МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Київський національний університет імені Тараса Шевченка
Інститут філології
Кафедра української мови та прикладної лінгвістики

УКРАЇНСЬКОМОВНИЙ МЕДІА ДИСКУРС СУЧАСНОЇ ПОП-КУЛЬТУРИ (СТАТИСТИЧНА ПАРАМЕТРИЗАЦІЯ ЗА ДОПОМОГОЮ ІНСТРУМЕНТІВ КОРПУСНОЇ ЛІНГВІСТИКИ)

Кваліфікаційна робота
освітнього ступеня «бакалавр»
за спеціальністю 035 «Філологія»,
спеціалізацією 035.10 «Прикладна лінгвістика»,
галузі знань 03 «гуманітарні науки»
ОПП «Прикладна (комп'ютерна) лінгвістика та англійська мова»
Іллі Гришина

Науковий керівник:
к. філол. н., доц. Оксана ЗУБАНЬ

Рецензент:
к.філол.н., доц. Святослав Шевель

Київ – 2021

ВСТУП

ВСТУП

У наш час мовознавчі дослідження все частіше спираються на статистичні методи. Раніше статистика слугувала лише для перевірки та схвалення лінгвістично обґрунтованих висновків, які засновувались на вже існуючій методології. Але сьогодні активна використання статистичних показників стає звичною для лінгвістів. Статистика допомагає визначити патерни організації створення підсистем мову і мовлення, адже: *«... статистичні методи і характеристики допомагають глибше проникнути в закони мови і мовлення і можуть бути основою або відправною точкою для встановлення таких закономірностей, яких без цих методів не вдалося б побачити»* [6, с.3].

Яскравим прикладом є стилістика, оскільки результати досліджень у цій мовознавчій галузі є менш репрезентативними без аналізу кількісних показників. Наприклад, цінність та об'єктивність наукового дослідження функціональних та авторських стилів буде поставлена під сумнів, а зроблені висновки вважатимуться необ'єктивними. Уособленням розвитку статистичної стилістики у вітчизняному мовознавстві була поява монографії «Статистичні параметри стилів»[37]. Праця була створена зусиллями відділу структурно-математичної лінгвістики Інституту мовознавства ім. О.О.Потебні НАН України. На сторінках книги В. Перебийніс пояснює, чому статистичні методи є важливими для стилістичних досліджень: *«Можливість використання статистичних методів у стилістиці ґрунтується на тому, що всякий матеріал мовлення (тобто текст) є результатом добору певних одиниць із загальнонародної мови. Добір цей залежить не лише від теми висловлювання, але й від форми її викладу (поетичний чи драматичний твір, художня чи наукова проза, науково-популярний чи науковий виклад і т.д.), від законів і канонів стилю чи жанру, від особистих уподобань автора і, нарешті, від тих законів, за якими будується мовлення, а також від законів мови. Мова диктує*

свої закони кожному, хто нею користується, і ступінь підсвідомого чи свідомого засвоєння цих законів мовцем позначається на якості добору і розташуванні мовного матеріалу в мовленні» [37, с.23].

Вираженням плідної праці і ґрунтовних досліджень науковців відділу структурно-математичної лінгвістики Інституту мовознавства ім. О. О. Потебні НАН України стали такі визначні лінгвістичні роботи: збірник «Вопросы статистической стилистики» [6], «Частотний словник сучасної української художньої прози» [49], «Частотні словники та їх використання» [41] та ін. Мовознавчий аналіз, що міститься у таких книгах, як «Система афіксального словотворення сучасної української мови» [26], «Морфемна структура слова» [33], «Словник афіксальних морфем української мови» [45], «Основи морфеміки сучасної української мови» [25], «Суфіксальна підсистема сучасної української літературної мови: будова та реалізація» [23] наглядно відображає те, як статистична методологія удосконалює дослідження у сфері морфемної системи української мови.

Науково-матеріальна спадщина сучасної української комп'ютерної лексикографії становить значну кількість електронних словників різних сфер. Деякі з них містять лінгвістичний портал (www.mova.info) лабораторії комп'ютерної лінгвістики Інституту філології Київського національного університету імені Тараса Шевченка: Відкритий словник новітніх термінів; Українсько-італійський граматичний словник дієслів; серія частотних словників (Частотний словник художньої прози, Частотний словник сучасної української публіцистики, Частотний словник наукового стилю, Частотний словник сучасної української поетичної мови та ін.); Словник української мови (Академічний тлумачний словник (1970 – 1980); Електронний граматичний словник української літературної мови (словозміна); Система електронних навчальних словників "Глоса" (англо-український, українсько-англійський навчальний словник); Труднощі англійського слововживання для українців; Тезаурус комп'ютерної лексикографії та ін. [31]

Ці словники приносять користь як науковцям, так і звичайним користувачам: вони слугують інформаційно-довідковою базою і основою для проведення значних лінгвістичних досліджень. Проблема полягає в тому, що більшість словників не мають функції автоматичного аналізу тексту. Так на перший план виходять ті електронні лінгвістичні продукти, які здатні надавати інформацію про мовні одиниці і аналізують текст у режимі онлайн, а також мають пошукову систему. Одним з таких продуктів є Корпус української мови [30], створений на базі лабораторії комп'ютерної лінгвістики Інституту філології Київського національного університету імені Тараса Шевченка. Цей корпус має низку важливих атрибутів:

1) формалізований опис мовних одиниць різних рівнів тексту [1];

2) лінгвістичні бази даних для анотування тексту та параметризації на морфологічному, морфемному та синтаксичному рівнях його структури;

3) рубрику "Статистика текстів", яка надає доступ до стилістичної параметризації корпусу;
4) можливість укласти частотні словники лексем і словоформ онлайн на базі корпусу; на сьогодні на базі корпусу існує більше 29 000 ЧС лексем та словоформ.

На основі описаних електронних словників можна ефективно класифікувати елементи українського тексту в режимі онлайн. Здобуті статистичні дані дозволяють краще опанувати принципи функціонування одиниць української мови у багатьох стилях, повноцінно проаналізувати атрибути ідіостилів українських письменників. Прикладом слугує аналіз статистичних показників морфемного рівня структури поетичного тексту Т. Шевченка [18; 20] - кількісні та статистичні характеристики морфемних структур слів, розкривають закономірності структуру тексту ідіостилю Шевченка.

Високий рівень кастомізації інформації, можливість проаналізувати українськомовний текстовий матеріал на кількох граматичних та стилістичних рівнях в Корпусі української мови є передвісниками появи лексикографічної системи нового покоління. Вона має бути універсальною з однаковим порогом входження як для викладача, письменника, публіциста, так і для мовознавця. Система нового покоління має задовольняти потреби в отриманні довідки з української мови та у володінні масштабною лінгвістичною базою знань.

Початок XXI ст. є хронологічною точкою відліку для розвитку української корпусної лінгвістики в царині комп'ютерної лінгвістики. Приблизно у той же час закріплюються терміни "корпус текстів" та "корпусна лінгвістика"[3]. Сьогодні у режимі вільного доступу в інтернеті наявні такі українськомовні корпуси: Корпус української мови (КУМ) [30], Генеральний регіонально анотований корпус української мови (ГРАК) [5], Корпуси текстів української мови [28], Браунський корпус української мови (БрУК) [4]. Існують також два закритих корпуси: Український національний лінгвістичний корпус [50] та Комп'ютерний фонд інновацій (КФІ) [24]. Першим володіє Український мовно-інформаційний фонд НАН України, а другим - Інститут української мови НАН України. Корпуси використовуються у дослідженнях галузей лексикології, лексикографії, структурно-математичної лінгвістики та для укладання словників [21].

Значущість напрямів розвитку сучасної корпусної лінгвістики у зв'язку із потребою аналізу великих масивів текстової інформації в мережі інтернет визначає **актуальність теми бакалаврського дослідження.**

Мета дослідження - провести статистичну параметризацію українськомовних медіатекстів на базі автоматично укладеного комп'ютерного частотного словника.

Поставлена мета вимагає виконання таких **завдань**:

1) глибинний огляд медійних сфер, у яких представлена сучасна українськомовна популярна культура;
2) формування текстових вибірок і створення корпусу медійних текстів;
3) автоматична обробка назв текстових файлів для усунення критичних помилок в інтерфейсі Python;
4) уніфікація кодування текстових файлів;
5) усунення помилок токенізації;
6) створення частотного словника словоформ і лем для загальної бази даних і для кожної тематичної вибірки;
7) створення конкордансу медійних текстів;
8) створення бази даних тонального сентимент-аналізу медійних текстів;
9) виведення статистики для загальної бази даних і тематичних словників;
10) лінгвістичний аналіз статистичного матеріалу.

Об'єкт дослідження - українськомовний медіатекст.

Предмет дослідження - лексична та граматична (частиномовна) статистична структура українськомовного медіатексту. **Теоретичне значення роботи** - систематизація та погляд на майбутній розвиток прийомів дослідження у сфері корпусної лінгвістики, статистичної лексикографії та автоматичного сентимент-аналізу.

Практичне значення роботи - створений корпус медійних текстів українськомовної поп-культури та його статистична параметризація надає нові фактичні дані про стан українськомовного дискурсу в сучасній медійній сфері.

Методи дослідження: метод комп'ютерного лексикографічного моделювання, статистичні методи та метод кількісних підрахунків, методи об’єктно-орієнтованого програмування (мова програмування Python), система керування базами даних SQL.

Матеріал дослідження: медіатексти, які складають зональну вибірку обсягом 197 877 слововживань. Матеріал було зібрано із: 1) Офіційного YouTube-каналу “Телебачення Торонто”[48], 2) офіційного YouTube-каналу “СЛУХ”[46], 3) офіційного сайту “СЛУХ медіа”[47], 4) сторінок “Простими словами”[43] та “MINCULTPYVIT”[63] на подкаст-платформах “Apple Podcasts” та “Soundcloud”, 5) офіційного сайту та Instagram-профілю Євгена Клопотенка[29],[59], 5) сторінки фільму “Мої думки тихі” з медіасервісу “Megogo”[32]; 6) сайту The Village UA[75].

Інформаційна база дослідження — мова програмування Python, бібліотека pymorphy2[65], nltk[64], український тональний словник “tone-dict-uk[15]” і система керування базами даних **SQLite**[71]**.**

У першому розділі **“Теоретичні засади опрацювання текстових даних інструментами корпусної лінгвістики”** проаналізовані етапи розвитку корпусної лінгвістики і статистичної лексикографії, а також дана оцінка ролі цих галузей для сучасного мовознавства.

У другому розділі **“Створення корпусу українськомовних медіатекстів та автоматична статистична параметризація текстових даних”** був описаний процес створення корпусу українськомовних медіатекстів, укладання частотного словника, словника-конкордансу на базі корпусу і бази даних тонального сентимент-аналізу українськомовних медіатекстів.

У третьому розділі **“Статистичні параметри медіатекстів поп-культури”** текстовим вибіркам корпусу була дана оцінка на основі статистичних та психолінгвістичних формул.

Структура і обсяг роботи. Робота складається із вступу, трьох розділів, висновків, списку використаної літератури (81 позиція), 21 додатку , всього 98 сторінок (разом із титульною сторінкою та додатками).

РОЗДІЛ 1. ТЕОРЕТИЧНІ ЗАСАДИ ОПРАЦЮВАННЯ ТЕКСТОВИХ ДАНИХ ІНСТРУМЕНТАМИ КОРПУСНОЇ ЛІНГВІСТИКИ

1.1. Історія розвитку та методологія корпусної лінгвістики

Корпусна лінгвістика - один із найперспективніших напрямків прикладного і теоретичного аспектів сучасного мовознавства. Галузь є відносно новою, оскільки її активний розвиток пов'язаний з технічним прогресом 60-х років минулого століття. У 1983 році виходить наукова праця "Corpus Linguistics: Recent Developments in the Use of Computer Corpora in English Language Research"[58]. Це стало причиною входження терміну "корпусна лінгвістика" в широкий науковий ужиток. Сьогодні корпусна лінгвістика тісно пов'язана з такими розділами мовознавства: *"лексикографія, стилістика, судова лінгвістика, перекладознавство, соціолінгвістика тощо"* [55, с. 2-3].

Але чим є корпусна лінгвістика у загальному сенсі? Насамперед корпусна лінгвістика встановлює основні принципи побудови лінгвістичних корпусів, використовуючи для цього сучасні технології. Це допомагає створити методику збирання мовних явищ у писемних та усних текстах та їх автоматичного або автоматизованого аналізу. Практичним результатом досліджень є корпус текстів – масив мовних даних, створений для вирішення лінгвістичних завдань. *"Корпус має бути вагомим за обсягом, уніфікованим, містити чітку структуру і розмітку, бути представленим в електронному вигляді"* [16, с.3].

Які ознаки характерні для корпусного аналізу?

- дослідження реальних моделей мовної реалізації у природних текстах;
- основа аналізу – великі за обсягом тексти;
- залучення комп'ютерних технологій для аналізу матеріалу;
- використання квантитативних, квалітативних та аналітичних методик.

"Корпусний матеріал дозволяє не лише оптимізувати і об'єктивізувати лінгвістичні дослідження, але й забезпечує новий погляд на традиційні

поняття” [44, с.185]. Корпусні дослідження та результати аналізу даних *“сприяють переоцінці звичних підходів до вивчення мови та ряду лінгвістичних теорій”* [61, с.1].

Станом на сьогодні існує два головних напрями дослідження корпусної лінгвістики. “*Перший спрямований на теорію та практику створення корпусів, а саме: обсяг і призначення корпусу, типологія і параметризація, алгоритми підбору базових одиниць, структурування тощо. Другий напрям стосується саме вивчення мови на основі корпусних досліджень*” [27, с.12].

Але треба пам’ятати, що процес укладання корпусу потребує імплементації двох напрямів одночасно. Це зумовлено тим, що, з одного боку, корпус є вихідним матеріалом для дослідження, а з іншого боку – результатом діяльності. Таким чином лінгвістичний корпус є об’єктом корпусної лінгвістики. А предметом є теоретичні засади та практичні системи створення та використання корпусів природної мови [17]. Мета цього розділу мовознавства полягає в описі мовної системи та розробці унікального відображення мовного матеріалу в самому корпусі.

У лінгвістичних корпусних дослідженнях першочергове значення має реальний текст. Це зумовлено тим, що теоретичним підґрунтя корпусної лінгвістики є структуралізм. Мова має вивчатися лише на основі писемних та усних текстів [12]. Інтерпретація тексту, характерна для корпусної лінгвістики, відрізняється від звичної в мовознавстві: текст є не мовною одиницею, а дійсністю, що піддається науковому спостереженню.

Існують ще такі особливості трактування тексту:

а) текст – єдиний цілісний вихідний пункт для науковця;

б) теорія має вичерпно описати текст за допомогою аналізу чи послідовного розподілу;

в) оскільки матеріали природної мови можуть бути дуже об’ємними, дозволено використовувати текстові вибірки та членувати тексти на такі категорії: твори окремих авторів, окремі праці, розділи, параграфи, на складні речення та слова.

Погляд на мову як на цілісне соціальне явище є основою корпусної лінгвістики. Мова є соціальним феноменом із проявами у тексті, а більшість текстів є мовленнєвими актами.

Незважаючи на те, що корпусна лінгвістика має свою історію розвитку і є беззаперечно актуальною наукою, у лінгвістичному дискурсі відбувається дискусія з приводу того, чи можна вважати корпусну лінгвістику самостійною наукою. Такі вчені, як Т.МакЕненрі та Р.Ксіао *“переконані, що корпусна лінгвістика є лише методологією”* [62, с.7]. Науковці мотивують це тим, що корпусна лінгвістика лише досліджує мовний матеріал, і як принцип для вивчення мови має свою теоретичну базу, але цього підгрунтя недостатньо, щоб бути самостійною науковою теорією. На противагу, традиційні галузі мовознавства детально досліджують окремі аспекти мовної системи.

Не зважаючи на авторитетну позицію зарубіжних науковців, переважна кількість вітчизняних лінгвістів відстоюють свої погляди. На їх думку, галузь має:

“1) власний предмет вивчення;

2) об'єкт вивчення;

3) мету дослідження;

4) терміноапарат і теоретичну базу;

5) власні прийоми і практичне значення.

Це дає підстави наполягати на тому, що корпусна лінгвістика є самостійною дисципліною” [14, с.8,12; 2, с.112-117].

Важливим є погляд Володимира Олександровича Плунгяна – члена-кореспондента РАН, завідувача відділу корпусної лінгвістики Інституту російської мови ім. В.В. Виноградова РАН, професора МДУ ім. М.В. Ломоносова. Він переконаний: *“корпус – це не просто новий і потужний інструмент: за використанням корпусу стоїть певна ідеологія, основні тенденції якої зародилися ще в класичній філології XIX століття, але значно інтенсифікувалися в останні десятиліття.*

1) основна увага прикута саме до тексту, як до інструмента комунікації, а не до його фрагментів – слова чи речення;

2) враховується квантитативний компонент мови, примат фактору квантитативних відношень в еволюції мови та структури її правил;

3) значущість синхронічної варіативності мови: не існує єдиної визначеної структури засобів вираження змісту, лише різні варіації, залежні від соціальних, психологічних та біологічних факторів;

4) приділяється увага діахронічній варіативності мови, її змінам у часі”[42, с. 7-20].

Для того, щоб правильно обробити лінгвістичні дані і змоделювати функціонування мови у різних галузях, умовах та ситуаціях, імплементувати потрібні аспекти корпусного аналізу в інших мовознавчих дисциплінах, обов’язково треба використовувати комп’ютерні засоби. Сучасні технології є невід’ємною часткою процесів автоматичного перекладу, добування інформації з природних текстів, створення практичних інструментів для взаємодії людини і комп’ютера тощо.

Алгоритм корпусного аналізу складається з трьох кроків:

1) категорійний аналіз, що ідентифікує мовні дані;

2) використання статистичних методів для підпорядкування мовних даних;

3) інтерпретація результатів.

Як зазначає Тьюберт, *“інтерпретація потребує інтелектуальної діяльності науковця, оскільки інтерпретація не може бути алгоритмізована. Таке зведення мови до набору процедур і є головною відмінністю між корпусною та комп’ютерною лінгвістикою”* [72, с.113]. Але це не означає, що галузі мовознавства не доповнюють одна одну. Спеціалізовані комп’ютерні програми для аналізу та оброблення об’ємних масивів мовних даних, які використовуються у корпусній лінгвістиці, - парсери, тегери, конкордансери, є практичним доробком лінгвістики.

Чим сучасний корпус текстів відрізняється від простої збірки текстів в електронній формі? Для цього треба заглибитися у деталі того, як мовознавці визначають саме поняття корпусу текстів.

На думку Й. Асмуссена, *“корпус є словесною єдністю, у якій текст укладається для подальшого лінгвістичного аналізу. Алгоритм для класифікації і структуризації текстів підпорядковується загальній логічній ідеї. Тексти мають бути репрезентативною вибіркою для конкретного використання мови”* [54, с. 123]. Ми можемо виділити такі дистинктивні ознаки, які допомагають відрізнити лінгвістичний корпус текстів від звичайної збірки текстів:

- Репрезентативність. Корпус містить велику кількість мовних феноменів, які підлягають подальшому лінгвістичному аналізу.
- Автентичність. Тексти створені реальними носіями мови для мовної комунікації.
- Відібраність. Вибірку створюють за чіткими параметрами, які відповідають формату корпусу та меті його створення.
- Збалансованість. Для збалансованості тексти корпусу потребують розрізнення та пропорційності жанрів.

- Машиночитаність. Корпус адаптований до комп'ютерної обробки, у ньому легко обробляються великі масиви даних. Електронні корпуси збагачені метатекстовою розміткою, а вплив людського фактору на результат аналізу даних неможливий.

О.Демська-Кульчицька класифікує корпуси за шістьма типами дихотомії :

I. Повнотекстові корпуси містяться тексти повного обсягу, а у фрагментарних корпусах відповідно уривки.

II. Метою дослідницьких корпусів є створення нових наукових гіпотез. Ілюстративні корпуси підтверджують вже існуючі теорії.

III. В дослідницьких корпусах тексти представлені у вигляді цілісних об'єктів. Тексти інтерпретаційних корпусів є довідковими системами.

IV. Діахронні корпуси репрезентують мову впродовж тривалого періоду. Синхронні корпуси репрезентують мову на певному часовому відрізку.

V. За допомогою динамічних корпусів можна відстежити зміни у мові. Статичні корпуси фіксують мовну ситуацію на конкретному проміжку часу.

VI. У загальномовних корпусах відображена національна мова. "Спеціалізовані корпуси мають розв'язувати нетипові лінгвістичні задачі" [13, с. 156-157].

Вважається, що формування корпусної лінгвістики як самостійної дисципліни розпочалось у 60-х роках минулого століття. Саме тоді з'явилися перші комп'ютерні корпуси. Е.Тогніні-Бонеллі розділяє історію розвитку на такі етапи:

"Етап 1(від середини 60-х до початку 80-х років XX століття).

Під час цього періоду формувалися знання про створення й підтримку масштабних корпусів – до 1 млн. слів. Але у цей період електронні корпуси були відсутні, тому текст треба було набирати вручну.

Етап 2 (1980-ті – 2000-ні рр). Цей етап розділяють на два періоди по 10 років:

а) у 1980-х з'явились перші сканери. Відтепер об'єм корпусу міг сягати 20 млн. слововживань;

б) у 1990-ті відбувся новий стрибок у розвитку комп'ютерних технологій. Через це у електронному форматі з'явились великі за обсягом тексти.

Етап 3 (початок 2000-х – наш час) - період створення оригінальних онлайн-текстів без фізичного аналогу. Цей етап визначається появою не бачених

раніше умов для створення корпусів необмеженого розміру" [76, с. 16-17].

Найбільш визначним корпусом, що виник під час першого етапу розвитку, справедливо вважають Браунівський корпус. Його укладанням займались вчені Браунівського університету - Г.Кучер та Н.Френсіс. Сам корпус містив 500 текстів з сумарним обсягом приблизно 1 млн. слів. Тексти відбирались у 1961 році серед найпопулярніших жанрів друкованої прози в США. Серед особливостей корпусу можна визначити наявність таких переваг, як: походження і склад текстів, жанрова різноманітність, можливість обробити тексти на комп'ютері.

Знаковим корпусом другого етапу є Британський національний корпус. Обсяг корпусу становив 100 млн. слів – це робить його першим мега-корпусом. Унікальність та історична цінність корпусу полягає в тому, що згодом за його стандартом були створені корпуси інших європейських мов. Наприклад, національні корпуси іспанської, італійської, чеської, хорватської мов.

Переважна більшість сучасних корпусів є фундаментальними [22]. Сьогодні технології досягли настільки великого рівня розвитку, що процес укладання необмежених за слововживаннями онлайн-корпусів став відносно нескладним. Основні зусилля науковці докладають до проведення наукових досліджень із метою розширення теоретичної бази корпусної лінгвістики.

Активно розвивається галузь лексичної граматики, лексикографії, когнітивної лінгвістики, стилістики, перекладознавства, фразеології.

Хоча в Україні корпусне мовознавство перебуває на етапі становлення, вітчизняні лінгвісти мають важливі доробки, якими можна пишатися. Академік НАН В.А.Широков керує розвитком Українського національного лінгвістичного корпусу (УНЛК) [50]. Корпус є безцінним джерелом для вивчення сучасної української літературної мови. Він містить тексти українською літератури із початку XIX до XXI століття. Обсяг УНЛК становить 100 млн. слововживань. УНЛК допомагає вирішити важливі завдання української лінгвістики. На основі УНЛК створюються сучасні словники, розробляються перекладні системи для комп'ютерного опрацювання мови, створюються системи автоматичного редагування та аналізатори у галузі морфології, синтаксису і семантики. Але цей корпус не є доступним для широкого кола користувачів.

Активно розвивається Дослідницький корпус сучасної української мови, над яким працюють лінгвісти Лабораторії комп'ютерної лінгвістики Інституту філології КНУ ім. Тараса Шевченка під керівництвом Н.Дарчук. Тексти корпусу допомагають у роботі над багатьма питаннями сучасної лінгвістики. Сам корпус представлений у виді текстів в електронній формі. Структурно корпус є інформаційно-довідковою системою.

На історичному тлі розвитку корпусного мовознавства України один з найвизначніших внесків зробила Н. Дарчук. Лише за останні 13 років Дарчук стала авторкою або соавторкою п'яти наукових робіт у галузі корпусної лінгвістики:

1. "Комп'ютерна лінгвістика (автоматичне опрацювання тексту), 2008";
2. "Корпус украинского языка, 2010"[7];
3. "Дослідницький корпус української мови: основні засади і перспективи, 2010"[7];

4. “Комп'ютерне анотування тексту: результати і перспективи: монографія, 2013”[11];
5. “АГАТ-семантика: семантична розмітка Корпусу української мови, 2016”[10].

Сьогодні у режимі вільного доступу в інтернеті наявні такі українськомовні корпуси: Корпус української мови (КУМ) [30], Генеральний регіонально анотований корпус української мови (ГРАК) [5], Корпуси текстів української мови [28], Браунський корпус української мови (БрУК) [4]. Існує Комп'ютерний фонд інновацій (КФІ) [24]. Першим володіє Український мовно-інформаційний фонд НАН України, а другим - Інститут української мови НАН України. Корпуси використовуються у дослідженнях галузей лексикології, лексикографії, структурно-математичної лінгвістики та для укладання словників [21].

Поміж усіх зазначених корпусів саме Корпус української мови, який постійно вдосконалюється колективом комп'ютерної лінгвістики Інституту філології Київського національного університету імені Тараса Шевченка під керівництвом доктора філологічних наук Н. Дарчук, містить найбільш грунтовну і розгорнуту параметризацію.

Корпус має три структурно-функціональні зони, пов’язан між собою:

1) модуль-текст - містить українські тексти в електронній формі;

2) модуль-аналізатор - автоматично шукає мовні явища за допомогою спеціального програмного забезпечення;

3) модуль-словник, - структурує результати автоматичного аналізу тексту в електронних словниках, доступ до яких надається користувачам в інтернеті.

1.2. Статистична лексикографія: етапи розвитку та значущість для лінгвістичних досліджень

За визначенням відомої української науковиці В.Перебийніс *“статистична лексикографія – галузь лінгвістики, яка спеціалізується на проблемі укладання частотних словників. Частотні словники (далі ЧС) – словники, у яких кожній словоформі відповідає число, що вказує на частоту вживання словоформи у корпусі”* [40,с. 30]. Актуальність статистичної лексикографії і частотних словників спричинена потребою мовознавців у поглибленні розуміння того, як мовні одиниці функціонують у тексті.

XIX століття - початок відліку сучасної історії частотних словників. У ті часи їх методологія і будова були доволі примітивними: 1) частотні словники не були оптимальної довжини, об’єм вибірок іноді сягав мільйонів слововживань; 2) фактор однорідності взагалі не враховувався під час укладання вибірок. Однак, це було логічно - чітко встановлених правил організації вибірки просто не існувало, і мовознавці були переконані в тому, що більший обсяг вибірки у будь-якому разі означає кращу репрезентативність мови.

Не зважаючи на велику кількість помилок, частотні словники доволі швидко стали корисними для вирішення лінгвістичних проблем, а саме використовувались для завдань морфології та словотворення. Мовознавці усвідомили, що важлива граматична інформація може міститися наприкінці слова. Розпочалось укладання інверсійних ЧС із розташуванням слів за алфавітом, але з кінця слова. Наприклад, флексії та суфікси групувались разом з метою визначення їх активності чи пасивності у кінці слова.

З часом мовознавці зрозуміли, що ЧС дає можливість проаналізувати тексти різних стилів і жанрів, отримати персоналізовану статистику, тому почали укладати галузеві ЧС. Стрімкий розвиток технологій ознаменував появу ЧС, укладених за допомогою комп’ютерних систем. Створюються словники для десятків мов, і ця цифра постійно зростає.

Однією з найвизначніших подій для української лінгвістики стало створення спільними зусиллями Академії наук УРСР та Інституту мовознавства ім. О.О. Потебні “Частотного словника сучасної української художньої прози” у 1981 році. (41)

В.І. Перебийніс пише: *“Це словник лем і словоформ. У кожній статті є лема з усіма словозмінними формами у вибірці, сума яких становить частоту леми. Вибірка сягає 500 тисяч слововживань, по 20 тисяч із 25 творів 22 авторів. Вибірка є хронологічно однорідною, оскільки всі твори написані після 1945 року і вкладаються у часовий проміжок у 30 років”*[36, с.15].

Словник розділений на пряму та авторську мову з процентним співвідношенням 28,2% на 71,8%. Обсяг прямої мови становить 141290 слововживань, авторської – 358967. ЧС використовуються стандартні статистичні дані: відносна частота, середня частота на 1000 слововживань, абсолютна частота, міра коливання середньої частоти, кількість джерел і мінімальних підвибірок, які мають слово чи словоформу.

ЧС сучасної української художньої прози роблять унікальним дві характеристики: 1) частота слів і словоформ репрезентує два види художньої прози; 2) статистичні показники містять не тільки частоту, а й рівномірний розподіл усіх одиниць підрахунку по вибірці.

В.І. Перебийніс переконана: *“це робить ЧС надзвичайно корисним для багатьох галузей. Словник можна імплементувати у такі сфери:*

1) *методика викладання української мови, формування лексичного і граматичного мінімуму, який потрібен учням;*
2) *створення навчальних матеріалів з української мови;*
3) *визначення таких показників слова, як ступінь уживаності, поширеності й рівномірності для теоретичних студій з україністики;*
4) *зіставлення багатьох категорій лексики різних мов;*

5) *створення програм для аналізу граматичних та лексичних одиниць у об'ємних текстах"* (39, с. 44).

Від самого початку існувало 3 основні причини створення ЧС:

1) для встановлення вживаності слів у викладанні іноземної мови ;

2) для відбору лексичного мінімуму (слово з високою частотою вживання є більш цінним на ранньому етапі вивчення мови);

3) для вдосконалення стенографії.

У підручнику "Частотні словники та їх використання" зазначено: *"Науковці також використовують ЧС для вирішення проблем математичної лінгвістики. За допомогою рангових списків у ЧС ми бачимо, що тільки одне слово є найчастотнішим. Найчастотніше слово є одним зі статистичних параметрів, які допомагають проаналізувати лексичні та стилістичні відмінності у різних за жанрами текстах. На основі даних ЧС був відкритий закон Ципфа, згідно з яким частота слова відповідає номеру слова у списку спадності частоти. Наприклад, четверте за вживаністю слово буде траплятися в чотири рази рідше, ніж перше. За статистичною структурою тексту можна порівнювати авторські стилі, використовуючи індекс різноманітності, індекс винятковості, індекс концентрації тощо"* [41, с. 127].

Частотні словники корисні і для вирішення проблем загального мовознавства. Наприклад, зв'язок лексично і граматичної семантики допомагає дізнатись семантику слова: найвища частота називного відмінка іменника свідчить про те, що він означає живу істоту. Також за допомогою ЧС можна більше дізнатись про справжню діяльність організації, вивчаючи навіть ті документи, що є у відкритому доступі. В.І. Перебийніс та В.М. Сорокін у своїй роботі "Традиційна та комп'ютерна лексикографія" описали цікавий випадок, який відбувся на початку 2000-х, і пов'язаний з міжнародним альянсом НАТО[40]. Через особливості політичного клімату в Україні ставлення до

організації, як і до потенційної інтеграції в неї нашої батьківщини, у ті роки було суперечливим. На сьогоднішній день після агресії Росії на Донбасі та анексії Криму більшості населення зрозуміло, що мілітаристичний профіль НАТО навпаки є перевагою для України. Але в ті роки суспільство бентежило, чи є програмні заяви НАТО про соціально-економічну допомогу іншим країнам та глобальні миротворчі наміри лише ширмою для активного розвитку військового напряму альянсу. Тож лінгвісти легко проаналізували документи організації. Для дослідження був взятий корпус текстів НАТО обсягом в 300 000 слововживань. Вибірка містила офіційні документи НАТО за 2002 рік, дебати членів організації, документи про структуру НАТО і програмні заяви про напрямки діяльності для зв'язків з громадськістю. Потім було створено частотний словник для подальшого аналізу і знаходження відповіді на такі питання: 1) якою є тематика найчастотніших слів; 2) яка частка лексики з військовою тематикою серед усієї лексики.

Результати аналізу, відображені у ранговому списку, найчастотніші повнозначні слова: *1) military – 1772; 2) security – 1528; 3) defense – 1445; 4) country – 1405; 5) alliance – 1336; 6) cooperation – 926; 7) council – 907; 8)force – 871; 9) state – 850; 10) united – 801.*

Такі іменники, як **military, defense, force** та **security** мають військово-оборонну конотацію. З впевненістю можна сказати, що лише слова **country** та **state** не мають військової семантики. З огляду на те, що дані рангового списку все ще не дають інформації про контекст і цими результатами можна маніпулювати, ми можемо зробити перші обережні висновки, що дійсно - основний напрямок діяльності НАТО передусім пов'язаний з військовою тематикою, а не суспільно-політичною.

Отже, дослідження ЧС важливі не тільки для статистичної лексикографії чи математичної лінгвістики, а і для мовознавства в цілому. Такі фактори, як постійний розвиток технологій, повсюдний попит на обробку великих даних у

кожній діджиталізованій сфері суспільства і універсальність використання гарантує актуальність частотних словників у майбутньому.

Класифікація ЧС базується на таких параметрах*: "1) одиниці підрахунку, 2) обсяг вибірки, на якій укладено ЧС, 3) характер вибірки, 4) обсяг словника, 5) спосіб подачі матеріалу в ЧС, 6) статистичні характеристики одиниць ЧС"* [41, с. 5-14].

Розглянемо кожен з параметрів, які обгрунтовані в підручнику В.І.Перебийніс:

1) Одиниці підрахунку. Можуть виступати різні одиниці мови, але лише за тієї умови, що вони мають формальне визначення. Бо формальне визначення виключає суб'єктивізм. Формальне визначення окреслюється такими способами: 1) приписання будови одиниці - склади, морфеми, фонеми слова; 2) визначення меж слова, наприклад, пропуск між слововживаннями; 3) такі диференційні ознаки, як суфікс чи префікс, написання з великої літери; 4) оточення слова; 5) присутність визначеного певного коду - цифри, літери, їх комбінації.

За характером одиниць частотні словники поділяються на ЧС словоформ, лем, словосполучень, речень. Такі ЧС містять дані про словозміну, лексичні одиниці і допомагають розв'язувати теоретичні завдання, пов'язані з граматичню, функціональною і семантичної класифікацією лексики [38].

Також існує пласт "псевдо-словників", які не можуть бути словниками в класичному розумінні тому, що повноцінні лексичні одиниці не є одиницями підрахунку. До таких "псевдо-словників" належать ЧС фонемосполучень, складів, морфем, афіксів. Правильним було б назвати їх частотними списками.

Словосполучення також є одиницями підрахунку і становлять матеріал відповідного словника сполучуваності. Ці словники використовують у питаннях функціонування лексики і побудови тексту, але такі чс складно створювати. На

це є дві причини: 1) оскільки словосполучення трапляються рідко, треба обробити дуже велику кількість тексту, щоб отримати достатньо матеріалу; 2) межі словосполучення важко окреслити без потрібного контексту - іноді ланцюжок словосполучень можна легко сплутати з одним словосполученням.

2) Обсяг вибірки. Проблеми, які виникають із визначенням обсягу, полягають у тому, що неможливо охопити масив текстів кожного стилю чи жанру через їх величезний об'єм. Така робота буде дуже довгою і малоефективною. Дуже великі масиви називаються генеральню сукупністю і для укладання ЧС на їх основі треба відібрати ту кількість текстів, яка буде достатньою для отримання правильних статистичних результатів. *"Є декілька способів організації вибірки, які гарантують те, що всі тексти вибірки мають однакову ймовірність потрапити у масив для подальшого статистичного дослідження"* [36, с. 13-24].

Розрізняють чотири види вибірок за обсягом: 1) великі, що мають мінімум 1 млн. слововживань; 2) середні - містять від 999 тис. до 400 тис. слововживань; невеликі, від 399 до 100 тис. слововживань; 4) мікро-ЧС, у яких обсяг становить менша 100 тис. Незалежно від кількості слововживань, кожна вибірка має відповідати певним критеріям: не містити тексти різних стилів і жанрів, тексти мають бути близькі один хронологічно. Оскільки кожен функціональний стиль чи жанр має власні лінгвістичні характеристики, відповідно не можна об'єднувати різнорідні тексти - це не збагачує вибірку, а лише псує її чистоту і таким чином шкодить істинності статистичних результатів. Отже, вибірки укладені з таких текстів є статистично та лінгвістично хибними.

3) Характер вибірки. За цією характеристикою розрізняють механічні, випадкові і зональні вибірки. Вони відрізняються способом підрахунку тексту. Такі методи існують для того, щоб усі одиниці підрахунки опинились у вибірці і зробили її більш репрезентативною. Для механічної обирають певну послідовність, наприклад кожну 5-ту чи 10-ту сторінку. Механічною вибіркою

найзручніше аналізувати велику книгу. Для випадкової вибірки через таблицю випадкових чисел обирається уривок, на основі якого потім будуть вести підрахунки. Але найбільш розповсюдженою є зональна вибірка. Створюється список джерел, з якого вибираються ті, що підлягають обстеженню. Уривки тексту з кожного джерела мають бути однакової довжини, але іноді беруть і весь текст з джерела. Кожна вибірка має бути поділена на однакові за довжиною підвибірки з підрахованою частотою одиниці, незалежно від обраного методу організації. Одержані дані можна використати для визначення середньої частоти.

4) Обсяг ЧС. Показник залежить від кількості відображених у ньому одиниць. Словники поділяються на повні (нараховують усі одиниці, від найчастотніших до тих, які можна знайти лише один раз) та неповні (містять одиниці вище мінімального порога частоти).

5) Спосіб подання матеріалу в ЧС. Зазвичай, використовують 2 способи подання одиниць в ЧС: алфавітно-частотний та ранговий списки. В алфавітно-частотному списку лінгвістичні та статистичні дані мають стандартну структуру - одиниця, її граматична характеристика та абсолютна частота. Щодо інших словників, то спосіб подання матеріалу в них залежить від самої тематики ЧС. Якщо звернутися до ЧС української художньої прози, то ми побачимо, що в ньому приводяться атрибути авторської мови, мови персонажів, авторська і пряма мова об’єднується. Одиничні словоформи виведені в окремий список. [49].

6) Статистичні характеристики ЧС. Найуживанішою характеристикою, яку можна побачити у різноманітних ЧС є абсолютна частота. Це показник того, скільки разів слово зустрічається у вибірці. Також нерідко вираховують і відносну частоту - відсотковий коефіцієнт абсолютної частоти у заданій вибірці. Одиниці мають свій порядковий номер у спадному ряді частоти - це називається ранг одиниці, що відображений у ранговому списку. Крім вказівки на частоту є

також інформація про розподіл одиниць підрахунку за частотою. Вона наводиться у ранговому списку, у якому кожна одиниця одержує ранг, тобто число, яке показує її місце в ряду одиниць, розташованих за спадом частоти.

Правильно і ретельно аналізуючи статистичні дані ЧС, лінгвіст може отримати великий об'єм інформації: для словників авторських стилів можна вивести лексичний спектр; відобразити співзалежність реєстру і тексту різноманітних авторів, визначити темпи приросту нової лексики. Порахувати індекси різноманітності, винятковості та концентрації, які також будуються на основі різних видів пропорції реєстру до тексту тощо [41]. Стандартний алгоритм статистичного аналізу ЧС включає в себе: 1) поділ вибірки на мінімальні підвибірки однакової довжини; 2) визначення абсолютної частоти; 3) побудова варіаційного ряду абсолютних частот; 4) обчислення середньої частоти, міри коливання; 5) обчислення критерію Стьюдента. Виконання усіх етапів дозволить "витягти" максимум інформації з статистичних показників вибірки для подальшої роботи лінгвістів. [36].

У підсумку, ґрунтовний аналіз статистичних показників частотного словника поліпшує наукову базу і полегшує роботу лінгвістам, які збирають інформацію про стилі, досліджують особливості авторського стилю, встановлюють вплив письменників на певні стилі, займаються атрибуцією анонімних творів, встановлюють авторство творів тощо.

ВИСНОВКИ ДО РОЗДІЛУ 1

У цьому розділі було визначено усі основні етапи розвитку корпусної лінгвістики та статистичної лексикографії у світовій та вітчизняній лінгвістиці. Були розглянуті особливості укладання словників, проблеми, які можуть виникати у цьому процесі на шляху, з'ясовані шляхи до їх вирішення. Проаналізовані ознаки, що відрізняють корпусну лінгвістику від простого електронного словника. Висвітлена міжнародна лінгвістична дискусія з приводу

того, чи можна вважати корпусну науку окремою галуззю мовознавства. Була надана детальна характеристика частотного словника та його повна класифікація. Було розглянуто важливість корпусної науки та статистичної лексикографії для мовознавства та суспільства. Такі фактори, як постійний розвиток технологій, повсюдний попит на обробку великих даних у кожній діджиталізованій сфері суспільства є базисом для актуальності цих сфер у майбутньому.

РОЗДІЛ ІІ. СТВОРЕННЯ КОРПУСУ УКРАЇНСЬКОМОВНИХ МЕДІАТЕКСТІВ ТА АВТОМАТИЧНА СТАТИСТИЧНА ПАРАМЕТРИЗАЦІЯ ТЕКСТОВИХ ДАНИХ

2.1 Конвертація текстових матеріалів у текстові дані

Дослідження сучасної українськомовної поп-культури включає в себе створення корпусу медійних українськомовних текстів, укладання на основі корпусу загального частотного словника медійних текстів та тематичних частотних словників, тональний сентимент-аналіз медійних текстів. Підсумком роботи є лінгвістичний аналіз отриманих статистичних характеристик текстових вибірок. Метою дослідження є виявлення взаємовідношень, спільних та відмінних рис між медійними інтернет-платформами і жанрами, які формують сучасну українську популярну культуру;

Етапи дослідження:

1. Глибокий огляд сфер, у яких представлена сучасна українськомовна поп-культура - прочитування медійних текстів, перегляд відеоматеріалів на хостингу YouTube, прослуховування подкастів, перегляд фільмів, слідкування за українськомовними інфлуенсерами на персональних сайтах; збір відкритої статистики, яка свідчить про рівень популярності представників обраних медійних сфер.

2. Формування текстових вибірок - переведення відео та аудіоматеріалів у текстовий формат, створення корпусу медійних текстів, укладеного на базі матеріалів тематичних вибірок, створення метатекстової розмітки для корпусу.

3. Опрацювання текстових вибірок - створення програм мовою програмування Python для таких цілей:

1) автоматична обробка назв текстовий файлів для усунення багів при розпізнаванні;

2) уніфікація кодування текстових файлів;

3) усунення помилок токенізації;

4) створення частотного словника словоформ і лем для загальної бази даних і для кожної тематичної вибірки;

5) підрахунок статистики для загальної бази даних і кожної тематичної вибірки;

6) створення конкордансу медійних текстів;

7) створення бази даних тонального сентимент-аналізу медійних текстів;

8) виведення статистики для загальної бази даних і тематичних словників;

4. Лінгвістичний аналіз статистичного матеріалу – визначення статистичних параметрів лексики, обрахування індексу різноманітності, індексу винятковості, індексу концентрації, створення таблиці відносних частот частин мови, визначення психолінгвістичних статистичних маркерів тексту, побудова графіків та діаграм для репрезентативності підсумків дослідження.

Для того, щоб проаналізувати українськомовну популярну культуру, для початку треба визначитись з матеріалами дослідження. Було обрано такі зразки сучасного українськомовного контенту, який є доступним онлайн:

1) “Телебачення Торонто” - щотижневий гумористичний випуск новин переважно суспільно-політичної тематики. Головний ведучий - український тележурналіст Роман Вінтонів, який виступає в сатиричному образі Майкла Щура. За 9 років існування канал має 464 тис. підписників та 99,9 млн. переглядів; 2) “СПАЛАХ” - документальний серіал про розвиток різних сфер української культури на ютуб-каналі музичного медіа “СЛУХ”. За 3 роки існування канал має 62,7 тис. підписників та 5,4 млн. переглядів; 3) “The Village

Україна” - інтернет-видання для переважно здебільшого молодої та освіченої аудиторії. Основні теми - культурне і громадське життя, інфраструктура в найбільших містах України. Видання також відоме великою кількістю матеріалів про способи всебічного покращення якості життя, а саме: духовний та інтелектуальний розвиток, ментальне здоров’я, кар’єрні перспективи, можливості вирішення розповсюджених життєвих проблем. На сьогодні сумарно має 207 тис. підписників у Facebook та Instagram;

4) “СЛУХ” - інтернет-видання про сучасну українську та світову музику. На сьогодні сумарно має 54,8 тис. підписників у Instagram, Facebook та Telegram.

5) “Мої думки тихі” - українська трагікомедія 2019-го року, дебютна робота українського режисера Антоніо Лукіча. Стрічка вважається одним з найяскравіших представників нової хвилі сучасного українського кінематографу. Стрічку можна переглянути на українських медіасервісах та стримінгових платформах Європи, Америки та Азії. Отримав 12 нагород на вітчизняних та міжнародних фестивалях.

6) Персональний сайт та профіль в Instagram Євгена Клопотенка - культурного і громадського діяча, талановитого шеф-кухаря та кулінарного блогера, телеведучого та підприємця. Клопотенко відомий своїми успішними спробами провести реформи, спрямовані на модернізацію системи харчування в українських закладах освіти. Євген постійно намагається популяризувати традиційні українські страви (наприклад, шпундру чи верещаку), а його основна мета - “через кулінарію нагадати українцям їх ДНК”. Станом на сьогодні кількість його підписників в Instagram та YouTube становить 754 тис.

7) “Простими словами” - подкаст про психологію від письменника та журналіста Марка Лівіна і українського гештальт-психолога Іллі Полудьонного. За даними Apple Podcasts, “Простими словами” є одним із найпопулярніших подкастів в Україні.

8) “MINCULTPRYVIT” - подкаст про культуру, український та світовий кінематограф. Автор - молодий український режисер Наріман Алієв, чия повнометражна стрічка “Додому” отримала 2 номінації на 72-му Каннському кінофестивалі у 2019-му році.

Класифікація медійних елементів за способом споживання контенту/ медійними платформами:

1. Відео контент: “Телебачення Торонто”, “СПАЛАХ”, “Думки мої тихі”. Ці медійні елементи доступні для перегляду онлайн на сайті та застосунках відеохостингу “YouTube” та на офіційних українських стримінгових платформах.

2. Текстовий контент: “СЛУХ”, “The Village Україна”, офіційний сайт та профіль у соцмережі Instagram Євгена Клопотенка.

3. Аудіо контент: “Простими словами”, “MINCULTPRYVIT”. Доступні для прослуховування на міжнародних платформах для подкастів. Подкастинг - створення ток-шоу з ведучим та гостями у діджитал-форматі. Останні 5-10 років подкасти стають все більш популярними через зручність вживання контенту та сильним емоційним зв’язком між ведучим та слухачами. Подкасти переважно прослуховують на мобільних пристроях). (перефразувати)

Класифікація медійних елементів за тематикою:

1. Суспільство і політика: “Телебачення Торонто”;

2. Культура і мистецтво загалом: “The Village Україна”, “Телебачення Торонто”, “СПАЛАХ”;

3. Люди, місто: “The Village Україна”, “Телебачення Торонто”;

4. Музика: “СЛУХ”, “СПАЛАХ”;

5. Кінематограф: “MINCULTPYVIT”, “Мої думки тихі”;

6. Ментальне здоров’я: “Простими словами”, “The Village Україна”;

7. Кулінарія: Сайт та Instagram-профіль Євгена Клопотенка, “СПАЛАХ”.

8. Гумор: “Мої думки тихі”, “Телебачення Торонто”, “СПАЛАХ”.

Відеоматеріали “Телебачення Торонто”, “СПАЛАХ” були переведені у текстовий формат за допомогою функції автоматичного витягу тексту з відео у відеохостингу “YouTube”. Текстові матеріали подкастів “MINCUTPRYVIT”, “Простими словами” та фільму “Думки мої тихі” були отримані емпіричним методом – записувались вручну одразу після прослуховування/перегляду контенту. Було створено 4 корпуси текстів:

1) Корпус медійних українськомовних текстів онлайн-видань, куди увійшли матеріали “The Village Україна” та “СЛУХ” - **50 текстів, 80 251 слововживань ;**
2) Корпус медійних українськомовних текстів з офіційного сайту та Instagram-профілю Євгена Клопотенка - **150 текстів, 53 689 слововживань;**
3) Корпус медійних українськомовних текстів з відеоконтенту “Телебачення Торонто”, “СПАЛАХ”, “Мої думки тихі” - **15 текстів, 58 486 слововживань;**
4) Корпус медійних українськомовних текстів подкастів на матерілаха “MINCULTPRYVIT” та “Простими словами” - **2 текста, 5451 слововживань.**

Кожен корпус мав таку метатекстову розмітку: **джерело, дата, мова, жанр, тема, автор, заголовок, примітки, к-сть слів.** Через стилістичну і тематичну специфіку текстові матеріали із соцмереж Євгена Клопотенка та текст фільму “Мої думки тихі” були виділені в окремі бази даних для зручності подальшого аналізу. Першочергово текст фільму “Мої думки тихі” не був

відокремлений у власний тематичний корпус через замалий обсяг та наявність лише одного медійного елементу.

На основі корпусу медійних українськомовних текстів автоматично за допомогою програмного забезпечення, написаного мовою програмування Python, було створено такі аналітичні інструменти:

1) Частотний словник словоформ медійних українськомовних текстів у 6 екземплярах: 1 загальна БД і 5 тематичних БД;

2) Частотний словник лем медійних українськомовних текстів у 6 екземплярах: 1 загальна БД і 5 тематичних БД;

3) База даних сентимент-аналізу медійних українськомовних текстів;

4) Словник-конкорданс медійних українськомовних текстів.

Усі частотні словники словоформ мають таку структуру **(див. Малюнок 2.1.):**

- Перша колонка – id словоформи, що призначається автоматично, тип даних integer primary key – ціле первинне число, що має залишатись незмінним;
- друга колонка – словоформа (wordform), тип даних varchar(255) – рядок змінної довжини до 255 символів;
- третя колонка - абсолютна частота (frequency), тип даних integer – ціле число;
- четверта колонка - лема (lemma), тип даних varchar(255) – рядок змінної довжини до 255 символів;
- п’ята колонка - частина мови (part of speech);

- шоста колонка - відносна частота (relative_frequency);

	id	wordform	frequency	lemma	pos	rel_freq
22	253	сказати	82	сказати	VERB	0.0003242824
23	260	каже	79	казати	VERB	0.0003124184
24	263	три	78	терти	VERB	0.0003084638
25	266	їсти	77	їсти	VERB	0.0003045091
26	275	викладіть	75	викласти	VERB	0.0002965998
27	281	думаю	73	думати	VERB	0.0002886905
28	291	вирішив	72	вирішити	VERB	0.0002847358
29	308	перемішайте	70	перемішати	VERB	0.0002768265
30	321	подобається	67	подобатися	VERB	0.0002649625

Малюнок 2.1. Фрагмент БД словника словоформ у середовищі програми SQLite

Усі частотні словники лем мають структуру: дані про id леми, абсолютну частоту, частину мови, відносну частоту **(див. Малюнок 2.2).**

- Перша колонка – id леми , що призначається автоматично, тип даних integer primary key – ціле первинне число, що має залишатись незмінним;
- друга колонка – лема (lemma), тип даних varchar(255) – рядок змінної довжини до 255 символів;
- третя колонка - абсолютна частота (frequency), тип даних integer – ціле число;
- четверта колонка - частина мови (part of speech);
- п’ята колонка - відносна частота (relative_frequency);

	id	lemma	frequency	pos	rel_freq
31	403	сталий	67	ADJF	0.0002649625
32	424	сирий	63	ADJF	0.0002491438
33	437	схожий	62	ADJF	0.0002451892
34	447	сучасний	61	ADJF	0.0002412345
35	454	оливковий	59	ADJF	0.0002333252
36	459	покроковий	59	ADJF	0.0002333252
37	467	єдиний	58	ADJF	0.0002293705
38	468	хороший	58	ADJF	0.0002293705
39	476	детальний	57	ADJF	0.0002254158
40	481	ніжний	56	ADJF	0.0002214612
41	496	важливий	55	ADJF	0.0002175065

Малюнок 2.2. Фрагмент БД словника лем у середовищі програми SQLite

Таблиці створені за допомогою функції db.execute:

```
db.execute('''CREATE TABLE wordform_freq
        (id INTEGER PRIMARY KEY,
         wordform VARCHAR(255),
         frequency INTEGER);''')
db.execute('''CREATE TABLE lemma_freq
        (id INTEGER PRIMARY KEY,
        lemma VARCHAR(255),
        frequency INTEGER);''')
database.commit()
```

Абсолютна частота автоматично вираховується за допомогою функції FreqDist: fd = FreqDist(words)

```
for word in fd.most_common():
        db.execute('INSERT INTO wordform_freq (wordform, frequency) VALUES(?, ?);' (word[0], word[1]))  database.commit()
```

Відносна частота вираховується за допомогою функції rel_freq**(див. Додаток 2):**

```
  for r in rows:
      db.execute('UPDATE lemma_freq SET pos = ?, rel_freq = ? WHERE id = ?;',
            (morph.parse(r[1])[0].tag.POS,
             round(r[2]/word_sum, 10),
             r[0]))
```

Лематизація здійснюється за допомогою функції MorphAnalyzer.parse (morph.parse) бібліотеки pymorphy2:

lemmas = [morph.parse(word)[0].normal_form for word in words]

print(list(lemmas)[:100])

fd = FreqDist(lemmas)

Кодифікація грамем у pymorphy2 згідно з даними порталу OpenCorpora [35]:

NOUN – іменник; **ADJF –** прикметник повна форма; **ADJS –** прикметник скорочена форма; **COMP –** ступінь порівняння; **VERB –** особова форма дієслова; **INFN –** інфінітив; **PRTF –** дієприкметник повна форма; **PRTS –** диєприкметник коротка форма; **GRND –** дієприслівник; **NUMR –** числівник; **ADVB –** прислівник; **NPRO –** узагальнено-предметний займенник; **PRED –** предикатив; **PREP –** прийменник; **CONJ –** сполучник; **PRCL –** частка; **INTJ -** вигук.

База даних сентимент-аналізу текстів має структуру:

- Перша колонка - id, що призначається автоматично, тип даних integer primary key – ціле первинне число, що має залишатись незмінним;
- друга колонка – id текстового файлу (fileid), тип даних varchar (50) – рядок змінної довжини до 50 символів
- третя колонка – оцінка тональності (score), тип даних integer – ціле число;
- четверта колонка – визначення позитивності/негативності/нейтральності тексту (sentiment), тип даних char – символьний тип даних з максимальною довжиною у 3 символи.

БД сентимент-аналізу українськомовних медійних текстів було створено **(див. Додаток 6)** за допомогою українського тонального словника, який знаходиться у відкритому доступі на порталі GitHub[15]. За словами авторів словник містить 3442 слів української мови, які мають не нейтральну тональність. Тональність лексики визначали за усередненням оцінок різних експертів. Словник створювався за допомогою ML-моделі з використанням векторів слів word2vec та lex2vec. Слова зведені до базової граматичної форми, прислівники змінені на спільнокореневі прикметники. Формат даних — tab-separated з наступними колонками: слово, дискретна тональність (з діапазону: -2, -1, 0, 1, 2), усереднена тональність.

Парсинг тонального словника відбувався з файлу 'tone-dict-uk.tsv':

```
with open('tone-dict-uk.tsv', 'r', encoding='utf-8') as file:

lines = file.read().split('\n')

tones = dict()

for line in lines:

line = line.split('                                                    ')

tones[line[0]] = int(line[1])

with open('tone_dict.pickle', 'wb') as file:

pickle.dump(tones, file)
```

Алгоритм не враховує правила, а зважає лише на дані словника. Алгоритм ігнорує слова, яких немає в словнику. Якщо оцінка слова більше, ніж 2, то алгоритм визначає його позитивним. Якщо менше, ніж 2 – відповідно слово має негативний сентимент. В інших випадках слову надається нейтральна характеристика **(див. Малюнок 2.3)**.

	id	fileid	score	sentiment
1	34	klop/klop_33.txt	-8	neg
2	35	klop/klop_34.txt	-3	neg
3	49	klop/klop_48.txt	-5	neg
4	87	klop/klop_86.txt	-5	neg
5	89	klop/klop_88.txt	-4	neg
6	101	movies/mov_00.txt	-16	neg

Малюнок 2.3. Фрагмент БД сентиментів у середовищі програми SQLite

Створені частотні словники й таблиця тональності текстів зберігаються у файлах бази даних (.db). Для кожного жанру текстів був створений окремий файл .db, і також окремий файл .db для всього корпусу. Для створення цих баз даних використовували SQLite. SQL є структурованою мовою запитів, яка використовується для запитів реляційної системи баз даних. SQL є стандартом, який визначає, як створюється реляційна схема, вставляються або оновлюються дані у зв'язку між таблицями, запускаються та зупиняються транзакції тощо. SQL - це мова запитів, яка використовується іншими базами даних SQL. SQL не є самою базою даних. Основними компонентами SQL є мова визначення даних (DDL), мова керування даними (DML), мова контролю даних (DCL) [56]. Але для завдань частотного словника, функціоналу SQLite достатньо.

Для швидкого доступу до бази даних використовуємо програмне забезпечення SQLiteStudio. SQLite - це програмне забезпечення, яке надає реляційну систему управління базами даних. Програма SQLite є легкою у налаштуванні, адмініструванні баз даних та необхідних ресурсів. Серед основних функцій SQLite можемо перерахувати автономність, безсерверність, нульову конфігурацію. Тобто SQLite навіть не потребує повноцінного встановлення на комп'ютер [52]. SQLite також має схожі можливості з базами даних високого класу - наприклад, вміє проводити повнотекстову індексацію та володіє підтримкою даних JSON. Дані застосунків, як правило, заповнені в напівструктуровані формати, такі як YAML або XML, можуть зберігатися як

таблиці SQLite, що дозволяє легше отримати доступ до даних та швидше їх обробити.

SQLite також забезпечує швидкий та потужний спосіб зберігання даних конфігурації програми. Замість аналізу формату файлу, такого як YAML, юзер може використовувати SQLite як інтерфейс для цих файлів - часто це набагато швидше, ніж оперувати ними вручну. SQLite може працювати з даними в пам'яті або зовнішніми файлами (наприклад, файлами CSV), ніби вони є власними таблицями бази даних, забезпечуючи зручний спосіб запиту цих даних [80].

Розвиток інтернету та безлічі соціальних мереж за останні 10-15 років, без сумніву, призвели до експоненціального збільшення нових жанрів мовної комунікації:

- незалежні онлайн-видання;
- ютуб-канали;
- персональні сайти;
- блоги у соц.мережах Instagram, Twitter, Facebook;
- аудіоподкасти.

Але який вплив занурення онлайн-комунікації у наше життя мало на розвиток та поширення українськомовної культури? Завдяки інтернету протягом останніх років ми постійно стикаємося з безліччю нових і яскравих українськомовних діячів культури, публіцистів, журналістів, письменників, музикантів, режисерів. Інтернет дає цим неординарним особистостям творчу свободу, якою вони користуються для розвитку україномовної культури і підвищення її популярності серед молодого покоління українців.

Саме позитивний культурний вплив на поглиблення проукраїнського дискурсу в різних суспільних сферах серед молоді та сучасної цільової аудиторії онлайн-медіа став головним критерієм вибору об’єкту дослідження.

“Телебачення Торонто” у кожному випуску популяризує нові українськомовні жарти і меми; проект “СПАЛАХ” створює документальні огляди історії розвитку різних сфер української культури; “The Village Україна” регулярно повідомляє про появу нових українських письменників, дизайнерів, фільмів тощо; “СЛУХ” у своїх музичних оглядах і статтях акцентує увагу саме на українській музиці, повідомляє про появу нових музичних талантів, випускає матеріали про легендарні гурти вітчизняної індустрії; подкаст “Простими словами” детально розглядає психологічні проблеми, притаманні українцям; гостями подкасту “MINCULTPRYVIT” часто стають молоді українські актори та режисери; стрічка “Мої думки тихі” з ніжністю і гумором зображає культурні особливості жителів регіонів Західної України. Медійна діяльність Євгена Клопотенка спрямована на усвідомлення українцями свого генетичного коду і самоідентифікації через популяризацію традиційної української кухні.

Отже, інтернет став одним із найбільших впливів на розквіт української мови останнім часом. Очевидно, соцмережі та медійні онлайн-платформи є майданчиком для експериментів, творення нової лексики та відкриття давно забутої; він також забезпечує платформу для людей, котрі не поглинаються зайвими канцеляризмами та академізмом, а намагаються знайти новаторський, креативний підхід до використання української мови у своїх творчих цілях.

При зборі корпусу українськомовних медійних текстів кожен текст зберігали в окремий документ у форматі .docx. Програмне забезпечення з таким форматом працювати не може. Тому вручну переконвертовували кожен файл із текстом у формат .txt. Для базової версії корпусу була створена таблиця **(див. Додаток 21)** у форматі .xlsx з метатекстовою розміткою **(див. Малюнок 2.4)**, що містить дані про ***джерело, дату, мову, тематику, авторство, додаткові примітки (наприклад, рекламний характер публікації) і кількість слововживань*** у кожному тексті. ***У результаті отримали корпус обсягом 217***

текстових файлів, розподілених по папках відповідно до своїх джерел. Загальний обсяг текстової вибірки становить 197 877 слововживань.

№	джерело	дата	мова	жанр	тема	автор	заголовок	примітки	к-сть слів
1	https://www.the-v	10.11.2020	укр	стаття	діти	Александр Дмит	Мам, скажи, що		2724
2	https://www.the-v	22.08.2020	укр	стаття	діти	Наті Авдєєва	11 фільмів, які в		513
3	https://www.the-v	10.02.2021	укр	стаття	секс, стосунки	Юлія Беба	«Це ж сором і н		2554
4	https://www.the-v	11.01.20201	укр	стаття	секс, стосунки	Юлія Беба	«Це взагалі нор		694
5	https://www.the-v	24.11.2020	укр	інтерв'ю	особистості	Ярослав Друзю	Праворадикали		4260
6	https://www.the-v	17.02.2021	укр	інтерв'ю	кіно, музика	Ярослав Друзю	Яким вийшов но		1738
7	https://www.the-v	27.11.2020	укр	стаття	кіно	The Village Spec	Як Джонні Депп	промо	1778

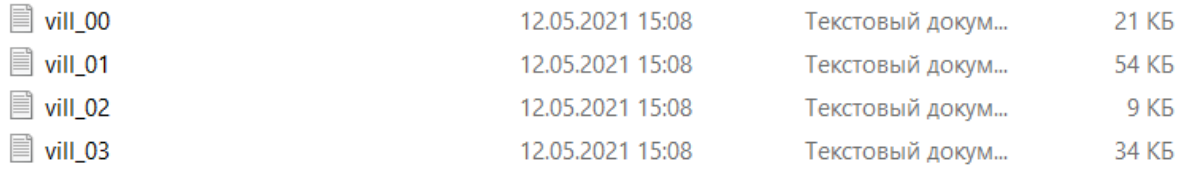
Малюнок 2.4. Фрагмент БД корпусу українськомовних текстів онлайн-видань

Для створення загальної бази даних для всього корпусу, відкривати в коді файли кожен окремо було б незручно й неефективно. Перед нами було два шляхи: вручну скопіювати вміст кожного файлу в один великий текстовий файл з усіма текстами, або піти розумним шляхом і скористатися інструментами оброблення текстів природними мовами, які знаходяться у вільному доступі. Зупинилися на бібліотеці nltk. NLTK - це потужний набір бібліотек і програм Python, який забезпечує користувача зібранням різноманітних алгоритмів природних мов. Він безкоштовний, простий у використанні, добре задокументований та перебуває у відкритому доступі. Бібліотеки NLTK мають широке коло прихильників, тому у разі виникнення будь-яких проблем завжди можна звернутися до тематичних форумів. NLTK складається з найпоширеніших алгоритмів, таких як tokenizing, part-of-speech tagging, stemming, sentiment analysis, topic segmentation, та named entity recognition. NLTK допомагає комп'ютеру аналізувати, попередньо обробляти та розуміти написаний текст [73].

І оскільки корпус - це сукупність текстових документів, для алгоритмів Python корпус є лише купою текстових файлів у каталозі. Тому створити власний корпус текстів на базі бібліотек Python цілком можливо. NLTK визначає список каталогів даних або шляхів у nltk.data.path, тому користувацький корпус має знаходитись в межах цієї директорії. Але для доступу до корпусу nltk.data.load не потрібен, бо дані будуть обробляти класи CorpusReader [66]. Аби ідентифікатори файлів, коли ми відкриватимемо їх за допомогою PlaintextCorpusReader бібліотеки nltk, були короткими і не було конфліктів/багів через кириличні символи/пробіли в назвах файлів, створили програмне забезпечення **(див. Додаток 5),** яке автоматично перейменовувало всі текстові файли в усіх папках корпусу.

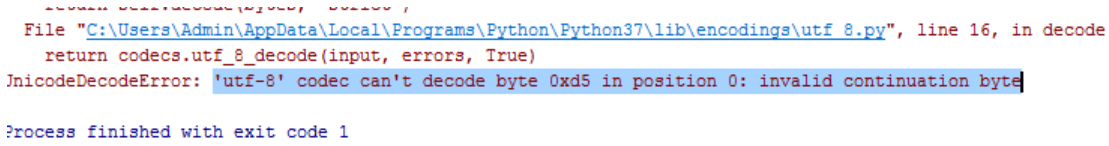
```
for n in range(len(dir)):
    if n < 10:
        os.rename(os.path.join(path, dir[n]),
                  os.path.join(path, 'tor_0' + str(n)+ '.txt'))
```

Нові найменування файлів створювалися за шаблоном: 3-4 латинські символи на позначення джерела тексту, нижнє підкреслення, двоцифрове число на позначення порядкового номеру файлу ***(див. Малюнок 2.5).***



Малюнок 2.5. Результат роботи програми автоматичного перейменування файлів

Наступним етапом попередньої обробки була уніфікація кодування кожного текстового файлу, адже перші спроби відкрити корпус за допомогою PlaintextCorpusReader не були успішними **(див. Малюнок 2.6)**.



Малюнок 2.6. Помилки під час запуску програми

З невідомих причин текстові файли мали різні кодування: одні - UTF-8 (кодування яке підходить для PlaintextCorpusReader), інші - ANSI. ANSI та UTF-8 є форматами кодування. ANSI - загальноприйнятий однобайтовий формат, що використовується для кодування латинського алфавіту; тоді як UTF-8 - це формат змінної довжини (від 1 до 4 байт), який може кодувати всі можливі символи[78]. За допомогою іншої частини коду, автоматично змінили кодування кожного файлу, який був закодований в ANSI, на UTF-8 **(див. Додаток 3).**

```
for n in range(len(dir)):
    with open(path+r'\\'+dir[n], 'r', encoding='ansi') as file:
        text = file.read()
    with open(path+r'\\'+dir[n], 'w', encoding='utf-8') as file:
        file.write(text)
```

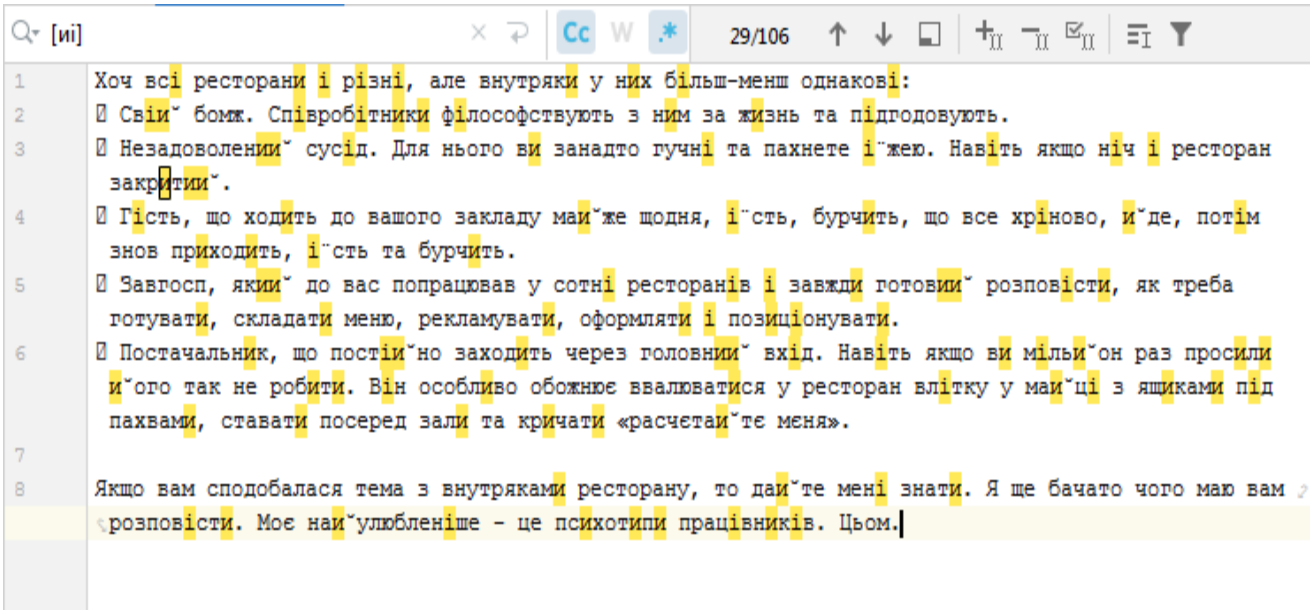
Наступна проблема, яка виникла на шляху зчитування корпусу за допомогою програмного забезпечення, стосувалася поділу текстів на токени (слововживання і знаки пунктуації). Токенізація, яка використовується в nltk, виконувалася з помилками через, як виявилося, наявність нестандартних символів у текстах. Тими символами, що створювали проблеми, були літери "й" і "ї".

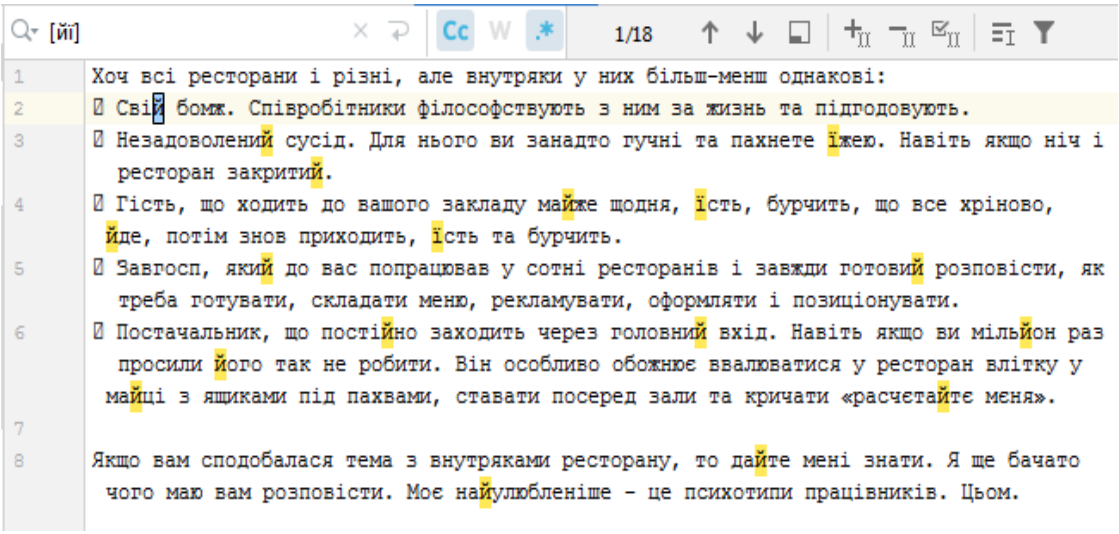
Справа в тому, що існує два способи репрезентації цих символів у комп'ютерному тексті. Є цілісні символи "й" і "ї". А є комбінації символів, які

відображаються як цілісні "й" і "ї", але насправді такими не є. Натомість це пари символів: й = и +˘ , ї = і +¨

За допомогою ще одного алгоритму **(див. Додаток 4)** ми автоматично виправили ці випадки, замінивши ці комбінації символів на цілісні символи "й" та "ї" (для верхнього і нижнього регістру) **(див. Малюнки 2.7 та 2.8).**

```
for n in range(len(dir)):

    with open(local_path+r'\\'+dir[n], 'r', encoding='utf-8') as file:

        text = file.read()

    text = text.replace('й', 'й')

    text = text.replace('ї', 'ї')

    text = text.replace('Й', 'Й')

    text = text.replace('Ї', 'Ї')
```





Малюнки 2.7, 2.8. Варіація тексту після обробки.

Нарешті корпус вдалося завантажити без помилок.

2.2. Автоматичне укладання електронного частотного словника українськомовних медіатекстів

Для роботи з корпусами nltk має такі методи і функції:

- .raw() надає вміст файлу без лінгвістичної обробки;
- .sentences() та .words - дає на вихід усі слова і речення в тексті;
- .categories() - визначає категорії файлів зі списку;
- .fileids() - дає на вихід список ідентифікаторів файлу.

Перший крок:

Створення та наповнення великими БД ЧС словоформ і лем.

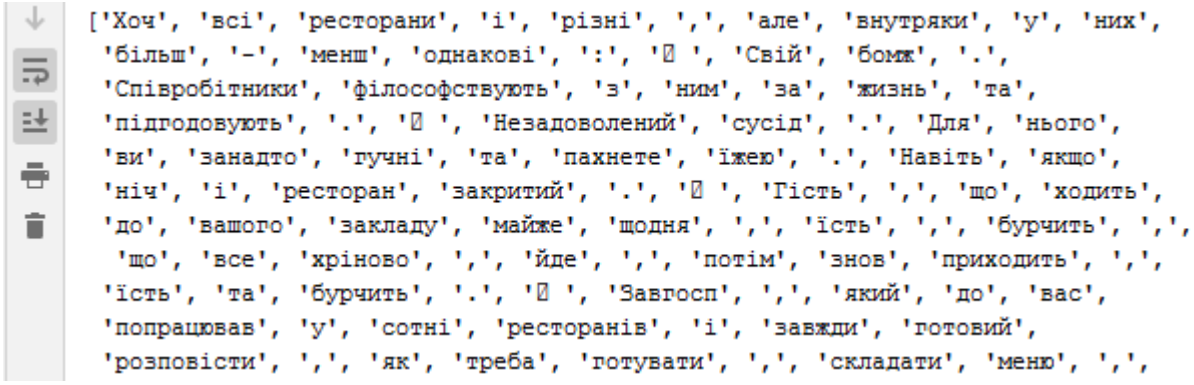
1) Для отримання списку токенів усіх текстів використали вбудований метод nltk для корпусів - .words(). Звели всі токени до нижнього регістру **(див. Малюнок 2.9).**

words = corpus.words()

print(list(words)[:100])

words = [word.lower() for word in words if word.replace("'", '').isalnum()]

fd = FreqDist(words)



```
['Хоч', 'всі', 'ресторани', 'і', 'різні', ',', 'але', 'внутряки', 'у', 'них',
 'більш', '-', 'менш', 'однакові', ':', '□ ', 'Свій', 'бомж', '.',
 'Співробітники', 'філософствують', 'з', 'ним', 'за', 'жизнь', 'та',
 'підгодовують', '.', '□ ', 'Незадоволений', 'сусід', '.', 'Для', 'нього',
 'ви', 'занадто', 'гучні', 'та', 'пахнете', 'їжею', '.', 'Навіть', 'якщо',
 'ніч', 'і', 'ресторан', 'закритий', '.', '□ ', 'Гість', ',', 'що', 'ходить',
 'до', 'вашого', 'закладу', 'майже', 'щодня', ',', 'їсть', ',', 'бурчить', ',',
  'що', 'все', 'хріново', ',', 'йде', ',', 'потім', 'знов', 'приходить', ',',
 'їсть', 'та', 'бурчить', '.', '□ ', 'Завгосп', ',', 'який', 'до', 'вас',
 'попрацював', 'у', 'сотні', 'ресторанів', 'і', 'завжди', 'готовий',
 'розповісти', ',', 'як', 'треба', 'готувати', ',', 'складати', 'меню', ',',
```

Малюнок 2.9. Список токенів усього тексту

2) Для того, аби згрупувати словоформи за частотою, використали вбудований функціонал nltk - клас FreqDist (Frequency Distribution). Наповнили першу таблицю словоформами і їх частотами**(див. Додаток 7).**

for word in fd.most_common():

db.execute('INSERT INTO wordform_freq (wordform, frequency) VALUES(?, ?);',

(word[0], word[1]))

database.commit()

3) Лематизували список словоформ корпусу **(див. Малюнок 2.10.)** за допомогою бібліотеки pymorphy2. *"Pymorphy2 - це морфологічний аналізатор української мови. Він використовує великі кодовані лексикони, побудовані на основі даних OpenCorpora та LanguageTool. Має набір лінгвістично вмотивованих правил, щоб забезпечити морфологічний аналіз та генерацію слів, що не містяться у словниках , але спостерігаються в інших документах . Аналізатор реалізований на мові програмування Python з додатковими*

розширеннями C ++. Акцент робиться на простоті використання, документації та розширюваності. Пакет поширюється за ліцензією з відкритим кодом, що заохочує його використання як в академічній, так і в комерційній сферах" [60, с.1].

```
lemmas = [morph.parse(word)[0].normal_form for word in words]

print(list(lemmas)[:100])

fd = FreqDist(lemmas)
```

```
['хоч', 'весь', 'ресторан', 'і', 'різний', 'але', 'внутряк', 'у', 'вони',
 'більш', 'менш', 'однаковий', 'свій', 'бомж', 'співробітник',
 'філософствувати', 'з', 'він', 'за', 'жизнь', 'та', 'підгодовувати',
 'незадоволений', 'сусід', 'для', 'він', 'ви', 'занадто', 'гучний', 'та',
 'пахнути', 'їжа', 'навіть', 'якщо', 'ніч', 'і', 'ресторан', 'закритий',
 'гість', 'що', 'ходити', 'до', 'ваш', 'закласти', 'майже', 'щодня', 'їсти',
 'бурчати', 'що', 'весь', 'хріново', 'йти', 'потім', 'знов', 'приходити',
 'їсти', 'та', 'бурчати', 'завгосп', 'який', 'до', 'ви', 'попрацювати', 'у',
 'сотня', 'ресторан', 'і', 'завжди', 'готовий', 'розповісти', 'як', 'треба',
 'готувати', 'складати', 'мень', 'рекламувати', 'оформляти', 'і',
 'позиціонувати', 'постачальник', 'що', 'постійно', 'заходити', 'через',
 'головний', 'вхід', 'навіть', 'якщо', 'ви', 'мільйон', 'раз', 'просити',
```

Малюнок 2.10. Лематизована лексика

4) Так само застосували на ньому клас FreqDist, погрупувавши леми за абсолютною частотою. Наповнили другу таблицю лемами і їх частотами.

```
for lemma in fd.most_common():
    db.execute('INSERT INTO lemma_freq (lemma, frequency) VALUES(?, ?);',
               (lemma[0], lemma[1]))
```

Другий крок:

Додавання до існуючих таблиць колонок pos, rel_freq + lemma для словника словоформ. Додали до старих таблиць потрібні колонки. У кожному рядку по черзі заповнили ці колонки потрібною інформацією. Для лематизації знову використали pymorphy2. Там, де бібліотека частину мови визначити не змогла (невідоме слово або знак пунктуації) в БД частина мови позначалася кодом UNKN.

Третій крок:

Повторюємо перший та другий крок окремо для кожної тематичної підвибірки за допомогою циклу.

Результат:

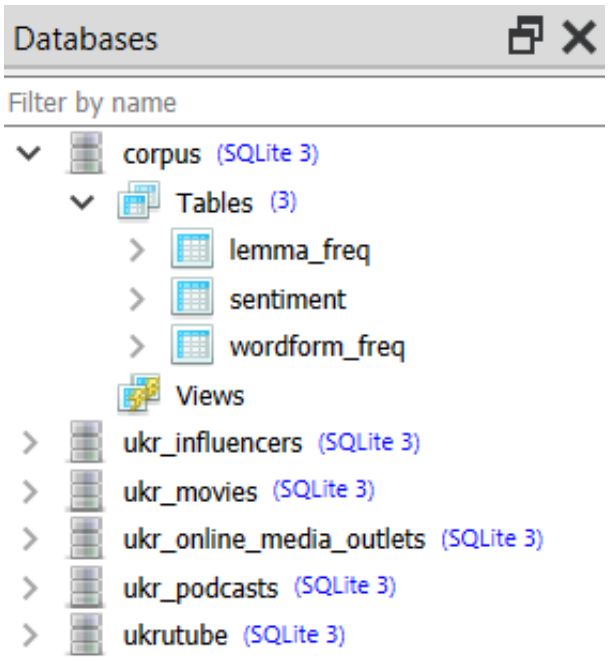
На вихід маємо такі файли **(див. Додаток 20) (див. Малюнки 2.11, 2.12.)**:

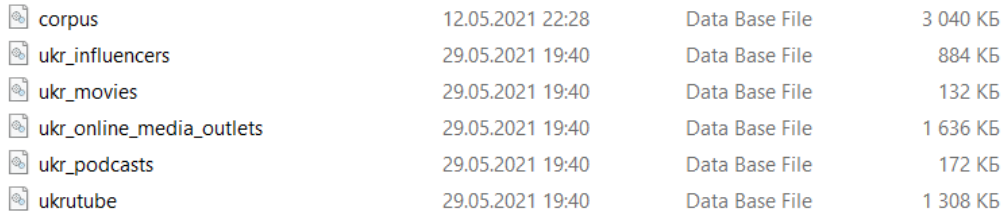
1) corpus.db - загальна БД усього корпусу;
2) ukr_influencers.db - БД текстів з офіційного сайту та Instagram-профілю Євгена Клопотенка;
3) ukr_movies.db - БД тексту стрічки “Мої думки тихі”;
4) ukr_online_media_outlets - БД текстів онлайн-видань “СЛУХ” та “The Village Україна”;
5) ukr_podcasts - БД текстового варіанту випусків подкастів “MINCULTPRYVIT” і “Простими словами”;
6) ukrutube - БД текстового варіанту відеоматеріалів “Телебачення Торонто” та “СПАЛАХ” на хостингу YouTube.

За допомогою формул *SELECT SUM(frequency) FROM lemma_freq;* та *SELECT SUM(frequency) FROM wordform_freq* у редакторі SQL дізнались кількісні характеристики кожної БД **(див. Додаток 2.10 та 2.11):**

1) corpus.db - 19981 лем/36696 словоформ;
2) ukr_influencers.db -6201 лем/10792 словоформ;

3) ukr_movies.db -1111 лем/1439 словоформ;
4) ukr_online_media_outlets 11632 лем/19893 словоформ;
5) ukr_podcasts -1382 лем/1933 словоформ;
6) ukrutube -9899 лем/15697 словоформ.





Малюнки 2.11. і 2.12. БД в інтерфейсі SQLite та файловому вигляді

2.3. Автоматичний сентимент-аналіз українськомовних медіатекстів

Дослідження в галузі лінгвістики, комунікації та психології висвітлили те, як ми проявляємо себе, розуміємо інших і на що впливає вираження суб'єктивності; як ми пов'язуємо емоції та думку з певними лінгвістичними аспектами, такими як конкретні слова чи синтаксичні патерни; і як ми можемо класифікувати мовні вирази відповідно до типу думки, яку вони передають.

Зосередимось на тому, що називається сентиментом, вираженням суб'єктивності як позитивної чи негативної думки. Багато методів дослідження цього поняття тісно пов’язані з такими галузями, як вивчення та класифікація емоцій, оскільки вони виражаються в мові.Теоретичний інтерес до вивчення суб’єктивності та її оцінки супроводжувався протягом останніх років підвищеною увагою до того, як ми висловлюємо думку в інтернеті. Сентимент-аналіз став важливим інструментом глибинного дослідження сугестивних засобів у всесвітній популярній культурі, зокрема в медійних текстах. Це відкрило поле аналізу настроїв у цифрових науках та комп'ютерній лінгвістиці, де суб’єктивність, думка та оцінка фіксуються для різних цілей. Цю сферу досліджень також називають "збиранням думок", через зацікавленість дослідників у видобуванні Big Data[70].

Сентимент-аналіз - це процес визначення того, чи є фрагмент тексту позитивно, негативно чи нейтрально забарвленим. Система сентимент-аналізу поєднує в собі обробку природної мови (NLP) та техніки машинного навчання, щоб зважено та чітко визначити тональність різних тем і категорій у реченні чи фразі[68].

Сентимент-аналіз може використовуватися для того, щоб проаналізувати відгуки споживачів, коли їх настрої суто негативні. Так само можна переглянути позитивні коментарі клієнтів, щоб з’ясувати, чому саме ці клієнти люблять певний сервіс/продукт. Тільки після успішного проведення цього дослідження бренд-менеджери та маркетологи можуть зосередитись на подальшому розвитку продукту.

Іншим застосуванням сентимент-аналізу є моніторинг настроїв публікацій у соціальних мережах. Під час оголошення Brexit інструмент соціальних мереж прогнозував, що варіанти "залишитися" в опитуванні були неправильними, за шість годин до того, як про це розповсюдилися новини. [74,79]

Типи сентимент-аналізу

Моделі сентимент-аналізу фокусуються на полярності (позитивні, негативні, нейтральні), а також на почуттях та емоціях (злі, щасливі, сумні тощо), терміновості (терміново, не на часі) і навіть намірах (зацікавлені та незацікавлені). Розглянемо найбільш розповсюджені типи аналізу [69],[77]:

1) Fine-Grained (Тонкодисперсний аналіз). Ця модель аналізу допомагає отримати точний показник полярності. Ви можете дослідити текст за такими категоріями полярності: дуже позитивно, позитивно, нейтрально, негативно або дуже негативно. Такий вид сентимент-аналізу корисний для вивчення відгуків та оцінок. Для шкали оцінок від 1 до 5 ви можете вважати показник 1 дуже негативним, а 5 - дуже позитивним. Для шкали від 1 до 10 можна вважати 1-2 дуже негативними, а 9-10 дуже позитивними оцінками[57].

2) Aspect-Based (Аналіз на основі аспектів). Цей вид аналізу визначає реальне ставлення людини, яке криється за її повідомленням. Наприклад, уявіть, що ви створюєте програмне забезпечення і отримали відгук клієнта, у якому сказано: "Софт довго обробляє дані на операційних системах з архітектурою X64". За допомогою аналізу на основі аспектів ви можете визначити, що рецензент залишив негативний коментар стосовно програмного забезпечення[51].

3) Emotion Detection (Розпізнавання емоцій). Як випливає з назви, цей тип аналізу розрізняє емоції. Сентимент-аналіз виявляє гнів, смуток, щастя, розчарування, страх, занепокоєння, паніку тощо. Системи виявлення емоцій зазвичай використовують лексикони - сукупність слів, що передають певні емоції. Деякі вдосконалені класифікатори також використовують продвинуті алгоритми машинного навчання. Наприклад: ***"His character development is the second season is about to kill me"***. Цей рядок може виражати почуття розчарування. Подібне речення - ***"The suspense in the last couple of episodes is***

straight up killing me!" - має зовсім інше, позитивне значення, оскільки в англійській мові метафоричний вираз ***"is killing me"*** має значення ***"я в повному захваті від чогось".*** Такі суперечності у виявленні емоцій якісно розрізняє саме цей вид аналізу[69].

4) Intent Analysis(Аналіз намірів). Правильне визначення намірів споживача може заощадити компанії час, гроші та зусилля. Дуже часто бренди агресивно рекламують продукцію тим потенційним споживачам, які не планують щось купувати найближчим часом. Точний аналіз намірів допомагає визначити справжні наміри людини і заощадити свій час[69].

5) На основі машинного навчання. Частіше за все для цього підходу використовують штучні нейронні мережі, а сам процес проходить у два етапи - навчання та прогнозування. На етапі навчання модель тренують співвідносити текст з конкретним тегом на основі контекстуальних ознак. На другому етапі модель дає оцінку тональності на основі даних, отриманих на тренуванні.

6) Підхід на основі правил та словника. Система працює на основі словника тональної лексики та правил, створених людиною. Для роботи цього методу треба зробити токенізацію, лематизацію, сегментацію на речення, визначити частини мови, провести синтаксичний аналіз речень.

7) Підхід на основі словника. Для цього методу використовується обмежений набір тональних слів та словник-тезаурус, наприклад, WordNet [81]. На першому етапі вручну збирається невелика кількість слів, потім до них добираються синоніми та антоніми у тезаурусі. Це відбувається за допомогою синсетів у WordNet. Процедуру можна повторювати доти, доки у словниках наявні синоніми чи антоніми.

Для визначення тональності текстів використовували тональний словник “tone-dict-uk.tsv”[15].

Перед застосуванням словника безпосередньо у програмі, зчитали його як текстовий файл, попарсили, створили з нього змінну типу "словник", і зберегли у вигляді байтів за допомогою бібліотеки pickle, щоб кожного наступного разу при здійсненні тонального аналізу не довелося парсити файл зі словником наново.

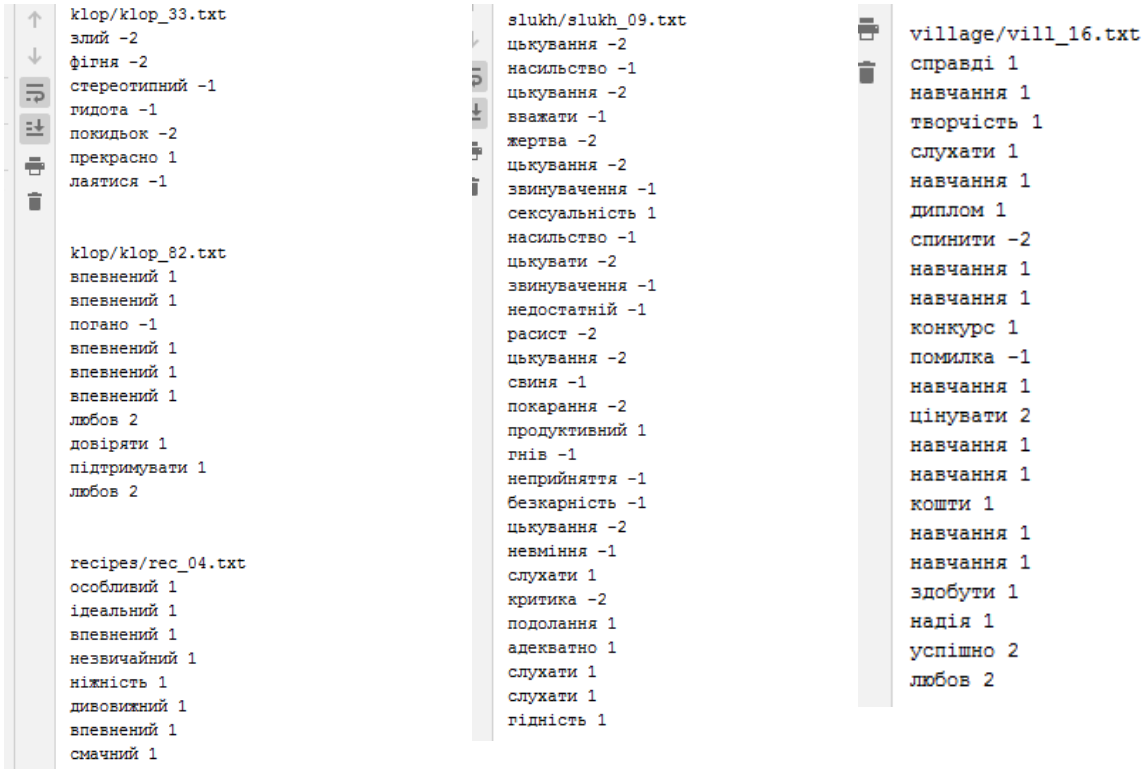
```
with open('tone-dict-uk.tsv', 'r', encoding='utf-8') as file:

lines = file.read().split('\n')

tones = dict()

for line in lines:

line = line.split('	')

tones[line[0]] = int(line[1])

with open('tone_dict.pickle', 'wb') as file:

pickle.dump(tones, file)
```

Pickle в Python в основному використовується для серіалізації та десеріалізації структури об'єктів Python. Іншими словами, це процес перетворення об’єкта Python у байтовий потік для зберігання у файлі / базі даних, підтримання стану програми в сеансах або транспортування даних через мережу. Байтовий потік може бути використаний для відтворення оригінальної ієрархії об’єктів. Весь цей процес схожий на серіалізацію об’єктів у Java або .Net[70].

У тональному словнику словам приписані додатні і від'ємні бали відповідно до їхнього сентименту. Тональність текстів корпусу визначалася таким чином, що сентимент тексту = сума сентиментів (балів у словнику) кожного слововживання тексту. Варто зазначити, що тексти які мали сумарний

сентимент в межах [2: -2] вважалися нейтральними **(див. Малюнки 2.13, 2.14, 2.15)**.

Результат:



Малюнки 2.13, 2.14, 2.15. Отримання текстів позитивної, негативної та нейтральної тональності.

2.4. Автоматичне укладання словника-конкорданса українськомовних медіатекстів

Для конкордансу знову використали вбудований функціонал nltk. Щоб кожного разу не читати текстові файли корпусу за допомогою PlaintextCorpusReader, бо це займає час, список зі всіма токенами корпусу

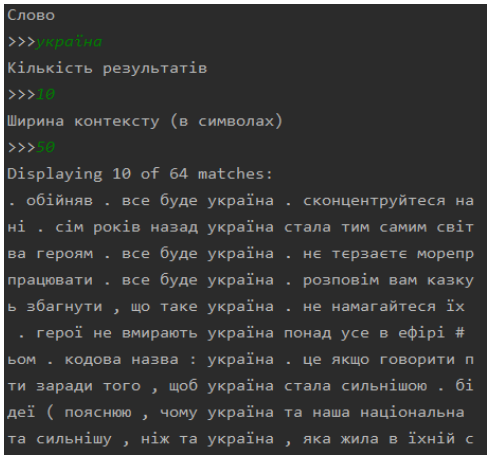
зберегли у файлі ccd_text.pickle . При запуску конкорданса текст корпусу зчитується саме з цього файла **(див. Додаток 1)**.

```
with
open(r'C:\Users\User\Desktop\ДИПЛОМ\КОРПУСПРАКТИЧНАЧАСТИНА\ukr_media\ccd_text.pickle', 'rb') as file:

words = pickle.load(file)
```

Зчитаний текст програма переконвертовує в клас Text (клас бібліотеки nltk). Бо клас Text має спеціальний метод concordance(), який і формує конкорданс**(див. Малюнок 2.16).**

```
while True:

    try:

        word = input('\nСлово\n>>>')

        if word == 'exit': break

        n = input('Кількість результатів\n>>>')

        if n == 'exit': break

        w = input('Ширина контексту (в символах)\n>>>')

        if w == 'exit': break

        text.concordance(word,

        lines=int(n),

        width=int(w)

        )
```

Результат:



Малюнок 2.16. Демонстрація роботи конкордансу

ВИСНОВКИ ДО РОЗДІЛУ 2

Цей розділ присвячений процесу створення корпусу українськомовних медіатекстів, укладання електронного частотного словника українських медіатекстів та бази даних тонального сентимент-аналізу українських медіатекстів. Був описаний вплив української мови на сучасний медіа дискурс в інтернеті, окреслені найвпливовіші сучасні медійні жанри/онлайн платформи. Детально проаналізовані такі елементи сучасної українськомовної поп-культури, як:

- проект “Телебачення Торонто”;
- документальний онлайн-серіал “СПАЛАХ”;
- онлайн видання “The Village Україна” та “СЛУХ”;
- українські подкасти “MINCULTPRYVIT” і “Простими словами”;
- медійна діяльність українського шеф-кухаря Євгена Клопотенка;
- стрічка “Мої думки тихі”.

Був створений корпус українськомовних текстів сучасної поп-культури, що налічує 217 текстів і сумарним обсягом у 197 877 слововживань. На основі корпусу був укладений частотний словник лем і словоформ, створені бази

даних загального корпусу і підвибірок, база даних тонального сентимент-аналізу українських медійних текстів та конкорданс медійних текстів.

РОЗДІЛ 3. СТАТИСТИЧНІ ПАРАМЕТРИ МЕДІАТЕКСТІВ ПОП-КУЛЬТУРИ

3.1. Статистичні параметри лексики

Індекс різноманітності Є однією з найпопулярніших статистичних характеристик для стилістичної оцінки тексту. Рахується за формулою V/N (N - довжина тексту, V - обсяг реєстру). Індекс різноманітності свідчить про лексичну різноманітність і "багатство" лексики. Важливим аспектом є залежність індексу різноманітності від обсягу тексту. Як зазначає Н.Дарчук, *"обсяг тексту збільшується швидше, ніж обсяг реєстру. Зі збільшенням тексту слова, що раніше зустрічалися, повторюються, а приріст нових слів зменшується, тому індекс падає. Це означає, що порівнювати за цим показником можна лише ті тексти, що мають відносно схожий обсяг.*

Індекс винятковості *- співвідношення кількості слів з частотою 1 до обсягу тексту (V1/N). Високий показник вказує на прагнення автора тексту вживати якомога менше повторів і намагання збагатити текст образною, колоритною лексикою.*

Індекс концентрації - *має однакову з індексом винятковості формулу, але протилежне значення. Для визначення цієї характеристики до уваги беруть лексику із частотою більше 10. Невелика кількість високочастотної лексики свідчить про розмаїтість тексту"* [41, с. 149-152].

Для розрахунків було використано формули індексів для лем і словоформ, які потім були введені у спеціальний редактор у програмному інтерфейсі SQLite:

1) Індекс різноманітності словоформ (див. Додаток 15)

```
WITH sum AS (SELECT SUM(frequency) AS sum FROM wordform_freq),
    wfs AS (SELECT COUNT(wordform) AS unique_wfs FROM wordform_freq)
```

```
SELECT wfs.unique_wfs*1.0 / sum.sum AS wordform_diversity FROM sum, wfs;
```

2) Індекс різноманітності лем (див. Додаток 16)

```
WITH sum AS (SELECT SUM(frequency) AS sum FROM lemma_freq),
    lms AS (SELECT COUNT(lemma) AS unique_lms FROM lemma_freq)
SELECT lms.unique_lms*1.0 / sum.sum AS lemma_diversity FROM sum, lms;
```

3) Індекс винятковості словоформ (див. Додаток 11)

```
WITH wfs1 AS (SELECT COUNT(wordform) AS wfs1 FROM wordform_freq WHERE frequency == 1),
    total AS (SELECT SUM(frequency) AS total FROM wordform_freq)
SELECT wfs1.wfs1*1.0 / total.total AS unique_wordforms FROM wfs1, total;
```

4) Індекс винятковості лем (див. Додаток 12)

```
WITH lms1 AS (SELECT COUNT(lemma) AS lms1 FROM lemma_freq WHERE frequency == 1),
    total AS (SELECT SUM(frequency) AS total FROM lemma_freq)
SELECT lms1.lms1*1.0 / total.total AS unique_lemmas FROM lms1, total;
```

5) Індекс концентрації словоформ (див. Додаток 13)

```
WITH wfs1 AS (SELECT COUNT(wordform) AS wfs1 FROM wordform_freq WHERE frequency > 10),
    total AS (SELECT SUM(frequency) AS total FROM wordform_freq)
SELECT wfs1.wfs1*1.0 / total.total AS concentration_wfs FROM wfs1, total;
```

6) Індекс концентрації лем (див. Додаток 14)

```
WITH lms1 AS (SELECT COUNT(lemma) AS lms1 FROM lemma_freq WHERE frequency > 10),
```

```
    total AS (SELECT SUM(frequency) AS total FROM lemma_freq)

SELECT lms1.lms1*1.0 / total.total AS concentration_lms FROM lms1, total;
```

У результаті було отримано такі показники:

Загальна вибірка (див. Малюнки 3.17 і 3.18):

- Індекс різноманітності словоформ 0.15
- Індекс різноманітності лем 0.08
- Індекс винятковості словоформ 0.08
- Індекс винятковості лем 0.04
- Індекс концентрації словоформ 0.0092
- Індекс концентрації лем 0.0096

Вибірка текстів Євгена Клопотенка:

- Індекс різноманітності словоформ 0.16
- Індекс різноманітності лем 0.09
- Індекс винятковості словоформ 0.09
- Індекс винятковості лем 0.04
- Індекс концентрації словоформ 0.011
- Індекс концентрації лем 0.012

Вибірка тексту стрічки “Мої думки тихі”:

- Індекс різноманітності словоформ 0.3
- Індекс різноманітності лем 0.2
- Індекс винятковості словоформ 0.2
- Індекс винятковості лем
- 0.14
- Індекс концентрації словоформ
- 0.009
- Індекс концентрації лем 0.010

Вибірка текстів українськомовних подкастів:

- Індекс різноманітності словоформ 0.28
- Індекс різноманітності лем 0.2
- Індекс винятковості словоформ 0.19
- Індекс винятковості лем 0.12
- Індекс концентрації словоформ 0.010
- Індекс концентрації лем 0.013

Вибірка текстів онлайн-видань:

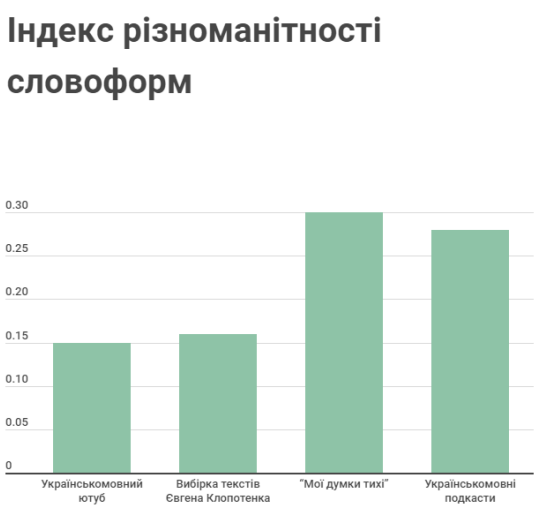
- Індекс різноманітності словоформ 0.2
- Індекс різноманітності лем 0.1
- Індекс винятковості словоформ 0.12
- Індекс винятковості лем 0.06
- Індекс концентрації словоформ 0.010
- Індекс концентрації лем 0.011

Вибірка текстів українськомовного ютубу

- Індекс різноманітності словоформ 0.22
- Індекс різноманітності лем 0.14
- Індекс винятковості словоформ 0.14
- Індекс винятковості лем 0.08
- Індекс концентрації словоформ 0.009
- Індекс концентрації лем 0.011

Оскільки для порівняння індексів різноманітності доречно обирати лише вибірки з відносно схожою кількістю слововживань, то було обрано чотири вибірки:

- Вибірку текстів Євгена Клопотенка - 53 869 слововживань;
- вибірку текстів українськомовного ютубу - 54840 слововживань;
- вибірку тексту стрічки “Мої думки тихі”- 3646 слововживання;
- вибірку текстів українськомовних подкастів - 5451 слововживань.



Малюнки 3.17 і 3.18 – діаграми індексу різноманітності

Після порівняння статистичних результатів, можна зробити такі висновки:

1. Лексика з подкастів є менш різноманітною, ніж лексика з фільму. Ймовірніше, це пов’язано з тим, що у подкастах присутній пласт психологічної термінології, яка повторюється у випусках. Зі свого боку персонажі фільму проживають неповторні події, що викликають у них яскраві емоції. Відповідно, це відображається на лексичному складі.

2. Однією з основних рис медійної діяльності Євгена Клопотенка є його грайлива схильність до вигадування дотепних форм звичайної української лексики. Особливо це стосується опису продуктів їжі. Відповідно, ця характеристика знайшла відображення в індексі різноманітності словоформ. Однак через обмеженість кулінарної тематики вибірка текстів українськомовного ютубу, яка охоплює декілька культурних сфер, має кращі показники у різноманітності лем.

Для розгляду показників індексів винятковості та концентрації було обрано вже усі вибірки.

Найліпшу міру винятковості як для лем, так і для словоформ, очікувано мають вибірки подкастів та фільму. Це зумовлено тим, що за своєю формою

подкаст є записом чи онлайн-трансляцією живої бесіди. Щодо вибірки фільму, то у сценарії стрічки прописані максимально життєві та природні діалоги. Невимушена, позбавлена штампів та канцеляризмів, жива бесіда виявилась більш винятковою, ніж написані за взірцем медійні тексти. Що стосується індексу концентрації, то різниця між показниками настільки несуттєва, що це не дає змоги зробити переконливі висновки.

3.2. Статистичні параметри частин мови

Для обрахування відносної частоти іменників, прикметників та дієслів**(див. Додатки 8,9,10)** у кожній вибірці була обрана така команда для SQLite: SELECT pos, SUM(rel_freq) FROM lemma_freq GROUP BY 1; Дані були згруповані в таблицю **(див. Таблицю 3.1**):

Частини мови	Загальна вибірка	Вибірка текстів Євгена Клопотенка	Вибірка тексту стрічки “Мої думки тихі”	Вибірка текстів онлайн-видань	Вибірка текстів українськомовних подкастів	Вибірка текстів українськомовного ютубу
іменник	0.27	0.25	0.17	0.24	0.19	0.23
прикметник	0.07	0.05	0.03	0.06	0.03	0.05
дієслово	0.10	0.09	0.09	0.08	0.08	0.08

Таблиця 3.1. Відносні частоти частин мови на вибірках

Велика кількість дієслів характерна для живого мовлення і для опису подій, що можна побачити у вибірці кінострічки. Також висока частотність дієслів присутня у вибірці текстів Євгена Клопотенка через те, що він регулярно описує процес приготування їжі. Частотність іменників та прикметників у вибірках онлайн-видань та українськомовного ютубу характеризується тим, що для публіцистичного стилю характерна образність та номіналізація дієслів.

3.3. Психолінгвістичні статистичні маркери тексту

Для виявлення психолінгвістичних маркерів у медійних текстах були використані показники, які описав В.П. Самохвалов **(див. Таблицю 3.2)**:

1) Коефіцієнт Трейгера, що свідчить про емоційну стабільність мовця. Показником є відношення обсягу дієслів до прикметників. Оптимальною ознакою є 1. Вищий результат може свідчити про емоційну напругу, нервовість, нестабільність, рекурентну компульсивну поведінку і тривоги. Також може вказувати на активну та енергійну особистість. Якщо у результаті отримується менше 1, то це може вказувати на тривожність, вагання, невпевненість[67].

2) Коефіцієнт визначеності дії **(див. Додаток 18)** характеризує ступінь соціалізації мовця. Для обчислення береться відношення кількості дієслів до іменників у вибірці [67].

3) Коефіцієнт агресивності **(див. Додаток 17)** сигналізує про агресивність мовлення. Рахується через співвідношення кількості дієслів разом з віддієслівними формами до усієї кількості слів. Взірцем є коефіцієнт 0,6[67].

Частини мови & формули	Загальна вибірка	Вибірка текстів Євгена Клопотенка	Вибірка тексту стрічки “Мої думки тихі”	Вибірка текстів онлайн-видань	Вибірка текстів українсько мовних подкастів	Вибірка текстів українсько мовного ютубу
к-сть лем	252866	67115	4714	101889	6930	72198
дієслово	26595	6397	470	8414	587	5968
Дієслівні форми	896	253	34	360	21	269
прикметник	17712	3200	153	5575	215	4014
іменник	67985	17949	871	26753	1421	17659
Коефіцієнт Трейгера	1.5	1.99	3.07	1.5	2.7	1.49
Коефіцієнт визначеності дії	0,4	0, 37	0,54	0,31	0, 41	0,33
Коефіцієнт агресивності	0,1	0,09	0,1	0,08	0,09	0,09

Таблиця 3.2. Статистичні показники та коефіцієнти

Аналіз текстових вибірок на наявність психолінгвістичних маркерів надав декілька цікавих інсайтів, як-от суттєво підвищений коефіцієнт Тейгера чи мінімальна ознака агресивності у кожній підвибірці. Незважаючи на те, що зазвичай коефіцієнт Трейгера **(див. Додаток 19)** характеризує здебільшого

негативні емоції, у нашому випадку це швидше є ознакою експресивності. Лише високі “3,07” та “2,7” у вибірці кінофільму та подкастів є наслідком сильних драматичних моментів стрічки та обговорення аспектів ментального здоров’я і життєвих переживань у подкасті. Підвищений показник визначеності дій у вищеназваних вибірках виникає через живий характер спілкування у фільмі та подкастах.

ВИСНОВКИ ДО РОЗДІЛУ 3

У цьому розділі був проведений додатковий лінгво-статистичний та психолінгвістичний аналіз статистичного матеріалу. Були визначені і обраховані такі статистичні параметри лексики, як: 1) індекс різноманітності; 2)індекс винятковості; 3) індекс концентрації; 4) коефіцієнт Трейгера; 5) коефіцієнт визначеності дії; 6) коефіцієнт агресивності. Була побудована таблиця відносних частот частин мови та діаграми для репрезентативності підсумків дослідження.

Основні цікаві особливості лексики, виявлені у ході статистичного аналізу: 1) підвищений коефіцієнт Трейгера в усіх вибірках; 2) мінімальний коефіцієнт агресивності в усіх вибірках; 3) різний рівень винятковості для словоформ і лем; 4) результати текстових вибірок подкастів та стрічки “Думки мої тихі” схожі між собою і відрізняються від інших для багатьох статистичних маркерів.

ВИСНОВКИ

Метою роботи було проведення статистичної параметризації українськомовних медіатекстів на базі автоматично укладеного комп'ютерного частотного словника.

Весь процес роботи можна поділити на такі етапи:

I. Опанування теоретичного матеріалу таких галузей лінгвістики, як корпусна лінгвістика і статистична лексикографія. Знайомство з концепцією автоматичного сентимент-аналізу і пошук інформації про український тональний словник. Глибинний огляд медійних сфер, у яких представлена сучасна українськомовна популярна культура.
II. Формування текстових вибірок і створення корпусу медійних текстів. Автоматична обробка текстових файлів для усунення критичних помилок в інтерфейсі Python, створення єдиного кодування. Фінальне виправлення всіх програмних помилок.
III. Створення частотного словника словоформ і лем для загальної бази даних і для кожної тематичної вибірки. Створення конкордансу медійних текстів і бази даних тонального сентимент-аналізу медійних текстів. Виведення статистики для загальної бази даних і тематичних словників;
IV. Подальший лінгвістичний аналіз статистичного матеріалу за допомогою індексів різноманітності, винятковості і концентрації, обрахування відносної частоти частин мови на вибірках. Застосування психолінгвістичних маркерів тексту - коефіцієнту Трейгера, коефіцієнту визначеності дії, коефіцієнту агресивності.

Після детального аналізу лексики українськомовних текстів сучасної української поп-культури стає очевидним той факт, що, хоча медійні тексти і належать до різних стилів, значної різниці у статистичних показниках між

вибірками не простежується. Основною ознакою усіх медійних текстів є елементи живого мовлення, відхід від штампованої лексики, частка якої була істотною у публіцистичному стилі у недалекому минулому. Це є ознакою суспільних змін, які відбуваються протягом 7-8 останніх років у суспільстві - українці прагнуть до свободи самовираження, усвідомити свою ідентичність і, нарешті, усвідомити, що означає “бути українцем”. Прямим наслідком зміни парадигми є банальне покращення рівня володіння українською мовою. Це призводить до природної відмови від стандартизованого медійного стилю і використання більш невимушеної лексики.

Логічним наступним кроком буде викладення баз даних медійних текстів у вільний доступ. Мовознавці і просто користувачі можуть порівнювати проаналізовані медійні тексти з інформацією про тексти іншого стилю. Подальші дослідження мають бути зосереджені на більш глибокому статистичному та психолінгвістичному аналізі лексики. Використані 3 розділі роботи індекси надають багато інформації про особливості стилю, проте не характеризують його у повній мірі. Радше вони слугують базовим аналізом лексики.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Алексієнко Л.А. Методика створення автоматизованої системи морфемно-словотвірного аналізу (АСМСА) слів української мови / Л.А. Алексієнко, Н.П. Дарчук, О.М. Зубань // Наукова спадщина професора С.В.Семчинського. Збірник наукових праць. – К., 2001, - Ч. 1. – С. 38 – 49.; 2.

2. Баранов А.Н. Введение в прикладную лингвистику / А.Н. Баранов. – Москва, 2001. – 358 с

3. Бобкова Т. В. До визначення корпусної лінгвістики в сучасному мовознавстві / Т. В. Бобкова // Наукові записки Національного університету "Острозька академія". Серія : Філологічна. - 2014. - Вип. 45. - С. 3-6. [Електронний ресурс] - Режим доступу: http://nbuv.gov.ua/UJRN/Nznuoaf_2014_45_3

4. Браунський корпус української мови. [Електронний ресурс]/Режим доступу - https://r2u.org.ua/corpus

5. Генеральний регіонально анотований корпус української мови (ГРАК) / М. Шведова, Р. фон Вальденфельс, С. Яригін, М. Крук, А. Рисін, В. Старко, М. Возняк. — Київ, Осло, Єна, 2017-2019.

6. Головин Б.Н. Вопросы статистической стилистики Київ : Наукова думка, 1974. – 331 с.

7. Дарчук 2008: Дарчук Н. П. Комп’ютерна лінгвістика (автоматичне опрацювання тексту): підручник. Київ: Видавничо-поліграфічний центр “Київський університет”, 2008. 351 с.

8. Дарчук 2010: Дарчук Н. Корпус украинского языка. Prace Filologiczne. Warszawa, 2010. t. LXIII, S.99–108.

9. Дарчук 2010а: Дарчук Н. П. Дослідницький корпус української мови: основні засади і перспективи. Вісник Київського національного університету імені Тараса Шевченка. Серія: Літературознавство. Мовознавство. Фольклористика. Київ: Видавничо-поліграфічний центр “Київський університет”, 2010. № 21. С. 45–49.

10. Дарчук 2016: Дарчук Н., Зубань О., Лангенбах М., Ходаківська Я. АГАТ-семантика: семантична розмітка Корпусу української мови. Українське мовознавство. Київ: Видавничо-поліграфічний центр “Київський університет”, 2016. Вип. 1 (46). С. 3–10.

11. Дарчук Н.П. Комп'ютерне анотування тексту: результати і перспективи: монографія / Н.П. Дарчук. – К., 2013. – 543 с.

12. Демська-Кульчицька О. Корпусна рецепція тексту / О. Демська-Кульчицька // Наукові записки. Т. 111. Сер. Філологічні науки. – 2010. – С.-3-6.

13. Демська-Кульчицька О. Один з аспектів морфологічної

анотації (до проблеми побудови тега) / О. ДемськаКульчицька // Українська мова. – 2004. – № 1. – С. 26-38.

14. Демська-Кульчицька О.М. Репрезентативність як ознака

текстового корпусу / О.М. Демська-Кульчицька. – Українська мова. – №3, 2005. – С. 100-107.

15. Дьомкин В., Чаплинський Д., Шеховцов С. Український тональний словник. [Електронне джерело]/Режим доступу: https://github.com/lang-uk/tone-dict-uk

16. Захаров В.П. Корпусная лингвистика: Учебно–метод.

пособие. / В.П. Захаров – СПб., 2005. – 48 с.

17. Захаров В.П., Богданова С.Ю. Корпусная лингвистика:

учебник для студентов гуманитарных вузов / В.П. Захаров,

С.Ю. Богданова. – Иркутск: ИГЛУ, 2011. – 161 с.

18. Зубань О. М. Особливості морфемної будови слів у поетичних текстах Т. Шевченка (на матеріалі Корпусу української мови) / О.М. Зубань // Українське мовознавство. - № 44/1. - К., 2014. - С. 123 - 133.

19. Зубань О.М. Параметризована база даних як інструмент дослідження корпусу текстів / О.М. Зубань // Лексикографічний бюлетень: Збірник наукових праць. – Вип.13. – К., 2006. – С. 37 – 43.

20. Зубань О.М. Стилеметричні ознаки морфемних структур слів у поетичному мовленні Т. Шевченка (на матеріалі Корпусу української мови) / О.М. Зубань // Мовні і концептуальні картини світу. - Вип. 48. - К., 2014. - С. 165 - 179.

21. Карпіловська 2019: Карпіловська Є.А. Здобутки академічної структурної та математичної лінгвістики у моделюванні українського слова. *Українська мова: науково-теоретичний журнал*. Київ, 2019. № 1(69). С. 18–36.

22. Карпіловська Є. А. Вступ до прикладної лінгвістики: Комп’ютерна лінгвістика. – Донецьк: ТОВ Юго-Восток, 2006. – 188 с.

23. Карпіловська Є. А. Суфіксальна підсистема сучасної української літературної мови: будова та реалізація / Інститут мовознавства ім. О.О.Потебні НАН України. - К., 1999. - 297с.

24. Карпіловська Є.А., Кислюк Л.П., Клименко Н.Ф., Критська В.І., Пуздирєва Т.К. та ін. Активні ресурси сучасної української номінації – Київ, 2013. – 414 с.

25. Клименко Н.Ф. Основи морфеміки сучасної української мови.; Ін-т змісту і методів навчання, Київський університет ім. Тараса Шевченка, 1998 – 183с.

26. Клименко Н.Ф. Система афіксального словотворення сучасної української мови». — К.: Наукова думка, 1973. — 187 с.

27. Копотев М., Мустайоки А. Современная корпусна русистика / М. Копотев, А. Мустайоки // Инструментарий русистики: корпусные подходы. – Хельсинки, 2008. – С. 7-24.

28. КТУМ: Корпуси текстів української мови. [Електронний ресурс]/Режим доступу - http://corpora.donnu.edu.ua/

29. Кулінарні рецепти від Євгена Клопотенка.[Електронний ресурс]/Режим доступу – https://soundcloud.com/the_village_ukraine/sets/mincultpryvit

30. КУМ:Корпус української мови. [Електронний ресурс] /Режим доступу: http://www.mova.info/corpus.aspx

31. Лінгвістичний портал mova.info [Електронний ресурс] /Режим доступу: http://www.mova.info/

32. Мої думки тихі.[Електронний ресурс]/Режим доступу - https://megogo.net/ua/view/8224985-mo-dumki-tihi.html

33. Морфемна структура слова : монографія / Т. О. Грязнухіна,

34. Н. Ф. Клименко, Л. І. Комарова, М. П. Муравицька, М. М. Пещак ; відп. ред. М. М. Пещак. – К. : Наукова думка, 1979. – 336 с.

35. Открытый корпус. [Електронне джерело]/Режим доступу: http://opencorpora.org/

36. Перебийніс В. І. Статистичні методи для лінгвістів : посібник / Перебийніс В. І.— Вид. 2, випр. і допов. — Вінниця : Нова Книга, 2013 — 176 с.

37. Перебийніс В. І. Статистичні параметри стилів / Валентина Перебийніс. – К.:Наукова думка, 1967. – 240 с.

38. Перебийніс В.І. Лексика, граматика, фонетика //Мова, людина, світ. До 70-річчя професора М.Кочергана. Збірник наукових статей. – К.: Вид. центр КНЛУ, 2006. – 352 с.

39. Перебийніс В.І. Математична лінгвістика: Навчальний посібник. — К. : Вид. центр КНЛУ, 2014. — 125 с.

40. Перебийніс В.І., Сорокін В.М. Традиційна та комп'ютерна лексикографія — Навч. посібник. —К.: Вид. центр КНЛУ, 2009. — 218 с.

41. Перебийніс В.С., Муравицька М.П., Дарчук Н.П. Частотні словники та їх використання / В.С. Перебийніс, М.П. Муравицька, Н.П. Дарчук – К.: Наукова думка, 1985. – 204 с.

42. Плунгян В.А. Корпус как инструмент и как идеология: о

некоторых уроках современной корпусной лингвистики /

В.А. Плунгян // Русский язык в научном освещении. – №2

(16), 2008. – С.7-20

43. Простими словами.[Електронний ресурс]/Режим доступу - https://podcasts.apple.com/ua/podcast/%D0%BF%D1%80%D0%BE%D1%81%D

1%82%D0%B8%D0%BC%D0%B8-%D1%81%D0%BB%D0%BE%D0%B2%D0%B0%D0%BC%D0%B8/id1487856534?l=uk

44. Рычкова Л.В. Праблема састауных аб'ектау у корпусах славянскімоу і лінгвістычных базах дадзеных / Л.В.Рычкова // Мовознауства. Літаратура. Культуралогія. Фальклярыстыка. XIII Міжнародныї з'езд славыстау.1Даклады беларускай дзлегацыі. – Мінськ, 2003. – С. 184-195.

45. Словник афіксальних морфем української мови / Н. Ф. Клименко, Є. А. Карпіловська, В. С. Карпіловський, Т. І. Недозим; Ін-т мовознавства ім. О. О. Потебні НАН України. – Київ: [б.в.], 1998. – 440 с.

46. СЛУХ.[Електронний ресурс]/Режим доступу - https://www.youtube.com/channel/UCf8Du8VB0xG5z8N68zbAayQ

47. СЛУХ-онлайн-медіа.[Електронний ресурс]/Режим доступу - https://slukh.media/

48. Телебачення Торонто.[Електронний ресурс]/Режим доступу - https://www.youtube.com/channel/UCF_ZiWz2Vcq1o5u5i1TT3Kw

49. Частотний словник сучасної української художньої прози : в двох томах / [редакційна колегія, В.С. Перебийніс (голова) ... та ін.]/Київ: Наукова думка, 1981. -1716 с.

50. Широков А. І. // Вісник Національної академії наук України. - 2011. - № 7. - С. 47-49. - [Електронний ресурс]/Режим доступу: http://nbuv.gov.ua/UJRN/vnanu_2011_7_13

51. A Comprehensive Guide to Aspect-based Sentiment Analysis. [Електронне джерело]/Режим доступу: https://monkeylearn.com/blog/aspect-based-sentiment-analysis/

52. About SQLite. [Електронне джерело]/Режим доступу: https://www.sqlite.org/about.html

53. Agrawal. Understanding Python pickling and how to use it securely. [Електронне джерело]/Режим доступу: https://www.synopsys.com/blogs/software-security/python-pickling/

54. Asmussen J. Korpuslinguistische Verfahren zur Optimierung

lexikalisch-semantischer Beschreibungen / J. Asmussen //

Sprachkorpora – Datenmengen und Erkenntnisfortschritt (Hrsg.

von W. Kallmeyer, G. Zifonun). Institut fur Deutscher sprache.

Jahrbuch 2006. Berlin – N.J.: Walter de Gruyter, 2007 – S. 123-

151.

55. Baker P., Hardie A., McEnery T. Glossary of Corpus

Linguistics / P. Baker, A. Hardie, T. McEnery. – Edinburgh

University Press, 2006 – 192 p.

56. Differences between SQL and SQLite. [Електронне джерело]/Режим доступу: https://www.geeksforgeeks.org/differences-between-sql-and-sqlite/

57. Fine-grained Sentiment Analysis in Python (Part 1). [Електронне джерело]/Режим доступу: https://towardsdatascience.com/fine-grained-sentiment-analysis-in-python-part-1-2697bb111ed4

58. Geoffrey Leech, Roger Garside. Recent developments in the use of computer corpora in English language research - June 1983.Transactions of the Philological Society 81(1):23 – 40. [Електронний ресурс]/Режим доступу:

https://www.researchgate.net/publication/229748696_Recent_developments_in_the_use_of_computer_corpora_in_English_language_research

59. Ievgen Klopotenko (@klopotenko).[Електронний ресурс]/Режим доступу – https://www.instagram.com/klopotenko/?hl=uk

60. M. Korobov. Morphological Analyzer and Generator for Russian and Ukrainian Languages. [Електронне джерело]/Режим доступу: https://arxiv.org/pdf/1503.07283.pdf

61. MacEnery T., Hardie A. Corpus Linguistics: Method, Theory

and Practice / T. MacEnery, A. Hardie. – Cambridge University

Press, 2012. – 294 p.

62. McEnery T., Xiao R., Tono Y. Corpus-based Language Studies:

an Advanced Resource Book / T. McEnery, R. Xiao, Y. Tono. –

London: Routledge, 2006. – 386 p.

63. MINCULTPRYVIT.[Електронний ресурс]/Режим доступу – https://soundcloud.com/the_village_ukraine/sets/mincultpryvit

64. Natural Language Toolkit — NLTK 3.6.2 documentation. [Електронний ресурс]/Режим доступу: https://www.nltk.org/

65. pymorphy2 — Python 3.9.5 documentation. [Електронний ресурс]/Режим доступу:

https://pymorphy2.readthedocs.io/en/stable/

66. Python text processing with NLTK 2.0: creating custom corpora. [Електронне джерело]/Режим доступу:

https://hub.packtpub.com/python-text-processing-nltk-20-creating-custom-corpora/

67. Samohvalov, V. P. (2002). Psihiatrija [Psychiatry]. Rostov-na-Donu: Feniks

68. Sentiment Analysis Explained. [Електронне джерело]/Режим доступу: https://www.lexalytics.com/technology/sentiment-analysis

69. Sentiment Analysis: A Definitive Guide. [Електронне джерело]/Режим доступу: https://monkeylearn.com/sentiment-analysis/

70. Sentiment Analysis: An Overview from Linguistics. [Електронне джерело]/Режим доступу: https://www.annualreviews.org/doi/pdf/10.1146/annurev-linguistics-011415-040518

71. sqlite3 — DB-API 2.0 interface for SQLite databases — Python 3.9.5

documentation. .[Електронний ресурс]/Режим доступу: https://docs.python.org/3/library/sqlite3.html

72. Teubert W. Corpus linguistics and lexicography / W. Teubert //

Text Corpora and Multilingual Lexicography ed. by W.

Teubert – John Benjamins Publishing Company – Amsterdam/

Philadelphia, 2007 – P. 109-134.

73. Text Analytics for beginners using NLTK. [Електронне джерело]/Режим доступу:https://www.datacamp.com/community/tutorials/text-analytics-beginners-nltk

74. The Twitris sentiment analysis tool by Cognovi Labs predicted the Brexit hours earlier than polls. [Електронне джерело]/Режим доступу:

https://techcrunch.com/2016/06/29/the-twitris-sentiment-analysis-tool-by-cognovi-labs-predicted-the-brexit-hours-earlier-than-polls/?guccounter=1

75. The Village UA. [Електронне джерело]/Режим доступу: https://www.the-village.com.ua/

76. Tognini-Bonelli E. Theoretical overview of the evolution of

corpus linguistics / E. Tognini-Bonelli // The Routledge

Handbook of Corpus Linguistics / Edited by Anne O'Keeffe and

Michael McCarthy. – Routledge, 2010. – 681 p.

77. TYPES OF SENTIMENT ANALYSIS AND HOW BRANDS PERFORM THEM. [Електронне джерело]/Режим доступу: https://www.analyticsinsight.net/types-of-sentiment-analysis-and-how-brands-perform-them/

78. What is the difference between the Unicode, UTF8, UTF7, UTF16, UTF32, ASCII, and ANSI encodings? [Електронне джерело]/Режим доступу: https://stackoverflow.com/questions/700187/unicode-utf-ascii-ansi-format-differences

79. What We Can Learn From the Postmortem on #Brexit Social Media. [Електронне джерело]/Режим доступу: https://www.dmnews.com/marketing-channels/social/article/13035644/what-we-can-learn-from-the-postmortem-on-brexit-social-media

80. Why you should use SQLite [Електронне джерело]/Режим доступу: https://www.infoworld.com/article/3331923/why-you-should-use-sqlite.html

81. Word Net. [Електронне джерело]/Режим доступу: https://wordnet.princeton.edu/

ДОДАТКИ

Додаток 1

Програма для пошуку по конкордансу

```
from nltk.text import Text
import pickle


with
open(r'C:\Users\User\Desktop\ДИПЛОМ\КОРПУСПРАКТИЧНАЧАСТИНА\ukr_media\ccd_text.pickle',
'rb') as file:
    words = pickle.load(file)

text = Text(words)

print('\nКонкорданс корпусу українських медійних текстів\nщоб вийти, введіть exit')

while True:
    try:
        word = input('\nСлово\n>>>')
        if word == 'exit': break

        n = input('Кількість результатів\n>>>')
        if n == 'exit': break

        w = input('Ширина контексту (в символах)\n>>>')
        if w == 'exit': break

        text.concordance(word,
                         lines=int(n),
                         width=int(w)
                         )

    except:
        print('Помилка')
```

Додаток 2

Програма, що вираховує відносну частоту для ЧС

```
import sqlite3, pymorphy2


morph = pymorphy2.MorphAnalyzer(lang='uk')

dbs = ['corpus.db',
       'ukr_influencers.db',
       'ukr_movies.db',
       'ukr_online_media_outlets.db',
       'ukr_podcasts.db',
       'ukrutube.db']

for dbn in dbs:

    database = sqlite3.connect('dicts/'+dbn)
    db = database.cursor()

    db.execute('ALTER TABLE lemma_freq ADD COLUMN pos CHAR(4);')
    db.execute('ALTER TABLE lemma_freq ADD COLUMN rel_freq DECIMAL(6,5);')
    db.execute('ALTER TABLE wordform_freq ADD COLUMN lemma VARCHAR(255);')
    db.execute('ALTER TABLE wordform_freq ADD COLUMN pos CHAR(4);')
    db.execute('ALTER TABLE wordform_freq ADD COLUMN rel_freq DECIMAL(6,5);')


    rows = db.execute('SELECT id, lemma, frequency FROM lemma_freq;').fetchall()
    word_sum = db.execute('SELECT SUM(frequency) FROM lemma_freq;').fetchall()[0][0]

    for r in rows:
        db.execute('UPDATE lemma_freq SET pos = ?, rel_freq = ? WHERE id = ?;',
                   (morph.parse(r[1])[0].tag.POS,
                    round(r[2]/word_sum, 10),
                    r[0]))

    db.execute('UPDATE lemma_freq SET pos = "UNKN" WHERE pos IS NULL;')
    database.commit()


    rows = db.execute('SELECT id, wordform, frequency FROM wordform_freq;').fetchall()
    word_sum = db.execute('SELECT SUM(frequency) FROM wordform_freq;').fetchall()[0][0]

    for r in rows:
        db.execute('UPDATE wordform_freq SET lemma = ?, pos = ?, rel_freq = ? WHERE id
= ?;',
                   (morph.parse(r[1])[0].normal_form,
                    morph.parse(r[1])[0].tag.POS,
                    round(r[2]/word_sum, 10),
                    r[0]))

    db.execute('UPDATE wordform_freq SET pos = "UNKN" WHERE pos IS NULL;')
    database.commit()
    database.close()
```

Додаток 3

Програма, що змінює кодування з ansi на utf-8

```
import os


path = r'..\ukr_media_corpus'
folders = os.listdir(path)

for folder in folders:
    dir = os.listdir(path+r'\\'+folder)

    for n in range(len(dir)):
        with open(path+r'\\'+dir[n], 'r', encoding='ansi') as file:
            text = file.read()
        with open(path+r'\\'+dir[n], 'w', encoding='utf-8') as file:
            file.write(text)
```

Додаток 4

Програма, що змінює символи

```
import os


path = r'..\ukr_media_corpus'
folders = os.listdir(path)

for folder in folders:

    local_path = path+r'\\'+folder
    dir = os.listdir(local_path)

    for n in range(len(dir)):

        with open(local_path+r'\\'+dir[n], 'r', encoding='utf-8') as file:
            text = file.read()

        text = text.replace('й', 'й')
        text = text.replace('ї', 'ї')
        text = text.replace('Й', 'Й')
        text = text.replace('Ї', 'Ї')

        with open(local_path+r'\\'+dir[n], 'w', encoding='utf-8') as file:
            file.write(text)
```

Додаток 5

Програма, що змінює назви файлів

```
import os


path = r'..\ukr_media_corpus\toronto'
dir = os.listdir(path)

for n in range(len(dir)):
    if n < 10:
        os.rename(os.path.join(path, dir[n]),
                  os.path.join(path, 'tor_0' + str(n)+ '.txt'))
    else:
        os.rename(os.path.join(path, dir[n]),
                  os.path.join(path, 'tor_' + str(n)+ '.txt'))
```

Додаток 6

Програма для сентимент-аналізу

```
from nltk.corpus import PlaintextCorpusReader
import pymorphy2, sqlite3, pickle, os


morph = pymorphy2.MorphAnalyzer(lang='uk')

corpus_root = os.path.dirname(os.path.realpath(__file__)) + r'\ukr_media_corpus'
corpus = PlaintextCorpusReader(corpus_root, '.*')

database = sqlite3.connect('dicts/corpus.db')
db = database.cursor()

db.execute('''CREATE TABLE sentiment
             (id INTEGER PRIMARY KEY,
              fileid VARCHAR(50),
              score INTEGER,
              sentiment CHAR(3));''')

# # Парсинг файла тонального словника, зберігання його в pickle
# with open('tone-dict-uk.tsv', 'r', encoding='utf-8') as file:
#     lines = file.read().split('\n')
#
# tones = dict()
# for line in lines:
#     line = line.split('  ')
#     tones[line[0]] = int(line[1])
#
# with open('tone_dict.pickle', 'wb') as file:
#     pickle.dump(tones, file)


with open('tone_dict.pickle', 'rb') as file:
    tones = pickle.load(file)

# files_print = ['klop/klop_33.txt', 'klop/klop_82.txt', 'recipes/rec_04.txt',
#                'recipes/rec_23.txt', 'slukh/slukh_09.txt', 'village/vill_16.txt']
files = corpus.fileids()

for file in files:

    # if file not in files_print:

    score = 0
    for word in corpus.words(file):
        try:
            score += tones[word]
        except KeyError:
            continue

    if score > 2: sentiment = 'pos'
    elif score < -2: sentiment = 'neg'
    else: sentiment = 'neu'
    db.execute('INSERT INTO sentiment (fileid, score, sentiment) VALUES(?, ?, ?);',
               (file, score, sentiment))
```

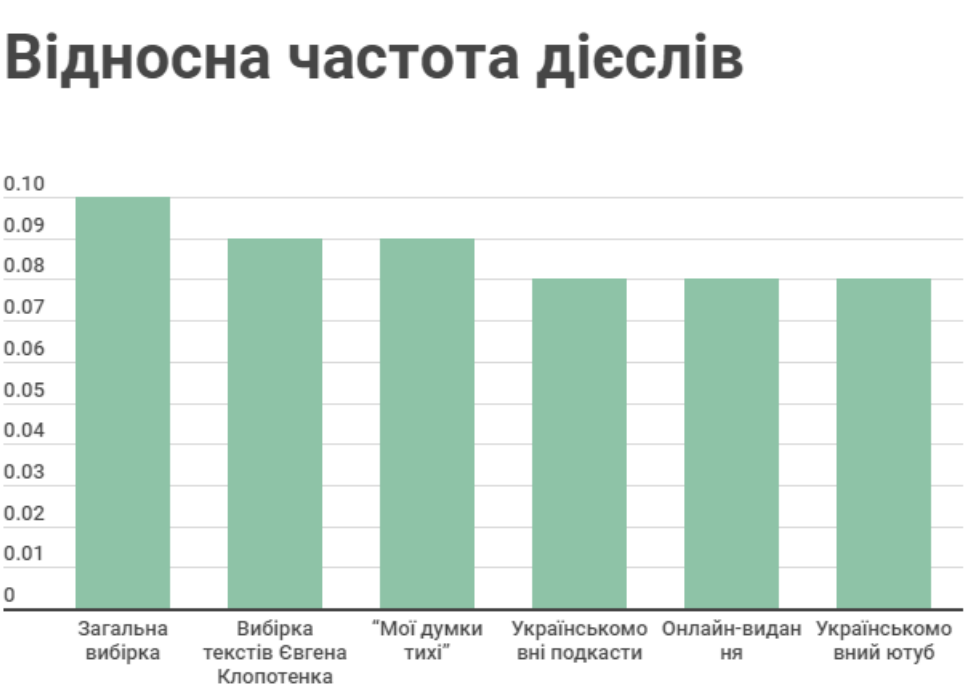
Додаток 7

Програма для частотного словника

```
from nltk.corpus import PlaintextCorpusReader
from nltk import FreqDist
import pymorphy2, sqlite3, os


morph = pymorphy2.MorphAnalyzer(lang='uk')

corpus_root = os.path.dirname(os.path.realpath(__file__)) + r'\ukr_media_corpus'
corpus = PlaintextCorpusReader(corpus_root, '.*')


database = sqlite3.connect('dicts/corpus.db')
db = database.cursor()

db.execute('''CREATE TABLE wordform_freq
             (id INTEGER PRIMARY KEY,
              wordform VARCHAR(255),
              frequency INTEGER);''')
db.execute('''CREATE TABLE lemma_freq
             (id INTEGER PRIMARY KEY,
             lemma VARCHAR(255),
             frequency INTEGER);''')
database.commit()


words = corpus.words()
print(list(words)[:100]) # заскрінити показати як виглядає токенізований текст
words = [word.lower() for word in words if word.replace("'", '').isalnum()]
fd = FreqDist(words)

for word in fd.most_common():
    db.execute('INSERT INTO wordform_freq (wordform, frequency) VALUES(?, ?);',
               (word[0], word[1]))
database.commit()


lemmas = [morph.parse(word)[0].normal_form for word in words]
print(list(lemmas)[:100]) # заскрінити показати як виглядає лематизований текст
fd = FreqDist(lemmas)

for lemma in fd.most_common():
    db.execute('INSERT INTO lemma_freq (lemma, frequency) VALUES(?, ?);',
               (lemma[0], lemma[1]))
database.commit()
database.close()
```

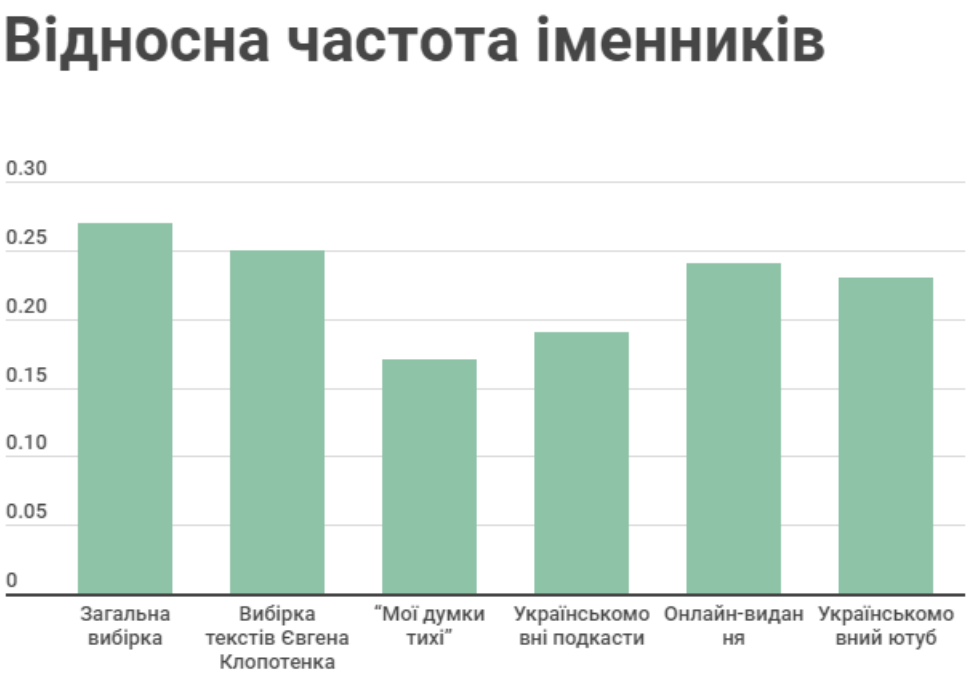
Додаток 8

Діаграма відносної частоти дієслів



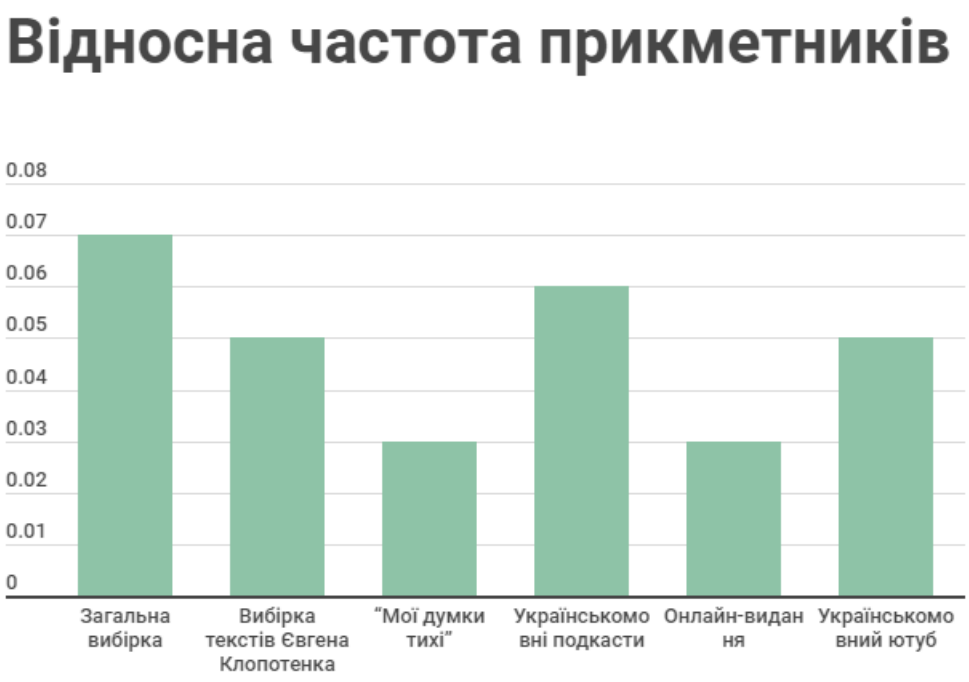
Додаток 9

Діаграма відносної частоти іменників



Додаток 10

Діаграма відносної частоти прикметників



Додаток 11

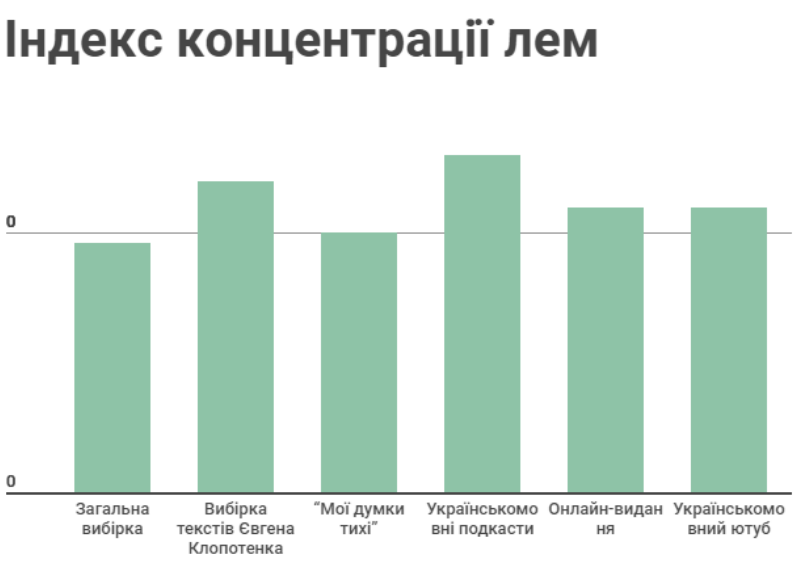
Діаграма індексу винятковості лем

Додаток 12

Діаграма індексу винятковості словоформ

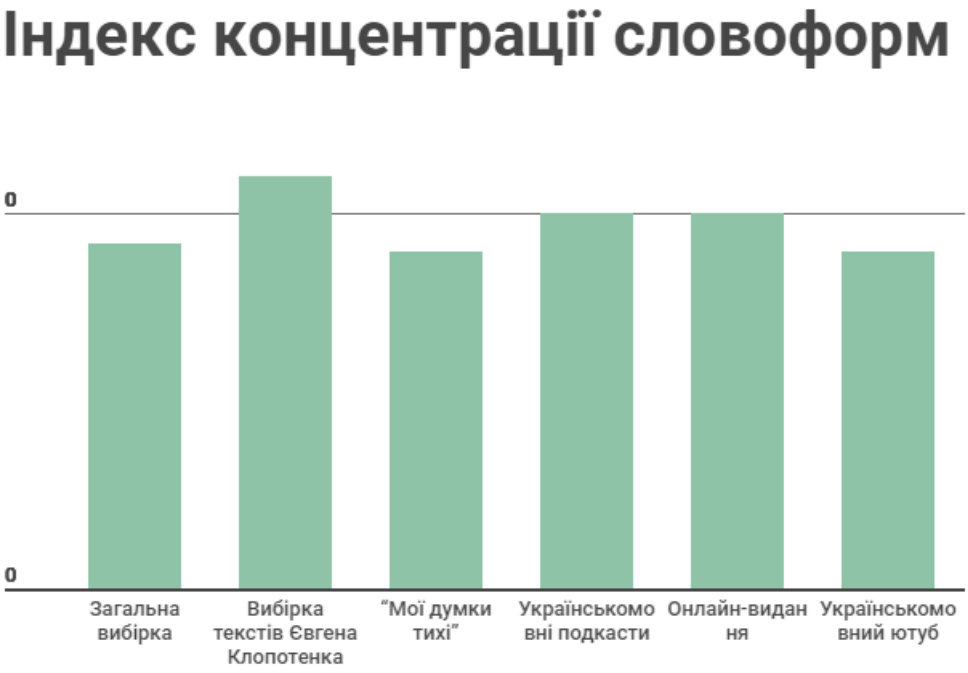
Додаток 13

Діаграма індексу концентрації лем



Додаток 14

Діаграма індексу концентрації словоформ

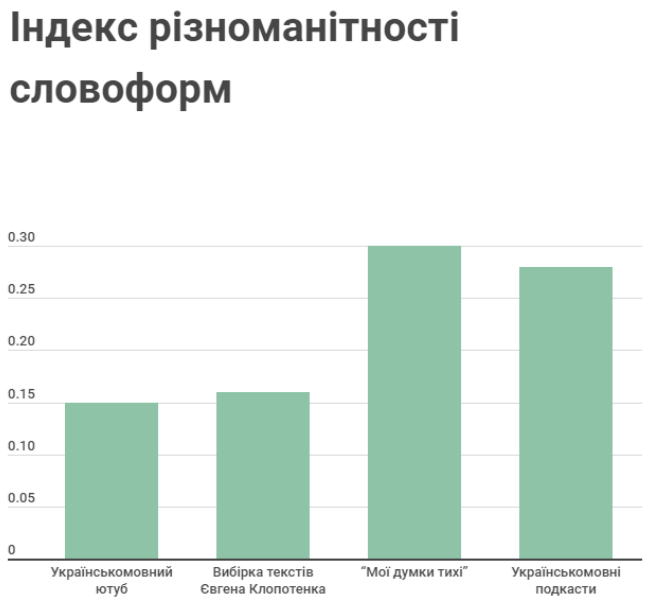


Додаток 15

Діаграма індексу різноманітності лем

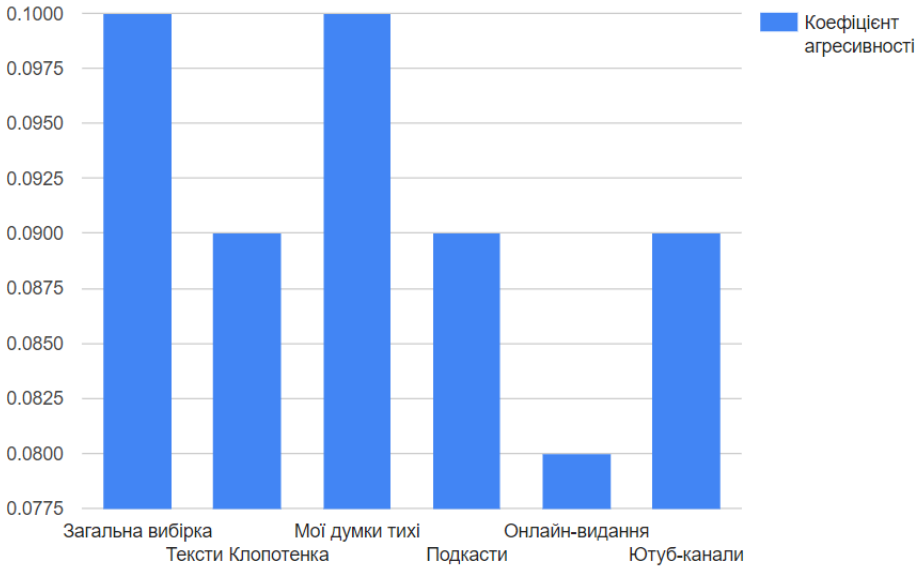
Додаток 16

Діаграма індексу різноманітності словоформ



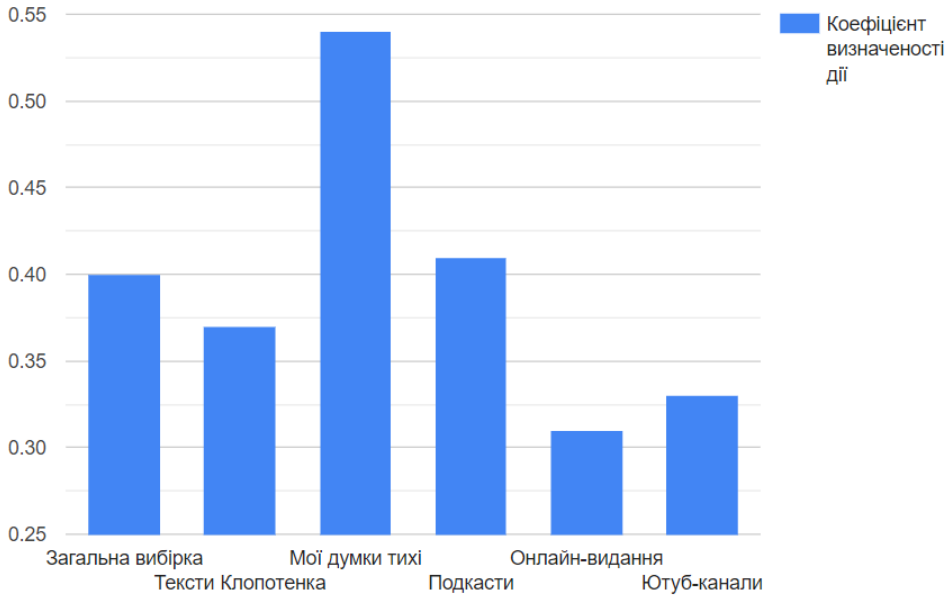
Додаток 17

Діаграма коефіцієнту агресивності



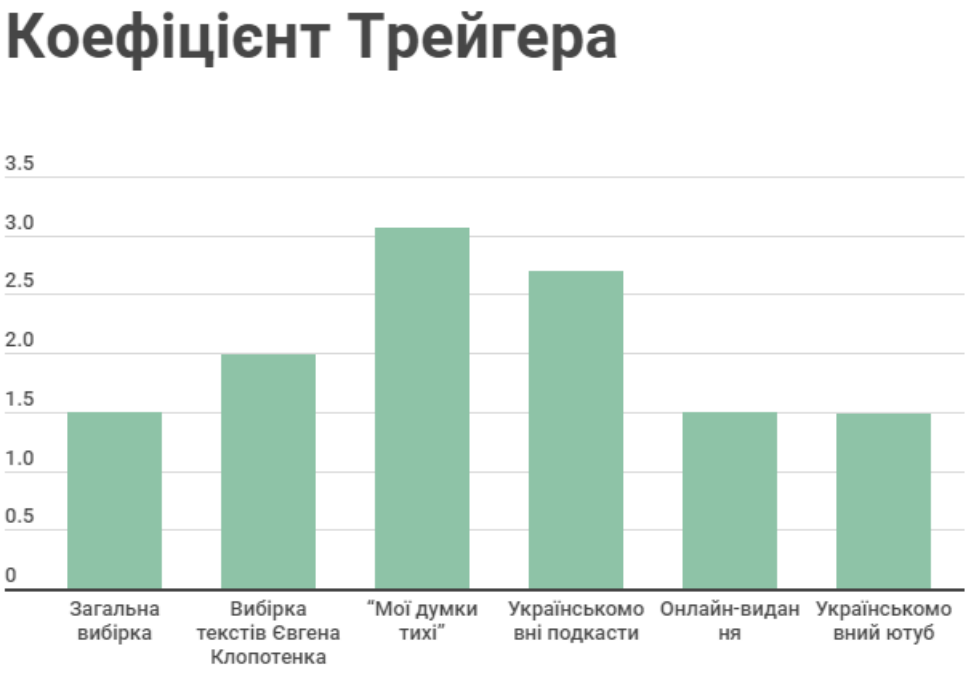
Додаток 18

Діаграма коефіцієнту визначеності дій



Додаток 19

Діаграма коефіцієнту Трейгера



Додаток 20

Посилання на файли баз даних українськомовних медійних текстів

https://drive.google.com/drive/folders/1yk4tMJmHLPVzc8VNA18rS73o9WLNuv4p?usp=sharing

Додаток 21

Посилання на файли корпусних таблиць українськомовних медійних текстів з метатекстовою розміткою

https://drive.google.com/drive/folders/18CC2JhFOsjF-8HRc5WyPBMGFPV-aLCJb?usp=sharing