



Т. П. Дяк, Ю. І. Грицюк

Національний університет "Львівська політехніка", м. Львів, Україна

ВИКОРИСТАННЯ КОРПУСНИХ ІНСТРУМЕНТІВ ДЛЯ ВИЯВЛЕННЯ КЛЮЧОВИХ СЛІВ У СТІЛЕЦЬКИХ І ПОВСТАНСЬКИХ ПІСНЯХ ЯК ЖАНРУ ФОЛЬКЛОРНОГО ДИСКУРСУ

У сучасному інформаційному суспільстві аналіз текстових матеріалів і визначення їх ключових особливостей мають велике значення в різних галузях науки, зокрема в корпусній лінгвістиці. Встановлено, що незважаючи на великий потенціал застосування методологій корпусів у різних галузях дослідження, досі існує потреба їхнього опанування для практичного застосування. Наявність великомасштабних комп'ютеризованих корпусів текстів, які було вдосконалено завдяки кращій цифровій інфраструктурі та технологічним досягненням, що відбуваються в епоху інформації, забезпечує базис для лінгвістичних досліджень. Проаналізовано спеціалізоване програмне забезпечення з потужними функціями оброблення та аналізу корпусів текстів, потрібних для здійснення лінгвістичних досліджень та його практичне використання у різних дослідженнях. Також з'ясовано, що не відсутність ефективних, потужних статистичних алгоритмів або алгоритмів машинного навчання, а доступ до них дослідників є вузьким місцем у розвитку підходів на основі корпусів і суміжних дисциплін. Наведено результати вивчення можливостей і методів використання корпусних інструментів для виявлення та аналізу ключових слів текстів щодо корпусної лінгвістики. Такі програмні інструменти, як корпусний менеджер AntConc та вебсистема Sketch Engine, мають важливе значення, надаючи можливість здійснювати різноманітні лінгвістичні дослідження, серед яких аналіз жанрових особливостей текстів. Дослідження проведено на підставі корпусу текстів, який нараховує 35 українських стрілецьких і повстанських пісень. Проаналізовано лексико-семантичні особливості ключових слів, встановлено їхні ролі в аналізі мови та детально вивчено функціонал корпусних інструментів для їхнього пошуку та аналізу. Запропоновано результати аналізу методів та інструментів, використаних для аналізу текстів стрілецьких і повстанських пісень, визначення ключових слів, виявлення основних тематичних і лінгвістичних ознак досліджуваних пісень. Для всебічного аналізу ключових слів використано функції *Collocates*, *N-Grams* та *Word List* у корпусному менеджері AntConc, а також функцію *Keywords* у вебсистемі Sketch Engine. Виявлено, що серед ключових слів найбільшу частоту вживання мають такі частини мови, як ви́гуки, сполучники і частки, що притаманно для фольклорних пісень. Ключові слова, подані іменниками, змальовують родинні зв'язки, військові будні та особисті почуття вояків. Досить значною є частка прикметників і дієслів. Також наявна велика кількість слівосформ з пестливо-зменшувальними суфіксами у піснях цього жанру, що вказує на ніжне ставлення до описаних об'єктів. Отримані результати дослідження є важливим внеском у вдосконалення корпусної лінгвістики та комплексне використання програмних інструментів корпусного менеджера AntConc та вебсистеми Sketch Engine для аналізу ключових слів.

Ключові слова: корпусна лінгвістика; корпусний менеджер; цільовий та еталонний корпуси; жанрова особливість текстів; вебсистема; частотний аналіз.

Вступ / Introduction

У гібридній війні Росії проти України мистецтво є оплотом надії та одночасно зброєю у боротьбі за душі і серця людей. Кожна війна як екзистенційна криза спричиняє створення нових мистецьких творів і відродження пластів культури, які походять з визвольних воєн за свободу та незалежність. Важливим завданнями у нинішній гібридній війні Росії проти України є, серед іншого, підняття бойового духу бійців. На тлі війни свідчимо важливість національного, воєнного фольклору. Зокрема пісня Українських січових стрільців "Ой у лузі червона калина" у 2022 році набула нової хвилі популярності після виконання Андрієм Хливиюком (Бум-

бокс) у перші дні російського вторгнення та кавер-версії культового гурту "Пінк Флойд" під назвою "Heu, Heu, Rise Up!". Вона стала гімном війни з початком повномасштабного вторгнення Росії в Україну. Національний, воєнний фольклор, який зародився близько 100 років тому, знову став популярним серед мешканців усіх регіонів України. Ця тенденція вказує на потяг українців до своїх коренів, до своєї героїчної спадщини. Кордоцентризм українців проявляється саме в тому, наскільки вони відновлюють у собі потяг до фольклору, адже він не є чимось нижчегартісним, архаїчним і віджилим, порівняно з сучасними віяннями.

Чеський історик Мірослав Грох [17] наприкінці минулого століття виокремив такі три етапи національно-

Інформація про авторів:

Дяк Тетяна Петрівна, канд. пед. наук, доцент, кафедра прикладної лінгвістики. Email: tetiana.p.dyak@lpnu.ua;

<https://orcid.org/0000-0003-0919-9792>

Грицюк Юрій Іванович, д-р техн. наук, професор, кафедра програмного забезпечення. Email: yurii.i.hrytsiuk@lpnu.ua;

<https://orcid.org/0000-0001-8183-3466>

Цитування за ДСТУ: Дяк Т. П., Грицюк Ю. І. Використання корпусних інструментів для виявлення ключових слів у стрілецьких і повстанських піснях як жанру фольклорного дискурсу. Науковий вісник НЛТУ України. 2024, т. 34, № 7. С. 60–71.

Citation APA: Dyak, T. P., Hrytsiuk, Yu. I. (2024). Application of corpus tools to identify keywords of Ukrainian rebel songs as a genre of the folklore discourse. *Scientific Bulletin of UNFU*, 34(7), 60–71. <https://doi.org/10.36930/40340708>

го відродження як академічний, культурний та політичний. Саме на академічному етапі дослідники збирали фольклор, і згодом на його базі вибудовували національну історію. Чи українська нація могла постати без фольклору? Фольклор втілює такі важливі завдання, як інтеграція, а саме поєднання людей у спільноту, та диференціація, тобто відокремлення цієї спільноти від іншої. Він стоїть біля витоків створення нації, формує її, надає їй культурної ідентичності, й забезпечує здатність протидіяти зовнішнім викликам.

Однак, фольклорний дискурс дотепер не отримав належного рівня наукових розвідок і детального опису. Складність дослідження фольклорного дискурсу полягає в тому, що він посідає особливе місце у житті соціуму, адже є складним феноменом, який не відповідає винайденим людством дихотоміям: інституціональний, персональний, усний, писемний. Саме потреба детальнішого аналізу зумовлює актуальність нашого дослідження.

У сучасному інформаційному суспільстві аналіз текстових матеріалів і визначення ключових особливостей мають велике значення в різних галузях науки, особливо в галузі корпусної лінгвістики. Ефективними програмними інструментами для проведення такого аналізу є корпусний менеджер AntConc та вебсистема Sketch Engine (англ. *Engine* – двигун). Ці інструменти надають можливість проведення різноманітних лінгвістичних досліджень, зокрема й аналіз жанрових особливостей текстів.

Актуальність обраної теми нашого дослідження виявляється в контексті відродження інтересу до фольклорного жанру загалом, особливо до українських стрілецьких і повстанських пісень, а також до важливості розвитку методів аналізу текстів у лінгвістичній науці. Вивчення ключових слів і характерних особливостей фольклорного жанру за допомогою програмних інструментів корпусного менеджера AntConc та вебсистеми Sketch Engine відкриває можливості для більш глибокого розуміння цього унікального літературного жанру, що є актуальним напрямом для досліджень у сучасному академічному середовищі.

Об'єкт дослідження – визначення ключових особливостей стрілецьких і повстанських пісень як символу національної ідентичності та духовної величі українського народу.

Предмет дослідження – методи і засоби аналізу фольклорного жанру, зокрема – текстів українських стрілецьких і повстанських пісень, що дають змогу визначити ключові слова та характерні його особливості.

Мета роботи – проаналізувати та охарактеризувати ключові слова у корпусі текстів стрілецьких і повстанських пісень як потужного ресурсу в утвердженні патріотизму і символу нескореності українського народу, що дасть змогу простежити відображення ідеї героїзму його національно-визвольної боротьби за незалежність та морально-ідейних цінностей людини.

Для досягнення зазначеної мети визначено такі *основні завдання дослідження*:

- визначити суть, поняття та роль програмних інструментів корпусного менеджера AntConc та вебсистеми Sketch Engine у дослідженнях корпусної лінгвістики, що дасть змогу дослідити механізм їх функціонування;
- проаналізувати поняття ключових слів, методів їх визначення та значення для лінгвістичних досліджень, що дасть можливість більш детально висвітлити особливості

ті практичного застосування корпусних механізмів у лексикологічних дослідженнях, які проводять у вітчизняному мовознавстві;

- дослідити стрілецькі та повстанські пісні як фольклорний жанр та його лінгвістичний контекст, що допоможе оцінити їх роль та значення у важливі історичні періоди для сучасного українського народу;
- використати методики аналізу частоти появи ключових слів в українських стрілецьких і повстанських піснях за допомогою програмних інструментів корпусного менеджера AntConc та вебсистеми Sketch Engine, що уможливить їхнє використання для аналізу та визначення ключових слів у текстах фольклорного жанру;
- проаналізувати ключові слова в українських стрілецьких і повстанських піснях, що забезпечить глибоке розуміння їх будови та суті;
- виявити та охарактеризувати основні тематичні та лінгвістичні ознаки українських стрілецьких і повстанських пісень, їхнє змістово-тематичне наповнення за допомогою аналізу частоти появи ключових слів.

Аналіз останніх досліджень та публікацій. Проведений нами аналіз відкритої періодики свідчить про постійний дослідницький інтерес багатьох науковців до проблематики фольклорного жанру та її висвітлення лінгвістами у історичній ретроспективі.

Вагомі сучасні наукові дослідження присвячені вивченню літературних жанрів, зокрема фольклорного жанру. Термін "фольклор" вперше використав 1846 року англійський археолог та письменник Вільям Томс, який винайшов цей термін як заміну термінології того часу "популярні старожитності" або "популярна література". Український учений-енциклопедист М. Максимович, який по праву вважається зачинателем української фольклористики, видавши у 1827 році свою книгу "Малоросійські пісні" та кілька збірників українських народних пісень. Видатний внесок у дослідження українського фольклору також зробили І. Франко та В. Гнатюк. Фольклорні жанри привертати увагу багатьох інших збирачів, видавців, *дослідників*, зокрема М. Костомарова, М. Драгоманова, П. Куліша, П. Чубинського, Б. Грінченка.

Фольклорний дискурс привертає увагу науковців своєю особливістю зображення картини світу, формою її характером вираження, адже зберігає особливу ціннісну інформацію, яка регламентується екзистенційними засадами соціуму. У фольклорному дискурсі описано певний алгоритм поведінки члена колективу у певних типових ситуаціях, алгоритм певних емоційних станів, пропонуються норми та стандарти статусно-рольової поведінки в естетично оброблених текстах. "Комунікант у фольклорному дискурсі є носієм певної соціальної ролі" [31, с. 11].

Повстанські пісні належать до фольклорного жанру, хоча і є значною мірою специфічними і розмістити їх у звичній схемі жанрово-тематичної класифікації не завжди можливо [32]. Українська повстанська пісня, за словами відомого дослідника національно-визвольного руху Григорія Дем'яна, – феноменальний духовний скарб нації, який послідовно і впевнено продовжує збагачуватися. Він окреслив основну масу цих пісень військово-повстанськими. До повстанських зараховуємо також і ті, що творяться, побутують на синхронному рівні та стосуються боротьби за відстоювання незалежності України від віковичного московського окупанта [10].

Сучасні розвідки жанру повстанських пісень зосереджені на застосуванні новітніх програмних інструментів для проведення лінгвістичних досліджень, особливо

у царині корпусної лінгвістики, яка охоплює всю різноманітність використання мови в усіх сферах мовної взаємодії, уможливує використання корпусів текстів у граматичних і лексикологічних дослідженнях. Визначення поняття корпусу висвітлюють такі провідні лінгвісти, як Р. Вільямс та М. Фінеган [14, 34]. Вони розглядають корпус як подане в машиночитаному форматі збирання текстів, відібраних відповідно до чітких мовних критеріїв. Українська мовознавиця О. Демська-Кульчицька визначає текстовий корпус як перетворену на електронну форму репрезентативну вибірку текстів природної мови, призначену для наукового та практичного її вивчення [9]. Особливостям укладання та використання корпусів текстів у лінгвістичних розвідках присвятили свої праці І. Кульчицький, О. Левченко, М. Касьяненко [19, 21, 22].

Серед популярних напрямів досліджень корпусної лінгвістики є частотні переліки і переліки ключових слів. Визначення теми та певних стилістичних властивостей тексту із застосуванням частотного аналізу ключових слів, основні корпусні методи їхнього пошуку привертають увагу сучасних дослідників [5, 19, 36].

Вважають [28], що поняття "ключовість" втілює якість, притаманну словам, групам слів і фразам, яка не залежить від мови, а радше від конкретного тексту чи текстів, у які вони внесені. По суті, ключові слова набувають значення в контексті всього тексту або корпусу текстів з урахуванням соціальних і культурних його особливостей. К. Габрієлатос, Дж. Кулпепер та Дж. Деммен у своїх працях [8, 15] обґрунтовують роль вивчення ключових слів поетичного тексту для систематизації та розкриття різноманітних особливостей мови, стилю та семантики поезії.

У корпусній лінгвістиці ключове слово має кількісну основу: це термін для слова, статистично характерного для тексту чи набору текстів. Ключове слово зазвичай вказує на значуще слово з заголовка або документа, яке використовують як показник вмісту. Однак, у корпусних лінгвістичних дослідженнях це поняття визначено як слово, "частота якого є надзвичайно високою порівняно з деякою нормою" [28, с. 135].

Ключові слова важливі для дослідження мови народу та його ідеології. Багато дослідників черпають натхнення з ідей ключових слів Реймонда Вільямса, які він визначає як терміни, які несуть соціокультурні значення [34, с. 6]. Ключові слова в дослідженні Вільямса були визначені на підставі суб'єктивного судження про соціокультурні значення попередньо визначеного переліку слів (наприклад, мистецтво, бюрократія, освіта, добробут, насильство та багато інших). Вони є заздалегідь визначеними зарезервованими словами, які використовують в програмуванні, які мають особливе значення для компілятора [34].

Ключове слово зазвичай вказує на значуще слово із заголовка або документа, яке використовують як показник його вмісту. Однак, у корпусних лінгвістичних дослідженнях це поняття визначено як слова, які є значно більш чисельні у певному корпусі порівняно з іншим корпусом [28]. Зокрема, хоча вони не можуть бути найважливішими словами в даному тексті або корпусі (головно тому, що їх важливість або ключовість визначена суто статистично), ключові слова часто є корисним способом охарактеризувати текст твору або його жанр і можуть бути використані для аналізу лексико-граматичних особливостей тексту в корпусі.

В останні десятиліття актуальним стало виділення ключових слів у тексті за допомогою статистичних методик. І. Ділай та М. Ділай у своїй праці [11] зазначають їх важливість для ідентифікації ключових концептів у дискурсах, для встановлення типової лексики жанру чи варіанта мови, чи її лексичного розвитку протягом часу.

Залежність від різниці лексичних частот у двох корпусах є характерною ознакою ключових слів [5]. Їх обчислюють за допомогою двох переліків слів, один з тексту, який потрібно дослідити, а інший із більшого довідкового корпусу, який діє як контрольний корпус або надає фонові дані для розрахунку ключових слів. Отже, статистична ключовість, виявлена в розподілі слів, пов'язана з їхньою відмінністю між цільовим і еталонним переліком, довідковим корпусом. Отже, ці статистичні ключові слова потрібно інтерпретувати разом із відмінними ознаками цільового та довідкового корпусу.

Виконаний аналіз наявних наукових праць свідчить про те, що деякі проблемні питання корпусної лінгвістики залишаються недостатньо вивченими, особливо використання корпусних інструментів для аналізу ключових слів. Враховуючи активне наукове обговорення ролі корпусних методик у сучасних лінгвістичних дослідженнях, виникає доцільність у більш детальному висвітленні особливості практичного застосування корпусних механізмів у лексикологічних дослідженнях, які проводять у вітчизняному мовознавстві.

Матеріали та методи дослідження. Для досягнення зазначених вище цілей застосована наступна методологічна парадигма: пошукове дослідження як початкове вивчення теоретичних особливостей корпусного менеджера AntConc і вебсистеми Sketch Engine та їх застосування в лінгвістичних дослідженнях; послідовна вибірка для систематизації та класифікації текстів українських стрілецьких і повстанських пісень для подальшого аналізу ключових слів; описове дослідження для розкриття провідних ідей та тематичних особливостей жанру стрілецьких і повстанських пісень; системний метод дослідження особливостей фольклорних творів; індуктивний метод як спосіб простежування окремих фактів і явищ для встановлення загальних правил і законів для формування підходу до встановлення істини, до з'ясування загальноновизнаного значення. Також послуговувалися такими методами як зіставний лінгвістичний аналіз, метод компонентного, семантичного та контекстуального аналізу, а також, трансформаційний та лінгвостатистичний методи.

Дослідження проведені на підставі корпусу текстів 35 українських стрілецьких і повстанських пісень, що налічує такі визначні твори, як "Ой у лузі червона калина", "За Україну", "Гей степами", "Ми йдемо вперед", "Йде січове військо", "Не сміє бути в нас страху", "Там, під Перемишлем" та інші.

Результати дослідження та їх обговорення / Research results and their discussion

У наш час створення корпусів і корпусно-базовані дослідження стали невід'ємною частиною лінгвістичних розвідок. Корпусна методологія стає елементом діяльності лінгвістів у різних областях. Вивчення ключових слів поетичного тексту з використанням корпусного підходу також може бути цікавим і продуктивним методом лінгвістичного аналізу. Наведемо деякі особ-

ливості корпусного підходу до визначення ключових слів [8, 15, 24]:

- *частотний аналіз* (англ. – *Frequency Analysis*) аналіз частоти вживання слів у тексті, допомагає виявити слова, які найчастіше трапляються і можуть бути ключовими для розуміння теми чи ідеї поезії;
- *аналіз поєднань слів* (англ. – *Word Combination Analysis*) вивчення поєднань слів або колокацій, може виявити ключові слова, які часто трапляються разом та утворюють основу для образів або метафор в поезії;
- *контекстуальний аналіз* (англ. – *Contextual Analysis*) розглядає контекст, в якому вживаються певні слова, для розуміння їхнього значення та ролі у тексті, що допомагає виявити ключові слова, які мають особливу семантичну вагу;
- *лексико-семантичний аналіз* (англ. – *Lexico-Semantic Analysis*) вивчає семантику слів і їхній контекстуальний зв'язок, що допомагає виявити слова, які важливі для розкриття теми, образів або символіки поезії;
- *семантичний аналіз* (англ. – *Semantic Analysis*) використання методів семантичного аналізу, дає можливість виявити слова, які мають близьке значення до ключових понять або концепцій у тексті;
- *асоціативний аналіз* (англ. – *Associative Analysis*) вивчає асоціації, які викликають певні слова у читачів, що можуть мати сильні асоціації, які відображають основні ідеї або емоції поезії;
- *синтаксичний аналіз* (англ. – *Syntactic Analysis*) досліджує синтаксичну структуру речень та взаємозв'язки між словами, базується на їхньому розміщенні у реченні та їхній синтаксичній ролі;
- *аналіз контексту* (англ. – *Contextual Analysis*) врахування контексту, у якому вживаються слова;
- *стилістичний аналіз* (англ. – *Stylistic Analysis*) аналіз стилістичних особливостей тексту, таких як метрика, ритміка, рима та інші.

Використання зазначених методів дає можливість систематично визначити ключові слова у поетичному тексті та розкрити їхню роль у передачі теми, настрою та ідей поезії.

Корпусний підхід до вивчення ключових слів поетичного тексту може допомогти систематизувати та розкрити різноманітні особливості мови, стилю та семантики поезії, що робить його цікавим інструментом для лінгвістичного аналізу [8].

Програмними інструментами корпусних досліджень є корпусний менеджер AntConc та вебсистема Sketch Engine. Саме ними послуговувалися під час здійснення нашого дослідження. Інструмент Sketch Engine є вебсистемою, яка слугує для вивчення того, як працює мова. За допомогою його алгоритмів аналізують корпуси автентичних текстів для миттєвого визначення типових чи рідкісних, незвичайних слововживань у мові. Ця система також розроблена для програм аналізу або розпізнавання тексту. Вебсистема Sketch Engine містить понад трильйон слів у 800 готових до використання корпусах на понад 100 мовах, кожен із яких має розмір до 80 мільярдів слів, щоб забезпечити справді репрезентативну вибірку мови. Це дає можливість лінгвістам, лексикографам, перекладачам досліджувати великі корпуси текстів і створювати складні запити. За допомогою нього можемо виявити контексти, у яких трапляється певне слово, або ж якими є ключові слова певного тексту чи корпусу. Характерною особливістю цього інструмента є поєднання статистичних даних із лінгвістичними критеріями для вилучення ключових слів і термінів. Зокрема, завдяки інструменту Concor-

dance можемо виявити розширений контекст слова чи терміна, як просто поєднання слів, так і цілі речення, в яких трапляється зазначене слово. Використовуючи вебсистему Sketch Engine, можна формувати переліки колокацій на необхідному лексичному рівні, а також виокремлювати перелік із зазначенням частоти кожної колокації у корпусі та значення зв'язку між ключовим словом і його колокацією [23].

Іншим важливим програмним інструментом у лінгвістичних дослідженнях є корпусний менеджер AntConc, який полегшує точний аналіз і ефективне дослідження лінгвістичних даних. Будучи започаткованим завдяки інноваційним зусиллям розробників Лоуренса Ентоні (англ. *Laurence Anthony*) та його команди, корпусний менеджер AntConc з'явився у 2002 році як необхідний інструмент для лінгвістичних досліджень, спочатку задуманий для задоволення педагогічних потреб студентів на курсі науково-технічного письма у Вищій школі інженерії Університету Осаки [2]. На тлі бурхливого розвитку корпусних досліджень мови корпусний менеджер AntConc швидко привернув увагу завдяки зручному інтерфейсу та потужним функціональним можливостям. Заснований на методології узгодження KWIC (англ. *Key Word in Context*), корпусний менеджер AntConc надає користувачам засоби для легкої навігації лінгвістичними корпусами, уможливаючи всебічний аналіз мовних моделей і їх вживання.

Корпусний менеджер AntConc заповнив важливу нішу у сфері корпусної лінгвістики завдяки своєму інтуїтивно зрозумілому графічному користувацькому інтерфейсу, оскільки він розроблений з використанням інструментарію PERL/TK і забезпечує безперебійну роботу в середовищах Windows, Linux і Unix, що вказує на його адаптивність і універсальність.

Корпусний менеджер AntConc AntConc посідає важливе місце в галузі мовного аналізу, зокрема завдяки своїй ролі в полегшенні дослідження текстових даних. Один з інструментів корпусного менеджера AntConc, *KeyWord List Tool* задовольняє фундаментальну вимогу дослідників оцінювати значущість слів шляхом порівняльного аналізу з еталонним корпусом. Для визначення міри ключовості слів (англ. *Keyness Measure*) використовуються статистичні показники, такі як χ^2 -квадрат (англ. *Chi-Squared*) або логарифмічну правдоподібність (англ. *Log-Likelihood* – є натуральним логарифмом правдоподібності), що надає дослідникам цінну інформацію про лінгвістичні явища, характерні для конкретного корпусу [20]. Зокрема, інструмент *KeyWord List Tool* дає можливість виконати елементарне реферування тексту, а саме – згенерувати набір ключових слів тексту або, при завданні переліку лем, – ключових лем тексту.

У нашому дослідженні було використано функції корпусного менеджера AntConc та вебсистеми Sketch Engine для аналізу 35 українських стрілецьких і повстанських пісень, які становлять 4203 токенів, 2997 слів і 245 речень. Кількість лем у цих піснях становить 893 (див. рис. 1).

Функція *Collocates* корпусного менеджера AntConc дає можливість зробити миттєвий статистичний аналіз синтаксично-семантичних єдностей, у яких один лексичний елемент поєднується з іншою лексичною одиницею. Тобто, було виконано аналіз тексту за словами, що знаходяться ліворуч та праворуч від досліджуваного елемента і формують стійке словосполучення. Отже,

встановлюємо колокації слів "військо" та "Україна" у стрілецьких і повстанських піснях (рис. 2 і 3).

Важлива функція *Word List* корпусного менеджера AntConc – інструмент підрахунку і подання всіх слововживань, які є у корпусі, як упорядкованого переліку. Це дає можливість швидко знайти слова, які найчастіше трапляються в досліджуваному корпусі. Інакше кажучи, це інструмент складання частотного словника конкретного корпусу (рис. 4).

Інша важлива функція *N-Grams* корпусного менеджера AntConc також пов'язана з пошуком у завантажених файлах груп слів (кластерів) заданої довжини. З її допомогою скануємо весь корпус на наявність класте-

рів довжини N (наприклад, 1 слово, 2 слова, ...). Це дає можливість знаходити загальні вирази в корпусі. На відміну від функції *Clusters*, тут йдеться не про пошук груп слів навколо ключового елемента, а про будь-які поєднання слів, що знаходяться поруч від них. Це дає можливість знаходити найпоширеніші у тексті словосполучення (рис. 5 та 6), наприклад для слів "військо" та "Україна" у стрілецьких і повстанських піснях. З цих рисунків також бачимо, що іменник "Україна" найчастіше поєднується із дієсловами ("зажурилася", "повстала", "здрігнеться"), тоді як іменник "військо" поєднується тільки з іменниками та дієсловами ("окопи", "пісня", "йде").



Рис. 1. Кількісний показник слів, tokenів і лем у текстах досліджуваних пісень у функціях вебсистеми Sketch Engine / Quantitative indicator of words, tokens and lemmas in the lyrics of the studied songs in the functions of the Sketch Engine web system

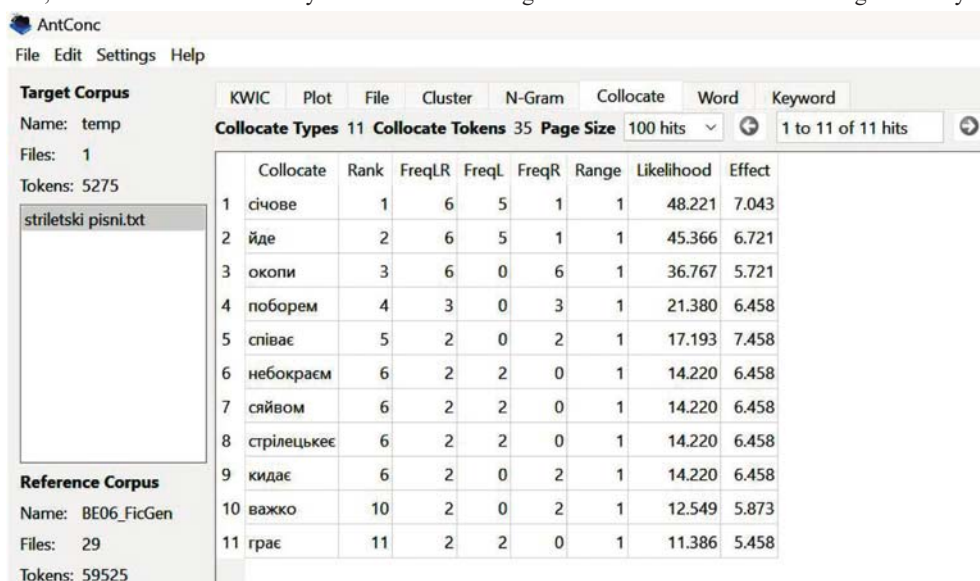


Рис. 2. Визначення колокацій до слова "військо" у стрілецьких і повстанських піснях за допомогою функції *Collocates* корпусного менеджера AntConc / Identifying collocations to the word ВІЙСЬКО/ ARMY in rifle and rebel songs using the *Collocates* function of the AntConc corpus manager



Рис. 3. Визначення колокацій до слова "Україна" у стрілецьких і повстанських піснях за допомогою функції *Collocates* корпусного менеджера AntConc / Identifying collocations to the word УКРАЇНА / UKRAINE in rifle and rebel songs using the *Collocates* function of the AntConc corpus manager

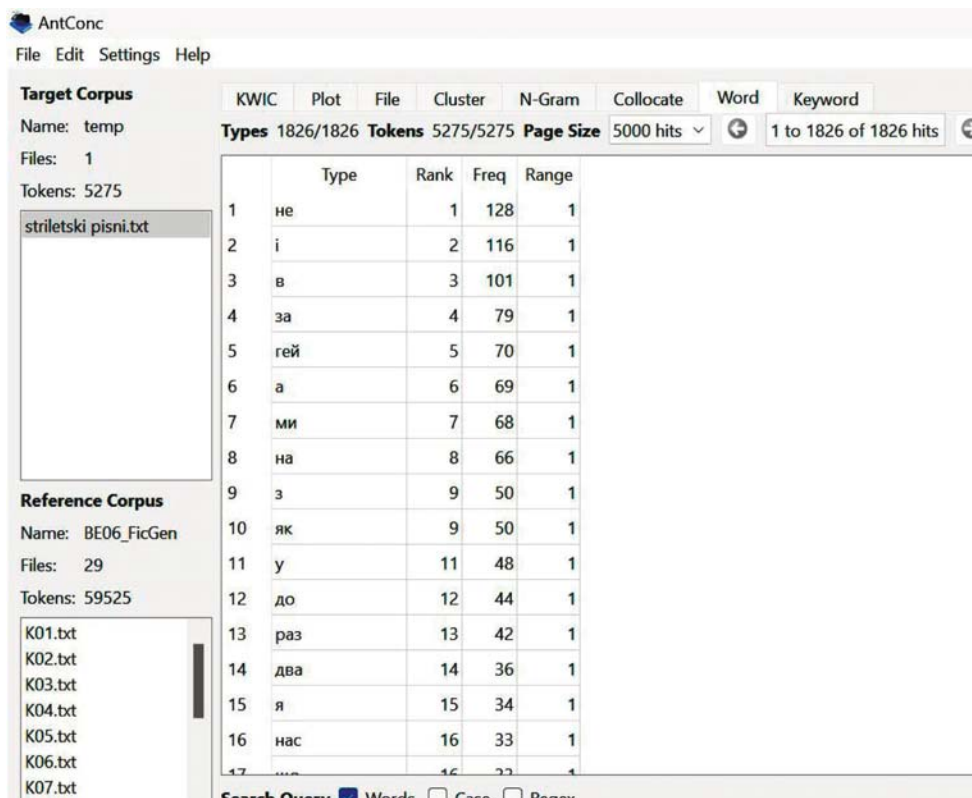


Рис. 4. Визначення переліку слів з їх частотою у стрілецьких і повстанських піснях за допомогою функції *Word List* корпусного менеджера AntConc / Finding word frequencies of list of words in rifle and rebel songs using the *Word List* tool function of the AntConc corpus manager



Рис. 5. Визначення словосполучень для слова "військо" у стрілецьких і повстанських піснях за допомогою функції *N-Grams* корпусного менеджера AntConc / Identifying collocations to the word ВІЙСЬКО/ ARMY in rifle and rebel songs using the *N-Grams* tool function of the AntConc corpus manager



Рис. 6. Визначення словосполучень для слова "Україна" у стрілецьких і повстанських піснях за допомогою функції *N-Grams* корпусного менеджера AntConc / Identifying collocations to the word УКРАЇНА / UKRAINE in rifle and rebel songs using the *N-Grams* tool function of the AntConc corpus manager

Отже, варто зробити висновки, що корпусний менеджер AntConc дає можливість легко та швидко проводити корпусні дослідження, адже цей програмний інструмент надзвичайно простий та зрозумілий у використанні. Для того, щоб почати працювати з корпусним менеджером AntConc, не потрібне тривале вивчення інтерфейсу програми, формул запитів, тощо. Ця програма

дає можливість працювати з будь-яким текстовим файлом формату *.txt, а не тільки з анотованими текстами, внесеними в традиційні корпуси. Користувач сам укладає корпус для аналізу за допомогою корпусного менеджера AntConc будь-якого обсягу, будь-якою мовою і будь-якої кількості файлів, попередньо зібраних у форматі *.txt в одну спільну папку. Це надзвичайно зручно,

оскільки дає можливість виконати статистичний аналіз практично будь-якого текстового матеріалу у цифровому форматі – текстовому файлі. Якщо ж досліджуваний текст подано на папері, то його потрібно спочатку оцифрувати – відсканувати і розпізнати.

Понад це, сервіс корпусного менеджера AntConc дає можливість аналізувати кілька файлів або повністю, як єдиний корпус, або паралельно, як окремі файли. Це уможливило здійснювати порівняльний аналіз кількісних показників у різних текстах.

Оскільки реферативний корпус української мови недоступний у корпусному менеджері AntConc, тому для нашого дослідження було використано онлайн вебсистему Sketch Engine (рис. 7).

Ключові слова – це слова (елементи з однією лексемою), які трапляються частіше в основному корпусі, ніж у довідковому. Їх можна використовувати для того, щоб визначити, що характерно для одного створеного корпусу (фокус-корпусу) або його підкорпусу в порівнянні з іншим корпусом з бази вебсистеми Sketch Engine

(референтним корпусом) або його підкорпусом. Таке порівняння можна також проводити між двома підкорпусами одного корпусу або між усім корпусом та одним із його підкорпусів.

Ключові слова можна виявити за допомогою функції *Keywords* вебсистеми Sketch Engine. Як правило, за довідковий корпус вибирають найбільший корпус мови. Користувач може встановити інший корпус або підкорпус як довідковий корпус/підкорпус.

Результат поділено на ключові слова (одиничні слова) та терміни (багатослівні елементи) і відображаються разом із посиланнями на речення як у фокус-корпусі, так і в референтному корпусі.

З допомогою функції *Keywords* вебсистеми Sketch Engine було видобуто ключові слова та термінологічні словосполучення у корпусі. На рис. 7 подано ключові слова корпусу, а на рис. 8 – термінологічні словосполучення цих ключових слів нашого корпусу. Наприклад, "Гей" має асоціації з "Гейта" та "Гейтиська", "остріг" має асоціації з "січовий", "стрілець" та "окоп", тощо.

Lemma	Lemma	Lemma	Lemma	Lemma	Lemma
1 вйю	24 остріг	47 кар	70 рідня	93 манджати	116 ненька
2 малин	25 усусус	48 шелестіти	71 біліти	94 галичаночка	117 взимі
3 рибчина	26 вороженько	49 журитися	72 панянонець	95 проводу	118 топір
4 дівчино	27 вкраїна	50 позір	73 ку_у_уль	96 вірненько	119 застрілитися
5 чорненький	28 приставати	51 здалека	74 млі-і-вати	97 рум'яненький	120 пірвати
6 вісьта	29 окоп	52 грізний	75 гине-гиня	98 присти	121 циганочка
7 гейта	30 січовий	53 страшити	76 нивонка	99 сумненький	122 чом
8 уха-ха	31 брова	54 небокрай	77 мілійони	100 гураган	123 всесильний
9 кладочка	32 розвеселити	55 обложити	78 лицарями	101 буйнесенький	124 кайдан
10 гей	33 перемишель	56 молоденький	79 м'ятою-рутий	102 земелька	125 забитися
11 оченята	34 перемишльома	57 братчик	80 зтеп	103 надніпрянщина	126 рейти
12 узесенький	35 хехати	58 стрілець	81 горонка	104 супротивитися	127 чепурненький
13 стависько	36 творче	59 хитатися	82 гад	105 хлопець-молодка	128 обкрутити

Рис. 7. Визначення ключових слів у корпусі українських стрілецьких і повстанських пісень за допомогою функції *Keywords* вебсистеми Sketch Engine / Identifying key words in the corpus of Ukrainian rifle and rebel songs using the *Keywords* function of the Sketch Engine web system

Term	Term	Term	Term
1 уста малин	18 люба дівчино	35 перемишель голови	52 ціла зима
2 кар оченята	19 дівчина серце	36 вражий німець	53 чорний брова
3 чорненький брова	20 човен по воді	37 хвиля весло	54 місячне сяйво
4 славне товариство	21 стрільці січовії	38 пане сотник	55 далека дорогий
5 залізний остроги	22 старий дуб	39 твердий хлопець	56 червона калина
6 грізний залізний	23 оченята чорний	40 яра пшеничка	57 братчик рани
7 грізний залізний остроги	24 деннеє зло	41 сонце сяйвом	58 рана себе
8 лицар грізний	25 дубочок січовик	42 стрілецький звичай	59 дівчата сльози

Рис. 8. Визначення термінологічних словосполучень у корпусі українських стрілецьких і повстанських пісень за допомогою функції *Keywords* вебсистеми Sketch Engine / Identifying multiword terms in the corpus of Ukrainian rifle and rebel songs using the *Keywords* function of the Sketch Engine web system

Остріг	Остріг Січовий Стрілець Окоп Фортеця Військо Оборона Бойовий Восний Укріплення
Славний	Славний Похвальний Відомий Відзначений Чудовий Помітний Героїчний Чесний Величний

Рис. 9. Асоціативні мережі українських стрілецьких і повстанських пісень / Word association networks of Ukrainian rebel songs

Внаслідок верифікації ключових слів з текстів стрілецьких і повстанських пісень, було виявлено частоту вживання вигуків і сполучників. Вони мають важливе значення у структурі пісень, адже дають можливість підсилити зв'язок між оспіваними ідеями та образами. Найбільш частотними виявилися вигуки "ой", "гей", "уха-ха", які коротко виражають почуття та емоції. Часте вживання часток і сполучників "аж", "то", "би", "і", "за", "на", "у", "в", "до", які не мають семантико-граматичних ознак слова, може свідчити про суб'єктно-об'єктні, причинно-наслідкові зв'язки між предметами і явищами, описаними у піснях. Ці частки й сполучники можуть вказувати на нюанси міркувань, контрастів у мисленні, а також на обґрунтування або відхилення від певних ідей чи поглядів, що відображаються в піснях.

Серед ключових слів виділяються іменники, що безпосередньо пов'язані з родинними зв'язками, такі як "рідня", "братчик", "сину", "ненька", "вдовиця", і які створюють атмосферу єдності та спорідненості. Проникливі особисті почуття людини ілюструють такі ключові іменники, як "дівчина", "рибчина", "оченята", "нічка". Приналежність до військового та козацького осередку та військові будні стрільців відображають наступні ключові іменники: "лицар", "січовик", "остріг", "вороженьки", "стрілець", "окоп", "сотник", "вороженько", "позір". Патріотичні почуття змальовує ніжне звернення "земленька", "нивонька". Потрібно зазначити, що характерною ознакою досліджуваних пісень є часте вживання назв місць рідного краю, зокрема "Перемишль", "Галичина", "Наддніпрянина", а також релігійних об'єктів і понять, як от "славень", "боже", "пом'янути". Вони допомагають створити атмосферу та контекст для подальшого розгортання подій.

Емоційний стан висвітлюється через призму ліричних мотивів, особливо мотиви туги, смутку прощання передаються у дієсловах та прикметниках, а саме "зажуритися", "полюбити", "пом'янути", "погибати", "в'янути", "жертвуючи", "сумнений". Простежуємо також застосування художнього засобу рефрену (повтору) у дієсловах, – "гине-гине", звертає особливу увагу на певні дії та вказує на мелодійність художнього твору. Варто зазначити також доволі велику кількість прикметників у ключових словах стрілецьких і повстанських пісень. Образ стрільця змальовується прик-

метниками "січовий", "молоденький", "рум'яненко", "сумнений", а також прислівником "вірньо".

Художні засоби, притаманні стрілецьким і повстанським пісням, такі як вигуки, повтори, метафора, анафора, пестливо-зменшувальні слова, є характерними ознаками української соціально-побутової народної пісні. Вживання пестливо-зменшувальних суфіксів у піснях цього жанру вказує на ніжне ставлення до описаних об'єктів, як от дівчини, матері, батьківщини. Особисті почуття виспівуються зворушливими у своїй простоті та правдивості словами.

На підставі вилучених ключових слів можна побудувати асоціативні мережі українських стрілецьких і повстанських пісень (рис. 9), позаяк вони дають більш повний результат аналізу асоціацій для кожного ключового слова, відображаючи різноманітні особливості та контексти, пов'язані з кожним із них.

Аналіз ключових слів в українських стрілецьких і повстанських піснях виявив ряд загальних тематичних і стилістичних особливостей, що відображають глибину та різноманітність цього жанру. Він підтверджує важливість відображення суспільно-історичних подій, героїчної боротьби народу, внутрішніх переживань та саможертвості у художньому вираженні у піснях.

Обговорення результатів дослідження. Корпусна лінгвістика – галузь мовознавства, предметом дослідження якої є принципи та методи формування корпусів текстів, а також розроблення комп'ютерних систем для їхнього опрацювання. У лінгвістиці корпус – підібрана й оброблена за певними правилами сукупність текстів, які використовують як базу для дослідження мови. Їх застосовують для статистичного аналізу і перевірки статистичних гіпотез, підтвердження лінгвістичних правил цієї мови. Загалом корпус автори розглядають як модель мовної системи, реалізовану в текстах різних функціональних можливостей.

В роботі [35] наведено результати використання інтегрованого корпусного інструменту Wordless із багатомовною підтримкою для вивчення мови, літератури та перекладу. Автор вважає, що це безкоштовна кросплатформна комп'ютерна програма з відкритим вихідним кодом зі зручним графічним інтерфейсом, який спеціально розроблено для задоволення потреб нетехнічних користувачів. Його кінцевою метою є усунення всіх непотрібних технологічних бар'єрів для використання передових технологій дослідниками в галузі корпусних досліджень. Наразі існує тільки невелика кількість науковців, які присвятили себе проблемі розроблення спеціалізованих інструментів для аналізу корпусу, серед яких найбільш поширені комп'ютерні програми WordSmith Tools [29], AntConc [3] і #LancsBox X [6] для запитів одномовних корпусів і ParaConc [4] для паралельного узгодження. Безсумнівно, ці інструменти надзвичайно полегшили процес оброблення та аналізу корпусів за останні 20 років, і велика кількість досліджень на підставі корпусів була б неможливою без новаторського внеску, зробленого цими інструментами та їх розробниками. Тим не менш, враховуючи численні досягнення та значні прориви, досягнуті після нового тисячоліття в багатьох галузях інформатики, особливо під час оброблення природної мови NLP (англ. *Natural Language Processing*) і вибух штучного інтелекту AI (англ. *Artificial Intelligence*) з 2010-х років і далі, що призводить до експоненційного зростання машинного навчання та алго-

ритмів глибокого навчання, доводиться з жалем визнати, що розроблення корпусних інструментів не встигає за темпами швидкої ітерації комп'ютерних технологій.

В роботі [27] наведено результати застосування анотованого корпусу Kāvi для пенджабі поезії з виявленням емоцій на підставі "Навраси". Автори вважають, що оскільки вірші написані з численними емоціями та почуттями, то комп'ютерний лінгвістичний аналіз для виявлення емоцій у віршах є дуже складним завданням. Основна мета їхньої роботи полягала в тому, щоб виявити емоції в поезії пенджабі на підставі різних особливостей, присутніх у поезії пенджабі. Це новаційний підхід до виявлення емоцій у віршах пенджабі. Пенджабський поетичний корпус "Kāvi" був анотований вручну на підставі індійської концепції "Navrasa". У цьому корпусі 948 поезій було класифіковано за 9 станами емоцій, як представлено в "Наврасі", а саме "каруна", "шрінгар", "хася", "раудра", "вір", "бхаянак", "вібхата", "адбхут" і "шаанті". Цей корпус був анотований вручну, а показник Каппа Флейса (англ. *Kappa Fleiss*) був розрахований для узгодження між анотаторами. Різні функції (лінгвістичні, поетичні та статистичні) використовувалися для побудови класифікатора, і з цими функціями експериментували з двома методами машинного навчання, Наївний Байес (англ. *Naive Bayes*, NB) і метод опорних векторів SVM (англ. *Support Vector Machine*). Було виявлено, що з точністю 70,02 % метод SVM покращив загальну точність завдання класифікації емоцій. Окрім цього, було виявлено, що поетичні особливості перевершують лінгвістичні особливості для завдання виявлення емоцій.

В роботі [7] наведено результати застосування розширеного методу TF-IDF для покращення вилучення ключових слів у традиційних дослідженнях на підставі корпусу, насамперед корпусу змін клімату. Автор стверджує, що вилучення ключових слів передбачає застосування алгоритмів оброблення природної мови NLP (англ. *Natural Language Processing*) або моделей, розроблених у сфері аналізу тексту. Вилучення ключових слів є загальноприйнятою технікою, яку використовують для дослідження лінгвістичних шаблонів у лінгвістичному полі корпусу, а тест логарифмічної правдоподібності Даннінга (англ. *Dunning's Log-Likelihood Test*, LLT) давно інтегрований у програмне забезпечення корпусу як модель NLP на підставі статистики. Хоча попередні дослідження підтвердили широку застосовність вилучення ключових слів у дослідженнях на підставі корпусів, LLT має певні обмеження, які можуть вплинути на точність вилучення ключових слів у таких дослідженнях. У своєму дослідженні автори узагальнили обмеження LLT, які містять втручання порівняльного корпусу, усунення граматичних і загальних слів, розгляд підкорпусної релевантності, гнучкість у виборі ознак і адаптованість до різних цілей дослідження. Щоб усунути ці обмеження, у автори запропонували розширений метод термінової частоти – інверсної частоти документа TF-IDF (англ. *Term Frequency-Inverse Document Frequency*). Для перевірки застосовності запропонованого методу як цільового корпусу було використано 20 високоцитованих дослідницьких статей про зміну клімату з бази даних WOS (англ. *Web of Science*) і проведено порівняння з традиційним методом. Експериментальні результати показали, що запропонований метод може ефективно подолати обмеження традиційного ме-

тоду та продемонстрували доцільність і практичність внесення алгоритму TF-IDF у відповідне дослідження на підставі корпусу.

В роботі [25] наведено результати використання електронних корпусів для теоретичної лінгвістики, передусім тематичні дослідження з синтаксису німецької мови. Автор вважає, що корпусна лінгвістика вимагає прикладів речень і як емпіричної основи для розвитку теорій, і як контрприкладів до попередніх узагальнень. На додаток до отримання таких прикладів шляхом інтроспекції, електронні корпуси можна використовувати для пошуку прикладів, які мають відношення до конкретної теоретичної проблеми. Цей другий варіант дослідники рідко використовують в генеративній лінгвістиці, можливо через те, що не повністю усвідомлено, що таке використання корпусів у принципі не залежить від фундаментальних методологічних питань, що розділяють емпіриків і раціоналістів. Дослідження автора ілюструє приклади з синтаксису німецької мови, як пошук у корпусах може допомогти знайти теоретично відповідні приклади. Такі приклади особливо цікаві тим, що вони демонструють широку варіацію потенційно релевантних параметрів. Тематичні дослідження вказують на те, як лінгвістичну термінологію, використану для виділення відповідного явища, можна реконструювати з точки зору емпіричних властивостей, які доступні безпосередньо або через анотації в корпусі.

В роботі [1] наведено результати застосування корпусу FASSILA для виявлення фейкових новин на алжирському діалекті та аналізу настроїв. Автори стверджують, що у контексті мов з низьким ресурсом алжирський діалект AD (англ. *Algerian Dialect*) стикається з проблемами через відсутність анотованих корпусів, що перешкоджає його ефективній обробці, зокрема в програмах машинного навчання ML (англ. *Machine Learning*), які покладаються на корпуси для навчання та оцінювання. У своєму дослідженні автори описують процес розроблення спеціалізованого корпусу для виявлення фейкових новин FN (англ. *Fake News*) і аналізу настроїв SA (англ. *Sentiment Analysis*) в AD під назвою FASSILA. Цей корпус складається з 10 087 речень, що охоплюють понад 19 497 унікальних слів AD, розглядає істотну нестачу лінгвістичних ресурсів мови та охоплює сім різних областей. Дослідники пропонують схему виявлення FN і анотації SA, що детально описує збір, очищення та маркування даних. Чудова угода між анотаторами вказує на те, що схема анотації створює високоякісні та послідовні анотації. Подальші експерименти з класифікації з використанням моделей на підставі BERT і ML представлені, демонструючи багатообіцяючі результати та висвітлюючи шляхи для подальших досліджень. Наразі набір даних є у вільному доступі, щоб полегшити майбутній прогрес у цій галузі.

В роботі [30] наведено результати автоматичного виявлення з корпусу "Оксюморон". Оксюморон (англ. *Oxymoron*) – мовне явище, у якому пара протилежних або антонімічних слів поєднується для передачі нового значення. Іноді його використовують для вираження образності, іронії чи риторики в тексті. Цьому питанню приділялося відносно менше уваги в сферах лінгвістики та обчислювальних дисциплін. Оксюмори відіграють значну роль у різних програмах оброблення мови. Автори є першопрохідцями у дослідженні оксюморонів у бенгальській мові. Корпусне дослідження оксюморону

є фундаментальним питанням, яке досі не досліджувалося. Запропоновано систему автоматизованого розпізнавання оксюморонів із заданого корпусу. Частотний аналіз, семантична подібність і словник антонімів були використані для розпізнавання оксюморонів у корпусі. Система досягла багатообіцяючих результатів під час тестування на бенгальському корпусі та виявила 308 різних оксюморонів. Описова статистика на підставі корпусу вимірювалася у двох різних корпусах. Найпоширеніші оксюморони ранжуються за їх частотою. Їх помітна присутність підкреслює важливість бенгальської мови. Їхнє дослідження мало на меті дослідити фундаментальні питання, що стосуються оксюморонів, таких як автоматичне виявлення оксюморонів у корпусі, описова статистика щодо оксюморонів у різних мовах, а також процес їх створення та створення. Окрім цього, було зроблено спроби витягти оксюморони з великих мовних моделей за допомогою нульових підказок, але результати не були такими багатообіцяючими порівняно із запропонованою авторами системою.

В роботі [33] наведено результати корпусного аналізу газетного дискурсу щодо угоди AUKUS. Автори вважають, що угода AUKUS (аббревіатура, розшифровується як "Australia, the United Kingdom, and the United States") – стратегічний договір між Австралією, Сполученим Королівством і Сполученими Штатами, передусім була спрямована на сприяння придбання Австралією восьми атомних підводних човнів у США та Великобританії. Ця угода призвела до раптового розірвання попереднього контракту з французькою державною Naval Group. У своєму дослідженні автори розглядають мову, яку використовували під час висвітлення угоди AUKUS у газетах різних англomовних та азіатських країн. Використовуючи комбінацію аналізу настроїв (Crossley та ін., 2017) і дослідження дискурсу за допомогою корпусу (Partington, 2013; Gillings та ін., 2023), автори зосередили увагу на виявленні ключових мовних моделей, тем і настроїв, вбудованих у дискурс. Їхні висновки вказують на загальну позитивну оцінку AUKUS в англomовних ЗМІ на відміну від негативних зображень у китайських публікаціях. Окрім цього, аналіз мовних компонент, таких як прикметники, іменники та дієслова, виявляє глибинні складності та суперечливі точки зору в самому англomовному дискурсі. Застосовуючи дослідження дискурсу за допомогою корпусу CADs (англ. *Corpus-Assisted Discourse Studies*), автори виявили контекстуальні та лінгвістичні фактори, які формують ці різноманітні точки зору.

В роботі [26] наведено результати аналізу настроїв спеціального корпусу мовлення, де підтверджено концепцію NLP (англ. *Neuro-Linguistic Programming*). Автори продемонстрували інноваційний спеціальний корпус мовлення гінді, присвячений оцінюванню позитивної інтенсивності їхньої мови. Завдяки ретельному збору зразків голосу людей із різних демографічних груп цей корпус добре розроблений для акустичних досліджень у аналізі настроїв. Від учасників було отримано кілька речень, що охоплюють спектр настроїв (надзвичайно позитивні, позитивні, нейтральні, негативні). Отриманий корпус складається з 225 аудіофайлів. Щоб гарантувати послідовність і врахувати нюанси, засновані на знаннях предметної області, настрої та їх інтенсивність (висока, середня, низька) були класифіковані за експертною анотацією. Автори використовували ком-

плексне програмне забезпечення PowerBI від Microsoft, щоб отримати уявлення про емоційний ландшафт корпусу. Використовуючи дослідницький аналіз і аналіз настроїв, корпус було проаналізовано, щоб виявити лінгвістичні особливості, що вказують на позитивну інтенсивність у мові гінді. Їхнє дослідження покращує розуміння вираження позитиву мовою гінді та надає цінні ресурси для лінгвістичних досліджень у межах мови гінді. Помічено, що позитивні настрої мають найвищу кількість – 58, а найменшу – дуже негативну – 27. Основний внесок їхньої роботи полягає у створенні спеціального набору даних мовою гінді, який, наскільки нам відомо, є першим у своєму роді. Окрім цього, автори не натрапили на жоден набір даних, зосереджений на інтенсивності позитивності. Ці внески, безумовно, будуть корисними для майбутніх дослідників у дослідженні позитивних впливів.

Загалом отримані нами результати дослідження підтверджують висновки інших науковців щодо ефективності використання корпусних інструментів для виявлення ключових слів у стрілецьких і повстанських піснях як жанру фольклорного дискурсу. Однак, важливою перевагою нашого підходу є збалансованість між точністю та швидкістю, яку забезпечують такі програмні інструменти, як корпусний менеджер AntConc та вебсистема Sketch Engine, що дало змогу більш детально висвітлити особливості практичного застосування корпусних механізмів у лексикологічних дослідженнях, які проводять у вітчизняному мовознавстві.

Отже, внаслідок виконаної роботи можна сформулювати такі наукову новизну та практичну значущість результатів дослідження.

Наукова новизна отриманих результатів дослідження – набула подальшого розвитку методика застосування таких програмних інструментів, як корпусний менеджер AntConc та вебсистема Sketch Engine, яка на відміну від наявних підходів, дає змогу більш детально висвітлити особливості практичного застосування корпусних механізмів у лексикологічних дослідженнях, які проводять у вітчизняному мовознавстві, забезпечуючи тим самим досягнення більш ґрунтовних і переконливих результатів.

Практична значущість результатів дослідження – удосконалену методику дослідження можна використати для аналізу соціальної складової пісень, притаманної жанру фольклору, яскраво відображену у стрілецьких і повстанських піснях як потужний ресурс утвердження патріотизму і символу нескореності українського народу, що дасть можливість наступним дослідникам простежити ідею героїзму його національно-визвольної боротьби за незалежність та, водночас, морально-ідейні цінності людини.

Висновки / Conclusion

Проаналізовано та охарактеризовано ключові слова у корпусі текстів стрілецьких і повстанських пісень як потужний ресурс в утвердженні патріотизму і символу нескореності українського народу, що дало змогу простежити відображення ідеї героїзму його національно-визвольної боротьби за незалежність та морально-ідейні цінності людини. За результатами проведеного дослідження можна зробити такі основні висновки.

1. З'ясовано, що створення та використання корпусів текстів є основою для сучасних лінгвістичних досліджень.

Застосування корпусних інструментів дає можливість миттєво обробити величезний обсяг мовного матеріалу і відібрати всі можливі у конкретному корпусі приклади вживання елементів, необхідних для відповідного аналізу. У розпорядження лінгвіста надаються об'єктивні кількісні дані, забезпечуючи досягнення більш ґрунтовних і переконливих результатів.

2. Встановлено, що використання електронних текстових корпусів може підвищити об'єктивність проведеного лінгвістичного дослідження за рахунок розширення його емпіричної бази, що дає можливість прийти до більш точних висновків, а також дає можливість більш детально висвітлити особливості практичного застосування корпусних механізмів у лексикологічних дослідженнях, які проводять у вітчизняному мовознавстві.
3. Доведено, що такі програмні інструменти, як корпусний менеджер AntConc та вебсистема Sketch Engine, стають незамінними помічниками у корпусних лінгвістичних дослідженнях, полегшуючи деякі особливості вивчення та проведення всебічного аналізу лінгвістичних даних. Застосування цих інструментів у сфері оброблення текстової інформації у корпусній лінгвістиці є важливим для аналізу великої кількості даних і вирішення різних завдань.
4. Виявлено, що важливою особливістю вебсистеми Sketch Engine є можливість пошуку ключових слів і термінологічних поєднань у корпусах за допомогою функції *Keywords*, що дає можливість дослідникам проводити власні розвідки. Водночас, корпусний менеджер AntConc надає можливість автоматизувати пошук кластерів, тобто знаходити загальні вирази в корпусі за допомогою функції *N-Grams*. За допомогою такої функції як *Word List* можна формувати переліки частотності вживання слів, а функція *Collocates* допомагає встановити та проаналізувати словосполучення з інтенсифікаторами.
5. Здійснено пошук і підрахунок ключових слів у стрілецьких і повстанських піснях, проаналізовано частотність та контекст вживання словоформ і словосполучень у них. Здійснений аналіз виявив, що соціальна складова, притаманна жанру фольклору, яскраво відображена у досліджуваних піснях. Українська ідентичність простежується у ключових словах та контексті стрілецьких і повстанських пісень. Однак, потрібно визнати, що поєднання корпусного менеджера AntConc та вебсистеми Sketch Engine у нашому дослідженні надало змогу всебічного вивчення мовних одиниць, допомогло розширити розуміння мови та текстів стрілецьких і повстанських пісень і, як наслідок, сприяло збагаченню лінгвістичних знань.

References

1. Abdedaïem, Amin, Dahou, Abdelhalim Hafedh, Cheragui, Mohamed Amine, & Mathiak, Brigitte. (2024). FASSILA: A Corpus for Algerian Dialect Fake News Detection and Sentiment Analysis. *Procedia Computer Science*, Vol. 244, 397–407. <https://doi.org/10.1016/j.procs.2024.10.214>
2. Anthony, L. (2011). AntConc: A Learner and Classroom Friendly, Multi-Platform Corpus Analysis Toolkit. URL: https://www.researchgate.net/publication/33010037_AntConc_A_Learner_and_Classroom_Friendly_Multi-Platform_Corpus_Analysis_Toolkit#fullTextFileContent
3. Anthony, L. (2024). AntConc (Version 4.2.4) [software], Waseda University. URL: <https://www.laurenceanthony.net/software/antconc/>
4. Barlow, M. (2024). ParaConc (Version 1.0) [software], Athelstan. URL: <https://www.paraconc.com/>
5. Brezina, V. (2018). Statistics in Corpus Linguistics: A Practical Guide. Cambridge: Cambridge University Press. <https://doi.org/10.1080/09296174.2019.1646069>
6. Brezina, V., & Platt W. (2024). #LancsBox x [software], Lancaster University. URL: <http://lancsbox.lancs.ac.uk>
7. Chen, Liang-Ching. (2024, September). An extended TF-IDF method for improving keyword extraction in traditional corpus-based research: An example of a climate change corpus. *Data & Knowledge Engineering*, Vol. 153, article ID 102322. <https://doi.org/10.1016/j.datak.2024.102322>
8. Culpeper, J., & Demmen, J. (2015). Keywords. In: D. Biber & R. Reppen (Eds.). *The Cambridge Handbook of English Corpus Linguistics* (Cambridge Handbooks in Language and Linguistics, pp. 90–105). Cambridge: Cambridge University Press. <https://doi.org/10.1007/9781139764377.006>
9. Demska-Kulchitska, O. (2002). What is new in the science of language? The culture of the word, Vol. 61, 70–74. [In Ukrainian]. URL: <https://ekmair.ukma.edu.ua/items/4f12fb83-3589-44e9-b397-fc766e1a115>
10. Demyan, H. (1992). Ukrainian songs of 1940–2000 years (historical and folklore research), Vol. 3, 10–42. [In Ukrainian]. URL: https://shron1.chtyvo.org.ua/Ukraina_v_mynulomu/Vypusk_03.pdf?PHPSESSID=gmlcntace2ee6js51ffhaolm4
11. Dilai, I. P., & Dilai, M. P. (2022). The role of frequency in linguistic research. *Transcarpathian Philological Studies*, 21(1), 146–151. <https://doi.org/10.32782/tps2663-4880/2022.21.1.27>
12. Dyak, T. P., Hrytsiuk, Y. I., & Horvat, P. P. (2022). The problem of fake news detection on Internet websites. *Scientific Bulletin of UNFU*, 32(6), 78–94. <https://doi.org/10.36930/40320612>
13. Fenii, N. S., & Hrytsiuk, Y. I. (2020). Automation of the process of classification of text news from internet sites by neural network methods. *Scientific Bulletin of UNFU*, 30(4), 123–133. <https://doi.org/10.36930/40300421>
14. Finegan, E. (2015). Language: its structure and use. Stamford, CT: Cengage Learning, 611 p. URL: https://www.academia.edu/44833419/Language_Its_Structure_and_Use
15. Gabrielatos, C. (2018). Keyness analysis: nature, metrics and techniques. In: Taylor, C., & Marchi, A. (Eds.). *Corpus Approaches to Discourse: A critical review*. Oxford: Routledge. 225–228. URL: https://www.academia.edu/34302240/Gabrielatos_C_2018_Keyness_Analysis_nature_metrics_and_techniques_In_Taylor_C_and_Marchi_A_eds_Corpus_Approaches_to_Discourse_A_critical_review_Oxford_Routledge_225_258?email_work_card=view-paper
16. Holoshchuk, S. L. (2022). Corpus linguistics: current state and research prospects. *Transcarpathian Philological Studies*. Uzhhorod: Publishing House "Helvetica", Vol. 1, issue 21, 249–252. [In Ukrainian]. URL: http://zfs-journal.uzhnu.u.edu.ua/archive/21/part_1/47.pdf
17. Hroch, M. (2000). Social preconditions of national revival in Europe: a comparative analysis of the social composition of patriotic groups among the smaller European nations. Columbia University Press, 220 p. URL: <https://pdfarchived.net/list/social-preconditions-of-national-revival-in-europe-miroslav-hroch-4900260>
18. Hrytsiuk, Y. I., & Dyak, T. P. (2021). The use of internet technologies in educational process in higher education institutions. *Scientific Bulletin of UNFU*, 31(1), 137–146. <https://doi.org/10.36930/40310123>
19. Kasyanenko, M., Lebedev, K. M., & Petrenko, P. M. (2009). Principles of building a corpus of texts of different functional styles. *Scientific Bulletin of Lesya Ukrainka Volyn National University*, 6, 25–28. [In Ukrainian]
20. Kilgariff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, 6(1), 97–133. <https://doi.org/10.1075/ijcl.6.1.05kil>
21. Kulchytsky, I. M. (2015). Technological aspects of compiling text corpora. *Text corpora data in linguistic research*. Monograph. Edited by O. P. Levchenko. Lviv: Lviv Polytechnic Publishing House, 29–45. [In Ukrainian]
22. Kunanets, N., Levchenko, O., & Hadzalo, A. (2018). The Application of AntConc Concordancer in Linguistic Researches, 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), Lviv, Ukraine, 144–147. <https://doi.org/10.1109/STC-CSIT.2018.8526591>
23. Learn how language works. (2024). Sketch Engine. URL: <https://www.sketchengine.eu/#blue>

24. McEnery, T., & Hardie, A. (2011). *Corpus Linguistics : Method, Theory and Practice*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511981395>
25. Meurers, W. Detmar. (2005, November). On the use of electronic corpora for theoretical linguistics: Case studies from the syntax of German. *Lingua*, Vol. 115, issue 11, 1619–1639. <https://doi.org/10.1016/j.lingua.2004.07.007>
26. Panickar, Suja, Sinha, Rimjhim, Chawla, Vidhi, Singh, & Londhe, Omkar. (2024). Sentiment Analysis of Custom Speech Corpus: A proof of concept for NLP. *Procedia Computer Science*, Vol. 244, 220–228. <https://doi.org/10.1016/j.procs.2024.10.195>
27. Saini, Jatinderkumar R., & Kaur, Jasleen. (2020). Kāvi: An Annotated Corpus of Punjabi Poetry with Emotion Detection Based on "Navrasa". *Procedia Computer Science*, Vol. 167, 1220–1229. <https://doi.org/10.1016/j.procs.2020.03.436>
28. Scott, M. (2010). Problems in investigating keyness, or cleaning the undergrowth and marking out trails. In: M. Bondi & M. Scott (Eds.), *Keyness in Texts*, 43–58. <https://doi.org/10.1075/scl.41.04.sco>
29. Scott, M. (2024). WordSmith tools (Version 9) [software], Lexical Analysis Software. URL: <https://lexically.net/wordsmith/>
30. Senapati, Apurbalal. (2024). Oxymoron: An Automatic Detection from the Corpus. *Procedia Computer Science*, Vol. 244, 49–56. <https://doi.org/10.1016/j.procs.2024.10.177>
31. Serazhim, K. S. (2002). Discourse as a sociolinguistic phenomenon: methodology, architectonics, variability: [based on materials from contemporary newspaper journalism]. Monograph. Kyiv, 392 p. [In Ukrainian]
32. Sokil, H. (2021). Conceptual principles of research of rebel songs. *Ethnographic notebooks*, 4(160), 953–961. <https://doi.org/10.15407/nz2021.04.953>
33. Trnavac, Radoslava, & Tenorio, Encarnacion Hidalgo. (2024, December). Breach of pacta sunt servanda: A corpus-assisted analysis of newspaper discourse on the AUKUS agreement. *Applied Corpus Linguistics*, Vol. 4, issue 3, article ID 100108. <https://doi.org/10.1016/j.acorp.2024.100108>
34. Williams, R. (1976). *Keywords: A vocabulary of culture and society*. New York: Oxford University Press. URL: https://books.google.com.ua/books?id=KnNWD9EYCGgC&pg=PA102&hl=uk&source=gsbs_toc_r&cad=2#v=onepage&q&f
35. Ye, Lei. (2024, December). Wordless: An integrated corpus tool with multilingual support for the study of language, literature, and translation. *SoftwareX*, Vol. 28, article ID 101931. <https://doi.org/10.1016/j.softx.2024.101931>
36. Zhukovska, V. (2020). Linguistic corpus as a new information and research tool of modern linguistics. Scholarly notes of V. I. Vernadsky TNU. Series: Philology. *Social communications*, 31(70), issue 3, Part 1, 113–119. <https://doi.org/10.32838/2663-6069/2020.3-1/20>

T. P. Dyak, Yu. I. Hrytsiuk

Lviv Polytechnic National University, Lviv, Ukraine

APPLICATION OF CORPUS TOOLS TO IDENTIFY KEYWORDS OF UKRAINIAN REBEL SONGS AS A GENRE OF THE FOLKLORE DISCOURSE

Pieces of folk art depicting national liberation struggles have been studied by modern scholars only during the years of independence, and is caused by recent revival of interest in the folklore genre. Therefore, relevance of this topic is manifested not only in the cognitive study and elucidation of features of the important period in the history of the Ukrainian people for its national liberation and its portrayal in folklore, but also it is a powerful resource in establishing patriotism, a symbol of insubordination, moral and ideological values. Our research deals with song genres, focusing on the conceptual principles of studying rebel songs. The proposed article is aimed at analyzing the issues of studying rebel song in the modern period, considering the main research approaches, clarifying the distinguishing features of studied songs through the analysis of key words, their frequency and semantic peculiarities. The paper draws attention to the artistic features of rebel songs, clarifying issues related to the features of rebel songs as markers of specific historical events. In this context, Sketch Engine tools and the corpus manager AntConc play a crucial role, providing the opportunity to conduct various linguistic studies, including the analysis of key words for revealing genre features of song texts. The research is based on the corpus of Ukrainian rebel songs, containing 35 pieces. An attempt is made to compare application of different corpus-based text analysis tools. Emphasis is put on defining the essence, concept and role of corpus tools, which will allow investigating the mechanism of their functioning. In the course of research, we applied such AntConc tools as *Collocates*, *N-Grams* and *Word List*, and also *Keywords* tool in Sketch Engine. The results obtained suggest that among the *Keywords*, the most frequently used parts of speech are exclamations, conjunctions and particles, which is typical for folk songs. Key words represented by nouns depict family ties, military everyday life and personal feelings of soldiers. The share of adjectives and verbs is quite significant. A large number of word forms with affectionate-diminutive suffixes in songs of this genre is observed as well, indicating a tender attitude towards the described objects. Therefore, the obtained research results are an important contribution to the development of corpus linguistics and the comprehensive use of AntConc and Sketch Engine corpus tools for keyword analysis. The concept of complex research of key words of rebel songs is considered as a necessary prerequisite for holistic comprehensive identification of important genre peculiarities of Ukrainian folklore.

Keywords: corpus linguistics; corpus manager; target and reference corpus; genre feature of texts; web system; frequency analysis.