30	газдовитий	7	0,84	лексичний	Ґрещук

31	плай	7	0,84	етнографічний	Шкрумеляк
32	плай	7	0,84	етнографічний	Шкрумеляк
33	барвінок	6	0,72	лексичний	Шкрумеляк
34	коц	6	0,72	етнографічний	Шкрумеляк
35	гражда	6	0,72	етнографічний	Шкрумеляк
36	опришок	5	0,60	етнографічний	Шкрумеляк
37	позлітка	4	0,48	лексичний	Шкрумеляк
38	будз	4	0,48	етнографічний	Шкрумеляк
39	бараболя	3	0,36	лексичний	Шкрумеляк
40	барабулі	3	0,36	лексичний	Шкрумеляк
41	борлак	3	0,36	лексичний	Шкрумеляк
42	гайдук	3	0,36	лексичний	Шкрумеляк
43	громадити	3	0,36	лексичний	Шкрумеляк
44	забавка	3	0,36	лексичний	Шкрумеляк
45	ненька	3	0,36	лексичний	Шкрумеляк
46	маржина	2	0,24	етнографічний	Шкрумеляк
47	вельон	1	0,12	етнографічний	Шкрумеляк

48	постоли	1	0,12	етнографічний	Шкрумеляк
49	збанок	1	0,12	лексичний	Шкрумеляк
50	запопадливий	1	0,12	лексичний	Шкрумеляк
51	кацавейка	1	0,12	етнографічний	Шкрумеляк

В.2 Гуцулізми в текстах Тараса Прохаська

№	Діалектизм	Абс. частота	Частка в уживаннях, %	Тип	Джерело
1	дідо	30	18,40	лексичний	Шкрумеляк
2	ліжник	18	11,04	етнографічний	Шкрумеляк
3	пивниця	11	6,75	лексичний	Шкрумеляк
4	будз	9	5,52	етнографічний	Шкрумеляк
5	опришок	8	4,91	етнографічний	Шкрумеляк
6	смерека	7	4,29	лексичний	Шкрумеляк
7	черес	7	4,29	етнографічний	Шкрумеляк
8	коц	7	4,29	етнографічний	Шкрумеляк
9	фіранка	7	4,29	лексичний	Шкрумеляк
10	ватра	6	3,68	етнографічний	Шкрумеляк

11	збанок	5	3,07	лексичний	Шкрумеляк
12	баняк	4	2,45	лексичний	Шкрумеляк
13	жереп	4	2,45	лексичний	Шкрумеляк
14	гражда	3	1,84	етнографічний	Шкрумеляк
15	вуйко	3	1,84	лексичний	Шкрумеляк
16	віжлунка	3	1,84	етнографічний	Шкрумеляк
17	афини	2	1,23	лексичний	Шкрумеляк
18	шпацер	2	1,23	лексичний	Шкрумеляк
19	маржина	2	1,23	етнографічний	Шкрумеляк
20	газда	2	1,23	лексичний	Грещук
21	дідич	2	1,23	лексичний	Шкрумеляк
22	дідич	2	1,23	лексичний	Шкрумеляк
23	дупло	2	1,23	семантичний	Шкрумеляк
24	дримба	1	0,61	етнографічний	Шкрумеляк
25	заворинє	1	0,61	лексичний	Шкрумеляк
26	гогодзи	1	0,61	лексичний	Шкрумеляк
27	гайдук	1	0,61	лексичний	Шкрумеляк

28	гадєрниця	1	0,61	єтнографічний	Шкрумеляк
29	пугарчик	1	0,61	лексичний	Шкрумеляк
30	спиж	1	0,61	лексичний	Шкрумеляк
31	кєчка	1	0,61	єтнографічний	Шкрумеляк
32	трембітар	1	0,61	єтнографічний	Шкрумеляк
33	трембіта	1	0,61	єтнографічний	Шкрумеляк
34	бриндза	1	0,61	єтнографічний	Шкрумеляк
35	файно	1	0,61	лексичний	Шкрумеляк
36	бартка	1	0,61	єтнографічний	Шкрумеляк
37	бартка	1	0,61	єтнографічний	Шкрумеляк
38	бахур	1	0,61	лексичний	Шкрумеляк
39	барвінок	1	0,61	лексичний	Шкрумеляк
40	баба	1	0,61	лексичний	Шкрумеляк