«Маркери дискурсу»

(шифр)

«АНАЛІЗ МЕДІА ДИСКУРСУ НА БАЗІ КОРПУСУ»

Галузь: Прикладна лінгвістика

2020/2021

АНОТАЦІЯ

Актуальність теми полягає в тому, що дискурс – це явище, яке, на жаль, і далі залишається в Україні й українознавстві радше обговорюваним, аніж аналізованим, тому виявлення, аналіз частотності дискурсивних маркерів та створення корпусу – це крок до майбутньої розробки парсеру українського дискурсу.

Мета роботи: відобразити результати застосування статистичних методів роботи з дискурсом засобами середовища мови програмування Python та порівняти природу та частоту появи дискурсивних маркерів у розмовному та письмовому медіадискурсі.

Об’єкт роботи – методи створення корпусу та методи аналізу дискурсу в середовищі програмування PyCharm.

Предметом роботи є дискурсивні маркери кореферентності у корпусі українського медіадискурсу.

Для реалізації визначеної мети є наступні задачі:

1. Визначити поняття корпусів в лінгвістиці.
2. Огляд існуючих корпусів української мови.
3. Ознайомитися із поняттями дискурс та дискурсивні маркери.
4. Огляд лексикону українських маркерів дискурсу.
5. Побудувати корпус українського медіадискурсу.
6. Реалізувати програмний продукт за допомогою бібліотек та модулів Python для роботи з текстами.

Як теоретичну базу для проекту було використано літературу, статті та сайти, в яких було описано підхід до програми. Практичною базою є у середовище програмування PyCharm.

Робота складається зі вступу, трьох розділів, висновків, списку літератури, який містить 12 джерел. Обсяг роботи – 30 сторінок. У вступі описується актуальність задачі роботи з українським медіадискурсом та

аналізу і виявлення кореферентних дискурсивних маркерів. У першому розділі описано поняття корпусної лінгвістики, поняття корпусу та огляд сучасних корпусів української мови. У другому розділі описано поняття дискурсу, його структура та виявлено проблеми його автоматичної обробки. Розглядаються дискурсивні маркери, наводяться приклади та оглядається лексикон маркерів українського дискурсу. У третьому розділі наводиться опис створення та структура корпусу заданої тематики, опис алгоритму. Надано візуалізацію та опис результатів виконання програми.

Робота була опублікована в матеріалах четвертої міжнародній конференції з прикладної лінгвістики та інтелектуальних систем у 2020 році: «Using NLP Python Tools in Methods of Content Analysis» / Computational linguistics and intelligent systems: proceedings of the 4nd International conference (2), 2020, p. 240-242. Робота була випробувана на Всеукраїнській олімпіаді з Прикладної Лінгвістики у 2019 році.

Ключові слова: корпусна лінгвістика, корпус текстів, медіадискурс, дискурсивні маркери, кореферентні маркери, скрейпінг веб-сторінок, парсинг xml-документів, python

ЗМІСТ

ВСТУП

Дискурс - складне комунікативно-когнітивний явище, до складу якого входить не тільки сам текст, а й різні екстралінгвістичні фактори (знання світу, думки, ціннісні установки), які відіграють важливу роль для розуміння і сприйняття інформації. Елементами дискурсу служать викладені події, учасники цих подій, обставини, що супроводжують події, фон, оцінка учасників події і тому подібне.

Дискурсивні маркери – це лінгвістичні елементи, що функціонують на рівні дискурсу і є незалежними від його основних складових – частин мови. Найчастіше дискурсивні маркери репрезентують слова або словосполучення, які кодують значення та співвідносяться з частками, сполучниками, прислівниками і вставними словами, фразами, конструкціями. У корпусному підході основна функція дискурсивних маркерів полягає в структурно-смисловій організації тексту, оформленні та впорядкуванні суджень, що пов’язують окремі текстові фрагменти.

В даній роботі описано принцип роботи з дискурсом та дискурсивними маркерами, сучасні системи для скрейпінгу веб-сторінок для формування корпусу, парсингу xml-файлів та застосування статистичних методів при обробці тексту. Завданням роботи було створити корпус українського медіадискурсу, автоматично проаналізувати та виявити кореферентні дискурсивні маркери у корпусі.

Для роботи з дискурсом було обрано мову Python, що сприяє написанню надійного програмного продукту. Вона легка у вивченні, ідеально підходить на роль другої мови програмування. Мова Python має відкритий безкоштовний ресурс, на ній легко створювати різні пакети (бібліотеки), що використовується іншими програмістами. На даний момент Python є однією

з найпопулярніших мов програмування та на ряду з Java всебічно розвивається на ринку програмістів.

1. КОРПУСИ УКРАЇНСЬКОЇ МОВИ

1.1 Визначення терміну “Корпусна лінгвістика”

Корпусна лінгвістика – галузь мовознавства, предметом дослідження якої є принципи та методи формування корпусів текстів, а також розроблення комп’ютерних систем для їхнього опрацювання. Під **корпусом текстів** розуміється значний за обсягом, представлений в електронному вигляді, уніфікований, структурований, розмічений, філологічно компетентний масив мовних даних, створений для вирішення конкретних лінгвістичних завдань.

Корпус, також, розглядають як модель мовної системи, реалізовану в текстах різних функціонал. стилів, структури, тематики, призначення, території та часу їх створення. Залежно від призначення виділяють фундаментальні і пошукові (дослідницькі) корпуси, для роботи з якими створюють спеціальні текстові процесори – системи опрацювання інформації про мовні одиниці морфологічних, синтаксичних, логіко-семантичних структур рівнів тексту. Такі системи зорієнтовані на анотацію (розмічення) тексту, яка передбачає формалізоване виділення у ньому мовних одиниць певних типів. [6]

Передусім, корпусна лінгвістика як галузь прикладного мовознавства займається визначенням загальних принципів побудови, обробки та експлуатації даних лінгвістичних корпусів (корпусів текстів) із використанням сучасних комп’ютерних технологій, розробленням методики збору реальних мовних явищ – писемних та усних текстів, а також способів їх збереження та аналізу.

1.2 Аналіз корпусів української мови

Сучасні мовознавчі дослідження активно послуговуються корпусним методом і для усіх європейських мов, серед яких не забуваємо про слов‘янські, існує по кілька корпусів текстів різного типу, обсягу, структури, наповнення та призначення. Й узагальнивши понад сорокалітню

історію становлення корпусної лінгвістики та побудови корпусів, мовознавці визнали, що створення корпусу є обов‘язком щодо рідної мови. [7]

Національні корпуси мають здебільшого фундаментальний характер, їхній обсяг сягає 200 і більше мільйонів слововживань. Такими є, наприклад, Британський національний корпус (British Naional Corpus). Корпуси текстів різних типів та призначення мають естонську, італійську, китайську, німецьку, польську, російську, сербську, словенську, словацьку, турецьку, українську і чеську мови.

В Україні перші корпуси україномовних текстів з’явилися у 1990-х роках, зокрема корпус науково-реферативних та публіцистичних текстів, на базі якого співробітники відділу структурно-математичної лінгвістики (нині у складі Інституту української мови НАНУ, Київ) створили систему автоматичного орфографічного контролю тексту «Рута» та систему машинного російсько-українського і українсько-російського перекладу «Плай», а також корпус українських художніх, публіцистичних і наукових текстів, сформовний у лабораторії комп’ютерної лінгвістики Київського університету. Над укладанням фундаментальних національних корпусів україномовних текстів працюють науковці Українського мовно-інформаційного фонду НАНУ (Київ) та Інституту української мови НАНУ. Від 2009 при Міжнародному комітеті славістів працює Комісія з корпусної лінгвістики на чолі з М. Лазинським (Варшава), яка координує діяльність розробників корпусів слов’янських мов. Від 1996 виходить «International Journal of Corpus Linguistics».

На сьогодні в існують наступні корпуси:

- Лінґвістичний портал *Mova.info*. Корпус української мови [1] пропонує пошук тільки по підкорпусах (разом поки недоступно). Також розробники зґенерували частотні словники по кількох розділах та авторах[2].

- Корпус української мови *ГРАК* (Генеральний реґіонально анотований корпус української мови). У корпусі зібрано близько 25 тисяч текстів обсягом до 200 млн. слів. Є розвинутий пошук по всіх текстах, частотний словник та інші корисні додатки.[3]
- Цікавий величезний корпус українських інтернет-текстів зібрано в Лейпцизькому університеті. За розміром майже на порядок перевищує попередні. В результатах видачі подає приклади, сполучуваність, навіть тривимірний граф сполучуваності.[4]
- Ще один величезний корпус інтернет-текстів Лабораторія української (близько 3 млрдів слів). Утім, невідомо, чим викликані мовні кумедності опису сторінки.[5]
- Загальномовний (або національний) неанотований та несистематизований корпус української мови. Містить 6,6 Гб україномовних текстів з електронної бібліотеки «Чтиво».[6]

2. АНАЛІЗ ТА ОБРОБКА МЕДІАДИСКУРСУ

2.1 Визначення терміну «Дискурс»

Із появою такого нового концепту в науці, як “дискурс”, учені почали активно розробляти й методики його дослідження. І хоча дискурс – явище вже не нове, а скоріше, доволі поширене у сфері гуманітаристики, методи й методологія дослідження різних типів дискурсу досі є недостатньо розробленими. Пов’язано це, насамперед, з тим, що за останні 20–25 років значно розширились межі тлумачення терміна “дискурс”.

Дискурс – єдність мовлення та ситуації, в якій воно відбувається. Включає як перебіг мовлення, так і його передумови, обмеження та результати, позамовний контекст і невисловлені цілі й наміри, які супроводжують акт мовлення. Таке тлумачення дискурсу є найбільш вагомим для аналізу медіадискурсу, тому що в цьому сенсі можна розглядати відношення мови та ідеології, досліджувати умови створення тексту в межах дискурсу. [8]

2.2 Аналіз дискурсу

Дискурс-аналіз - це сукупність аналітичних методів інтерпретації різного роду текстів чи висловлювань як продуктів мовленнєвої діяльності людей, здійснюваної в конкретних суспільно-політичних обставин і культурно-історичних умовах. Тематичну, предметну і методичну специфіку таких досліджень покликане підкреслити саме поняття дискурсу, під яким розуміється соціально обумовлена і культурно закріплена система раціонально організованих правил слововживання і взаємини окремих висловлювань в структурі мовної діяльності.[9]

Однією з основних проблем в НЛП є обробка дискурсу - побудова теорій і моделей того, як висловлювання злипаються, утворюючи узгоджений дискурс. У нашому дослідженні дискурс-аналіз представлятиме собою виявлення формальних характеристик дискурсу (маркерів кореферентних зв’язків, що пов’язують між собою аргументи дискурсу і

формують закінчену думку). Критичний дискурс-аналіз буде застосований для порівняння частотності використання маркерів кореферентних зв’язків у письмовому та усному медіа-дискурсі.

2.3 Структура дискурсу

Когерентний (зв’язний) дискурс повинен володіти такими властивостями:

- Співвідношення когерентності (зв’язку) між висловлюваннями.

Дискурс був би послідовним, якби він мав значні зв'язки між своїми висловлюваннями. Ця властивість називається відношенням когерентності. Має бути якесь пояснення, щоб виправдати зв'язок між висловлюваннями.

- Відносини між сутностями.

Інша властивість, яка робить дискурс зв'язаним, полягає в тому, що повинні бути певні види відносин між сутностями. Такий вид послідовності називається когерентністю на основі сутностей.

- Структура дискурсу.

Важливе питання, що стосується дискурсу, полягає в тому, яку структуру повинен мати дискурс. Відповідь на це питання залежить від сегментації, яку ми застосували до дискурсу. Реалізувати дискурсивну сегментацію досить складно, але це дуже важливо для додатків пошуку інформації, підсумовування тексту і вилучення інформації.

1) Сегментація дискурсу без нагляду.

Клас сегментації дискурсу без нагляду часто представлений як лінійна сегментація. Ці алгоритми залежать від згуртованості, яка може бути визначена як використання певних лінгвістичних пристроїв для зв'язування текстових одиниць разом. З іншого боку, згуртованість лексики - це згуртованість, на яку вказують відносини між двома або більше словами в двох одиницях, наприклад використання синонімів.

2) Сегментація контрольованого дискурсу.

Попередній метод не має будь-яких помічених вручну сегментів. З іншого боку, сегментація контрольованого дискурсу повинна мати навчальні дані з маркуванням кордонів. У контрольованій сегментації дискурсу важливу роль відіграють дискурсивний маркер або ключові слова. Маркер дискурсу або ключове слово - це слово або фраза, які функціонують для позначення структури дискурсу.

- Узгодженість тексту.

Щоб досягти зв'язкового дискурсу, ми повинні зосередитися на зв'язності відносин. Ми беремо два терміни S0 і S1, щоб представити значення двох пов'язаних речень. З цього випливає, що стан, що затверджується терміном S0, може викликати стан, затверджується S1. Існують наступні відносини між висловлюваннями:

1) Два твердження показують результат відносин: «Рам був спійманий в вогні. Його шкіра згоріла.»
2) Два твердження можуть показувати відносини: «Рам бився з одним Шьяма. Він був п'яний.»
3) Два твердження є паралельні : «Рам хотів машину. Шиам хотів грошей.»
4) Два терміни показують розвиток відносини: «Рам був з Чандигарха. Шиам був з Керали». [8]

- Побудова ієрархічної структури дискурсу.

Когерентність всього дискурсу може також розглядатися ієрархічною структурою між відносинами когерентності. Наприклад, наступний уривок може бути представлений у вигляді ієрархічної структури (рис. 2.1):

S 1 - Рам пішов в банк, щоб внести гроші.

S 2 - Потім він сів на поїзд до магазину одягу Шиам.

S 3 - Він хотів купити одяг.

S 4 - У нього немає нового одягу для вечірки.

S 5 - Він також хотів поговорити з Шиам щодо його здоров'я.

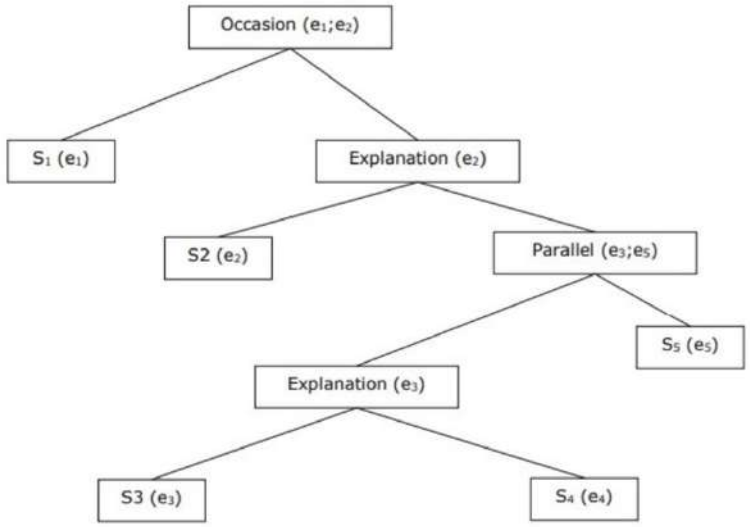


Рисунок 2.1 - Ієрархічна структура дискурсу

- Позначення зв’язку

Інтерпретація речень з будь-якого дискурсу є ще одним важливим завданням, і для досягнення цієї мети нам необхідно знати, про кого або про який об'єкт іде мова.

1) Референт - це сутність, яка згадується. Наприклад, в останньому наведеному прикладі Рам є референтом.
2) Corefer - коли два вирази використовуються для посилання на один і той же об'єкт, вони називаються corefers. Наприклад, «Рам і він - помічники.»
3) Антецедент - термін має ліцензію на використання іншого терміну. Наприклад, Рам є попередником посилання він.
4) Анафора і Анфор - це може бути визначено як посилання на сутність, яка була раніше введена у речення. Вираз, який здійснює посилання, називається анафоричним.
5) Модель дискурсу - модель, яка містить уявлення сутностей, які згадувалися в дискурсі, і відносини, в яких вони беруть участь.
6) Посилальний вираз - вираз природною мовою, що використовується для виконання посилання, називається

посилальним виразом. Наприклад, уривок, використаний вище, є посилальним виразом.

2.4 Граматичні типи посилальних виразів та маркери кореферентних зв'язків

- Невизначені іменникові фрази.

Такий вид посилання представляє об'єкти, які є новими для слухача в контексті дискурсу. Наприклад, у реченні: «Рам одного разу ходив, щоб принести йому трохи їжі.», - *йому* є невизначена вказівка.

- Певні фрази іменників.

На противагу вищесказаному, такого роду посилання представляють сутності, які не є новими або не можуть бути ідентифіковані для слухача в контексті розмови. Наприклад, у реченні: «Я читав «Таймс оф Індія».» - *Таймс оф Індія* - це певне посилання.

- Займенники / вказівні займенники

Це форма певного посилання. Наприклад, «Рам сміявся так голосно, як тільки він міг.» Слово, *він* представляє займенник, що посиється на ім'я.

- Імена.

Це найпростіший тип посилального виразу. Це може бути ім'я людини, організації та місця. Наприклад, в наведених вище прикладах Рам є виразом, що посилається на ім'я.

Завдання пошуку кореферентних зв'язків полягає в тому, щоб знайти посилальні вирази в тексті, що посилаються на одну і ту ж сутність. Простіше кажучи, це завдання пошуку виразів corefer. Набір виразів coreferring називається ланцюгом coreference. Наприклад, «Він, Головний менеджер і його підлеглі в офісі.», *його* - це посилання на займенник *він*. Вирішення кореферентності полягає у вирішенні займенникової анафори - знайти слово Рам, тому що Рам є попередником до займенників *він* та *його*. [9]

Маркери в українському дискурсі представлені підрядними та сурядними сполучниками та союзними прислівниками (і, а, або, що і т.д.). Вони демонструють як саме здійснене посилання між сутностями.

2.5 Дискурсивні маркери

Дискурсивні маркери – це мовні одиниці, які служать для встановлення певного експліцитного зв'язку між даним сегментом дискурсу і попереднім висловлюванням.

В цілому, з точки зору граматики дискурсивні маркери можна поділити на чотири класи – вставні слова (проте, таким чином, до речі), сполучники сурядного зв'язку (а, але, і, або), підрядного зв'язку (в той час, оскільки, хіба що) та прийменникові вирази (в результаті, через, незважаючи на, порівняно з).

Більш ґрунтовним є розрізнення дискурсивних маркерів за семантичними ознаками. Вважається, що дискурсивні маркери мають основне значення, яке вказує на тип зв'язку між висловлюваннями, або сегментами дискурсу, і, відповідно, їх можна поділити на додаткові, причинно-наслідкові, умовні, темпоральні й протиставні:

- Додаткові: також, і, а, та, крім того, далі (більше), відповідно, додатково (до), до того ж, до речі, зокрема, аналогічно, або, ні, ні ... ні, що більше.
- Причинно-наслідкові: відповідно, як результат, тому що, отже, з цієї причини, отже, для того, щоб, через, отже, таким чином, оскільки, так, так що.
- Умовні: припускаючи, навіть якщо, якщо, у такому випадку, інакше, передбачено, хіба що, за певних умов.
- Темпоральні: після цього / того, згодом, врешті-решт, по-перше, нарешті, спочатку, нарешті, далі, зараз, по-друге, потім, нарешті.

- Протиставні: хоч(а), але, навпаки, незважаючи, однак, у порівнянні, навпаки, замість цього, незважаючи на, тим не менше, навпаки, з іншого боку, швидше (ніж), все-таки, тоді як, поки.

Разом з тим, дискурсивний маркер може мати інші значення, які є прагматично-орієнтованими і розпізнаються лише у контексті комунікації. Наприклад, «так» / «таким чином» може вказувати не лише на причинно-наслідковий зв'язок між окремими висловлюваннями (а), але й на початок нової розмови (б), перевірку того, чи було правильно зрозуміле почуте (в), або означати, що для адресата отримана інформація є неважливою (г).

Для аналізу медіадискурсу у цій роботі було використано xml-документ із маркерами дискурсу в українській мові. XML-документ українських маркерів має наступну структуру. (рис. 2.2) [11]

```
<?xml version='1.0' encoding='UTF-8'?>
<dimlex>
  <entry id="c1" word="а">
    <english_equivalent>but</english_equivalent>
    <orths>
      <orth type="cont" canonical="1" onr="1o1">
        <part type="single">а</part>
      </orth>
    </orths>
    <syn>
      <cat>coord_conj</cat>
      <sem>
        <pdtb3_relation sense="contrast"/>
      </sem>
      <sem>
        <pdtb3_relation sense="conjunction"/>
      </sem>
      <sem>
        <pdtb3_relation sense="synchronous"/>
      </sem>
    </syn>
  </entry>
```

Рисунок 2.2 - Структура xml-документу українських дискурсивних маркерів

Кореневим елементом є елемент з іменем <entry>, в який у якості параметра «word», передається дискурсивний маркер. В єлементі <english_equivalent> розміщується єквівалент цього маркера англійською мовою. У єлемент <syn> у якості дочірнього єлементу <cat> передається граматична частина дискурсивного маркеру. У дочірньому єлементі <sem> перелічуються усі семантичні ознаки маркеру.

Цей документ зберігає усі дискурсивні маркери української мови, які будуть використані для сегментації текстів у корпусі медіадискурсу та пошуку дискурсивних висловів: когерентності та анафори.

3 СТВОРЕННЯ ТА АНАЛІЗ КОРПУСУ УКРАЇНСЬКОГО МЕДІАДИСКУРСУ

3.5Створення корпусу українського медіа дискурсу

У рамках цієї роботи було створено корпус українського розмовного та письмового медіадискурсу. Корпус включає в себе тексти з 2000 по 2020 роки. Корпус поділений на дві частини: розмовний медіадискурс та письмовий медіадискурс.

Частина із розмовним медіадискурсом налічує більше чотириста тисячі слів. До неї входять тексти, що поділяються ще на два типи: інтерв'ю та промови. Тексти з інтерв'ю були зібрані із онлайнових версій українських газет, журналів та інших інтернет видань. Тексти із промовами складаються із промов українських політиків (привітання із новим роком, плани на майбутне).

Частина із письмовим медіадискурсом налічує більше мільйона слів. Тексти у цьому розділі поділяються на дискурс із соціальної мережі «Фейсбук» та газетний дискурс. До неї входять тексти із українських газет «Кримська Світлиця», «Галнет» та «Українська правда». Значною частиною цього разділу є тексти із постів у Фейсбуці українських політиків, активістів, артистів та науковців. [12]

Файли для створення корпусу українського медіадискурсу були отриманні із корпусу ГРАК, що згадується у підрозділі 1.3 та за допомогою методу скрейпінгу веб-сторінок. Для того аби зібрати тексти статей із різноманітних інтренет ресурсів видавництв, необхідно використати метод скрейпінгу веб-сторінок за домогою бібліотеки BeautifulSoup та requests.

Усі тексти зберігаються в корпусі у форматі .txt із кодуванням utf-8. Приклад зберігання текстів у корпусі зображено на рис. 3.1. Приклад змісту файлів із кожного з чотирьох розділів корпусу можна бачити на рис. 3.2, рис. 3.3, рис. 3.4 та рис. 3.5.

Рисунок 3.1 – Корпус українського медіадискурсу

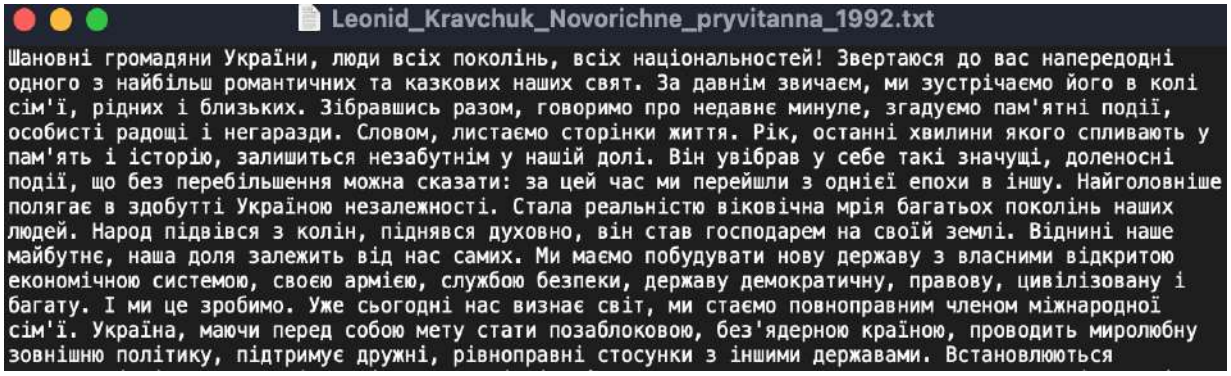


Рисунок 3.2 – Файл із розділу розмовний медіадискурс «промови»

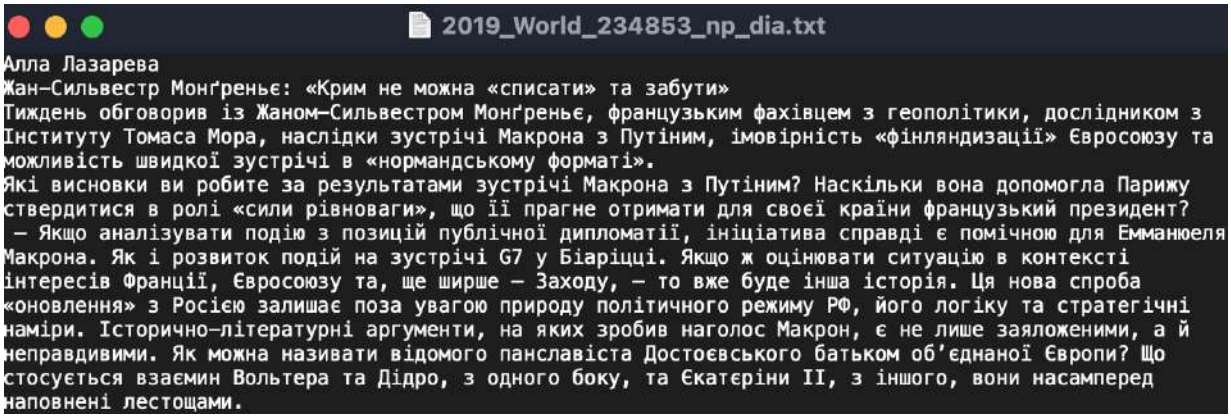
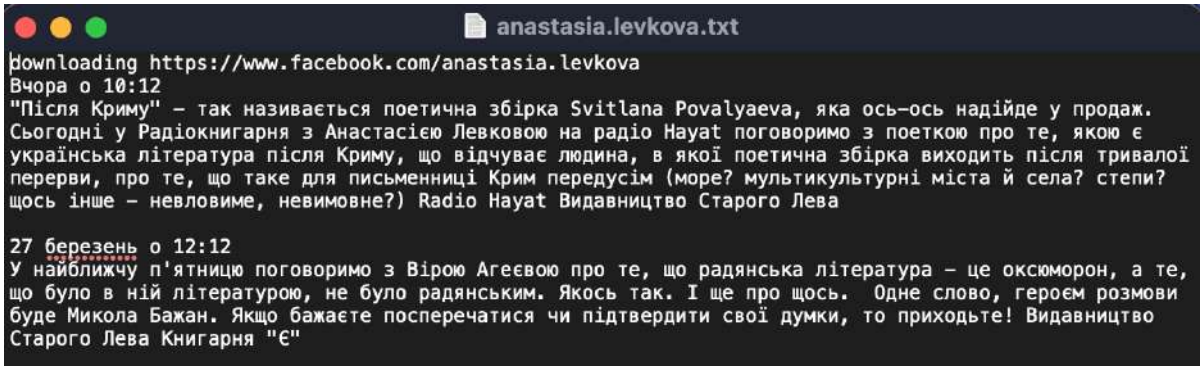
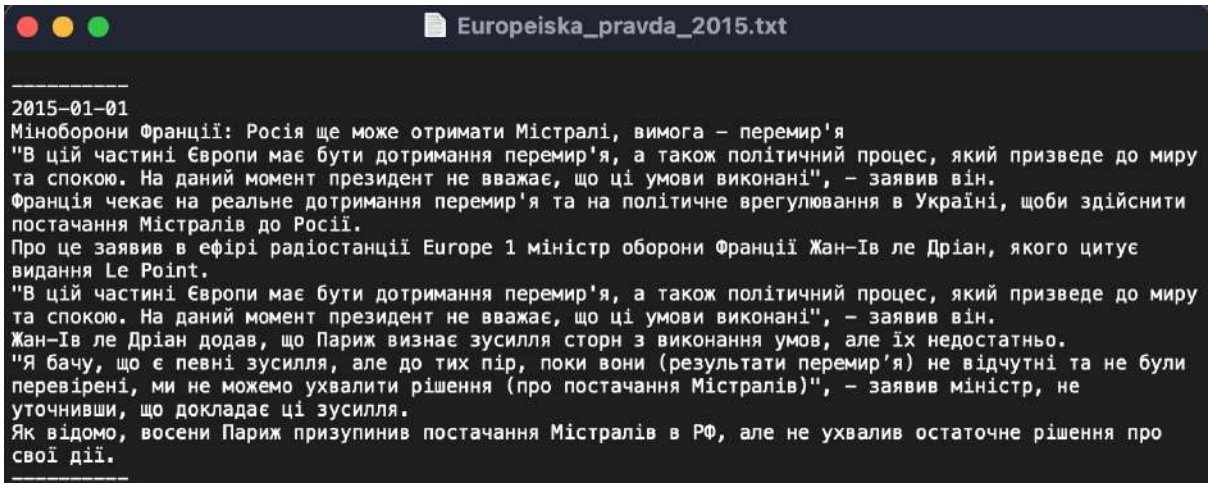


Рисунок 3.3 – Файл із розділу розмовний медіадискурс «інтерв'ю»



3.4 – Файл із розділу письмовий медіадискурс «Фейсбук»



3.5 – Файл із розділу письмовий медіадискурс «Газети»

Методом аналізу корпусу було обрано пошук маркерів дискурсу, що були згадані у підрозділі 2.4 та порівняння частоти їх появи у кожних розділах та підрозділах корпусу.

3.6Алгоритм роботи програми

1. Імпортувати необхідні бібліотеки та модулі для скрейпінгу.
2. Отримати список посилань на усі випуски газети певного року за допомогою методу get_all_links(url), використовуючи бібліотеки «BeautifulSoup», «requests» та «re».
3. Отримати список, що містить посилання на усі статті певного випуску за допомогою методу get_texts(url).

4. Отримати текст статті та записати його в файл, що збирає тексти статей за певний рік. Для цього використовуємо функцію downloading_texts(url).
5. Помістити усі файли із статтями у директорію корпусу, де містяться файли із газетним дискурсом.
6. Об’єднати усі файли із підрозділу «Промови» розділу розмовного дискурсу в один «speeches.txt» за допомогою модулю «os» та її функції walk().
7. Об’єднати усі файли із підрозділу «Інтерв’ю» розділу розмовного дискурсу в один «interviews.txt».
8. Об’єднати усі файли із підрозділу «Фейсбук» розділу письмового дискурсу в один «FB.txt».
9. Об’єднати усі файли із підрозділу «Газети» розділу письмового дискурсу в один «newspapers.txt».
10. Об’єднати усі файли із розділу письмового дискурсу в один «written.txt».
11. Об’єднати усі файли із розділу розмовного дискурсу в один «spoken.txt».
12. Зібрати маркери дискурсу із xml-документу у список «markers_list» та словник використовуючи модуль xml.dom.minidom.
13. Деякі маркери із xml-документу мають трикрапку в середині чи на кінці. Використовуємо регулярні вирази та редагуємо маркери для більш ефективного пошуку.
14. Записуємо відредаговані маркери у новий список edited_markers.
15. Рахуємо скільки раз зустрівся кожний маркер із списку у файлі «speeches.txt» використовуючи клас «Counter» модулю «collections».
16. Знаходимо відносну частоту появи певного маркеру на 1000 слів у корпусі, так як корпус поділяється на дві нерівні частини.

17. Створюємо список, елементами якого є списки, значеннями яких є: назва файлу («speeches»), маркер та його відносна частота появи у файлі «speeches.txt».
18. На основі створенного списку, формуємо датафрейм використовуючи модуль «pandas». [10]
19. Формуємо csv-файл «speeches.csv» із створенним датафреймом.
20. Повторюємо порядок дії (15-19) із файлами «interviews.txt», «FB.txt», «newspapers.txt», «written.txt» та «spoken.txt».
21. Для візуалізації результатів аналізу імпортуємо бібліотеку «matplotlib» та «seaborn».
22. Отримуємо датафрейм із файлу «speeches.csv».
23. Сортуємо датафрейм за стовпчиком «count».
24. Виводимо 10 маркерів, що зустрічаються найчастіше у файлі speeches.txt. (рис. 3.6)
25. На основі датафрейму формуємо гістограму. (рис. 3.7).

Відносна частота появи маркерів у файлі speeches.txt

	Unnamed: 0	file	word	count
21	21	speeches	і	401.787139
3	3	speeches	але	172.969523
0	0	speeches	а	171.214297
56	56	speeches	та	163.874262
78	78	speeches	як	159.725547
115	115	speeches	так	97.016116
98	98	speeches	й	96.218286
77	77	speeches	ще	87.282591
67	67	speeches	тому	76.112973
76	76	speeches	чи	69.092070

Рисунок 3.6 – Відсортований датафрейм за стовпчиком «count» на основі файлу speeches.txt

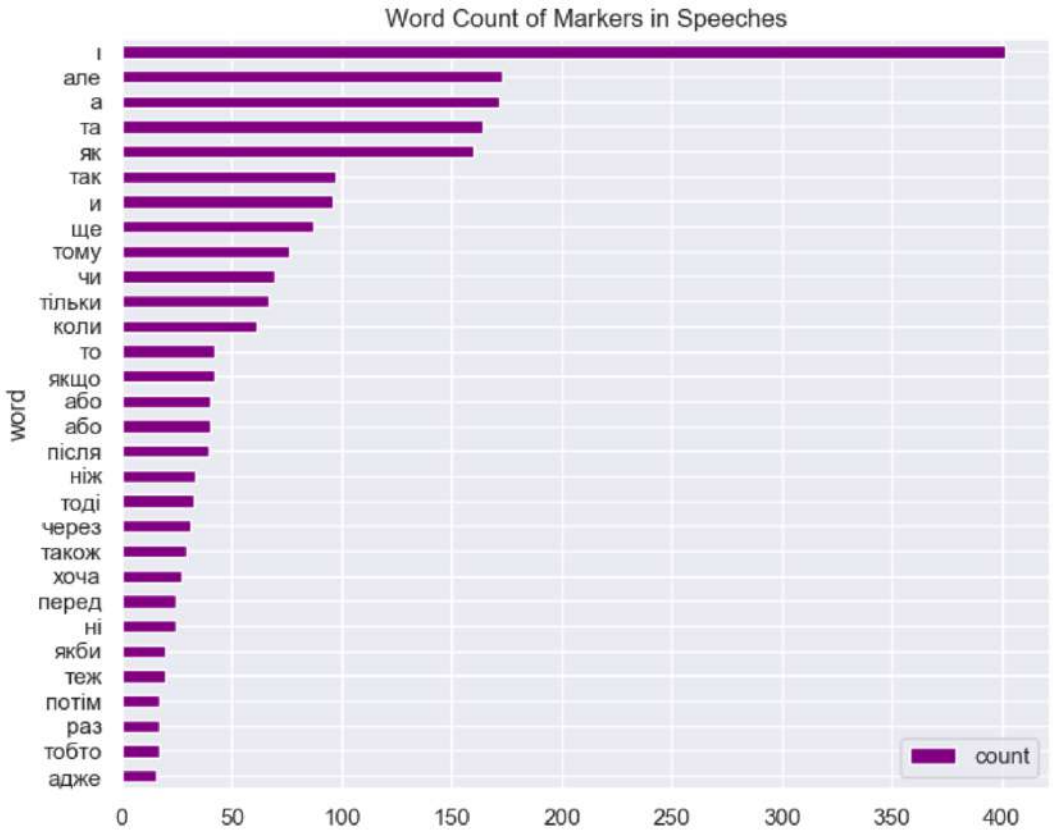


Рисунок 3.7 – Гістограма частотності появи маркерів у файлі speeches.txt

26.Повторюємо порядок дій (22-25) із файлом «interviews.csv». (рис. 3.8, рис. 3.9)

	Unnamed: 0	file	word	count
21	21	interviews	і	448.323662
0	0	interviews	а	272.308188
56	56	interviews	та	226.370084
78	78	interviews	як	225.789813
98	98	interviews	й	195.841393
3	3	interviews	але	171.889104
76	76	interviews	чи	133.268859
23	23	interviews	коли	133.268859
115	115	interviews	так	132.495164
67	67	interviews	тому	107.962605

Рисунок 3.8 - Відсортований датафрейм за стовпчиком «count» на основі файлу interviews.txt

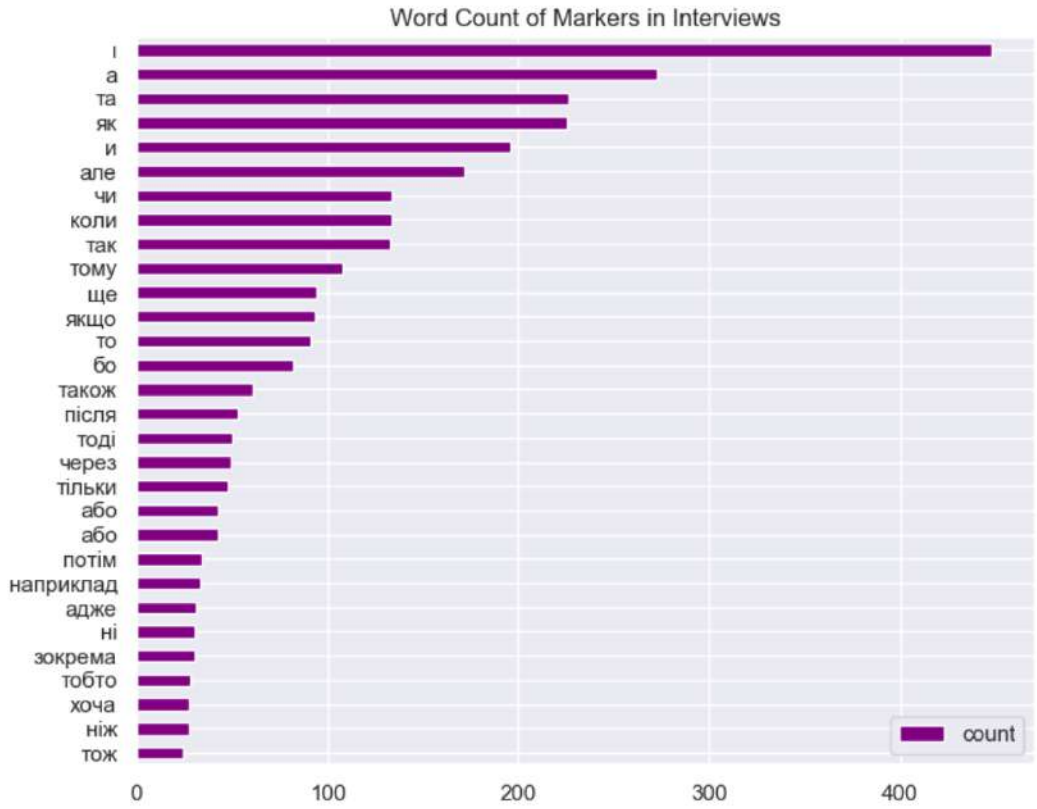


Рисунок 3.9 – Гістограма частотності появи маркерів у файлі interviews.txt

27. Для порівняння результатів аналізу файлів speeches.txt та interviews.txt об’єднуємо два датафреми та будуємо об’єднану гістограму. (рис. 3.10, рис. 3.11)

	Unnamed: 0	file	word	count
21	21	interviews	і	448.323662
21	21	speeches	і	401.787139
0	0	interviews	а	272.308188
56	56	interviews	та	226.370084
78	78	interviews	як	225.789813
98	98	interviews	й	195.841393
3	3	speeches	але	172.969523
3	3	interviews	але	171.889104
0	0	speeches	а	171.214297
56	56	speeches	та	163.874262

Рисунок 3.10 - Відсортований датафрейм за стовпчиком «count» на основі файлів speeches.txt та interviews.txt

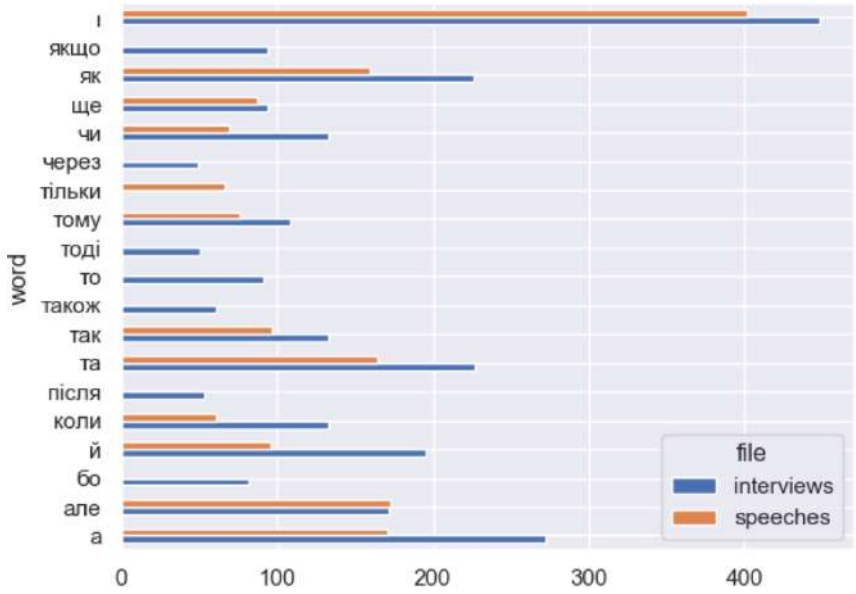


Рисунок 3.11 – Гістограма частотності появи маркерів у файлах speeches.txt та interviews.txt

28. Повторюємо дії 22-27 для файлів «FB_data.csv» та «Newspapers_data.csv». (рис. 3.12, рис. 3.13, рис. 3.14, рис. 3.15, рис. 3.16)

Відносна частота появи маркерів у файлі FB.txt

	Unnamed: 0	file	word	count
21	21	FB	і	134.645015
0	0	FB	а	96.018802
78	78	FB	як	62.739195
56	56	FB	та	48.875651
115	115	FB	так	48.577433
3	3	FB	але	48.286317
77	77	FB	ще	40.553974
63	63	FB	то	38.949282
98	98	FB	й	37.333939
23	23	FB	коли	32.725775

Рисунок 3.12 - Відсортований датафрейм за стовпчиком «count» на основі файлу FB.txt

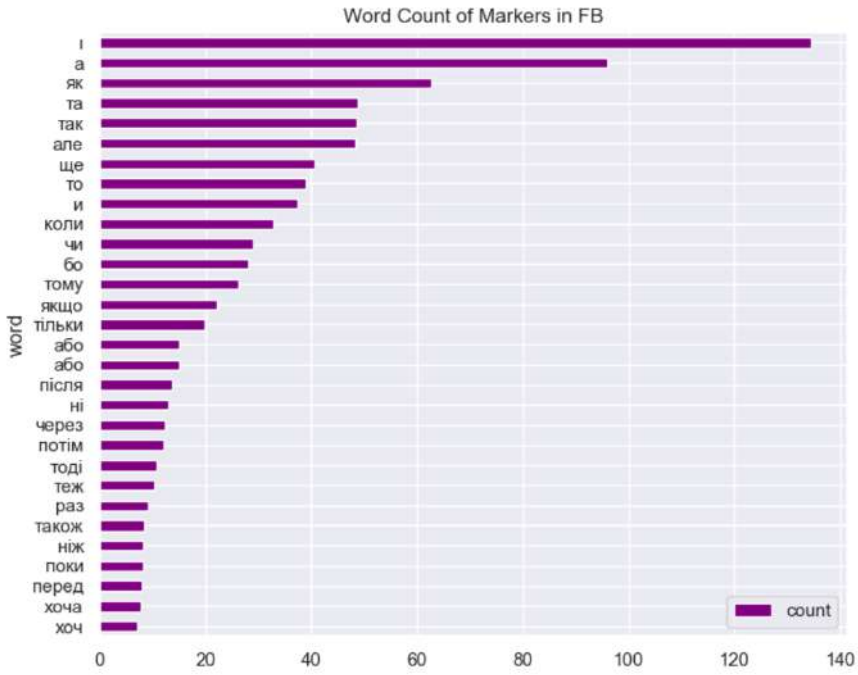


Рисунок 3.13 – Гістограма частотності появи маркерів у файлі FB.txt

Відносна частота появи маркерів у файлі newspapers.txt

	Unnamed: 0	file	word	count
21	21	newspapers	і	186.537867
56	56	newspapers	та	134.582248
78	78	newspapers	як	80.644187
59	59	newspapers	також	56.424996
0	0	newspapers	а	51.333885
54	54	newspapers	раніше	33.246764
3	3	newspapers	але	32.816880
67	67	newspapers	тому	27.718664
75	75	newspapers	через	26.730995
46	46	newspapers	після	26.585332

Рисунок 3.14 – Відсортований датафрейм за стовпчиком «count» на основі файлу newspapers.txt

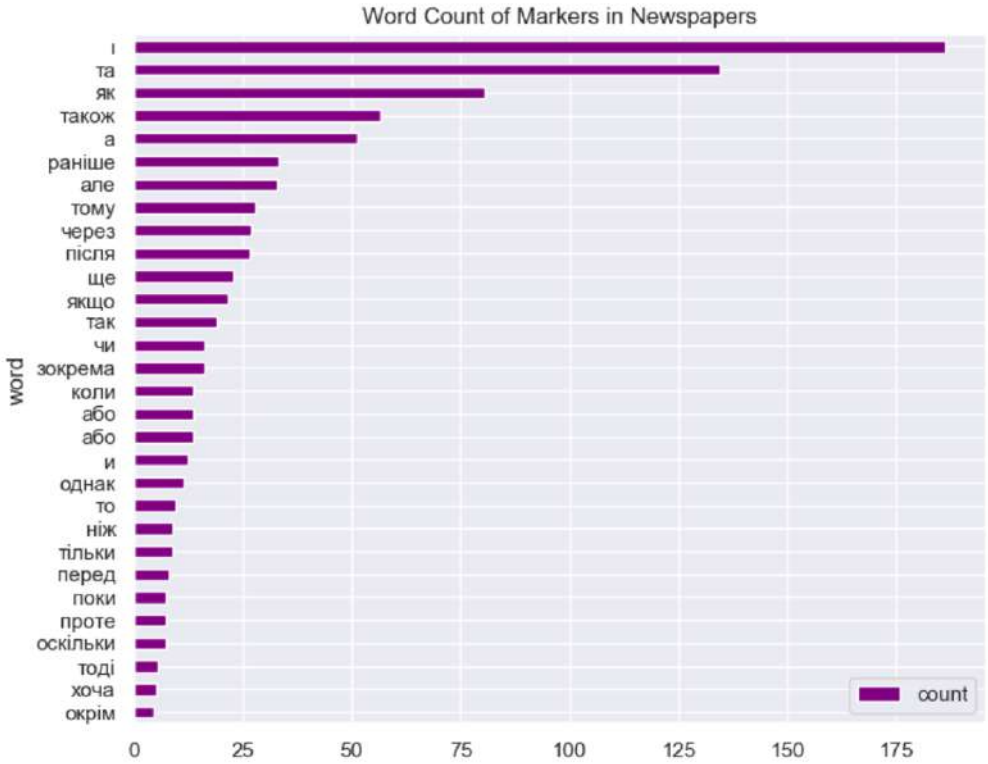


Рисунок 3.15 – Гістограма частотності появи маркерів у файлі newspapers.txt

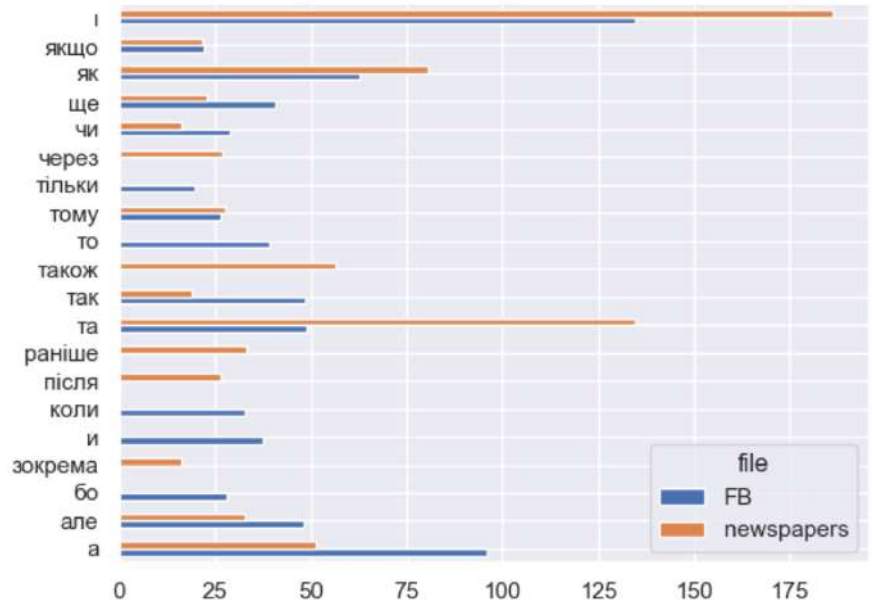


Рисунок 3.16 – Гістограма частотності появи маркерів у файлах FB.txt та newspapers.txt

29. Об’єднуємо чотири датафрейми та робимо результуючу гістограму по усім підрозділам корпусу. (рис. 3.17)

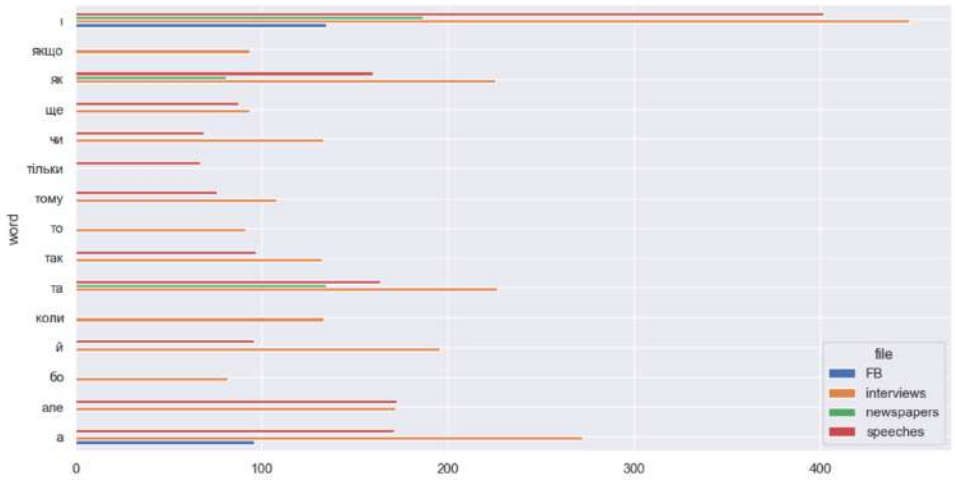


Рисунок 3.17 – Гістограма частотності появи маркерів у файлах FB.txt, newspapers.txt, speeches.txt та interviews.txt

30. Об’єднуємо датафрейми із файлів s_combined.csv та w_combined.csv, та отримуємо результуючу гістограму по двох розділах корпусу. (3.18)

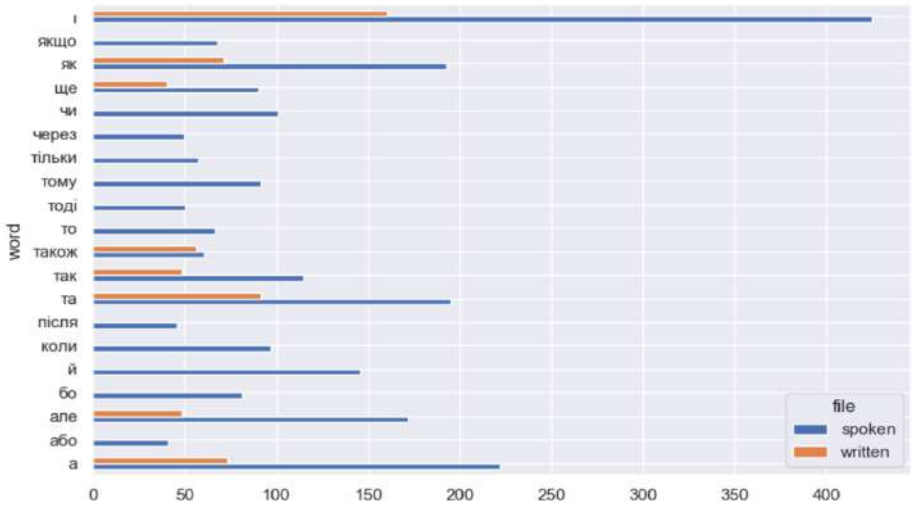


Рисунок 3.18 – Гістограма частотності появи маркерів у файлах spoken.txt та written.txt

3.7Результати аналізу

Дискурсивні маркери відображають те, як людина працює над текстом, про що вона думає і як сприймає кінцевого адресата, крім того, обізнаність кінцевого адресата в відповідній тематиці. Дискурсивні маркери включають в себе великий обсяг понять, а саме: взаємодії, комунікації, ієрархії, ввічливості, пам'яті, уваги тощо. Також, вони відображають інтенції адресанта і адресата, тлумачать дискурс з позиції мовця і слухача. [10]

Аналізуючи отримані результати у вигляді датафрейму можна зробити висновок, що дискурсивні маркери частіше зустрічаються в усному медіа-дискурсі ніж у письмовому.

Найбільшу відносну частоту появи у тексті мають маркери: «і», «а», «як», «та», «але». Найчастіше ці маркери об'єднують аргументи дискурсу позначаючи «зв'язок» між ними. Їх семантична ознака – додатковість, тому вони використовуються для продовження однієї і тієї ж думки.

Проведений аналіз довів твердження, що дискурсивні маркери більш притаманні розмовному дискурсу ніж письмовому. Через те, що думка, яку несе людина при усному спілкуванні є неконтрольованою та, частіше за все, спонтанною. В письмовій формі маркери ж з'являються менш частотно, що пояснюється тим, що автор письмового повідомлення завчасно формулює висловлення.

ВИСНОВКИ

В ході виконання роботи було розглянуто природу та структуру медіадискурсу. Приведено класифікацію дискурсивних маркерів та огляд їх використання в українській мові. Виконано пошук та аналіз частотності появи маркерів в українському медіадискурсі. Розглянуто методи парсингу xml-документів, методи скрейпінгу веб-сторінок та способи візуалізації даних, використовуючи бібліотеки та модулі Python. Проаналізовано основні статистичні методи підрахунку відносної частоти появи дискурсивних маркерів у корпусі текстів .

Було створено корпус текстів українського медіадискурсу, розроблено програмне забезпечення для підрахунку відносної частоти появи маркерів дискурсу у цьому корпусі та візуалізовано результати аналізу.

Робота була опублікована в матеріалах четвертої міжнародній конференції з прикладної лінгвістики та інтелектуальних систем у 2020 році: «Using NLP Python Tools in Methods of Content Analysis» / Computational linguistics and intelligent systems: proceedings of the 4nd International conference (2), 2020, p. 240-242. Робота була випробувана на Всеукраїнській олімпіаді з Прикладної Лінгвістики у 2019 році.

В майбутньому планується розширення корпусу українського медіадискурсу на основі цього корпусу та розробка програмного забезпечення для парсингу медіадискурсу української мови.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

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

2. (Частотні словники) [Електронний ресурс]. — Режим доступу: http://www.mova.info/carticle.aspx?l1=210&DID=5215

3. 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. — [Електронний ресурс]. — Режим доступу: http://www.parasolcorpus.org/bonito/run.cgi/first_form

4. Корпус українських інтернет-текстів [Електронний ресурс]. — Режим доступу: https://corpora.uni-leipzig.de/de?corpusId=ukr_mixed_2014

5. Лабораторія української [Електронний ресурс]. — Режим доступу: https://mova.institute

6. О. Демська-Кульчицька, «НАЦІОНАЛЬНИЙ КОРПУС УКРАЇНСЬКОЇ МОВИ: КОНЦЕПТУАЛЬНИЙ АСПЕКТ».

7. В. В. Павленко, “ДИСКУРСИВНИЙ АНАЛІЗ І КРИТИЧНИЙ ДИСКУРСИВНИЙ АНАЛІЗ МЕДІАДИСКУРСУ: ДО ПОСТАНОВКИ ПРОБЛЕМИ”

8. Дискурс український медій: ідентичності, ідеології, владні стосунки/ Володимир Кулик. – К.: Критика, 2010 – 656с.

9. Роль номінативних одиниць і дискурсивних маркерів у формуванні іміджу ООН в англомовному медіа-дискурсі/ Вісник Житомирського державного національного університету ім. І. Франка – 2016р, – № 2(84) – 141 – 145с.

10. Pandas Python documentation [Електронний ресурс]. — Режим доступу: https://pandas.pydata.org

11. Лексикон українських дискурсивних маркерів Електронний ресурс]. — Режим доступу: https://github.com/TScheffler/UK_DiMLex/blob/master/UK_DiMLex.xml

12. Газета «Кримська світлиця» Електронний ресурс]. — Режим доступу: http://svitlytsia.crimea.ua