МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
УКРАЇНСЬКИЙ КАТОЛИЦЬКИЙ УНІВЕРСИТЕТ

Гуманітарний факультет
Кафедра філології

МОВА ДІАСПОРИ Й МАТЕРИКОВОЇ УКРАЇНИ: ПОРІВНЯЛЬНИЙ КОРПУСНИЙ АНАЛІЗ (НА МАТЕРІАЛІ ПУБЛІЦИСТИКИ)

Студентки IV курсу
групи ГФІ-17/б
Христини Належатої

Науковий керівник:
кандидат філол. наук
Василь Старко

Львів 2021

ЗМІСТ

ВСТУП

Корпусна лінгвістика — це перспективна галузь сучасного мовознавства, що дає змогу швидко опрацьовувати величезні обсяги мовних даних. У світі ця галузь розвивається із 60-х років ХХ століття, а в Україні особливої уваги їй надають протягом останнього десятиліття: створюються і поповнюються корпуси, з'являються перші дослідження на основі корпусних даних тощо.

Зважаючи на відносно недовгий період зацікавлення корпусною лінгвістикою в Україні, є безліч ділянок, які можна вивчати, застосовуючи корпусні методи до вивчення української мови. На одному із цих напрямків — вивченні мови української діаспори — зосереджене наше дослідження.

Ми маємо на меті (1) підготувати тексти української діаспори для розміщення їх у Генеральному регіонально анотованому корпусі української мови (ГРАК); та (2) проаналізувати ці тексти, виокремивши їхні характерні особливості в порівнянні із відповідними публіцистичними текстами материкової України. Об'єктом дослідження є сучасне публіцистичне мовлення української діаспори, представлене в текстах діаспорної газети «Свобода». Предмет — характерні, статистично значущі особливості текстів української мови діаспори та материкової України на рівні морфології, словотвору, лексикології та семантики. Матеріалом для дослідження є тексти газети «Свобода» за 2016-2019 роки, а також, як база для порівняння, публіцистичні тексти материкової України тих самих років розміщені в корпусі ГРАК.

Наші завдання: (1) попереднє опрацювання текстів газети «Свобода» за 2016-2019 роки для додавання їх у корпус ГРАК, (2) створення співвідносних фокусного та референтного підкорпусів, (3) аналіз застосування методу ключових слів та пристосування його до особливостей нашого дослідження, (4) аналіз особливостей діаспорної публіцистики на основі створених підкорпусів з застосуванням виробленої методології, (5) виділення тих характерних мовних одиниць, які є ключовими для досліджуваних текстів.

У дослідженні застосований описовий метод (характеристика наявних корпусів української мови), метод корпусного аналізу (створення підкорпусів, проведення пошуків у корпусах та аналіз отриманих результатів), метод ключових слів (генерація списків ключових слів та виділення статистично значущих одиниць), метод критичного аналізу (аналіз та підсумування літератури на суміжну тематику, аналіз отриманих результатів при проведенні пошуку у корпусах), зіставний метод (порівняння отриманих результатів та виділення характерних особливостей фокусного підкорпусу).

Новизна дослідження полягає в підготовці та додаванні текстів у корпус ГРАК, а також — у застосуванні корпусних методів, зокрема обчислення статистичної значущості для досліджуваних явищ. Наскільки нам відомо, це дослідження є чи не першою спробою вивчити основні особливості мови діаспори із застосуванням корпусних методів.

Робота складається із трьох розділів. У першому розділі окреслено сферу досліджень корпусної лінгвістики, а в підрозділі 1.2 описано наявні корпуси української мови для визначення найвдалішого корпусу для проведення нашого дослідження. У другому розділі оглянуто основні мовознавчі праці про мову української діаспори та газети «Свобода» зокрема і подано їхні головні результати. Третій розділ — практичний — містить характеристику газети «Свобода» (підрозділ 3.1), опис процесу підготовки текстів газети до додавання у корпус (3.2), та створення досліджуваних підкорпусів (3.3). Підрозділ 3.4 містить визначення поняття ключовості, а у підрозділі 3.5 оглянуто найпопулярніші статистичні метрики, що можуть використовуватись для визначення ключових слів. Підрозділ 3.6 містить відбір та аналіз ключових слів, притаманних мові газети «Свобода», у порівнянні із текстами материкових ЗМІ. Основний текст роботи завершують висновки та список використаної літератури. В додатку 1 вміщено матеріали, що ілюструють процес редагування текстів для додавання їх у корпус ГРАК, в додатку 2 — частини автоматично згенерованих таблиць частотності слів у досліджуваних корпусах.

РОЗДІЛ І.
КОРПУСНА ЛІНГВІСТИКА ТА КОРПУСИ УКРАЇНСЬКОЇ МОВИ

У цьому розділі буде описано теоретичні засади поняття корпусної лінгвістики, а також наведені приклади конкретних корпусів української мови, зокрема ГРАК, на основі якого і буде проведене практичне дослідження мови періодики української діаспори.

1.1 Корпусна лінгвістика

Корпусна лінгвістика — це галузь комп'ютерної лінгвістики, що базується на вивченні мови з використанням корпусних даних та інструментарію.

Об’єктом корпусної лінгвістики є корпус. Корпус — це спеціально відібрана колекція текстів, чи їх фрагментів (частин, або окремих речень), що уможливлюють їх використання для проведення мовознавчих досліджень[1]. Якісно побудований корпус повинен бути репрезентативним та збалансованим. Рівень репрезентативності корпусу — це відповідність текстів його наповнення до різноманіття мови у її природному середовищі. Так, репрезентативний корпус писемної мови має містити різні письмові жанри як інформативних, так і художніх текстів, а також — спонтанні текстові повідомлення та листи[2]. Збалансованість корпусу визначається відповідністю пропорцій його наповнення різними жанрами до пропорцій, у яких ці жанри існують у мові. Відповідно, репрезентативний та збалансований корпус повинен відбивати всі характерні особливості мови у її кількісних та якісних варіаціях. Дослідження, проведені на основі цього корпусу мають бути показовими для всієї мови. Крім того, у корпусу є такі характеристики, як розмір, тип за часовим періодом охоплення (синхронний, діахронний чи моніторинговий), особливості текстів наповнення, та їх розмітки — ці

[1] Anatol Stefanowitsch. Corpus linguistics: A guide to the methodology. Berlin 2020, — с. 22
[2] Там само, с 28-36

характеристики розробники корпусу підбирають з урахуванням мети створення корпусу.

Побудова корпусу — це процес, що потребує попереднього планування та врахування багатьох факторів, а також людської праці та часу. При тому, готовий корпус є основою, на базі якої можна провести безліч досліджень із залученням набагато менших ресурсів, аніж ті, що необхідні для проведення аналогічних досліджень без використання методів корпусної лінгвістики. Висновки цих корпусних досліджень дають змогу охоплювати порівняно більші обсяги джерел, аніж ті що їх інтуїтивно добирають дослідники. Таким чином дослідження, проведені на основі корпусу претендують на більшу об'єктивність матеріалів, що є їхньою основою[3].

Пошук у корпусі може підтвердити, або спростувати очікувані результати того, чи іншого дослідження, дати науковцям ширшу картину дійсності, а отже — розуміння доцільності вибору тих чи інших стратегій чи припущень. Методи корпусної лінгвістики можна використовувати для підтвердження та ширшого впровадження висновків досліджень, що були проведені за використанням іншої методології, (особливо, коли мова йде про кількісні дослідження). Крім того, методів корпусної лінгвістики достатньо для того, аби робити самостійні обґрунтовані висновки та передбачення.

Серед найбільш впливових корпусів у світі — Brown Corpus of Standard American English[4]. Це синхронний корпус американської англійської 1961 року. Він складається із 500 текстів, кожен з яких містить 2 тисячі слів, що разом становить 1 мільйон слів. До Браунського корпусу входить 15 категорій текстів, кількість текстів у кожній категорії з огляду на збалансованість відрізняється. Дизайн корпусу став основою для «сім'ї браунських корпусів», серед яких британський

[3] Вікторія Жуковська. Вступ до корпусної лінгвістики: навчальний посібник. Житомир 2013. С 9-10
[4] The Brown Corpus. 1998. — Режим доступу: https://www1.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/corpora/list/private/brown/brown.html

LOB Corpus[5], корпус індійської англійської Kolhapur Corpus[6], Браунський український корпус (БрУК)[7] та інші.

Корпуси браунської сім'ї є відносно невеликими. Напротивагу їм існують набагато більші корпуси, зокрема — національні, серед яких British National Corpus (BNC)[8], Open American National Corpus (OANC)[9], National Corpus of Polish (Narodowy Korpus Języka Polskiego — NKJP)[10], Russian National Corpus (Национальный корпус русского языка) [11] та інші.

1.2 Корпуси української мови

1.2.1 Корпус української мови лінгвістичного порталу mova.info

Досить великим за обсягом є Корпус української мови лінгвістичного порталу mova.info[12] (близько 100 млн слів). Корпус розробили співробітники лабораторії комп'ютерної лінгвістики та кафедри сучасної української мови Інституту філології Київського університету імені Тараса Шевченка. Він містить законодавчі (2,5 млн слововживань), наукові (8,7 млн), фольклорні (1,3 млн) та поетичні (789 тисяч) тексти, найбільшими є розділи публіцистики (44,6 млн слововживань) та художньої прози (40,2 млн).

Кожен розділ має низку підрозділів, і користувач може здійснювати пошук, як у цілому корпусі, так і в окремій його частині (підкорпусі). Корпус української мови лінгвістичного порталу mova.info містить метатекстову розмітку (автор, рік видання, джерело, жанр, тип та хронотип (для художньої прози), сфера функціонування тексту, тематика), а також — інформацію про структурне

[5] Lob Corpus. — Режим доступу: https://www1.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/corpora/list/private/LOB/lob.html
[6] Manual Of Information To Accompany The Kolhapur Corpus Of Indian English, For Use With Digital Computers. 1986. — Режим доступу: http://korpus.uib.no/icame/manuals/KOLHAPUR/INDEX.HTM
[7] Браунський корпус української мови // GitHub. — Режим доступу: https://github.com/brown-uk/corpus
[8] British National Corpus (BNC). — Режим доступу: https://www.english-corpora.org/bnc/
[9] The Open American National Corpus. 2002–2015. — Режим доступу: https://www.anc.org/
[10] Narodowy Korpus Języka Polskiego. 2008-2012. — Режим доступу: http://nkjp.pl/
[11] Russian National Corpus. 2003–2021. — Режим доступу: https://ruscorpora.ru/old/en/index.html
[12] Корпус текстів української мови. 2003–2021. — Режим доступу: http://www.mova.info/corpus.aspx

розташування кожної словоформи (розділ, номер абзацу, речення та відповідного слова), її граматичну та лексико-граматичну та лексичну інформацію[13].

Головними обмеженнями використання корпусу від mova.info є можливість здійснювати пошук лише в одному обраному підкорпусі, але не в усіх текстах загалом. Крім того, корпус не показує статистичної інформації результатів пошуку, а отже важко надається до кількісних досліджень мови.

Корпус не містить діаспорних текстів, але може стати допоміжним джерелом для дослідження сучасної української публіцистики.

1.2.2 Корпуси проєкту «Лабораторія української»

«Лабораторія української» є спільним польсько-українським корпусним проєктом. У його межах збудовано корпус зі знятою омонімією розмічений руками (Золотий стандарт), обсягом 140 тисяч токенів. Тексти у ньому містять вручну розставлену розмітку — морфологічну та синтаксичних залежностей. Джерельна база корпусу — це статті, новини, дописи, підручники, листи, казки, художня проза; обмежень за часом створення тексту у добірці немає. Так, корпус, хоча і є невеликим, але містить дуже точні дані розмітки, що дає змогу використовувати його як золотий стандарт при створенні та перевірці автоматичних мовних аналізаторів[14].

Команда проєкту «Лабораторія української» також розробила автоматично розмічений корпус зі знятою омонімією Звідусіль[15] [16], що містить 3 мільярди токенів. Тексти, що входять до корпусу взяті з різноманітних джерел, що є у вільному доступі (переважно в інтернеті), зокрема, дописів користувачів в

---

[13] Наталія Дарчук. Дослідницький корпус української мови: основні засади і перспективи // Вісник Київського національного університету імені Тараса Шевченка. Серія: Літературознавство. Мовознавство. Фольклористика. – № 21. Київ 2010. С 45–49.

[14] Лабораторія української: Золотий морфосинтаксовий стандарт. — Режим доступу: https://mova.institute/%D0%B7%D0%BE%D0%BB%D0%BE%D1%82%D0%B8%D0%B9_%D1%81%D1%82%D0%B0%D0%BD%D0%B4%D0%B0%D1%80%D1%82

[15] Звідусіль // NoSketchEngine — Режим доступу: https://mova.institute/bonito/run.cgi/first_form?corpname=zvidusil

[16] Звідусіль: Про корпус. NoSketchEngine — Режим доступу: https://mova.institute/bonito/run.cgi/corp_info?corpname=zvidusil&struct_attr_stats=1&subcorpora=1

соцмережах, тощо. Для пошуку у корпусі, можна уточнити такі параметри, як підкорпус за джерелом, назву, автора, та час появи тексту, проте, крім цих даних, тексти не містять більше ніякої метаінформації. Доступна також статистична інформація про результати пошуків. Цінним для цього дослідження може бути можливість доступу до публіцистики Звідусіль, що містить майже три мільйони токенів[17]. Що ж до регіонального розрізнення між мовою материка та діаспори, то відповідної інформації корпус не містить.

Третім корпусним проєктом «Лабораторії української» є паралельні корпуси польської, англійської, французької, німецької, іспанської та португальської мов, вирівняні до української. Тексти, що входять до цих корпусів розмічено автоматично[18].

1.2.3 Браунський корпус української мови

Браунський корпус української мови все ще перебуває на стадії розробки. Його структура близька до структури Brown Corpus of Standard American English, з деякими змінами для кращої репрезентації стану сучасної української мови. До корпусу входять лише оригінальні неперекладні тексти авторів з материкової України, які зазнали редагування. Часовий період, який охоплює корпус — 2010-2018 роки. Тексти поділені на такі категорії: преса (25%), релігійна література (3%), професійно-популярна література (7%), «естетичні інформативні» тексти (7%), адміністративні документи (3%), науково-популярна література (5%), наукова література (10%), навчальна література (15%), художні тексти (25%)[19].

Всі тексти, які додають до корпусу мають бути вичитані. Тексти містять метадані про категорію тексту, автора, назву твору, джерело, інформацію про видання, довжину фрагмента, правопис та помилки, що містяться в фрагменті.

---

[17] Там само.
[18] Лабораторія української: Золотий морфосинтаксовий стандарт. — Режим доступу: https://mova.institute/%D0%B7%D0%BE%D0%BB%D0%BE%D1%82%D0%B8%D0%B9_%D1%81%D1%82%D0%B0%D0%BD%D0%B4%D0%B0%D1%80%D1%82
[19] Вимоги до текстів для БрУК // GitHub. — Режим доступу: https://github.com/brown-uk/corpus/blob/master/doc/vymohy_do_frahmentiv.md

Корпус є порівняно невеликим, але цінним з огляду на якість текстів, що до нього входять, а також — їхньої розмітки.

1.2.4 Проєкти спільноти lang-uk

Найбільшим проєктом спільноти lang-uk є корпус UberText, який містить понад 665 мільйонів токенів. Корпус складається з текстів трьох джерел: періодика (70%), Вікіпедія (15%) та художня література (15%), що зібрані автоматично. Для корпусу існує також окрема лематизована версія.

Крім того, цією ж спільнотою розроблений лематизований та токенізований корпус законів та нормативно-правових актів України, обсягом більше як 578 мільйонів токенів.

Також на базі текстів Браунського корпусу української мови спільнота lang-uk розробила корпус NER-анотацій. Він складається із 229 текстів, що містять всього 217 тисяч токенів з більш як 6 тисячами розмічених іменованих сутностей. Анотація була виконана вручну, за тегами «персона», «місце», «організація» та «різне»[20]. Корпус є корисним для тренування та тестування моделей із розпізнавання іменованих сутностей.

1.2.5 Українські вебкорпуси Лейпцизького університету

До корпусної колекції Лейпцизького університету входять корпуси для більш як 300 мов. Серед них і три українські корпуси, тексти для яких зібрані автоматично.

Корпус української мови 2014 року містить майже 2 мільярди токенів[21], до нього ввійшли тексти різних жанрів. Корпус новинних текстів 2018 року містить 300 мільйонів токенів[22]. Корпус інтернет-текстів 2019 року містить 1,2 мільярда

[20] Спільнота lang-uk. 2016–2017. — Режим доступу: https://lang.org.ua/uk/corpora/#anchor6
[21] Ukrainian (ukr_mixed_2014): Про корпус // NoSketchEngine — Режим доступу: http://cql.corpora.uni-leipzig.de/bonito/run.cgi/corp_info?corpname=ukr_mixed_2014
[22] Ukrainian (ukr_newscrawl_2018): Про корпус // NoSketchEngine — Режим доступу: http://cql.corpora.uni-leipzig.de/bonito/run.cgi/corp_info?corpname=ukr_newscrawl_2018

токенів[23]. Тексти цих корпусів не мають детальної метаінформації — є лише інформація щодо дати та джерела, звідки взятий кожен текст, проте сам корпусний менеджер дає змогу отримати широкий спектр статистичної інформації про тексти в корпусі та окремі конкорданси. Тож новинний корпус української мови 2018 року може стати корисним допоміжним джерелом для цього дослідження, але не підійде для розмежування мови діаспори та материкової України, бо не містить відповідних даних.

1.2.6 Корпус української мови бібліотеки «Чтиво»

Корпус бібліотеки «Чтиво» містить 600 мільйонів слів та складається із текстів, що були автоматично розпізнані зі сканів книжок бібліотеки. Він не вичитаний, і не містить анотацій, а пошук здійснюється дослівно, із можливістю задавати шаблони гнізд слів чи словоформ. Немає також можливості отримати статистичну інформацію про корпус та запити.

Отже, корпус української бібліотеки «Чтиво» не може бути надійним джерелом для проведення наукових досліджень мови.

1.2.7 Генеральний регіонально анотований корпус української мови (ГРАК)[24]

Генеральний регіонально анотований корпус української мови (ГРАК) є корпусом, що дозволяє проводити чи не найрізноманітніші види досліджень.

Остання на час написання — одинадцята — версія корпусу містить 728 мільйонів токенів із 93 тисяч документів.

[23] Ukrainian (ukr-ua_web_2019): Про корпус // NoSketchEngine — Режим доступу: http://cql.corpora.uni-leipzig.de/bonito/run.cgi/corp_info?corpname=ukr-ua_web_2019

[24] Генеральний регіонально анотований корпус української мови (ГРАК) / М. Шведова, Р. фон Вальденфельс, С. Яригін, А. Рисін, В. Старко та ін. – Київ, Львів, Єна, 2017-2021. — Режим доступу: http://uacorpus.org/

У корпусі містяться тексти, дата створення яких коливається від 1816 року[25], і до сучасності (проте, варто зауважити, що сучасних текстів у ГРАКу найбільше). Окрім дати створення є також розмітка за датою публікації[26]. Тож корпус дає змогу проводити як синхронні дослідження (що вивчають конкретний проміжок часу), так і діахронні (що вивчають історичний розвиток мови).

Приблизно половина текстів у корпусі містять регіональну розмітку: вказівку на регіональну приналежність автора чи перекладача (враховуючи всі регіони, де він перебував понад десять років). Для публіцистичних текстів (газета, журнал, або інформаційний сайт[27]) вказано місце видання. Тексти з України розмічені за регіонами, що відповідають адміністративному поділу держави, натомість для авторів з-за кордону вказано лише країну.[28]

Тексти у корпусі містять також інформацію про стиль: приблизно половина текстів належать до художнього стилю, великою є частка публіцистики. Крім того, є розмітка деяких жанрів, а для наукових текстів також — тематики та галузі тексту[29].

Приблизно третина текстів корпусу — це переклади з інших мов. Інформація про мову оригіналу вказана в розмітці[30]. Про авторів, чи перекладачів вказана така інформація, як рік народження, стать, а також регіональна приналежність[31]. Тексти ГРАКу містять інформацію про джерело, звідки їх взято: видавництво, чи інтернет-покликання, рік та місце видання[32].

---

[25] Maria Shvedova. The General Regionally Annotated Corpus of Ukrainian (GRAC, uacorpus.org): Architecture and Functionality // Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2020), vol. I: Main Conference. Lviv 2020. с 489–506.
[26] Генеральний регіонально анотований корпус української мови (ГРАК) / М. Шведова, Р. фон Вальденфельс, С. Яригін, А. Рисін, В. Старко та ін. – Київ, Львів, Єна, 2017-2021.\\ Розмітка текстів. Датування. — Режим доступу: http://www.parasolcorpus.org/Kyiv/ua/rozmitka-tekstiv/datuvannya
[27] Там само\\ Розмітка текстів. Відомості про медіа. — Режим доступу: http://www.parasolcorpus.org/Kyiv/ua/rozmitka-tekstiv/vidomosti-pro-media
[28] Там само\\ Розмітка текстів. Регіональна розмітка. — Режим доступу: http://www.parasolcorpus.org/Kyiv/ua/rozmitka-tekstiv/regionalna-rozmitka
[29] Там само\\ Розмітка текстів. Стилі, тематика і жанри. — Режим доступу: http://www.parasolcorpus.org/Kyiv/ua/rozmitka-tekstiv/stili-tematika-i-zhanri
[30] Там само\\ Розмітка текстів. Мова оригіналу. — Режим доступу: http://www.parasolcorpus.org/Kyiv/ua/rozmitka-tekstiv/mova-originalu
[31] Там само\\ Розмітка текстів. Відомості про автора тексту. — Режим доступу: http://www.parasolcorpus.org/Kyiv/ua/rozmitka-tekstiv/vidomosti-pro-avtora-tekstu
[32] Там само\\ Розмітка текстів. Джерело тексту. — Режим доступу: http://www.parasolcorpus.org/Kyiv/ua/rozmitka-tekstiv/dzherelo-tekstu

Наявність регіональної розмітки та розмітки за роками публікації та стилями дає змогу створити підкорпус діаспорної публіцистики та відповідний підкорпус текстів материкової України для їхнього подальшого дослідження та порівняння.

Більшість текстів у ГРАКу написані згідно із сучасним правописом, однак є й такі, що написані скрипниківкою чи желехівкою (правопис вказано у розмітці)[33].

Тексти корпусу розмічені морфологічним аналізатором. Він розпізнає близько 98% слів, і присвоює кожному із них лему та набір тегів. Морфологічний аналізатор базується на Великому електронному словнику української мови (ВЕСУМ)[34].

Обмеженням корпусу ГРАК є те, що у ньому не знята омонімія: слова, яким можуть бути присвоєні різні граматичні форми, містять всі можливі набори тегів. Це може призводити до викривлення результатів пошуку за граматичними формами[35].

Для найчастотнішої лексики в текстах (конкретні та абстрактні іменники, власні назви, прикметники, прислівники, дієслова) додано семантичну розмітку. (Частина тегів, що належать до семантики слова є також у ВЕСУМі, тому ці теги дублюються у морфологічній та семантичній розмітці[36].)

Інструмент роботи з корпусом NoSketchEngine дає можливість створювати підкорпуси із заданими параметрами (всі описані раніше властивості можна визначити при побудові підкорпусу). Це істотно розширює область потенційних досліджень, що можуть бути проведені на основі ГРАКу, у порівнянні з іншими корпусами української мови. Завдяки широкому спектру метаінформації, яка доступна до текстів корпусу, він дозволяє чи не найбільше специфікувати пошук та досліджувати сучасну українську мову у синхронії та діахронії, зважаючи на такі фактори, як регіональні особливості, інформація про джерела, авторів, стилі та

[33] Там само\\ Розмітка текстів. Правопис. — Режим доступу: http://www.parasolcorpus.org/Kyiv/ua/rozmitka-tekstiv/pravopis
[34] Там само\\ Розмітка текстів. Морфологічна розмітка. — Режим доступу: http://www.parasolcorpus.org/Kyiv/ua/rozmitka-tekstiv/morfologichna-rozmitka
[35] Там само\\ Розмітка текстів. Граматична омонімія. — Режим доступу: http://www.parasolcorpus.org/Kyiv/ua/rozmitka-tekstiv/gramatichna-omonimiya
[36] Там само\\ Розмітка текстів. Семантична анотація. — Режим доступу: http://www.parasolcorpus.org/Kyiv/ua/rozmitka-tekstiv/semantichna-anotaciya

жанри, правопис, мову оригіналу та редакторські втручання. Проте для проведення успішного дослідження варто пам’ятати про обмеження, які накладає корпусний метод та особливості побудови самого ГРАКу, критично оцінюючи результати, отримані при пошуку в корпусі.

Отже, хоча для досліджень української мови ще не розроблено універсального корпусу, науковці продовжують працю над низкою різноманітних корпусів, що різняться за характеристиками, а отже — придатні для розв’язання різних типів завдань, що постають перед сучасними лінгвістами. Нам здається, що серед корпусів української мови, що є у вільному доступі Генеральний регіонально анотований корпус має чи не найбільший потенціал для лінгвістичних досліджень. Це єдиний на сьогодні корпус, що, завдяки своїй розмітці, дає змогу виокремити діаспорні тексти та досліджувати їх[37], порівнюючи із текстами із материкової України. Корпусні менеджери, доступні для роботи із корпусом надають гнучкий інструментарій для проведення комплексного дослідження текстів.

[37] Maria Shvedova, Ruprecht von Waldenfels. Regional Annotation within GRAC, a Large Reference Corpus of Ukrainian: Issues and Challenges. 2021. — Режим доступу: http://ceur-ws.org/Vol-2870/paper4.pdf — с 12

РОЗДІЛ ІІ.
ФЕНОМЕН МОВИ УКРАЇНСЬКОЇ ЗАХІДНОЇ ДІАСПОРИ У НАЯВНИХ ДОСЛІДЖЕННЯХ

У цьому розділі увагу буде зосереджено на феномені мови української західної діаспори (насамперед — діаспори Сполучених Штатів Америки, мову публіцистики якої буде досліджено у наступному розділі). Розглянемо наявні дослідження на вказану тематику, а також окреслимо теоретичні засади, якими послуговуються мовці.

Аби говорити про особливу мову діаспори, варто спершу окреслити саме поняття «українська діаспора».

«Енциклопедія історії України» подає два визначення слова діаспора. «Це 1) частина певної етнічної спільноти (народу), що постійно проживає поза межами історичної батьківщини в іноетнічному чи інонаціональному середовищі на правах або зі статусом національно-культурної меншини; 2) етнічна меншина, яка зберігає зв'язок зі своєю батьківщиною, національні традиції тощо. Діаспора утворюється внаслідок дії тих чи інших чинників, які обумовлюють масове переселення частини населення з однієї країни в іншу»[38].

Як зазначає Олександр Тараненко, поняття «діаспора» замінило поняття «українська еміграція» у 80–90-х роках минулого століття. Його вживають до українців та їхніх нащадків, що проживають за територіальними межами України, насамперед — в Північній та Південній Америці, зокрема — у Сполучених Штатах Америки та Канаді, але також — у Західній Європі та Австралії[39]. На противагу поняттю діаспори з’явилось поняття материкової України, яке позначає саме територію України[40]. Цей поділ дає можливість розмежування мови, преси, літератури діаспори та материкової України.

---

[38] Олексій Ясь. Діаспора // Енциклопедія історії України, т. 2: Г-Д. Київ 2004. — Режим доступу: http://www.history.org.ua/?termin=Diaspora

[39] Олександр Тараненко. Мова української західної діаспори і сучасна мовна ситуація в Україні // Мовознавство, № 2-3. С. 63–99. — Режим доступу: http://nbuv.gov.ua/UJRN/MoZn_2013_2-3_7 – с 63-64.

[40] Там само, с 64.

За часом виїзду з України виділяють чотири хвилі української еміграції, яка і стала основою української діаспори.

Перша хвиля припадає на період між 70–ми роками XIX століття та початком Першої світової війни, і становить близько пів мільйона осіб з обох імперій, до складу яких входили українські землі[41]. Як зазначає Богдан Ажнюк, цю групу емігрантів складали здебільшого селяни та наймити (найчастіше — це неписьменні люди), які виїжджали аби заробити грошей та повернутися додому; часто вони не усвідомлювали своєї національної чи етнічної приналежності[42].

Друга хвиля еміграції була кількісно набагато меншою і відбулась у період між двома світовими війнами. Вона мала частково політичний характер — як і третя хвиля, яка відбулась після закінчення Другої світової війни та складалась в основному з осіб, що після війни перебували у таборах для переміщених осіб у Західній Європі, але також — біженців з Української РСР[43]. До складу другої та третьої хвилі еміграції входили вже добре освічені українці[44].

Починаючи із 90–х років минулого століття триває четверта хвиля української еміграції. Вона має соціально-економічний характер, і оцінена як найчисленніша з еміграцій, а також — найосвіченіша[45]. З четвертою хвилею пов'язують збільшення кількості мовців української мови в українській діаспорі США удвічі: від 7 до 14% [46].

Зважаючи на те, що українська мова емігрантів та українська мова на материковій Україні розвивались порізно, вони набували відмінних рис. Саме тому з'являється потреба говорити про українську мову діаспори та протиставляти їй материкову мову. Зокрема, дослідження та порівняння двох мовних варіантів може

[41] Леся Біловус. Україномовна періодика у національно-культурному житті української діаспори США (1991–2017 рр.). Тернопіль 2017. — с 9.
[42] Богдан Ажнюк. Мовна єдність нації: діаспора й Україна. Київ 1999. — с 46.
[43] Там само — с 46, 51.
[44] Леся Біловус. Україномовна періодика, — с 8-10.
[45] Там само, — с 10-11.
[46] Українці у США: кількість, міграція, заняття та уподобання. За матеріалами відкритої лекції Воловини О. «Українці в США та вплив міграції з України на діаспору в США (на базі офіційної статистики)» // Справжня варта, 21 листопада 2012. — Режим доступу : http://varta.kharkov.ua/articles/1090601.html

розповісти про особливості впливу історичних та культурних процесів, а також — іншомовного середовища на мовців української.

Існує низка досліджень, результати яких допоможуть нам зрозуміти особливості феномена мови діаспори, оцінити можливі фактори, що вплинули на різницю між мовою материкової України та діаспори, побачити, які зміни сталися із мовою діаспори у діахронії та як вона вплинула на материковий варіант української мови. З огляду на те, що це дослідження буде враховувати лише мову публіцистичних видань, особливу увагу буде звернено на дослідження діаспорної публіцистики, зокрема досліджуваної нами газети «Свобода».

Монографія Богдана Ажнюка «Мовна єдність нації: діаспора й Україна» є корисним джерелом, що пояснює причини відмінності діаспорного та материкового варіанта української мови, вона також містить низку прикладів з преси, що стануть корисними у нашому дослідженні.

Автор розглядає мову як головний індикатор національної ідентичності мовців, аналізує мовну свідомість та мовну поведінку у двомовному середовищі. У роботі також проаналізовано лексичний паралелізм та функціональну стратифікацію, також правописний дуалізм між материковим та діаспорним варіантами української. Свої тези дослідник ілюструє прикладами із періодичних видань, зокрема — «Свободи», а також результатами власного опитування мовців з діаспори.

Богдан Ажнюк зокрема виділяє процес «віднайдення» емігрантами першої хвилі своєї національної приналежності вже у Сполучених Штатах, і безпосередній зв'язок ідентифікації з мовою, що стала для емігрантів найлегшим способом підтримки контакту із земляками. Дослідник виділяє галицький інтердіалект, як такий, що був загально зрозумілим для представників різних регіонів, і спричинив нівелювання розбіжностей між їхніми діалектами[47]. Пише дослідник і про мовну інтерференцію: він пов'язує велику кількість запозичень із недостатньою початковою сформованістю української мови в діаспорі: відсутністю єдиного

[47] Богдан Ажнюк. Мовна єдність нації: діаспора й Україна — с 48.

варіанта, а також — недостатньою освіченістю та діалектною строкатістю перших емігрантів[48]. Про мовну політику публіцистичних видань Ажнюк говорить як про досить ліберальну — мову «Свободи» він описує як часом строкату, зокрема тому, що газета часто вміщує на своїх сторінках матеріали авторства широкого загалу своїх дописувачів, суттєво їх не правлячи[49] — «таким чином преса віддзеркалює певні тенденції в розвитку живого мовлення»[50] (тобто — мовлення своїх дописувачів, що відбите у їхніх текстах).

Згадує автор і про намагання діаспорян якомога більше віддалити українську мову від російської, що об'єднує і мовних пуристів, і прихильників запозичень[51]. Так, діаспорні мовці можуть віддавати перевагу словам, що не є однокореневими із російськими відповідниками, несвідомо вводячи у мову запозичення з інших мов, навіть коли у цьому немає потреби.

Щодо дослідження слововживання західною діаспорою, Богдан Ажнюк пише: «на наш погляд, визначальною рисою слововжитку українців діаспори, що віддзеркалює й відповідний стан мовної свідомості є поєднання «київських» і «діаспорних» рис».

Автор звертається і до питання правопису, зазначаючи, як і інші дослідники, що діаспорна традиція спирається на харківський правопис 1928–1929–го року (відомий також як скрипниківка) та правописний словник Голоскевича[52]. Про невідповідність правопису материкової України та того, що використовували в діаспорі дослідник пише так: «контрольоване владою унормування правопису на Україні, що було складовою частиною асиміляційної політики, викликало природний спротив більшості закордонного українства. Мова, зокрема правопис, також потрапили у сферу політичного протистояння… За альтернативу колонізаторському правописові було визнано так званий харківський правопис 1928–1929–го рр. та правописний словник Голоскевича»[53]. Сам автор визнає

[48] Богдан Ажнюк. Мовна єдність нації: діаспора й Україна — с 280.
[49] Там само, — с 280.
[50] Там само, — с 313.
[51] Там само, — с 309.
[52] Там само, — с 349-355.
[53] Богдан Ажнюк. Мовна єдність нації: діаспора й Україна — с 350-351.

основним недоліком харківського правопису певну непослідовність, що виникла як наслідок намагання його творців поєднати наддніпрянську та галицьку правописну традиції. Для того, аби виправити це, на думку Богдана Ажнюка, необхідна була єдина система освіти та авторитетний науковий центр, що дало б змогу вивчити та «загладити гострі кути» скрипниківки[54].

У своїй іншій статті «Мовні роздоріжжя української нації» Богдан Ажнюк висловлює думку про те, що зі здобуттям Україною незалежності, мова діаспори поступово наближається до материкового варіанта, якому все більше надаватимуть перевагу закордонні мовці, а отже і саме протиставлення діаспорного варіанта із літературною українською мовою материка стане неактуальним[55].

Подібну тезу висловлює і Л. Козачевська у своїй статті «Мова української діаспори як віддзеркалення мовної свідомості та мовної діяльності носіїв-білінгвів». Авторка наголошує, що більш наближеною до мови материка є писемна мова і мотивує це двома причинами. По-перше, великою кількістю емігрантів четвертої хвилі (що є носіями сучасного варіанта мови материка), залучених до створення писемних текстів закордоном, та, по-друге, свідомим намаганням діаспорян наблизити свою писемну мову до материкового варіанта[56].

У статті Олександра Тараненка «Мова української західної діаспори і сучасна мовна ситуація в Україні» розглянуто стан та статус мови західної української діаспори та протиставляє її українській материковій мовній традиції. Він постулює вплив першої на другу, виділяючи особливості мови діаспори на лексичному, словотвірному, синтаксичному, фонетичному та правописному рівнях.

Автор стверджує, що мовна практика діаспори є досить невпорядкованою, а позиції щодо правильного варіанта української літературної мови серед діаспорян різняться[57]. Науковець зазначає, що: «мова публічної діяльності основної частини

---

[54] Там само, — с 351-353.
[55] Богдан Ажнюк. Мовні роздоріжжя української діаспори. — Режим доступу: http://kulturamovy.univ.kiev.ua/KM/pdfs/Magazine53-54-6.pdf , — с 7.
[56] Л. Козачевська, О. Сидоренко. Мова української західної діаспори як віддзеркалення мовної свідомості та мовної діяльності носіїв-білінгвів // Мовні і концептуальні картини світу, вип. 46(2). Київ 2013, — с 148.
[57] Олександр Тараненко. Мова української західної діаспори і сучасна мовна ситуація в Україні // Мовознавство, № 2-3. С. 63–99. — Режим доступу: http://nbuv.gov.ua/UJRN/MoZn_2013_2-3_7 , — с 68.

західної діаспори в її правописній, граматичній основі та значною мірою в лексичному інвентарі та стилістичних особливостях прагне спиратися на мовні настанови й мовну практику 1920-х – початку 1930-х років у радянській Україні (періоду «українізації»), зокрема на «харківський» правопис 1928 р., та довоєнного періоду в Галичині, на Буковині, але водночас — у практиці різних друкованих видань, у численних рекомендаціях та застереженнях — зазнає різноманітних «редагувань», що й породжує хаотичність самої мовної практики»[58]. Також згадується, що діаспорний варіант мови спирається головно на південно-західне наріччя, та на варіант західноукраїнської УЛМ до приєднання її до СРСР, що також зумовлює відмінність його від материкової літературної мови, базованої на південно-східному наріччі[59].

Щодо повного злиття мовних традицій діаспори та материка — Тараненко не вбачає підстав стверджувати, що таке може відбутись, хоча і не заперечує певного зближення двох мовних варіантів[60].

Вивченням періодики західної діаспори займається Леся Біловус. У її монографії «Україномовна періодика у національно-культурному житті української діаспори США (1991–2017 рр.)»[61] проаналізовано національно-культурне життя діаспори та збереження її представниками своєї національної ідентичності крізь призму діяльності діаспорних періодичних видань. Авторка аналізує ціннісні орієнтації україномовної періодики США, її сприяння освіті, формуванню національної ідентичності та державотворенню.

У монографії окреслено роль періодики, як важливого культурно-національного та просвітницького чинника, що за відсутності певних соціальних інститутів став чи не головним консолідаційним фактором для українців США. Зокрема, відзначено відповідну роль «Свободи» у цьому процесі.

[58] Там само, — с 69.
[59] Там само, — с 69.
[60] Там само, — с 93-94.
[61] Леся Біловус. Україномовна періодика у національно-культурному житті української діаспори США (1991–2017 рр.). Тернопіль 2017.

Є й низка досліджень самої «Свободи». Це, зокрема, частина монографії Міхаеля Мозера «New Contributions to the History of the Ukrainian Language» під назвою «The «Mirror from Overseas»: The History of Modern Standard Ukrainian as Reflected in the North American Ukrainian Newspaper Svoboda (The Early Years: from 1893 to the 1930s)». У дослідженні автор зосереджується на ранньому періоді існування «Свободи», подаючи інформацію про редакторів, аналізуючи мовну строкатість та правописні зміни у номерах газети. Зокрема, проаналізовано як заголовки «Свободи» показували історію газети та її мови: поступовий перехід від етимологічного правопису до фонологічного (желехівки)[62] і далі. Міхаель Мозер детально аналізує появу та розгортання впливу нових правописних норм в газеті, зокрема — адаптацію мови «Свободи» до материкових стандартів[63], аж поки Україна не зазнала русифікаційної політики з боку СРСР: тоді діаспора продовжила розвивати норми довоєнної мовної практики материка[64]. Щодо харківського правопису, то автор не відзначує конкретного моменту, коли газета починає керуватися лише ним, а радше вказує на неможливість прийняти наступні правописні зміни, що відбулись на материку.

Присвятив свою монографію історії газети «Свобода» і її редактор Петро Часто. Праця із назвою «Вільне слово американської України» була видана із нагоди 120-ти річчя (!) видання газети. У хроніці-хрестоматії Петро Часто подає провідні теми, як впродовж свого існування висвітлювала «Свобода», таким чином показуючи й розвиток життя західної діаспори, висвітлений на шпальтах газети.

У передмові Леонід Рудницький так окреслює роль та значущість газети: «від часу свого першого числа, яке побачило світ 15 вересня 1893 року, «Свобода» була і залишається чимось більшим, ніж етнічна газета, ніж друкований орган Українського Народного Союзу. Вона є, в першу чергу, українським часописом, який відзеркалює українську свідомість у Сполучених Штатах Америки, а

[62] Michael Moser. New Contributions to the History of the Ukrainian Language. Edmonton — Toronto 2016. — с 414-418.
[63] Там само, — с 435-436.
[64] Там само, — с 438-439.

рівночасно є конструктивним, будуючим чинником у житті української громади на американській землі. З історичної точки зору, її пріоритетним завданням, як це висловлено у згаданому першому числі, було допомагати українському народові втримати його ідентичність, зберегти його віру, обряд, традиції та звичаї, і — а це преважливе — його мову»[65].

Отже, існує низка досліджень мови публіцистики західної української діаспори різних періодів. Вони дають змогу краще зрозуміти які саме чинники вплинули на ті чи інші особливості мови діаспори, та як вона розвивалось починаючи від першої хвилі еміграції українців. Дослідники досить часто використовують і матеріали діаспорної газети «Свобода», вказуючи на її впливовий статус та довгу історію функціонування у середовищі українців західної діаспори. У наступній частині цього дослідження для порівняння мовних варіантів також буде використано сучасні тексти газети «Свобода».

[65] Петро Часто. Вільне слово Американської України. Нью-Йорк — Ужгород 2012. — с 5-6.

РОЗДІЛ ІІІ.
КОРПУСНИЙ АНАЛІЗ ТА ПОРІВНЯННЯ ТЕКСТІВ ГАЗЕТИ «СВОБОДА» ТА ПУБЛІЦИСТИКИ МАТЕРИКОВОЇ УКРАЇНИ

У цьому розділі буде описана практична частина дослідження мови публіцистики української західної діаспори. Також буде пояснено деякі статистичні метрики та поняття, потрібні для розуміння того, як здійснювався аналіз текстів.

Для проведення дослідження було обрано Генеральний регіонально анотований корпус української мови: він містить необхідну розмітку, зокрема регіональну, а корпусні менеджери дають змогу задавати різноманітні параметри пошуку. Окрім того, в ГРАКу можна створити користувацькі підкорпуси з необхідним вмістом для того, аби провести порівняння мови діаспори та материка. Параметри створених підкорпусів буде описано в підрозділі 3.3.

Метою нашого дослідження є порівняння саме сучасної мови публіцистики у діаспорі та в Україні. Для того, аби мати змогу працювати із достатнім для дослідження корпусом сучасних діаспорних текстів, до виходу одинадцятої версії корпусу ГРАК було підготовано тексти діаспорної газети «Свобода» за 2016-2019 роки.

3.1 Газета «Свобода»

Газету «Свобода» було обрано, як впливове та репрезентативне, а також — найстаріше публіцистичне діаспорне видання українською мовою. Газету засновано у 1893 році, і дотепер вона виходить безперервно. Видання є офіційним органом Українського Народного Союзу.

Петро Часто пише: ««Свобода» — це досить точне, бо надпартійне, надконфесійне загальногромадське дзеркало, у якому відбилися і відбиваються ще й нині всі грані буття усіх чотирьох хвиль української еміграції за океан, всі процеси формування організованого українського життя на Північно-американському континенті і, врешті, процеси перетворення української еміграції

в діяспору»[66]. Без сумніву, газета відбиває і явища, притаманні українській мові діаспори, зокрема правописні, словотвірні та стилістичні її особливості.

У попередньому розділі ми згадували про дослідження мови на основі газети «Свобода», проте нам невідомі дослідження її мови із застосуванням корпусного методу. Відтак, ця робота стає першою спробою дослідження мови публіцистики української діаспори, та порівняння її із відповідним материковим варіантом із використанням можливостей комп'ютерної та корпусної лінгвістики.

3.2 Підготовка текстів «Свободи»

Підготовча частина дослідження складалась з опрацювання всіх текстів газети «Свобода» за 2016-2019 роки. Тексти були автоматично стягнуті з електронного архіву, розміщеного на офіційному сайті газети svoboda-news.com.

Початково кожен текст містився в окремому файлі у форматі html. Для того, щоб конвертувати файли у текстовий формат (txt), було використано програму «BootCat»[67]. Потім за допомогою інструменту «Combine files» програми «Total Commander» всі тексти за один рік було об'єднано в текстовий файл — так ми отримали чотири файли, кожен із яких містив тексти всіх статей газети «Свобода» за окремий рік. Вигляд текстів до попереднього опрацювання: Додаток 1, зображення 1.

Після об'єднання файлів тексти було очищено від частин, що не є безпосередньо частинами тексту статей. Для цього було використано інструментарій програми «Notepad++», яка дає змогу здійснювати масовий пошук та заміну елементів, використовуючи регулярні вирази, або звичайний пошук. Повторювані елементи, такі як «*Like Друк Email*», що є частиною зовнішнього інтерфейсу сайту «Свободи» було вилучено за допомогою масової заміни. Окрім того, ми вручну вилучили імена авторів статей та первинні джерела, які не є корисним елементом для текстів у корпусі. У текстах їх можна було знайти за

[66] Петро Часто. Вільне слово Американської України. Нью-Йорк — Ужгород 2012. — с 18.
[67] BootCat — Режим доступу: https://bootcat.dipintra.it/

знаком «|», наприклад: «*Світлана Орел* |», або «*Радіо Свобода* |» (див. Додаток 1, Рисунок 2, рядок 76). Подібним чином вилучено й посилання на авторів світлин, вміщених у статтях (наприклад: «*(Фото: Георгій Лук'янчук)*» (див. Додаток 1, Рисунокм2, рядок 69)) та посилання (див. Додаток 1, Рисунок 2, рядки 73-74).

У корпусі ГРАК публіцистичні тексти розділені між двома стилями: JOU (публіцистичний) та SPO (розмовний). Останній містить тексти інтерв'ю — вони відділені від решти, адже є зразками усного мовлення (хай і відредагованого). Зважаючи на цей принцип, із текстів «Свободи» вилучено всі інтерв'ю. Їх вміщено в окремі файли за роками, і при додаванні в корпус присвоєно мітку стилю SPO.

Після вилучення усіх згаданих елементів, що не відповідали формату корпусу, а також відокремлення текстів інтерв'ю, файли виглядали як у Додатку 1, на Рисунку 3, і були готові для додавання їх у корпус. Оновлений ГРАК-11, що містив оброблені тексти газети «Свобода» за 2016-2019 роки, вийшов 15 лютого 2021 року.

3.3 Створення підкорпусів

Для проведення дослідження було створено два підкорпуси.

Перший — фокусний підкорпус, який можна знайти в інтерфейсах корпусних менеджерів NoSketchEngine[68] та SketchEngine[69] для ГРАКу-11 за назвою «SVOBODA_2016_2019». Для його створення було використано такі параметри:

- DOC.MEDIANAME — «Свобода»;
- DOC.PUBLICATIONYEAR — «2016», «2017», «2018», «2019».

[68] Генеральний регіонально анотований корпус української мови (ГРАК): Grak v.11: Пошук \\ NoSketchEngine — Режим доступу: http://www.parasolcorpus.org/bonito/run.cgi/first?corpname=grac11&reload=1&iquery=&queryselector=iqueryrow&lemma=&phrase=&word=&char=&cql=&default_attr=word&fc_lemword_window_type=both&fc_lemword_wsize=5&fc_lemword=&fc_lemword_type=all&usesubcorp=&fsca_doc.author=&fsca_doc.translator=&fsca_doc.authTrans=&fsca_doc.born=&fsca_doc.title=&fsca_doc.date=&fsca_doc.mediaName=&fsca_doc.locCode=&fsca_doc.region=&fsca_doc.publicationCity=&fsca_doc.publisher=&fsca_doc.publication=&fsca_doc.uri=

[69] Генеральний регіонально анотований корпус української мови (ГРАК): GRAC V.11: dashboard // SketchEngine — Режим доступу: https://parasol.vmguest.uni-jena.de/grac_crystal/#dashboard?corpname=grac11

За заданими параметрами до корпусу увійшли тексти діаспорної газети «Свобода» за 2016-2019 роки, включаючи інтерв'ю. Обсяг підкорпусу текстів «Свободи» — 2 539 857 токенів (~ 1 983 852 слів).

Другий підкорпус — референтний — містить публіцистику України того ж часового проміжку. Назва підкорпусу в інтерфейсах корпусних менеджерів — «UA_NATIONAL_JOU_2016_2019». Параметри, за якими було побудовано корпус:

- DOC.PUBLICATIONYEAR — «2016», «2017», «2018», «2019»
- DOC.MEDIANAME — «Голос України|Дзеркало тижня|Жінка|Освіта|Світ|День|Порадниця|Сільські вісті|Слово|Українська правда|Україна молода|Український тиждень»

До списку видань підкорпусу увійшла низка загальнонаціональних газет та журналів, що представлені у корпусі ГРАК та вміст яких, на нашу думку, співвідносний із вмістом фокусного корпусу текстів газети «Свобода». До підкорпусу не ввійшли регіональні видання. Обсяг отриманого підкорпусу — 25 631 210 токенів (~ 20 020 229 слів).

3.4 Ключові слова

Першим етапом дослідження текстів «Свободи» стало визначення ключових слів (key words). Ключові слова — це ті слова, різниця частот яких між двома корпусами є не лише великою, а й статистично значущою[70] (тобто — не випадковою).

Для того, аби виділити ключові слова було використано інструмент «word list» у корпусному менеджері NoSketchEngine[71]. За його допомогою згенеровано порівняльні списки лем та слів за їхньою частотністю у двох корпусах. На вершинах списків опинилися ті слова чи леми, різниця частот яких між корпусами

[70] Costas Gabrielatos. Keyness Analysis: Nature, metrics and techniques // Corpus approaches to discourse: a critical review. Milton 2018. P. 225–258. — с 229.
[71] Генеральний регіонально анотований корпус української мови (ГРАК): Параметри словника частот. — Режим доступу: http://www.parasolcorpus.org/bonito/run.cgi/wordlist_form?corpname=grac11

є найяскравіше вираженою (див. Додаток 2, рис. 1). Списки впорядковані за показником величини ефекту «відношення зі згладжуванням». Детальніше про метрику йтиметься у наступному підрозділі.

3.5 Статистичні метрики

У цьому дослідженні буде використано два види статистичних метрик: (1) ті, що показують величину ефекту, та (2) ті, що показують статистичну значущість ефекту. Метрики величини ефекту допомагають виявити потенційні ключові слова, натомість метрики статистичної значущості показують, чи знайдені різниці частотностей не є випадковими.

3.5.1 Метрика відношення зі згладжуванням

Метрика величини ефекту, використана у дослідженні має назву відношення зі згладжуванням[72]. Вона показує, наскільки частотнішим є певне слово у фокусному корпусі в порівнянні з його частотністю в референтному корпусі. Відповідно, ми отримуємо можливість знайти слова, особливо притаманні одному з двох порівнюваних корпусів.

Формула для підрахунку показника відношення зі згладжуванням:

$$\frac{f\ pm_{\ rm\ focus}\ + N}{f\ pm_{\ rm\ ref}\ + N}$$

Де $f\ pm_{\ rm\ focus}$ — це відносна частотність слова (на мільйон) у фокусному корпусі; $f\ pm_{\ rm\ ref}$ — це відносна частотність слова (на мільйон) у референтному корпусі, а N — це коефіцієнт згладжування.

Коефіцієнт згладжування за замовчуванням встановлюється як 1. Коли ми генеруємо частотний список слів для порівняння їх у двох корпусах, то при значенні коефіцієнта згладжування 1 на вершині списку опиняються слова із низькою частотою вживання. Для того, аби отримати результати із більш

[72] Simple maths // SketchEngine — Режим доступу: https://www.sketchengine.eu/documentation/simple-maths/

частотними словами корпусів, коефіцієнт згладжування потрібно збільшувати до значення 10, 100, 1 000, 10 000 і т.д. Визначення найбільш влучного значення коефіцієнту згладжування залежить від параметрів порівнюваних корпусів та мети дослідників (отримати у згенерованому списку більш частотні чи більш рідкісні слова).

Також для отримання якісних списків слів за показниками відношення зі згладжуванням важливо, аби порівнювані корпуси були співвідносними за всіма характеристиками, окрім досліджуваної[73]. Відповідно, обидва створені нами підкорпуси містять публіцистичні тексти за той самий часовий проміжок, а різницею (яка мотивує це дослідження) є походження цих текстів із діаспори чи материкової України.

3.5.2 Метрики статистичної значущості

Показники величини ефекту показують, наскільки великою є різниця частотності двох досліджуваних одиниць. На противагу їм метрики статистичної значущості допомагають зрозуміти міру невипадковості цих даних. Іншими словами, метрики статистичної значущості дають змогу оцінити, наскільки ймовірними є отримані дані за умови правдивості нульової гіпотези. У нашому випадку нульова гіпотеза зводиться до твердження, що вживання мовної одиниці не залежить від того, чи її використовують у діаспорі чи в материковій Україні[74].

Показники метрик статистичної значущості співвідносні зі значеннями **p-значення (значення ймовірності)**. Що нижче p-значення, то більш статистично значущими є результати[75]. У корпусній лінгвістиці прийнято, що якщо p-значення менше 0.01[76], то є достатня ймовірність правдивості спостереженого явища за істинності нульової гіпотези.

[73] Adam Kilgarriff. Simple maths for keywords. — Режим доступу: https://www.sketchengine.eu/wp-content/uploads/2015/04/2009-Simple-maths-for-keywords.pdf , — с 1.
[74] Costas Gabrielatos. Keyness Analysis: Nature, metrics and techniques, — с 230.
[75] Там само, — с 231.
[76] Там само, — с 239.

До метрик статистичної значущості належать хі-квадрат (chi-square або χ^2), точний критерій Фішера (Fisher exact test), логаритмічна правдоподібність (log-likelihood), Баєсів інформаційний критерій (Bayesian Information Criterion) та інші. Далі буде подано коротку характеристику кожної з перерахованих метрик.

Метрику **хі-квадрат** (χ^2 або chi-square) використовують для порівняння двох чи більше взаємозамінних слів у принаймні двох корпусах. Тест хі-квадрат має два обмеження: (1) дані для підрахунку не мають містити нульових частот, та (2) не більше чверті досліджуваних даних можуть мати значення менше за п'ять[77]. Отже, метрика добре працює із великими вибірками слів.

Для малих вибірок застосовують **точний критерій Фішера** (Fisher exact test). Обмеження щодо застосування критерію полягає у складності розрахунку, адже його формула включає розрахунок факторіалів[78].

Метрика логаритмічної правдоподібності (log-likelihood, LL) дозволяє оцінити різницю частотності одного слова у двох корпусах.

Формула розрахунку LL така:

$$LL = 2\,(O_{11} * ln\,\frac{O_{11}}{E_{11}} + O_{21} * ln\,\frac{O_{21}}{E_{21}})$$

де O_{11}— це спостережувана частотність досліджуваного феномена (слова, речення, і т.п.) у фокусному корпусі; O_{21}— це спостережувана частотність досліджуваного феномена референтному корпусі ; E_{11}та E_{21}— це очікувані частотності досліджуваного феномена у фокусному та референтному корпусах відповідно, за умови, що немає різниці у частотності цього феномена між двома корпусами[79].

Для розрахунку LL використовуємо онлайн-калькулятор «Log-likelihood and effect size calculator»[80]. У таблицю вводимо абсолютні частоти досліджуваного феномену у фокусному та референтному корпусах, а також кількості випадків, у яких досліджуваний феномен міг з'явитись у відповідних корпусах (якщо

[77] Anatol Stefanowitsch. Corpus linguistics: A guide to the methodology, — с 177.
[78] Там само, — с 228-229.
[79] Vaclav Brezina. Statistics in Corpus Linguistics: A Practical Guide. Cambridge — New York 2018. — с. 84.
[80] Log-likelihood and effect size calculator — Режим доступу: http://ucrel.lancs.ac.uk/llwizard.html

досліджуємо частотність слова, то це — загальна кількість слів у корпусі)[81]. Для того, аби p-значення було меншим за 0.01, значення логаритмічної правдоподібності має бути більшим за 6.63, а меншим за 0.001 p-значення буде за LL 10.83[82].

Баєсів інформаційний критерій (БІК, англ. Bayesian Information Criterion, BIC) є спрощенням метрики під назвою Баєсів фактор. Для вирахування Баєсового інформаційного критерію при аналізі ключовості (міри того, наскільки ключовим є слово) необхідне значення логаритмічної правдоподібності для різниці частот досліджуваного слова у порівнюваних корпусах (LL) та сума обсягів порівнюваних корпусів (N)[83]:

$$BIC \approx LL - log\ (N)$$

Значення БІК інтерпретують, як кількісний показник ступеня переконливості доказів проти нульової гіпотези[84]. У випадку, якщо значення БІК менше за 2, вважається, що результати не варті особливої уваги. БІК від 2 до 6 позначає наявність достатніх доказів проти нульової гіпотези (у такому випадку p-значення = 0,00018, а LL = 13,98). Значення БІК від 6 до 10 позначає велику, а більше ніж 10 – дуже велику переконливість доказів проти нульової гіпотези.

БІК має перевагу над поширеною метрикою логаритмічної правдоподібності. При використанні останньої для великих за розміром корпусів може виявитися, що чи не всі слова є ключовими (мають значення LL більше за встановлений поріг). На противагу, значення Баєсового інформаційного критерію не залежать від розмірів корпусів, а отже є інструментом визначення ключовості з ширшою сферою застосування. Саме тому при відбиранні ключових слів, які є статистично значущими, ми будемо взоруватися на значення Баєсового інформаційного критерію.

---

[81] Tony McEnery, Andrew Hardie. Statistics in corpus linguistics. Cambridge 2012 // — Режим доступу: http://corpora.lancs.ac.uk/clmtp/2-stat.php

[82] Log-likelihood and effect size calculator — Режим доступу: http://ucrel.lancs.ac.uk/llwizard.html

[83] Andrew Wilson. Embracing Bayes factors for key item analysis in corpus linguistics // New Approaches to the Study of Linguistic Variability: Language Competence and Language Awareness in Europe, vol. 4. Frankfurt 2013. P. 3–11. – с 6.

[84] Там само,— с 6.

3.6 Відбір й аналіз ключових слів

У нашому дослідженні особливу увагу буде надано аналізу та порівнянню лексики двох корпусів. Це дасть змогу помітити відмінності як на рівні правопису, так і у слововживанні.

Для того, аби отримати список потенційно ключових слів, було використано інструмент word list корпусного менеджера NoSketchEngine. Щоб охопити максимально широкий спектр потенційно ключових слів було згенеровано списки слів і лем. Списки лем можуть містити непередбачувані результати у тих випадках, де існує омонімія лем, проте будуть більш точними при оцінюванні кількостей вживань слів (адже враховують зведені результати вживання різних відмінків, чисел тощо). Списки слів дадуть можливість побачити відмінності на рівні граматики (наприклад, вживання закінчень родового відмінка *-а*/*-я* чи *-у*/*-ю*).

Частотні списки було згенеровано також із вказуванням різного коефіцієнта згладжування: від 1 до 10 000. Таким чином нам вдасться охопити як більш рідковживані, так і дуже частотні слова корпусів.

Для того, аби знайти слова, притаманні референтному підкорпусу, але менш притаманні фокусному (тобто ті, яких, можливо, уникають у текстах фокусного підкорпусу), при генеруванні списку слів було замінено місцями референтний і фокусний корпуси. Так, у випадку, коли в інструменті обрано корпус текстів материка як фокусний, а текстів «Свободи», як референтний, нам вдається побачити слова, які в останньому трапляються рідше ніж в материкових текстах.

Найпоказовішим є частотний список лем для корпусу текстів «Свободи» із встановленим коефіцієнтом згладжування 1 (див. Додаток 2, рис. 1). Він показує слова, які відповідно до метрики відношення зі згладжуванням у діаспорних текстах є набагато частотнішими, ніж у материкових. У списку знаходимо особливо багато прикладів слів, правопис яких між корпусами відрізняється. Це можна пояснити тим, що газета «Свобода» (як і ще багато діаспорних видань)

послуговується харківським правописом 1928-1929 року, відповідно, і правописним словником Голоскевича.

Найбільший показник — у леми *відс.*. Це пояснюється тим, що газета «Свобода» не використовує знака відсотка: %, а натомість вживає скорочення *відс.*.

Цікавіше проаналізувати наступні леми списку.

Чи не найпомітнішою особливістю є велика кількість власних назв, зокрема тих, написання яких відрізняється від усталеного в Україні. Серед них: *Ню-Йорк, Европа, Ню-Джерзі, Вашінґтон, Філядельфія, Чикаго, Олександер, ЕС (на позначення ЄС — Європейський Союз), Фльорида, Еспанія, Евросоюз*. Варто розуміти, що велика кількість вживань власних назв міст та штатів у корпусі текстів «Свободи» зумовлена структурою її текстів: кожна новинна стаття на початку містить вказівку на місто чи штат, про подію в якому йтиметься. Ці вказівки написані великими літерами, коли йдеться про місто (*НЮ-ЙОРК, ЧИКАҐО, ВАШІНҐТОН*); у випадку коли вказано місто і штат перше написано великими літерами, а друге — звичайним чином (*КЕРГОНКСОН, Ню-Йорк*).

Проаналізуймо детальніше іменник *Ню-Йорк*. Слово зустрічається у фокусному корпусі «Свободи» 1 110 разів (437,03 на мільйон), а у референтному корпусі – 0 разів. У такому випадку БІК для цього слова становить 5 324,83, а LL — 5 341,74. Це дуже високі показники. Проаналізуймо паралельну форму *Нью-Йорк*. Вона трапляється у текстах фокусного підкорпусу лише двічі (0.79 на мільйон), і 941 (36.71 на мільйон) раз — у текстах «Свободи». Тоді БІК становить 141,92, і вказує на те, що ця форма притаманна текстам материкової України.

Цікаво, що назву міста набагато частіше вживають у текстах діаспори. Порівнявши відносні частоти слова Ню-Йорк для діаспори та, відповідно, Нью-Йорк для України, дізнаємось, що власну назву майже в 12 разів частіше вживають у текстах фокусного підкорпусу. А зумовлене це тим, що місто є важливим економічним, політичним та культурним центром, а також одним із найбільших міст Сполучених Штатів Америки. Відтак, у «Свободі», яка видається у США, місто згадують набагато частіше, аніж у публіцистичних текстах материкової України.

До переліку власних назв увійшло і слово *Европа* із його похідними *ЕС*, *Євросоюз*, прикметником *європейський*, *евро*, *Євромайдан*, *євроатлантичний*, *європеєць*, *Євробачення*, *Європарлямент*, *євроінтеграція*, а також — *західньоєвропейський*, *загальноєвропейський*, *євроінтеграційний*, *антиєвропейський*, *Єврокомісія*, *проєвропейський*, *східньо-європейський*. Такий варіант написання із використанням літери *е* пропонований у словнику Голосекевича[85]. Напротивагу цим словам у списку частотностей, де як фокусний корпус використано тексти материкової України, а як референтний — тексти Свободи, і коефіцієнт згладжування становить 1, на вершині опиняються слова, яких у материкових трапляються значно частіше, і серед них знаходимо слова *Європа*, *європейський*, *Євросоюз*, *євро*, і т.д. (див. Додаток 2, рис. 2).

Порівняймо кількість вживань слів із коренем *-европ-* в текстах «Свободи» і та слів із коренем *-європ-* у текстах референтного корпусу. Отримуємо такий розподіл: у «Свободі» 2 596 (1 022,10 на мільйон) випадків вживання кореня *-европ-* і 5 (1,97 на мільйон) випадків вживання кореня *-європ-*. У текстах материка — 72 (2,81 на мільйон) випадки вживання кореня *-европ-*, і 28 537 (1 113,40 на мільйон) випадків вживання кореня *-європ-*.

Варто зазначити, що результати мають невелику похибку, адже до конкордансів потрапила й невелика кількість слів які містять шукані корені, як от *невропатолог*.

Такий розподіл частотностей дає БІК 11 698,95 для кореня *-европ-* у текстах діаспори в порівнянні з текстами материка, і БІК 5 303,28 для кореня *-європ-* у текстах материка, порівнюючи їх з текстами з діаспори. За результатами бачимо, що згідно зі статистичними даними форма *Европа* із похідними сильно притаманна діаспорній публіцистиці, а форма *Європа* із похідними — публіцистиці материкової України.

[85] Григорій Голоскевич. Правописний словник. Київ, 1929. — Режим доступу: https://r2u.org.ua/data/%D0%9F%D1%80%D0%B0%D0%B2%D0%BE%D0%BF%D0%B8%D1%81%D0%BD%D0%B8%D0%B9%20%D1%81%D0%BB%D0%BE%D0%B2%D0%BD%D0%B8%D0%BA%20(1929).pdf , — с 154.

Згідно із Правописом 1928-1929 років, в іншомовних словах «дифтонг au й ou передаємо через ав (яв) і ов»[86]. У згенерованому списку слів знаходимо такі слова на підтвердження використання цього правила редакцією «Свободи»: *лавреат, Савт-Бавн-Брук, скавтський, авкціон, Апелбавм (Applebaum), авдиторія, Савдівська Аравія, скавт, Савт (South), Авшвіц, Кавка (Каука), авдієнція, лавреатка, Бравн, заввага, Бавнд, авдит, автизм*, тощо. Слово *автизм* (аутизм) трапляється 6 разів у текстах з діаспори й жодного разу — в референтному підкорпусі. Це дає йому БІК 11,97, що є достатнім, аби підтвердити невипадковість явища. Відповідно, інші перелічені слова теж мають достатній показник, аби бути притаманними текстам «Свободи» (адже вони вказані у порядку спадання їх показників відношення зі згладжуванням).

У списку частотностей за текстами діаспори (див. Додаток 2 рис. 1) у верхній частині опинилось слово *Вашінґтон* у формі, правопис якої відповідає словнику Голоскевича[87]. Високу частотність слова можна пояснити тим, що місто є столицею Сполучених Штатів Америки, а отже часто фігурує в новинах. *Вашінґтон* використовується як вказівка на місце, про яке йтиметься в окремій новині «Свободи» (тоді слово написане великими літерами), а також в самих текстах новин — у прямому значенні (наприклад: «*відкриття пам'ятника Т. Шевченкові у Вашінґтоні за часів Президента США Двайта Айзенгавера…*»), та як метонімія на позначення уряду (наприклад: «*Нібито Вашінґтон вже порозумівся з Москвою щодо українського питання…*»). Окрім того, поодинокі випадки вживання слова включають згадки американського державного діяча і президента Джорджа Вашінґтона, а також — посилання «Свободи» на новини із відомої американської газети «Вашінґтон Пост» («The Washington Post»).

Усього у фокусному підкорпусі знаходимо 962 випадки вживання слова (378,76 на мільйон) або його похідних (наприклад, *вашінґтонський*) і ще 8 випадків із літерою г замість ґ. Натомість у референтному підкорпусі такого правопису слова

[86] Український правопис, видання перше. Харків 1929. — Режим доступу: https://r2u.org.ua/data/other/%D0%9F%D1%80%D0%B0%D0%B2%D0%BE%D0%BF%D0%B8%D1%81-1928.pdf – параграф 70.
[87] Григорій Голоскевич. Правописний словник, — с 54.

немає. Такі дані дають показник БІК 4 612,60. Натомість українська преса використовує прийнятий варіант *Вашингтон* (1 326 випадки у референтному корпусі, або 51,73 випадки на мільйон).

Варте уваги те, що на вершині аналізованого списку опинилося слово *Київ*. Інтуїтивно здається, що в українській пресі назву столиці мають вживати не суттєво менше, аніж в діаспорній «Свободі». Для того, аби зрозуміти звідки різниця частотностей, здійснимо аналіз конкордансів підкорпусів.

Назва столиці України у всіх відмінках разом трапляється 3 018 разів (1 188,26 на мільйон) у текстах «Свободи» та 11 249 разів (438,88 на мільйон) в українській публіцистиці. Очевидно, що відносна частота суттєво відрізняється. Проте варто виключити ті випадки, де вживання слова зумовлене структурою статей (1 457 разів трапляється форма КИЇВ). Тоді у «Свободі» залишається 1 561 слововживання (786,85 випадків на мільйон). Значення БІК у такому випадку становить 125,89, а отже різниця залишається статистично значущою. Її можна пояснити частішим згадуванням у «Свободі» слова *Київ* у метонімічній функції, наприклад: «*Київ і Захід не визнають анексію*» або: «*Київ захищає Меджліс…*». Можна також припустити, що не всі українські ЗМІ, що увійшли до референтного корпусу є новинними, а отже вони рідше за «Свободу» згадують Київ.

Багато слів у згенерованих списках опинилися тому, що їхнє написання відрізняється від прийнятого в Україні. Це зумовлено використанням діаспорою Правопису 1928-1929 років та, відповідно, Правописного словника Голоскевича 1929 року.

Багато слів, що потрапили до списків ключових слів належать до іншомовних і є наслідком того чи іншого способу транслітерації чи транскрипції їх з оригіналу. Раніше ми описували передавання дифтонгів au та ou. Варто також виділити слова, що містять типове для грецької мови сполучення голосних (дифтонг) ία. В материковому варіанті української ці слова найчастіше містять сполучення *іа*, а в діаспорному — *ія*. Зі списку можна вилучити такі слова, що є прикладами цього розрізнення (подано у порядку, в якому вони розміщені в згенерованих списках): *діяспора, патріярх, матеріял, соціяльний, патріярхат,*

ініціятива, спеціяльний, асоціяція, територіяльний, меморіяльний, меморіял, ініціятор, діялог, матеріяльний, варіянт, ґалерія, фортепіяно, потенціял, олімпіяда, патріярший, спеціяліст, спеціяльно, діяспорний, авіяція, соціялістичний, піяніст, авіяційний, медія і так далі.

Всі ці слова відрізняються від материкового варіанта написанням і є притаманними публіцистичному стилю, тому їх можна назвати особливістю діаспорного варіанта української, базуючись на їхньому високому показнику (score) у списках, згенерованих за метрикою відношення зі згладжуванням (за яким і впорядковуються автоматично згенеровані списки). Цей показник у перелічених словах становить 267,2 для слова *діяспора* і спадає, становлячи 17,9 для слова *медія* (зазначимо, що коли цей показник близький до одиниці, то різниці у вживанні слів у корпусах немає).

Для того, аби переконатись, що частоти всіх перелічених слів є статистично значущими та притаманними корпусу діаспорних текстів, підрахуймо БІК для слова *медія*, яке має найменший показник метрики відношення зі згладжуванням серед них. У текстах діаспори слово у різних відмінках трапляється 45 разів (17,72 на мільйон), а в українській пресі — 24 рази (0,94 на мільйон). Баєсів інформаційний критерій для слова становить 115,02, а отже належність слова до мови діаспори не є випадковою (як і слів, що передують йому у списку).

Цікавою особливістю фокусного підкорпусу є вживання м’якого л в запозичених словах[88]. До них належать такі слова: *парлямент, Філядельфія, плян, парляментський, плянувати, Фльорида, декларація, бльок, плятформа, ісламський, капелян, клясичний, заплянований, доляр, плянета, капеля, плянуватися, плянування, забльокувати, мельодія, фолькльор, бльокувати, рекляма, лябораторія, бльокада, алькоголь, Голяндія, плякат, проклямація, деклямувати, голяндський, Філядельфійський, Нідерлянди* тощо. Зважаючи що слова перелічені у порядку спадання їхнього показника за відношенням по згладжуванні, підрахуємо БІК для останнього слова — *Нідерлянди*, аби зрозуміти, чи воно, а отже, і попередні слова

[88] Український правопис, видання перше, — параграф 54.

є статистично значущими. Слово трапляється 60 разів (23,62 на мільйон) у корпусі діаспори, і жодного разу у референтному корпусі. БІК становить 271,84, а отже перелічений список можна вважати притаманним діаспорним текстам.

Зі згенерованих списків можна виділити й вживання літери ґ там, де в материковому варіанті сучасної української мови вживають літеру г. Це слова *Вашінґтон, реґіон, конґрес, Чикаго, аґресія, делеґація, еміґрація, аґент, дириґент, конґресовий, аґенція, пропаґанда, делеґат, еміґрант, аґресор, нелеґальний, бриґада, неґативний, ліґа, колеґа, реґіональний, конґресмен, КҐБ (?), ґалерія, Аргентина, інтеґрація, інтеліґенція, енергія, міґрант, аґенство, ориґінальний, зрезиґнувати, ґвардія, ґарантія, аґрарний, врегулювання, аґресивний, редаґувати, ґарантувати, аргумент, леґенда, Гонґ-Конґ, міґрація, енергетичний, прогрес, енергетика, леґендарний* тощо.

Харківський правопис 1928-1929 років містить таке правило, щодо написання іншомовних слів: «Чуже h передаємо нашим г, щождо чужого g, то в новіших запозиченнях його треба передавати через ґ, у запозиченнях же засвоєних давніше, особливо з грецької мови, віддаємо нашим г»[89]. Проаналізуймо слова із коренем *-леґенд-* (слово легендарний має найменший показник за метрикою відношення зі згладжуванням у наведеному переліку: 21,8) трапляються у корпусі діаспори 122 рази, а у корпусі материкової України — всього раз. Це дає БІК 558,78, що є дуже високим показником.

Ще одна цікава особливість — написання імені *Олександер*. БІК для цього явища становить 2 254,55. Схожа тенденція й у слова *міністер*, БІК для якого становить 2 500,06. Обидва слова знаходимо у словнику Голоскевича саме в такій формі написання.

Прикладом відмінного написання слова у мовленні діаспори є прислівник *покищо*. У словнику Голоскевича знаходимо таку статтю: «*по́кищо, присл.; але поки що́ (що́сь). По́кищо нічо́го не дав. Поки що́ тра́питься*». Написання разом трапляється у фокусному підкорпусі 176 (69.30 на мільйон) разів, і лише раз — у

[89] Український правопис, видання перше, — параграф 55.

референтному. Таким чином БІК складає 817,91. Натомість роздільне написання *поки що* трапляється лише раз у текстах діаспори, і 3 252 рази у ЗМІ материка, і має БІК 584,26.

У списках ключових слів, згенерованих за словоформами знаходимо морфологічні особливості мови української західної діаспори, що проявляються у їх відмінюванні. Знаходимо форму родового відмінка *Ізраїля*, що притаманна фокусному корпусу, і трапляється там 214 (84,26 на мільйон) разів, і лише 7 (0,27 на мільйон) разів у текстах материка. БІК для цієї словоформи становить 952,16.

Варта уваги й форма відмінювання *Путіном*, притаманна текстам діаспори. БІК для неї становить 530,99. Напротивагу, форма орудного відмінка *Путіним* переважає у текстах материка, проте БІК для неї значно менший: 70,08. Форма *Путіним* на материку вживається відповідно до чинного правопису[90], а діаспора, очевидно, відмінює його відповідно до правопису 1929 року, який вказує як зразок відмінювання чоловічого прикметникового прізвища форму *Бутвином* (для прізвища *Бутвин*)[91].

Знаходимо у текстах «Свободи» й інший погляд на відмінювання деяких слів, зокрема, іншомовних, які заведено вважати невідмінюваними в сучасній українській мові. Яскравим прикладом є слово *бюро*, яке Григорій Голоскевич у своєму словнику фіксує як відмінюване. Провівши пошук можливих форм непрямих відмінків слова у фокусному підкорпусі знаходимо 147 (57.88 на мільйон) слововживань. Той самий запит знаходить лише 2 (0.08 на мільйон) результати у корпусі текстів материка. БІК для явища достатньо високий, і становить 669, 67.

Яскраво виражена і тенденція використовувати закінчення *-и* для іменників жіночого роду на приголосну та -ть[92]. Це, зокрема, словоформи *незалежости, области, діяльности, смерти, Гідности, участи, творчости, імени, цілісности, відповідальности, кількости, справедливости, державности, любови,*

[90] Український правопис. Київ 2019. — параграф 85.
[91] Український правопис, видання перше, — параграф 78.
[92] Там само, — параграф 25.

ідентичности тощо. Словоформа *ідентичности* трапляється у фокусному підкорпусі 78 (30.71 на мільйон) разів, і лише раз — у референтному. БІК для неї становить 347,92.

Ще одна особливість діаспорної мови — це наявність таких слів, як *роля, медаля, заля, візита, аналіза, оркестра*, тощо. Ці слова знаходимо у словнику Голоскевича. Пошуки в корпусі вимагають врахування омонімії словоформ (наприклад, родовий відмінок однини слова *візита — візити*, і називний відмінок множини слова *візит — візити*). Так, відкинувши можливі випадки омонімії, отримуємо такі результати. БІК для слова *роля* становить 1411,84, *медаля* — 550,95, *заля* — 549,18, *візита* — 368,08, *аналіза* — 161,15, *оркестра* — 354,47. Всі слова мають достатню статистичну значущість, аби сказати, що вони притаманні саме діаспорному підкорпусу.

Згенеровані списки слів містять деякі прикметники на *-овий*, не властиві материковій мові. І хоча в сучасній українській мові є такі прикметники (напр. *малиновий, зірковий, пластиковий*), у діаспорних текстах знаходимо і такі форми, що мало вживаються на материку (*конґресовий, кредитовий, стейтовий, пластовий, спортовий*).

Проаналізуймо вживання форми *спортовий*, та *пластовий*. Всі вони більш притаманні фокусному корпусу і мають БІК 530,99 та 1403,53 відповідно. Для порівняння підрахуймо БІК для прикметника, на *-овий*, що вважається нормативним на материку на материку: слова *зірковий*. Воно трапляється 6 разів у фокусному, та 430 разів у референтному підкорпусі, і має БІК 29,88, що хоч і не є надто великим показником, проте підтверджує його приналежність до корпусу материкових текстів, на відміну від форм *конґресовий, кредитовий, стейтовий, пластовий, спортовий.*

Наявність в мові діаспори прикметників на *-овий*, що не вживаються активно в мові материка можна пояснити тим, що, по-перше, творення прикметників на *-овий* було призупинене спробами радянської політики репресувати відповідний суфікс як запозичення з польської мови, що спричинило різке зменшення використання «небажаних» слів, хоч і не зникнення форм, що вже

активно використовувались в мові. По-друге, у текстах ЗМІ часто фігурують власні назви, а у діаспорних ЗМІ, зокрема *Український конґресовий комітет Америки* (УККА), *Українська кредитова спілка* тощо, що додає частотності для цих слів у корпусах.

Серед ключових слів також знаходимо слово *стейт*, та відповідний прикметник *стейтовий*. *Стейт* у діаспорних текстах вживається на заміну слова *штат*, а *стейтовий* — на позначення приналежності до штату, наприклад, «*стейтовий сенатор*» — сенатор штату, «*Пенсильванський стейтовий Університет*» — Університет штату Пенсильванія тощо. БІК для слів *стейт* та *стейтовий* разом становить 1097,39.

Ще одне слово, не притаманне українській мові материка — *кредитівка*. Воно позначає кредитну спілку, і зазначено у чернетці Російсько-українського словника складної лексики авторства Караванського як галицизм[93]. Наявність у мові діаспори галицизмів відзначає низка дослідників, зокрема Богдан Ажнюк, про доробок якого йшлося у попередньому розділі. Слово *кредитівка* трапляється 161 (63.39 на мільйон) раз у фокусному підкорпусі, і жодного — у референтному. БІК для нього становить 757,89.

Більш притаманне мові діаспори й вживання літери *т* на місці грецької літери θ (тети). Це насамперед слова *катедра* (університету), *мітґ, Атени, етер*. БІК для останнього становить 127,52.

Цікава тенденція і з вживанням слова *вояк*. Якщо в українських текстах материка його нечасто вживають до опису сучасних подій, то в «Свободі» воно активно використовується, коли йдеться про війну на сході України тощо. Це бачимо із таких прикладів текстів «Свободи»: «*поки українські вояки стоять неприступною стіною проти ворога на сході, країну тихцем, в напівтаємних політичних торгах, зраджує її власна влада*», — або : «*Військовий медик з Мишиґену, котрий кілька разів ризикував своїм життям за своїх побратимів у*

[93] Святослав Караванський. Російсько-український словник складної лексики. 2012. Кредитівка. — Режим доступу: https://r2u.org.ua/s?w=%D0%BA%D1%80%D0%B5%D0%B4%D0%B8%D1%82%D1%96%D0%B2%D0%BA%D0%B0&scope=all&dicts=all&highlight=on

В'єтнамській війні, став першим вояком кому Президент Дональд Трамп причепив Медалю Гонору». Показник БІК для слова вояк дуже високий і становить 1121,50. Напротивагу слову *вояк*, українські материкові ЗМІ частіше використовують слово *воїн*, що має БІК 250,30. Слово *солдат* теж більше використовується на материку, але його БІК 22,85 порівняно невисокий, хоча й задовольняє критерій статистичної значущості. А більш формальне слово *військовослужбовець* не є притаманним жодному із корпусів і має від'ємний БІК -16,51.

У частотних списках знаходимо слово *вислід*. За електронною версією «Російсько-українського словника складної лексики» Караванського, це галицизм, що є синонімом до слова *результат*[94]. У фокусному підкорпусі знаходимо такі приклади вживання: «*Отож висліди голосування за те, як оцінювати діяльність уряду за 2015 рік, були вповні передбачними*», або «*Служба безпеки України спромоглся, у висліді відчайдушньої операції, заарештувати і вивезти до Києва російського терориста Володимира Цемаха*». Пошук у корпусі й підрахунок БІК для кожного слова показує, що слово *вислід* більш притаманне діаспорній мові (БІК 998,27), а *результат* — материковій, з БІК 464,41.

Частіше в діаспорних текстах вживається і слово *летовище*, що утворене засобами української мови, напротивагу іншомовному слову *аеропорт*. Перше має БІК 554,32, і характерне для текстів діаспори, а друге — *аеропорт* — навпаки, більше вживається в материкових текстах і має БІК 267,92.

Цікаво, що в частотних списках опинилося слово *провадити*. У фокусному підкорпусі знаходимо дуже різноманітні приклади його використання: «*Зборами провадила президія у складі*», «*вона працювала на Кубі, а також провадила концертну діяльність, виступала на фестивалях*». У референтному корпусі слово провадити найчастіше використовується у словосполученні «*провадити політику*», наприклад: «*я з величезною повагою ставлюся до тих музикантів, які вирішили не виступати в Росії, принаймні поки вона провадить таку агресивну*

[94] Святослав Караванський. Російсько-український словник складної лексики. 2012. Вислід. — Режим доступу: https://r2u.org.ua/s?w=%D0%B2%D0%B8%D1%81%D0%BB%D1%96%D0%B4&scope=all&dicts=all&highlight=on

політику». БІК для слова провадити становить 334,68, і показує, що воно характерне саме для мови діаспори.

Цікаво, що й однокореневе до попереднього слово *провідник* теж вирізняється як притаманне фокусному підкорпусу. БІК для нього дуже високий і становить 2056,52. На противагу, паралельна форма *лідер* більш характерна для мови материка, і має БІК 1112,16. Проаналізувавши контексти вживання слова *провідник* можна побачити, що його вживають там, де українські ЗМІ більш схильні використовувати слово лідер. Наприклад, у фокусному підкорпусі знаходимо словосполучення: «*провідник радикалів Олег Ляшко*», і всього 6 разів, коли Олега Ляшка називають провідником, і жодного разу — лідером. Натомість у референтному корпусі — навпаки знаходимо 182 випадки вживання цього імені зі словом *лідер*, і жодного — зі словом *провідник*. Можливо, така преференція зумовлена небажанням діаспори використовувати кальки з англійської тоді, коли у слова є питомий відповідник в українській мові, тоді як українські ЗМІ часто несвідомо вживають англіцизми, і рідше — питомі форми таких слів.

Частотні списки дають змогу побачити й ті слова, які не видаються особливими для того чи іншого корпусу без використання методів корпусної лінгвістики. Одним із таких слів є дієслово *прибувати* у його доконаній формі *прибути*. З конкордансу фокусного підкорпусу бачимо, що діаспорні ЗМІ частіше вживають його у випадках, де їх колеги на материку обирають слово *приїхати, завітати*, рідше — *прийти*. Наприклад, у текстах «Свободи» знаходимо: «*відбувши строк покарання, вийшов з колонії в Астрахані й прибув в Україну*». БІК для слова становить 77,98.

Цікавою знахідкою згенерованих списків є і слова, що в материковій українській поступово були замінені іншими. Це, наприклад, слово *осідок*. Академічний тлумачний словник української мови 1970-1980 років пояснює слово як «*постійне місце перебування, розташування кого-, чого-небудь*» і подає синонім *резиденція*[95]. На жаль, слово не має великої частотності у досліджуваних корпусах

[95]Академічний тлумачний словник української мови (СУМ). 1970–1980. Осідок. — Режим доступу: http://sum.in.ua/s/osidok

(трапляється 25 разів у фокусному і 20 разів у референтному), проте його БІК 22,24 достатній, аби говорити про те, що його більше використовують в текстах діаспори. При цьому відповідні йому слова *резиденція* та *штаб-квартира* не можна назвати притаманними якомусь із корпусів, БІК для них становлять -3,45, і -10,37, відповідно.

Слово *оселя* теж статистично притаманне діаспорним текстам, його БІК становить 307,53. Так само і зі словом *осередок*, БІК якого 516,35. Слово *осередок* на материку може бути замінене словом *центр*, в іншому значенні — *частина, відділ, представництво* тощо. Наприклад, у «Свободі» знаходимо «*Важливим духовним осередком села був храм Чесного і Животворячого Хреста Господнього*», або «*у 1921-1922 роках створював повстанські осередки по селах Уманщини*».

Знаходимо й цікаве слово *залога*. «*Коли російська залога, яка вийшла зі стін фортеці, знищила Запорізьку Січ*», або ж «*Він 50 років служив на фльоті, очолював залоги пасажирських суден*», чи «*залога літака, на якому він був, не змогла собі дати ради з пасажиром, який поводився шалено*». З Академічного тлумачного словника української мови 1970-1980 років[96] та контексту бачимо, що слово *залога* може відповідати словам *засідка*, *охорона,* або ж *гарнізон*; *екіпаж*, чи *команда*. Тож слово *залога* є відповідником до вживаного на материку іншомовного слова *екіпаж*, і могло б його замінити, адже цілком вписується у реалії сьогодення («*Залога МКС тепер складається з американця, німця й росіянина*»). БІК для нього становить 270,88.

Ще одне слово, що майже не функціонує на материку, — *славень*. Слово нечасте у досліджуваних корпусах, воно трапляється 11 разів у фокусному і 5 разів у референтному підкорпусі, і його БІК становить 17,10. Його синонім — слово *гімн*, вживається в діаспорних текстах у формі *гимн*. БІК для нього становить 406,85. Цікаво, що таке написання не мотивоване нормами правопису 1928-1929 років, бо у словнику Голоскевича знаходимо форму *гімн*.

[96] Академічний тлумачний словник української мови (СУМ). 1970–1980. Залога. — Режим доступу: http://sum.in.ua/s/zalogha

Ще одна пара слів *оплески* та *аплодисменти*. Перше більш притаманне діаспорі та має БІК 66,98. Натомість слово аплодисменти, що є його відповідником, хоч і дещо частіше трапляється у текстах материка, не можна віднести до жодного із досліджуваних корпусів за статистичними даними, адже його БІК від'ємний і становить -14,33.

Крім того, у текстах діаспори частіше трапляється слово *совєтський* (180 разів, напротивагу 198 разам у референтному підкорпусі). Його вживають на позначення Радянського Союзу: «*Головна мета Президента Росії Володимира Путіна – це відновлення Совєтського Союзу*», і того, що з ним пов'язане: «*хто ризикував своїм життям у спротиві совєтському режимові у Литві*». БІК для цього слова становить 363,57. Водночас слово *радянський* не можна назвати притаманним тому чи тому підкорпусу. Його БІК становить -8.52.

Знаходимо у згенерованих списках слово *зрезигнувати*. Воно притаманне діаспорній мові й має БІК 387,33. Можна припустити, що слово прийшло в українську із польської, де є слово *(z)rezygnować*, було використовуване емігрантами із західної України, носіями галицького діалекту, і залишилось у мовленні діаспори донині. До того ж слово зафіксоване у словнику Голоскевича, що легітимізує його використання у діаспорі.

Отже, в цьому розділі ми зосередились на виділенні та аналізі ключових для підкорпусу мови західної української діаспори слів. Для цього було відібрано тексти газети «Свобода» за 2016-2019 роки, та підготовано їх до додавання у Генеральний регіонально анотований корпус української мови (ГРАК) (версія 11). Було створено фокусний та референтний підкорпуси, а опісля — згенеровано частотні списки слів та обрано статистичні метрики, що допомогли виділити ті ключові слова, що мають достатню статистичну значущість. Результати аналізу ключових слів показали, що більшість відмінностей між сучасною українською мовою публіцистики материка, та сучасною мовою публіцистики української західної діаспори спричинена використанням різних правописних систем. Відмінності проявляються також на рівні морфології (відмінювання окремих слів), словотвору, лексикології (вибір того чи іншого слова серед можливих

відповідників) та семантики (вживання слова у ширшому чи вужчому контексті, надання йому нових (відтінків) значень). Окрім того, у текстах діаспори збереглись деякі галицизми та полонізми, що їх вживали емігранти із західної України, приїхавши до США. Безперечно, проведений аналіз не є повним, адже він охоплює досить малий часовий період, і фокусується на текстах лише одного діаспорного видання. Наше дослідження використовувало метод ключових слів, а отже не охопило синтаксичного рівня мови. Проте результати нашої роботи можуть стати корисними для подальших та більших за обсягом досліджень української мови західної діаспори.

ВИСНОВКИ

У цій роботі нам вдалося досягти мети та виявити характерні особливості текстів публіцистики української західної діаспори. Для цього ми дослідили представленість діаспорних текстів у корпусах української мови, схарактеризували роль газети «Свобода» у житті західної діаспори та підготували її тексти за 2016-2019 роки до додавання у корпус ГРАК. Ми побудували фокусний та референтний підкорпуси, і на їх основі згенерували частотні списки потенційно ключових слів. Також нам вдалося дослідити застосування методу ключових слів, зокрема, найпопулярніші статистичні метрики, що можуть бути його частиною, та обрати оптимальний спосіб визначення статистично значущих ключових слів для нашого дослідження. Ми проаналізували значущі ключові слова та їх групи й запропонували пояснення причин їхньої ключовості.

В дослідженні охоплено морфологічний, словотвірний, та лексико-семантичний рівні мови й визначено основні характерні одиниці, що до них належать.

Наше дослідження має свої обмеження, адже воно зосереджене на текстах короткого часового проміжку і лише одного видання. Проте, наскільки нам відомо, ця робота — перше корпусне дослідження, в якому застосовано метрики статистичної значущості та метод визначення ключових мовних одиниць для вивчення української мови західної діаспори. Подальші корпусні дослідження, проведені на більших обсягах текстів, для більших часових проміжків та з врахуванням інших стилів можуть дати більш репрезентативні результати для характеристики мови української західної діаспори.

СПИСОК ДЖЕРЕЛ ТА ЛІТЕРАТУРИ

1. Богдан Ажнюк. *Мовна єдність нації: діаспора й Україна*. Київ 1999.
2. Вікторія Жуковська. *Вступ до корпусної лінгвістики: навчальний посібник*. Житомир 2013.
3. Л. Козачевська, О. Сидоренко. *Мова української західної діаспори як віддзеркалення мовної свідомості та мовної діяльності носіїв-білінгвів* // Мовні і концептуальні картини світу, випуск 46(2). Київ 2013.
4. Леся Біловус. *Україномовна періодика у національно-культурному житті української діаспори США (1991–2017 рр.)*. Тернопіль 2017.
5. Наталія Дарчук. Дослідницький корпус української мови: основні засади і перспективи // Вісник Київського національного університету імені Тараса Шевченка. Серія: Літературознавство. Мовознавство. Фольклористика. – № 21. Київ 2010. С 45–49.
6. Петро Часто. *Вільне слово Американської України*. Нью-Йорк — Ужгород 2012.
7. Anatol Stefanowitsch. *Corpus linguistics: A guide to the methodology*. Berlin 2020.
8. Andrew Wilson. *Embracing Bayes factors for key item analysis in corpus linguistics* // New Approaches to the Study of Linguistic Variability: Language Competence and Language Awareness in Europe, vol. 4. Frankfurt 2013. P. 3–11.
9. Costas Gabrielatos. *Keyness Analysis: Nature, metrics and techniques* // Corpus approaches to discourse: a critical review. Milton 2018. P. 225–258.
10. Maria Shvedova. *The General Regionally Annotated Corpus of Ukrainian (GRAC, uacorpus.org): Architecture and Functionality* // Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2020), vol. I: Main Conference. Lviv 2020. P. 489–506.
11. Michael Moser. *New Contributions to the History of the Ukrainian Language*. Edmonton — Toronto 2016.

12. Vaclav Brezina. *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge — New York 2018.

Електронні ресурси:

13. Богдан Ажнюк. *Мовні роздоріжжя української діаспори*. — Режим доступу: http://kulturamovy.univ.kiev.ua/KM/pdfs/Magazine53-54-6.pdf
14. Браунський корпус української мови // GitHub. — Режим доступу: https://github.com/brown-uk/corpus
15. Вимоги до текстів для БрУК // GitHub. — Режим доступу: https://github.com/brown-uk/corpus/blob/master/doc/vymohy_do_frahmentiv.md
16. Генеральний регіонально анотований корпус української мови (ГРАК) / М. Шведова, Р. фон Вальденфельс, С. Яригін, А. Рисін, В. Старко та ін. – Київ, Львів, Єна, 2017-2021. — Режим доступу: http://uacorpus.org/
17. Генеральний регіонально анотований корпус української мови (ГРАК): Grak v.11: Пошук \\ NoSketchEngine — Режим доступу: http://www.parasolcorpus.org/bonito/run.cgi/first?corpname=grac11&reload=1&iquery=&queryselector=iqueryrow&lemma=&phrase=&word=&char=&cql=&default_attr=word&fc_lemword_window_type=both&fc_lemword_wsize=5&fc_lemword=&fc_lemword_type=all&usesubcorp=&fsca_doc.author=&fsca_doc.translator=&fsca_doc.authTrans=&fsca_doc.born=&fsca_doc.title=&fsca_doc.date=&fsca_doc.mediaName=&fsca_doc.locCode=&fsca_doc.region=&fsca_doc.publicationCity=&fsca_doc.publisher=&fsca_doc.publication=&fsca_doc.uri=
18. Генеральний регіонально анотований корпус української мови (ГРАК): GRAC V.11: dashboard // SketchEngine — Режим доступу: https://parasol.vmguest.uni-jena.de/grac_crystal/#dashboard?corpname=grac11

19. Генеральний регіонально анотований корпус української мови (ГРАК): Параметри словника частот. — Режим доступу: http://www.parasolcorpus.org/bonito/run.cgi/wordlist_form?corpname=grac11
20. Звідусіль // NoSketchEngine — Режим доступу: https://mova.institute/bonito/run.cgi/first_form?corpname=zvidusil
21. Звідусіль: Про корпус. NoSketchEngine — Режим доступу: https://mova.institute/bonito/run.cgi/corp_info?corpname=zvidusil&struct_attr_stats=1&subcorpora=1
22. Корпус текстів української мови. 2003–2021. — Режим доступу: http://www.mova.info/corpus.aspx
23. Лабораторія української: Золотий морфосинтаксовий стандарт. — Режим доступу: https://mova.institute/%D0%B7%D0%BE%D0%BB%D0%BE%D1%82%D0%B8%D0%B9_%D1%81%D1%82%D0%B0%D0%BD%D0%B4%D0%B0%D1%80%D1%82
24. Олександр Тараненко. *Мова української західної діаспори і сучасна мовна ситуація в Україні* // Мовознавство, № 2-3. С. 63–99. — Режим доступу: http://nbuv.gov.ua/UJRN/MoZn_2013_2-3_7
25. Олексій Ясь. *Діаспора* // Енциклопедія історії України, т. 2: Г-Д. Київ 2004. — Режим доступу: http://www.history.org.ua/?termin=Diaspora
26. Спільнота lang-uk. 2016–2017. — Режим доступу: https://lang.org.ua/uk/corpora/#anchor6
27. *Українці у США: кількість, міграція, заняття та уподобання. За матеріалами відкритої лекції Воловини О. «Українці в США та вплив міграції з України на діаспору в США (на базі офіційної статистики)»* // Справжня варта, 21 листопада 2012. — Режим доступу : http://varta.kharkov.ua/articles/1090601.html

28. Adam Kilgarriff. *Simple maths for keywords*. — Режим доступу: https://www.sketchengine.eu/wp-content/uploads/2015/04/2009-Simple-maths-for-keywords.pdf
29. BootCat — Режим доступу: https://bootcat.dipintra.it/
30. British National Corpus (BNC). — Режим доступу: https://www.english-corpora.org/bnc/
31. Lob Corpus. — Режим доступу: https://www1.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/corpora/list/private/LOB/lob.html
32. Log-likelihood and effect size calculator — Режим доступу: http://ucrel.lancs.ac.uk/llwizard.html
33. Manual Of Information To Accompany The Kolhapur Corpus Of Indian English, For Use With Digital Computers. 1986. — Режим доступу: http://korpus.uib.no/icame/manuals/KOLHAPUR/INDEX.HTM
34. Maria Shvedova, Ruprecht von Waldenfels. *Regional Annotation within GRAC, a Large Reference Corpus of Ukrainian: Issues and Challenges*. 2021. — Режим доступу: http://ceur-ws.org/Vol-2870/paper4.pdf
35. Narodowy Korpus Języka Polskiego. 2008-2012. — Режим доступу: http://nkjp.pl/
36. Russian National Corpus. 2003–2021. — Режим доступу: https://ruscorpora.ru/old/en/index.html
37. Simple maths // SketchEngine — Режим доступу: https://www.sketchengine.eu/documentation/simple-maths/
38. The Brown Corpus. 1998. — Режим доступу: https://www1.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/corpora/list/private/brown/brown.html
39. The Open American National Corpus. 2002–2015. — Режим доступу: https://www.anc.org/
40. Tony McEnery, Andrew Hardie. *Statistics in corpus linguistics*. Cambridge 2012 // — Режим доступу: http://corpora.lancs.ac.uk/clmtp/2-stat.php

41. Ukrainian (ukr_mixed_2014): Про корпус // NoSketchEngine — Режим доступу: http://cql.corpora.uni-leipzig.de/bonito/run.cgi/corp_info?corpname=ukr_mixed_2014
42. Ukrainian (ukr_newscrawl_2018): Про корпус // NoSketchEngine — Режим доступу: http://cql.corpora.uni-leipzig.de/bonito/run.cgi/corp_info?corpname=ukr_newscrawl_2018
43. Ukrainian (ukr-ua_web_2019): Про корпус // NoSketchEngine — Режим доступу: http://cql.corpora.uni-leipzig.de/bonito/run.cgi/corp_info?corpname=ukr-ua_web_2019

Довідкова література:

44. Григорій Голоскевич. *Правописний словник*. Київ, 1929. — Режим доступу: https://r2u.org.ua/data/%D0%9F%D1%80%D0%B0%D0%B2%D0%BE%D0%BF%D0%B8%D1%81%D0%BD%D0%B8%D0%B9%20%D1%81%D0%BB%D0%BE%D0%B2%D0%BD%D0%B8%D0%BA%20(1929).pdf
45. *Український правопис*, видання перше. Харків 1929. — Режим доступу: https://r2u.org.ua/data/other/%D0%9F%D1%80%D0%B0%D0%B2%D0%BE%D0%BF%D0%B8%D1%81-1928.pdf
46. Святослав Караванський. *Російсько-український словник складної лексики*. 2012. — Режим доступу: https://r2u.org.ua/dicts/karavansky
47. Агатангел Кримський, Сергій Єфремов. *Російсько-український словник*. 1924–1933. // Електронна версія, А–П. Київ, 2007. — Режим доступу: https://r2u.org.ua/data/%D0%A0%D0%BE%D1%81%D1%96%D0%B9%D1%81%D1%8C%D0%BA%D0%BE-%D1%83%D0%BA%D1%80%D0%B0%D1%97%D0%BD%D1%81%D1%8C%D0%BA%D0%B8%D0%B9%20%D0%B0%D0%BA%D0%B0%D0%B4%D0%B5%D0%BC%D1%96%D1%87%D0%BD%D0%B8%D0%B9%20%D1%81%D0%BB%D0%BE%D0%B2%D0%BD%D0%B8%D0%BA%20%281924-33%29.pdf

48. *Академічний тлумачний словник української мови (СУМ)*. 1970–1980. — Режим доступу: http://sum.in.ua/

Додаток 1

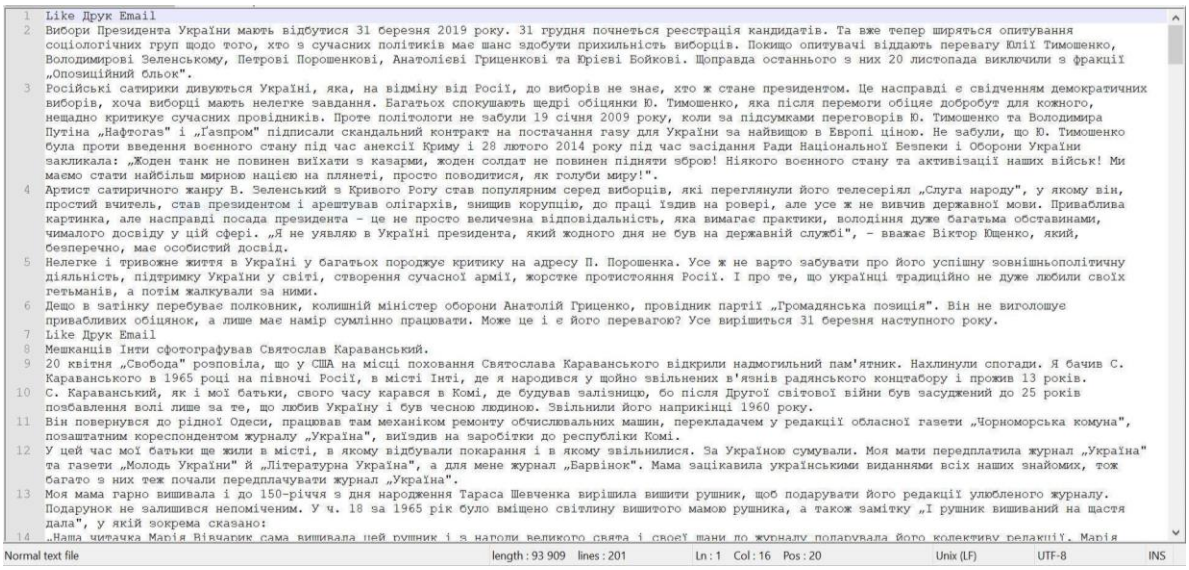

Рисунок 1. Вигляд файлу з текстами «Свободи» до оброблення.

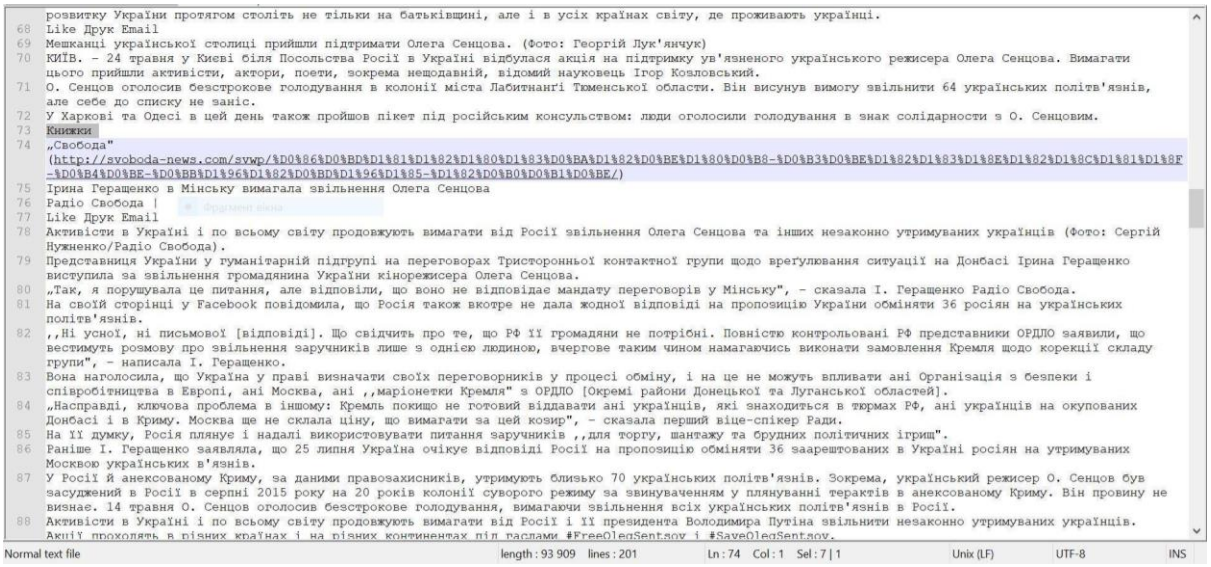

Рисунок 2. Елементи, які було вилучено з текстів.

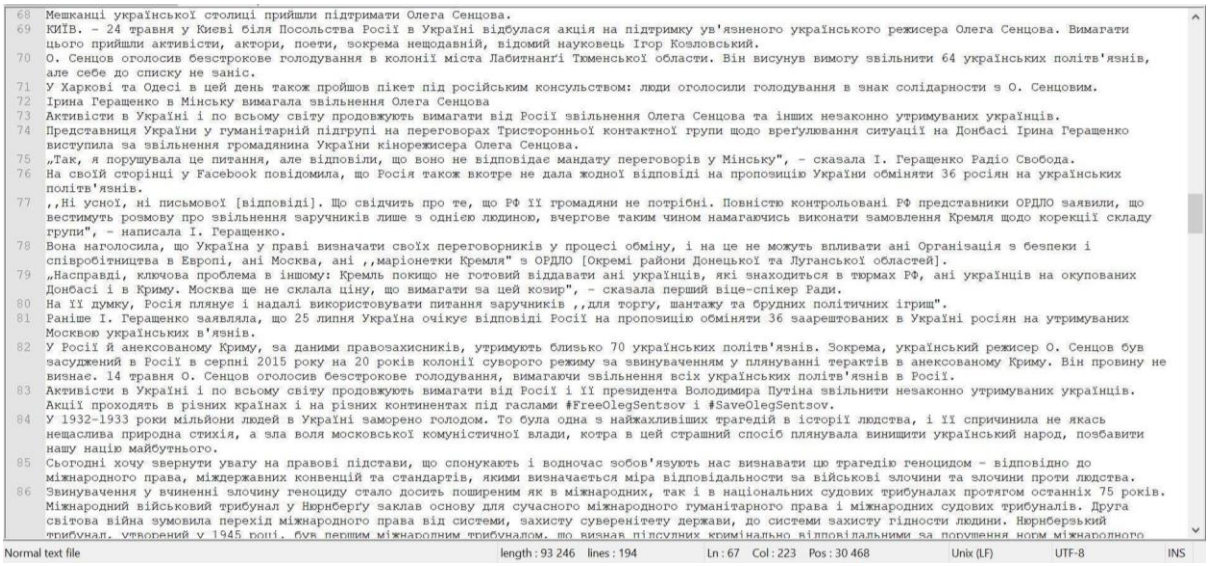

Рисунок 3. Готовий текст після оброблення.

Додаток 2

12	lemma	Freq	Freq/mill	Freq_ref	Freq_ref/mill	Score
13	----------					
14	|від.|	2439	960.3	0	0.0	961.3
15	|європейський|	1173	461.8	5	0.2	387.3
16	|Ню-Йорку|	745	293.3	0	0.0	294.3
17	|діяспора|	729	287.0	2	0.1	267.2
18	|патріярх|	702	276.4	1	0.0	267.0
19	|кий|Київ|	1457	573.7	31	1.2	260.1
20	|Європа|	1204	474.0	24	0.9	245.3
21	|д-р|	956	376.4	15	0.6	238.1
22	|Ню-Джерзі|	540	212.6	0	0.0	213.6
23	|парлямент|	510	200.8	0	0.0	201.8
24	|міністер|	524	206.3	1	0.0	199.5
25	|матеріял|	555	218.5	4	0.2	189.9
26	|ВАШІНҐТОН|	474	186.6	0	0.0	187.6
27	|УНСоюзу|	468	184.3	0	0.0	185.3
28	|соціяльний|	567	223.2	6	0.2	181.7
29	|патріярхат|	458	180.3	0	0.0	181.3
30	|Філядельфія|	430	169.3	0	0.0	170.3
31	|реґіон|	459	180.7	2	0.1	168.6
32	|конґрес|	832	327.6	25	1.0	166.3
33	|плян|	530	208.7	8	0.3	159.8
34	|ініціятива|	417	164.2	1	0.0	159.0
35	|спеціяльний|	406	159.9	2	0.1	149.2
36	|Чикаґо|	376	148.0	1	0.0	143.4
37	|Ню-Йорк|	361	142.1	0	0.0	143.1
38	|аґресія|	598	235.4	17	0.7	142.2
39	|Олександер|	492	193.7	11	0.4	136.2
40	|асоціяція|	348	137.0	1	0.0	132.8
41	|територіяльний|	329	129.5	1	0.0	125.6
42	|роля|	368	144.9	5	0.2	122.1
43	|парляментський|	288	113.4	0	0.0	114.4
44	|плянувати|	287	113.0	0	0.0	114.0
45	|делеґація|	284	111.8	0	0.0	112.8
46	|НЮ-ЙОРК|	280	110.2	0	0.0	111.2
47	|меморіяльний|	263	103.5	0	0.0	104.5
48	|клюб|	284	111.8	4	0.2	97.6
49	|архиєпископ|	258	101.6	3	0.1	91.8
50	|східній|	222	87.4	0	0.0	88.4
51	|Багряний|	450	177.2	27	1.1	86.8
52	|еміґрація|	257	101.2	5	0.2	85.5
53	|СУМ|	297	116.9	10	0.4	84.8
54	|Вашінґтоні|	212	83.5	0	0.0	84.5
55	|меморіял|	203	79.9	0	0.0	80.9
56	|прем'єр-міністер|	198	78.0	0	0.0	79.0
57	|св.|	1439	566.6	160	6.2	78.4
58	|проф.|	344	135.4	19	0.7	78.4
59	|аґент|	240	94.5	6	0.2	77.4
60	|СКУ|	737	290.2	74	2.9	74.9
61	|дириґент|	200	78.7	2	0.1	74.0
62	|посадник|	184	72.4	0	0.0	73.4
63	|катедра|	252	99.2	10	0.4	72.1
64	|конґресовий|	179	70.5	0	0.0	71.5
65	|кляса|	262	103.2	12	0.5	70.9
66	|аґенція|	261	102.8	12	0.5	70.7
67	|пропаґанда|	211	83.1	5	0.2	70.4
68	|делеґат|	174	68.5	0	0.0	69.5
69	|кредитовий|	208	81.9	5	0.2	69.4
70	|ЄС|	287	113.0	17	0.7	68.5

Рисунок 1, вершина частотного списку лем згенерованого для підкорпусу текстів Свободи (SVOBODA_2016_2019) у порівнянні із референтним підкорпусом публіцистики материкової України (UA_NATIONAL_JOU_2016_2019) і коефіцієнтом згладжування 1.

	lemma	Freq	Freq/mill	Freq_ref	Freq_ref/mill	Score
13	----------					
14	|ЄС|	8472	330.5	0	0.0	331.5
15	|долар|	6225	242.9	0	0.0	243.9
16	|Європа|європи|	6046	235.9	0	0.0	236.9
17	|європейський|	12375	482.8	3	1.2	221.8
18	|парламент|	7140	278.6	1	0.4	200.6
19	|ініціатива|	4935	192.5	0	0.0	193.5
20	|соціальний|	8819	344.1	2	0.8	193.1
21	|Європа|	6481	252.9	1	0.4	182.1
22	|планувати|	4190	163.5	0	0.0	164.5
23	|проект|	18973	740.2	11	4.3	139.0
24	|варіант|	3493	136.3	0	0.0	137.3
25	|законопроект|	3928	153.3	1	0.4	110.7
26	|план|	7435	290.1	5	2.0	98.1
27	|рейтинг|	3439	134.2	1	0.4	97.0
28	|спр|	2184	85.2	0	0.0	86.2
29	|неспр|	2182	85.1	0	0.0	86.1
30	|регіональний|	2177	84.9	0	0.0	85.9
31	|глобальний|	2160	84.3	0	0.0	85.3
32	|потенціал|	2125	82.9	0	0.0	83.9
33	|спеціальний|	3802	148.3	2	0.8	83.5
34	|регіон|	7367	287.4	7	2.8	76.8
35	|асоціація|	2637	102.9	1	0.4	74.5
36	|Євросоюз|	1870	73.0	0	0.0	74.0
37	|масштабний|	1865	72.8	0	0.0	73.8
38	|енергетичний|	1800	70.2	0	0.0	71.2
39	|клуб|	3236	126.3	2	0.8	71.2
40	|аграрний|	1743	68.0	0	0.0	69.0
41	|матч|	3759	146.7	3	1.2	67.7
42	|євро|	2376	92.7	1	0.4	67.2
43	|Іспанія|	1615	63.0	0	0.0	64.0
44	|парламентський|	2789	108.8	2	0.8	61.4
45	|агресор|	1521	59.3	0	0.0	60.3
46	|електроенергія|	1502	58.6	0	0.0	59.6
47	|матеріал|	5087	198.5	6	2.4	59.3
48	|діалог|	2083	81.3	1	0.4	59.0
49	|агресія|	3669	143.1	4	1.6	56.0
50	|енергетика|енергети	1368	53.4	0	0.0	54.4
51	|клас|	3524	137.5	4	1.6	53.8
52	|зарплата|	3510	136.9	4	1.6	53.6
53	|спеціаліст|	1311	51.1	0	0.0	52.1
54	|Вашингтон|	1283	50.1	0	0.0	51.1
55	|класичний|	1752	68.4	1	0.4	49.8
56	|енергія|	2234	87.2	2	0.8	49.3
57	|аргумент|	1712	66.8	1	0.4	48.6
58	|мітинг|	1205	47.0	0	0.0	48.0
59	|аудиторія|	1651	64.4	1	0.4	46.9
60	|паспорт|	1602	62.5	1	0.4	45.6
61	|запланований|	1122	43.8	0	0.0	44.8
62	|регулярно|	1119	43.7	0	0.0	44.7
63	|2017-й|	1560	60.9	1	0.4	44.4
64	|бригада|	2000	78.0	2	0.8	44.2
65	|спеціально|	1093	42.6	0	0.0	43.6
66	|інтеграція|	1082	42.2	0	0.0	43.2
67	|Маріуполь|	1488	58.1	1	0.4	42.4
68	|турнірний|	1055	41.2	0	0.0	42.2
69	|Нідерланди|	1055	41.2	0	0.0	42.2

Рисунок 2, вершина частотного списку лем згенерованого для підкорпусу текстів публіцистики материкової України (UA_NATIONAL_JOU_2016_2019) у порівнянні із референтним підкорпусом текстів Свободи (SVOBODA_2016_2019) і коефіцієнтом згладжування 1.