МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Київський національний університет імені Тараса Шевченка

Навчально-науковий інститут філології
Кафедра української мови та прикладної лінгвістики

Лінгвістична параметризація текстів з мережі Facebook (на прикладі новинних дописів 2021 та 2023 років)

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»
студентки IV курсу
Христини ДЯКУН

Науковий керівник
док. філол. н., Наталія ДАРЧУК

КИЇВ – 2023

ЗМІСТ

Вступ. 3
РОЗДІЛ 1. 5
Роль медіа у суспільстві та вплив змін у соціумі на мову новин 5
1.1. Роль медіа в сучасному житті суспільства 5
1.2. Лінгвістика українськомовних новинних текстів 8
1.3. Дослідження впливу військових подій на мову та стилістику текстів 9
Висновки до розділу 1 12
РОЗДІЛ 2. Створення корпусу медіатекстів та частотного словника 13
2.1. Метод створення текстового корпусу у проведенні лінгвістичного аналізу. 13
2.2. Основні підходи статистичної параметризації 16
2.3. Використання статистичної параметризації для виявлення змін у лінгвістичних особливостях 20
2.4. Створення корпусу текстів воєнного та передвоєнного періоду на матеріалах новин у Facebook 23
2.5. Частотний словник. Алгоритм роботи з частотним словником. 26
2.6. Створення частотних словників словоформ.(програмне забезпечення) 29
Висновки до розділу 2 37
Розділ 3. Граматична та лексична параметризація новинних текстів за двома вибірками 39
Висновки до розділу 3 45
Висновки 46
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ 47

Вступ.

Сучасні медіа є важливою частиною нашого життя. Медійні джерела надають інформацію, емоції, можливість спілкування, а також впливають на наше світосприйняття. Разом з тим важливо досліджувати, як змінюються медіа в час, який охоплює інтервал до повномасштабного вторгнення і під час воєнний дій.

Актуальність дослідження. Актуальність даної теми полягає в дослідження сучасних особливостей українськомовного медіапростору, а саме медіатекстів з інтернет-мережі Facebook ,адже саме формат онлайн поширення інформації є невід'ємною частиною сьогодення. Також актуальною темою являється виявлення змін у медіатекстах з початком повномасштабного вторгнення 24.02.2022.

Мета та завдання дослідження. Метою нашої роботи є дослідження змін, що відбуваються у новинному дискурсі на сучасному етапі, зокрема змін параметрів новинних текстів, розміщених у мережі Facebook до початку повномасштабного вторгнення Російської Федерації в Україну та в час активних воєнних дій.

Відповідно до мети були поставлені такі завдання:

- збір матеріалу дослідження –текстів новин, розміщених на сторінках українськомовних ЗМІ в мережі Facebook, за період квітень-травень 2021 року (до повномасштабного вторгнення) та квітень-травень 2023 року (в час активних воєнних дій);
- Оформлення матеріалу в корпуси у табличній формі
- написання програмного коду для створення частотних словників на основі зібраних текстів;
- завантаження баз даних для комфортного доступу;
- кількісний порівняльний аналіз граматичних та лексичних параметрів текстів.

Об’єктом дослідження є українськомовні медіатексти з соціальної мережі Фейсбук.

Предмет дослідження – лексико-граматичні особливості новинних медійних текстів на сучасному етапі.

Практичне значення отриманих результатів полягає в створенні частотних словників на базі корпусів текстів, що являється зручним методом для лінгвістичного аналізу українськомовних текстів, а також у дослідженні сучасних змін лексико-граматичних параметрів медійного мовлення.

Методологічною базою дослідження є методи ком’ютерної лінгвістики, зокрема метод укладання частотних словників для аналізу текстів; використання програмного забезпечення та бібліотек\ модулів - sqlite3, os, enum , collections.defaultdict , docx, Spacy.

Матеріалом дослідження є 100 українськомовних текстів, дібраних з офіційних сторінок українськомовних ЗМІ, що були опубліковані в мережі Facebook за період квітень-травень 2021 року та квітень-травень 2023 року.

Структура та обсяг дипломної роботи. Кваліфікаційна робота складається зі вступу, 3 розділів, висновків та списку використаної літератури та додатків.

РОЗДІЛ 1.
Роль медіа у суспільстві та вплив змін у соціумі на мову новин

1.1. Роль медіа в сучасному житті суспільства

У сучасному демократичному суспільстві гарантії стабільності та ефективного управління включають наявність розвинених, демократично організованих мас-медіа, які об'єктивно висвітлюють різні події та явища. Початковим завданням медіа є висвітлення актуальних подій та інформування населення, проте їх роль та вплив можуть змінюватися в залежності від типу та форми подання інформації.

Роль медіа в сучасному суспільстві уже досліджено у багатьох працях, зокрема працях Г.Почепцова [5], Д. Патрикаракоса [10], А.Досенко [15] та інших.

Згідно з теорією Джона Томпсона [12] про взаємодію мас-медіа та суспільства , він виділяє три види взаємодії, які залежать від рівня особистої участі та контролю:

- 1. Міжособистісна взаємодія: Включає особисті розмови та діалоги між людьми, такі як розмови на вечірках. В цьому виді взаємодії ключову роль відіграють індивідуальні фактори, такі як сприйняття, реакція та інтерпретація повідомлень.

- 2. Групова взаємодія: Це взаємодія, яка відбувається в групі людей, наприклад, в рамках дискусій, обговорень або спільного перегляду медійних вмістів. В цьому випадку вплив на взаємодію має не тільки індивід, але і соціальна динаміка в групі.

- 3. Масова взаємодія: Це взаємодія, яка відбувається через масові медіа, такі як телебачення, радіо, преса та Інтернет. Цей вид взаємодії має більший охоплення і може впливати на широку аудиторію. Однак, зауважується, що ця взаємодія не є однобічною, оскільки глядачі, читачі та слухачі активно сприймають інформацію, коментують її, розповідають далі, трансформують інтерпретацію і перерозподіляють її у своєму оточення.

Важливо відмітити, що Томпсон вважає, що мас-медіа не лише передають інформацію пасивним отримувачам, але створюють динамічну взаємодію, в якій глядачі, слухачі та читачі активно сприймають, інтерпретують і коментують повідомлення, впливаючи на свої навички, знання та досвід.

Сучасні медіа виконують багато різних функцій, таких як інформаційна, освітня, соціалізаційна та багато інших. Вони впливають на психологічний та соціальний стан людей, дозволяючи різним групам виражати свої думки та інтереси. Відповідно до концепції відомого канадського соціолога і культуролога Г.М.Маклюена ера мас-медіа і електронної інформації радикально змінює як життя людини, так і її саму [22].

Т. Г. Добросклонська, аналізуючи контекстне вживання словосполучення «медіамова», виокремлює три найбільш поширені визначення. По-перше, мова ЗМІ – це весь корпус текстів, які створюються і розповсюджуються засобами масової інформації. По-друге, – це стійка внутрішня мовна система, що характеризується певним набором лінгвостилістичних властивостей та ознак. По-третє, – це особлива знакова система змішаного типу з певним співвідношенням вербальних й аудіовізуальних компонентів, специфічних для кожного з засобів масової інформації: газет, радіо, телебачення, Інтернету [14].

На сьогоднішній день, електронні цифрові технології та Інтернет відіграють ключову роль у майбутньому комунікаційної сфери. Роль Інтернету в

сучасному суспільстві досліджується у працях Г.Почепцова [5], Д. Патрикаракоса [10], Р. Макнамі, [25] та інших.

Традиційні постачальники інформації, включаючи видавництва словників, енциклопедій та мас-медіа, визнають важливість адаптації до змін. Вони намагаються перейти на цифрову форму публікацій, оскільки кількість цифрового контенту, що публікується, значно перевищує кількість друкованих матеріалів.

Також електронні медіа надають можливість миттєвого доступу до інформації з будь-якого пристрою з Інтернет-підключенням. Вони дозволяють користувачам отримувати останні новини, розваги, спортивну інформацію, навчальний матеріал та багато іншого.

Важливим плюсом інтернет-видань являється те, що вони пропонують широкий спектр контенту, включаючи текстові статті, фотографії, відео, аудіо та інтерактивні елементи. Це дозволяє користувачам отримувати інформацію у форматі, який їм найбільше відповідає і сприймається краще.

Онлайн медіа надають можливість користувачам взаємодіяти з контентом та залучатися до дискусій. Вони можуть залишати коментарі, висловлювати свої думки та навіть створювати власний контент,що напряму впливає на залученість користувачів ,на відміну від традиційних видань.

Завдяки алгоритмам читач має можливість можливість персоналізувати вміст під власні потреби та інтереси. Користувачі можуть отримувати інформацію, яка найбільше їх цікавить, і уникати непотрібного контенту.

На відміну від паперових видань, електронні дозволяють залучати аудиторію з усього світу. Незалежно від вашого місця проживання, ви можете отримати доступ до контенту з будь-якої країни, дізнатися про світові події та спілкуватися з людьми з різних культур.

Електронні мас-медіа є важливим засобом комунікації в сучасному світі, де майже кожна людина має доступ до Інтернету. Вони перетворюють спосіб сприйняття та розповсюдження інформації, надаючи широкі можливості для спілкування, освіти, розваг та самовираження.

Facebook-одна з найпопулярніших соціальних мереж, яка на сьогоднішній день являється базою для обміну інформації (фото, відео, текстова комунікація) між двома або більше користувачами, є платформою, що використовується не лише для особистих цілей, а й для поширення актуальних новин на широкий загал.

1. Широке охоплення: Facebook має велику базу користувачів із мільярдами активних користувачів у всьому світі. Ця величезна аудиторія надає новинним організаціям і видавцям можливість охопити велику кількість людей і підвищити видимість їх вмісту.

2. Доступність і зручність: Facebook дозволяє користувачам отримувати доступ до статей новин, відео та іншого вмісту лише кількома клацаннями. Він забезпечує зручну платформу для користувачів, щоб бути в курсі останніх новин і подій, не відвідуючи кілька веб-сайтів або покладаючись виключно на традиційні медіаджерела.

3. Соціальний обмін та залучення: соціальний характер Facebook спонукає користувачів ділитися новинними статтями, коментувати публікації та брати участь в обговореннях. Це може сприяти поширенню інформації, заохочувати діалог і сприяти громадській активності, оскільки користувачі можуть висловлювати свою думку та взаємодіяти з іншими, хто має інші точки зору.

4. Персоналізована стрічка новин: алгоритмічна стрічка новин Facebook представляє користувачам вміст на основі їхніх інтересів, уподобань і попередніх взаємодій. Цей персоналізований підхід може допомогти людям знайти новини та теми, які відповідають їхнім конкретним інтересам, покращуючи загальний досвід споживання новин.

Однією з головних проблем, пов’язаних із Facebook як новинною платформою, є поширення дезінформації або «фейкових новин». Швидкий обмін і вірусність вмісту на платформі може призвести до поширення неточної або оманливої інформації, що може мати серйозні наслідки для окремих людей і суспільства.

Але Роджер Макнамі вважає, що: як соціальна мережа Facebook як соціальне явище завдає шкоди суспільству [25].

1.2. Лінгвістика українськомовних новинних текстів

Вивчення особливостей українськомовних текстів передвоєнного періоду є важливим завданням лінгвістичного дослідження, оскільки це допомагає розуміти та аналізувати мовні і стилістичні зміни, що відбуваються в контексті соціально-політичних, культурних і історичних подій. Дослідження такого типу може надати значну кількість інформації про розвиток та еволюцію української мови перед початком війни.

Тексти медійного дискурсу в Інтернеті широко вивчаються. Можна згадати праці О. Дзюбіної про сленгові неологізми соціальних мереж Twitter та Facebook (на матеріалі англійської мови) [13], Н. Коломієць про гіпертекст інтернет-новин (на матеріалі англійської мови) [20]. Про тексти, пов’язані з військовими конфліктами писав Д. Патрикаракос («Війна у 140 знаках. Як соціальні медіа змінюють військові конфлікти XXI століття») [10].

Одним з аспектів, який варто вивчити, є мовна ситуація та її зміни. Перед воєнними періодами можуть бути періоди стабільності і незмінності мовного середовища, але також можуть спостерігатися інтенсивні зміни під впливом соціально-політичних факторів.

Додатково до цього, вивчення особливостей українськомовних текстів включає аналіз мовної структури, лексики, граматики і стилістики. Це означає дослідження різних аспектів мови, таких як використання спеціальних слів, термінології, фразеологізмів, речень і організації тексту. Аналізуючи ці елементи, дослідники можуть виявити зміни у структурі мови та її використанні, що можуть бути пов'язані з політичними, соціальними і культурними факторами перед війною.

Також варто враховувати вплив зовнішніх змін у воєнний період. Під впливом політичних і культурних змін можуть поширюватися нові слова, фрази і вирази з інших мов. Дослідження таких запозичень та їх адаптації українською

мовою може вказувати на вплив інших культур на українську мову у воєнний період.

Вивчення особливостей українськомовних текстів у воєнний період також може включати аналіз мовленнєвої поведінки та практик. Це означає дослідження способів комунікації, використання мови в різних соціальних групах, виявлення стилістичних засобів, таких як образи, алегорії, іронія і т. д. Аналізуючи мовленнєву поведінку та практики, дослідники можуть зрозуміти специфіку комунікативної ситуації її вплив на мовлення.

Отже, вивчення особливостей українськомовних текстів у воєнний період включає аналіз мовної ситуації, мовної структури, використання лексики і граматики, вплив зовнішніх мов та мовленнєвої поведінки та практик. Це дає можливість отримати глибоке розуміння та об'єктивний аналіз розвитку української мови у певному відрізку часу , а також виявити зв'язок між мовою та соціально-політичними подіями, культурними змінами та історичним контекстом.

1.3. Дослідження впливу військових подій на мову та стилістику текстів

Дослідження впливу військових подій на мову та стилістику текстів є важливою складовою лінгвістичного аналізу українськомовних матеріалів під час воєнного періоду. Це дозволяє виявити зміни, які відбулися у способі сприйняття, вираження та використання мови під впливом військових подій, а також розкрити зв'язок між політичними, соціальними та культурними факторами і мовною діяльністю.

Одним з аспектів дослідження є зміна лексичного складу та вживання слів української мови. Воєнні події, соціальні зміни та нові реалії можуть спричинити появу нових термінів, фразеологізмів, військово-технічної лексики та інших специфічних слів. Дослідження лексичної сфери дозволяє виявити, які нові слова та вирази з'явилися в мовленні під впливом воєнних подій, а також як змінилося значення та вживання існуючих слів.

Стилістичні особливості текстів також є предметом дослідження при вивченні впливу військових подій на мову. Війна може впливати на тон, емоційну забарвленість, образність та інші аспекти стилістики текстів. Наприклад, зміна соціального контексту та переживань може призводити до зміни стилю мовлення, збільшення вживання метафор, агресивної лексики чи інших стилістичних засобів. Дослідження стилістичних відтінків українськомовних текстів допомагає розкрити специфіку мовлення воєнного періоду та встановити зв'язок між мовою, політичним контекстом та соціокультурними перетвореннями.

Враховуючи ці фактори, військовий час призводять до зміни мовленнєвого стилю та стилістики текстів, використання нових термінів та емоційно забарвлених висловів.

Застосування статистичних параметрів у дослідженні впливу військових подій на мову та стилістику текстів є важливим інструментом. Лексичне розмаїття та частота вживання слів є предметом аналізу для виявлення змін у лексиконі під час воєнного періоду. Через використання статистичних методів, таких як частотний аналіз та кластерний аналіз, можна визначити, які слова та фрази стали більш поширеними, а які втратили свою актуальність у текстах воєнного періоду. Також можна провести аналіз синтаксичних конструкцій, звертаючи увагу на зміни в структурі речень та організації текстів.

Важливим аспектом дослідження є також порівняння текстів до воєнного періоду та під час нього. Порівняння статистичних параметрів текстів з різних періодів дозволяє виявити та проаналізувати зміни, що відбулися в мовленні під впливом військових подій. Застосування статистичної параметризації допомагає зробити аналіз об'єктивним та обґрунтованим, надає можливість встановити закономірності та тренди в мовленні воєнного періоду.

Оцінка впливу повномасштабного вторгнення на українськомовні текстові матеріали є складним і багатогранним завданням, що передбачає аналіз та інтерпретацію широкого спектру текстів, створених під впливом воєнних подій. Цей процес включає в себе оцінку змін у мовному виразі, стилістичних

особливостях, лексичному складі, синтаксичних конструкціях, а також вивчення контекстуальних факторів та соціокультурних аспектів, що впливають на мовлення.

Дослідження впливу повномасштабного вторгнення на українськомовні текстові матеріали вимагає комплексного підходу, який включає аналіз мовних, соціокультурних, історичних та політичних аспектів. Цей процес дозволяє краще зрозуміти зміни у лінгвістичних особливостях, збагатити нашу знання про вплив воєнних подій на мову та стилістику текстів, а також сприяє розвитку наукових досліджень у галузі лінгвістики та соціолінгвістики.

Висновки до розділу 1

Вплив отримання інформації шляхом інтернет видань є досить неоднозначний. З одного боку, він може приносити інновації та комфорт. З іншого боку, він може призводити до спрощення та порушення мовних норм.

Дослідження особливостей українськомовних текстів в певний часовий відрізок передбачає собою аналіз різних аспектів мови. Це включає вивчення мовної ситуації на той час, аналіз мовної структури, аналіз використання лексики та граматики, вплив зовнішніх мов на українську мову, а також вивчення мовленнєвої поведінки та практик. Цей метод дає змогу отримати глибоке розуміння та об'єктивний аналіз етапу розвитку української мови, а також розкрити наслідковий зв'язок між мовою, соціально-політичними подіями та історичним контекстом того часу.

Також дослідження з лінгвістичного аналізу текстів під час воєнного періоду дають цінні знання про мовні особливості, зміни та комунікативні виміри цього періоду. Ці дослідження допомагають розуміти вплив воєнних подій на мовну систему та комунікацію, а також розвивають методи та підходи для більш глибокого аналізу воєнного дискурсу.

РОЗДІЛ 2. Створення корпусу медіатекстів та частотного словника

2.1. Метод створення текстового корпусу у проведенні лінгвістичного аналізу.

Від правильного вибору корпусу залежить якість та достовірність результатів дослідження. При виборі текстового корпусу для аналізу українськомовних текстів під час воєнного періоду необхідно враховувати кілька ключових аспектів. Важливо визначити період часу, який охоплює дослідження. Воєнний період може включати різні етапи, наприклад, початок війни, період активних бойових дій або післявоєнний час.

У проведенні дослідження ми спираємося на праці Н.Дарчук [7; 8; 9], В.Перебийніс [29; 30] та інших.

Необхідно враховувати обсяг та представленість текстів у корпусі. Відповідно до обсягу дослідження і доступних ресурсів можна визначити, скільки текстів включити у корпус. Крім того, необхідно звернути увагу на якість та доступність текстів у корпусі. Тексти повинні бути чіткі, читабельні та можливі для аналізу.

Доступність текстів також має важливе значення, оскільки це впливає на можливість отримати більше даних для аналізу. Після уважного врахування цих аспектів можна обрати відповідний текстовий корпус для лінгвістичного аналізу українськомовних текстів під час воєнного періоду. Вибраний корпус дозволить зосередитися на конкретних лінгвістичних явищах, стилістичних особливостях та впливі воєнних подій на мовлення. Це дозволить здійснити глибокий та об'єктивний аналіз відповідно до поставлених дослідницьких цілей.

В процесі вибору текстового корпусу для аналізу українськомовних текстів під час воєнного періоду, розглянуто такі аспекти.

Масштаб корпусу: визначення обсягу текстових файлів, який можна включити до дослідження. Це можуть бути окремі документи, повні текстові колекції або великі корпуси, залежно від обсягу та мети дослідження.

Вибраний текстовий корпус повинен відповідати поставленим дослідницьким питанням.

Важливо мати структуровану систему збереження та організації текстів в корпусі. Це може включати використання спеціальних програм або створення власної бази даних, яка дозволяє зберігати текстові дані та додаткову інформацію про них. Завершальний етап - лінгвістичний аналіз текстів в корпусі. Науковець має можливість провести дослідження та аналіз лексики , синтаксису та стилістичних особливостей.

У статистичній параметризації українськомовних текстів робота "Комп’ютерне анотування українського тексту: результати і перспективи" Наталії Петрівни Дарчук [7] є надзвичайно важливою та є великим внеском у галузь комп'ютерної лінгвістики, яка досліджує методи та результати комп'ютерного анотування українського тексту. Робота описує методи інформаційної обробки тексту, зазвичай використовувані українською мовою, такі як морфосинтаксичний аналіз, лематизація, тематичне моделювання та інші. Результати застосування цих методів на прикладі українських текстів також розглядаються в монографії.

У роботі також надається огляд потенційного розширення комп'ютерного анотування українського тексту та його застосування в галузях, таких як машинний переклад, автоматична обробка мовленнєвої інформації та інші. Завдяки цим перспективам, робота може бути корисною як для дослідників у галузі комп'ютерної лінгвістики, так і для фахівців, що працюють з українською мовою і зацікавлені у використанні комп'ютерних методів для обробки тексту.

Сучасне мовознавство все більше використовує математичні методи і комп'ютерні технології для переходу від описових до аналітичних методів досліджень. Сучасна філологічна освіта неможлива без опору на обчислювальну лінгвістику, яка допомагає осмислити наукові результати і

провести лінгвістичні експерименти з використанням числових методів, які запропоновані Б. Л. Ван дер Варденом.

Один з найпоширеніших і доступних кількісних методів аналізу тексту - це статистичний аналіз, який полягає у підрахунку кількості вживань окремих слів у заданому тексті. Цей метод широко застосовується для різних цілей, зокрема:

1. Статистична стилістика: Математично точне розрізнення літературних стилів і жанрів шляхом вивчення вживання окремих слів у тексті.
2. Атрибуція тексту: Встановлення авторства анонімних або підроблених текстів шляхом порівняння їх статистичних характеристик з відомими авторами.
3. Аналіз мовних одиниць: Опис поведінки різних мовних одиниць (букв, морфем, слів) у тексті, включаючи їх розподіл, сполучуваність та частоту вживання.
4. Вимірювання інформативності текстів: Визначення кількості інформації, яка міститься в тексті та його складових частинах.
5. Відновлення текстів та мов: Використання статистичного аналізу для відновлення текстів або мов за їх фрагментами.
6. Визначення рівня спорідненості мов і швидкості мовних змін: Дослідження статистичних характеристик текстів для визначення рівня спорідненості мов, а також для оцінки швидкості мовних змін і часу поділу різних мов.

Статистична параметризація є підходом і набором методів, які використовуються для аналізу лінгвістичних даних. Цей підхід ґрунтується на використанні статистичних показників, що дозволяють описати, категоризувати та аналізувати різні мовні явища. Статистична параметризація застосовується в лінгвістичному аналізі для отримання кількісних показників, які допомагають краще розуміти та описувати різні аспекти мови.

Один з важливих аспектів статистичної параметризації в лінгвістичному аналізі полягає в аналізі текстових корпусів. Корпус представляє собою велику колекцію текстів, і він служить основою для досліджень мовних

явищ. Застосування статистичних методів дозволяє виявляти регулярності, закономірності та залежності в мовленні шляхом аналізу значних обсягів даних. Наприклад, статистична параметризація дозволяє визначити частоту вживання певних слів, лексичних одиниць, конструкцій або мовних засобів у тексті. Крім того, вона дозволяє порівняти різні мовні варіанти та здійснити аналіз мовних особливостей.

Статистична параметризація також використовується для виявлення стилістичних та семантичних особливостей тексту. За допомогою статистичних методів можна виявити ключові слова, тематичні групи слів, тенденції в мовленні, аналізувати текстові структури та розкривати смислові зв'язки між різними частинами тексту. Це дозволяє зрозуміти контекст та значення, що несе текст, і виявити особливості комунікації.

Використання статистичної параметризації у лінгвістичному аналізі дозволяє проводити порівняльні дослідження різних текстів та мов. За допомогою статистичних методів можна порівнювати та аналізувати характеристики мови, стилістичні особливості і інші мовні параметри. Це дозволяє виявити схожості та відмінності між мовами та текстами, а також встановити вплив історичних, соціокультурних та інших факторів на мовні вияви.

2.2. Основні підходи статистичної параметризації

Статистична параметризація відіграє важливу роль у лінгвістичному аналізі текстів, надаючи можливість числово описувати та аналізувати різні аспекти мовних даних. Цей підхід використовує статистичні методи для оцінки та опису різних характеристик тексту, таких як лексичні, синтаксичні, семантичні та стилістичні особливості.

Один із ключових методів статистичної параметризації – це лексичний аналіз, який включає підрахунок та аналіз розподілу лексичних одиниць у тексті. Це можуть бути окремі слова, фрази або навіть більші мовні одиниці. Лексичний аналіз допомагає встановити частоту вживання конкретних слів, їх

розподіл за частинами мови, а також виявити ключові слова та терміни, що характеризують текст.

Інший метод – синтаксичний аналіз, який досліджує структуру речень та взаємозв'язки між словами у тексті. З використанням статистичних методів можна аналізувати частоту вживання різних синтаксичних конструкцій, таких як підрядні речення, прикладні конструкції або звороти. Синтаксичний аналіз допомагає розуміти синтаксичну структуру тексту та виявляти особливості, що можуть вказувати на специфіку мовлення або стиль автора.

Третій метод – семантичний аналіз, який фокусується на значенні та взаємозв'язках між словами та фразами у тексті. За допомогою статистичних показників можна аналізувати семантичну подібність слів, відношення антонімів та синонімів, а також визначати контекстуальні значення. Семантичний аналіз допомагає розуміти смислову структуру тексту та виявляти зв'язки між різними мовними одиницями.

Крім того, при статистичній параметризації застосовуються методи стилістичного аналізу, що вивчають особливості авторського стилю та стилістичних прийомів у тексті. Це можуть бути аналіз вживання метафор, епітетів, граматичних конструкцій або оцінних засобів. Стилістичний аналіз допомагає виявляти авторську індивідуальність, впливи певних літературних шкіл або стилістичні особливості певного періоду.

Усі ці підходи та методи статистичної параметризації дозволяють нам ретельно аналізувати українськомовні тексти під час воєнного періоду. Вони допомагають виявляти мовні тенденції, зміни у стилі та семантиці, комунікативні стратегії та інші особливості, які свідчать про вплив воєнних подій на мовлення та мовну культуру. Статистична параметризація відкриває шлях до глибшого розуміння лінгвістичних аспектів воєнного періоду і сприяє розвитку лінгвістичних досліджень у цій сфері.

Методи збору та обробки даних для статистичної параметризації українськомовних текстів включають ряд процедур та підходів, які допомагають отримати об'єктивні та надійні результати аналізу. Важливим етапом у процесі статистичної параметризації є збір та обробка даних, які дозволяють виявити та

описати лінгвістичні характеристики текстів у воєнний період. Нижче описані деякі методи, що використовуються в цьому контексті:

- Збір матеріалу: текстові дані можна збирати з різних джерел, таких як архіви, бібліотеки, електронні бази даних, веб-сторінки тощо. Для воєнного періоду можуть бути доступні офіційні документи, листування солдатів, пресові публікації, мемуари та інші джерела. Важливо, щоб корпус текстів був репрезентативним і мав різноманітність.
- При аналізі текстів можна вивчати мовленнєві одиниці, такі як слова, речення, фрази. Для вибору конкретних одиниць можна використовувати лінгвістичні категорії або морфологічні ознаки. Наприклад, можна досліджувати вживання певних слів, їх частотність, стилістичні особливості тощо.
- Створення колекції текстових даних дозволяє зібрати різні документи та джерела для подальшого аналізу. Корпус текстових даних можна створити за допомогою спеціальних програм або скриптів.
- Лексико-семантичний аналіз виявляє лексичні та семантичні особливості текстів. Для цього використовуються лексико-семантичні ресурси, такі як словники, тезауруси, енциклопедії, які допомагають визначити значення слів та їх взаємозв'язки.
- Статистичний аналіз дозволяє виявити закономірності та тенденції у текстах. Використовуються різні методи, наприклад, частотний аналіз, аналіз варіацій, класифікація, щоб отримати кількісні дані та визначити статистичні характеристики текстів.
- Машинна обробка даних допомагає автоматизувати процеси збору та аналізу текстових даних. За допомогою комп'ютерних програм та алгоритмів можна швидше і ефективніше обробляти великі обсяги текстової інформації і отримувати результати візуалізації та статистичного аналізу.
- Анотація: Лінгвістична анотація полягає у встановленні лінгвістичних ознак або міток для окремих текстових одиниць. Цей метод дозволяє відмітити морфологічні, синтаксичні або семантичні характеристики

тексту. Лінгвістична анотація допомагає стандартизувати та систематизувати дані для подальшого аналізу.

Аналіз статистичних параметрів, таких як різноманітність лексики, є надзвичайно важливою частиною лінгвістичного дослідження, що дозволяє вченим отримати об'єктивну інформацію про тексти та їх особливості. В контексті українськомовних текстів у воєнний час такий аналіз статистичних параметрів, включаючи варіативність вживання слів, частоту вживання слів, синтаксичні конструкції тощо, дозволяє розуміти мовлення того часу, виявляти його особливості та робити висновки щодо соціально-політичного контексту.

Одним з важливих аспектів аналізу статистичних параметрів - це різноманітність лексики. Воно відображає різноманітність вживання слів та їх варіативність у текстах. Аналізуючи різноманітність лексики, вчені мають змогу виявити основні тематичні групи слів, вживання специфічної термінології, наявність запозичених слів та термінів з інших мов та зміни у словниковому складі в період повномасштабного вторгнення.

Частота вживання слів є ще одним невід’ємним параметром, який допомагає виявити найрозповсюдженіші та найважливіші слова в українськомовних медіатекстах. Аналізуючи частоту вживання слів, можна виявити ключові концепти, поняття та теми, які переважали у соціальному житті того часу. Також ці дані допомагають виявити еволюцію вживання слів в залежності від соціальних та політичних змін.

Синтаксичні конструкції відображають способи організації речень та тексту. Аналізуючи синтаксичні параметри, філологи можуть виявити особливості синтаксичної структури українськомовних текстів перед воєнним періодом та в період воєнних дій. Наприклад, вони можуть з'ясувати, як змінилася вживання різних типів речень (простих, складних, складнопідрядних) або які синтаксичні засоби використовувалися для наголошення певних ідей чи посилення емоційного забарвлення текстів.

Для аналізу статистичних параметрів використовуються різні методи, такі як частотний аналіз, лексико-семантичний аналіз, стилістичний аналіз, синтаксичний аналіз та інші. Кожен з вище перерахованих методів має свої

особливості та переваги. Наприклад, частотний аналіз дозволяє виявити найпоширеніші слова та визначити їх вагомість у текстах, а лексико-семантичний аналіз допомагає розкрити значення та вживання слів у відповідному контексті.

2.3. Використання статистичної параметризації для виявлення змін у лінгвістичних особливостях

Використання статистичної параметризації є ефективним підходом для виявлення змін у лінгвістичних особливостях мовлення, що відбуваються під впливом різних факторів, у тому числі під час воєнного періоду. Цей підхід дозволяє об'єктивно та систематично аналізувати тексти з метою виявлення змін у використанні лексики, синтаксичних конструкцій, стилістичних засобів та інших мовних характеристик.

Один з основних методів статистичної параметризації - аналіз лексичного розмаїття, передбачає вивчення різноманітності та розподілу слів у текстах. Застосування статистичних методів, які включають лексичне розмаїття, типологічні міри та інші, дає змогу виявити зміни в уживанні слів, збагачення або зменшення лексичного запасу, а також особливості вживання нових слів, термінології та специфічної лексики, що характеризують воєнний період.

Крім цього, статистична параметризація може бути застосована для вивчення частоти вживання слів. Аналіз частоти вживання слів дозволяє виявити тенденції та зміни в мовленні, а також визначити ключові слова та тематичні акценти в текстах воєнного періоду. Частотний аналіз може розкрити емоційну насиченість, специфічну термінологію та інші особливості, які впливають на мовлення під час воєнного конфлікту.

Важливою складовою статистичної параметризації є аналіз стилістичних особливостей текстів. Це включає дослідження використання стилістичних засобів, риторичних фігур, метафор, алегорій та інших мовних засобів, що характеризують стиль та виразність мовлення. Виявлення змін у стилістичних особливостях текстів під час воєнного періоду може свідчити про

зміну емоційної тональності, вплив політичних чинників, стилістичну адаптацію до воєнного контексту та інші аспекти.

Застосування статистичної параметризації для виявлення змін у лінгвістичних особливостях українськомовних текстів перед воєнним періодом є важливим інструментом для дослідження впливу військових конфліктів на мову та стиль мовлення. Цей аналітичний підхід дозволяє виявити та проаналізувати мовні зміни, які сталися у зв'язку з військовими подіями, і сприяє глибшому розумінню мовної динаміки та соціокультурного контексту воєнного періоду.

Корпус текстів - це структурована, систематизована та програмно оброблена колекція різних видів текстів природної мови, що представляє різні варіанти і форми існування мови. Використання корпусного підходу передбачає використання комп'ютера та корпусних методик для аналізу та опису текстового матеріалу. Одним з головних переваг корпусного підходу є можливість отримання широкого спектру інформації швидко та безперешкодно, під час введення тексту в комп'ютер та обробки його як корпусу. Сьогодні дані корпусів масштабно використовуються в лексикографії, стилістиці, судовій лінгвістиці, лінгвістичній варіантології, перекладознавстві, соціолінгвістиці, методиці навчання і вивчення іноземної мови та в багатьох інших лінгвістичних дослідженнях [17].

Цей підхід дозволяє отримати різноманітну інформацію, таку як частотні словники, конкордансі, колокації, статистику вживання слів, пошук специфічних лінгвістичних чи семантичних залежностей, порівняльний аналіз мовних явищ та багато іншого. Використання корпусного підходу допомагає вивчати мову більш об'єктивно, аналізувати її характеристики та вдосконалювати алгоритми та моделі, які використовуються у галузі обробки природної мови.

Корпусний аналіз має кілька характерних ознак. Перш за все, це емпіричний підхід до аналізу мовних даних, де досліджуються реальні моделі мовної реалізації в природних текстах. Ключовою основою для аналізу є великі,

структуровані колекції природних текстів (корпуси). Для проведення досліджень залучаються комп'ютерні технології, що дозволяє ефективно обробляти та аналізувати лінгвістичний матеріал. Корпусний аналіз використовує як кваліативні, так і квантитативні аналітичні методики, пріоритет надається останнім, зокрема вивченню частоти вживання лінгвістичних одиниць та статистичним дослідженням сполучуваності. Результати корпусного аналізу не тільки допомагають зробити нові висновки про мову, але і визначають нові напрями досліджень, які раніше не привертали уваги дослідників. Корпуси текстів можуть містити тексти усного і писемного мовлення або бути змішаного типу.

Корпус текстів - це колекція текстових документів, яка використовується для досліджень у сфері обробки природної мови (Natural Language Processing, NLP) та інших суміжних галузях. В корпусі текстів можуть бути представлені різні типи документів, такі як статті, книги, веб-сторінки, соціальні медіа повідомлення і т. д.

Особливості корпусу текстів:

- Розмір: Корпус текстів може бути дуже різних розмірів - від невеликих збірок кількох документів до великих сукупностей з мільйонами документів.
- Тематика: Корпус текстів може бути сфокусованим на певну тематику або бути різноманітним з різними темами. Наприклад, корпус текстів може містити тільки спортивні статті, новини про політику або текстові дані з різних галузей.
- Мова: Корпус текстів може бути написаний на різних мовах. Корпуси текстів для англійської мови є найпоширенішими, але існують також корпуси для багатьох інших мов, включаючи українську.
- Формат: Корпуси текстів можуть бути представлені у різних форматах, таких як текстові файли, PDF, docx, бази даних або спеціалізовані формати для роботи зі збірками документів, такі як JSON або XML.
- Анотації: Деякі корпуси текстів можуть містити додаткову інформацію або анотації про тексти, такі як мітки класифікації, наголоси, посилання між текстами, переклади або іншу лінгвістичну інформацію. Це допомагає

покращити виконання алгоритмів обробки природної мови та додаткову обробку текстів.

- Доступність: Деякі корпуси текстів доступні для загального використання і можуть бути безкоштовними, тоді як інші можуть вимагати платну підписку або ліцензію.
- Використання: Корпус текстів використовується для різних цілей, таких як тренування моделей машинного навчання, валідація алгоритмів NLP, вивчення мови, аналіз соціальних мереж, інформаційний пошук і багато іншого.

На порталі Mova.info23 (Інститут філології Київського університету імені Тараса Шевченка) доступний Дослідницький корпус сучасної української мови. Основною метою корпусу є забезпечення інформаційно-довідкової системи, яка дозволяє відповідати на різні питання, пов'язані з українською мовою.

Корпус текстів є важливим інструментом для розробки та оцінки різних алгоритмів обробки природної мови. Вибір та підготовка корпусу текстів відіграють важливу роль у досягненні точності і ефективності роботи алгоритмів, що використовують ці дані.

2.4. Створення корпусу текстів воєнного та передвоєнного періоду на матеріалах новин у Facebook

Для подальшого створення частотного словника нам потрібно було створити два корпуси текстів. В один з яких ми помістили медіатексти зібрані за період квітень-травень 2021 року. В інший ми занесли матеріали за період повномасштабного вторгнення ,а саме квітень-травень 2023 року.

Вибір джерел:

Для даного дослідження було обрану соціальну мережу Фейсбук ,а саме медіатексти з офіційних сторінок українськомовних ЗМІ.(«СТБ», «1+1» та «Радіо Свобода»). Слід зазначити, що новинні тексти - це тексти, які дуже

залежать від поточних подій у країні та суспільстві, власне їх темою і є суспільне життя народу. Тому початок широкомасштабного вторгнення в Україну став причиною кардинальної зміни у характері новин.

Основою корпусів являються вручну оброблені тексти. Загалом було зібрано -100 текстів, кожен з яких є окремим дописом у мережі Facebook. А саме: 50 новинних текстів, що були опубліковані в період до початку повномасштабного вторгнення, та 50 постів в період після початку російського вторгнення на територію України. Також слід зазначити , що після збору матеріалу , ми вручну очищували тексти від непотрібних елементів , а саме видаляли емоджі, хештеги та HTML-теги. Зібрані матеріали були занесені до двох таблиць Excel: в одній було зібрано тексти до повномасштабного вторгнення, в іншій - тексти, опубліковані після початку повномасштабного вторгнення. Таким чином, ми отримали матеріал корпусу, на якому базується подальше дослідження.

Структура корпусу. У першому стовпчику вказано порядковий номер тексту, далі ми бачимо сам текст. Також в корпус ми внесли покликання на пост з соціальної мережі Facebook , дату його публікації та кількість слів у даному тексті. В шостому стовпчику ми вказали джерело публікації , а саме з яких із вищеперерахованих ЗМІ було взято новинний текст. Нижче прикріплене зображення фрагментів корпусу.

ДОВОЄННІ ТЕКСТИ					
1	"У «Слузі народу» прокоментували заяву руху «Чесно» про те, що Микола Тищенко «вимагав журналістські посвідчення, перевіряв на справжність, вимагав паспорти, проводив обшук людей з метою пошуку зброї» в рамках спостереження за підрахунком голосів на 87-му округзі. У «Чесно» додали, що він також «звинувачував Олександра Шевченка в сепаратизмі, вимагав вибачень і встати перед ним на коліна".	https://l.facebook.com/l.p	Дата публікації- 01.04.2021	55 слів	Джерело- сайт" Радіо Свобода"
2	"Тих, хто критикував соціалістичний лад, вирізняється свободою мислення, або хотів утекти з СРСР визнавали «неосудними» і піддавали примусовому психіатричному лікуванню. Дисиденти не так страшилися ув'язнення в тюрмі чи ГУЛАГу, як радянської каральної психіатрії, працівники якої часто говорили «мы тебя научим Родину любить".	https://l.facebook.com/l.p	Дата публікації- 24.04.2021	42 слова	Джерело- сайт "Радіо Свобода"
3	"Перед та післяпологову депресію переживає кожна друга українська жінка. Утім, більшість воліє мовчати про свій стан. Чим це небезпечно для мами та немовляти - знає журналістка Євгенія Цветанська. Вона сама пережила депресію і наважилася показати надто особисті кадри, аби зарадити іншим. Де шукати порятунку - особистий досвід у спецпроєкті "Сама, сама".	https://l.facebook.com/l.p	Дата публікації- 30.04.2021	51 слово	Джерело- сайт телеканалу "ТСН"

Фрагмент корпусу текстів до повномасштабного вторгнення

За таким самим принципом було укладено корпус новинних текстів за період квітень-травень 2023 року. Слід зазначити ,що тексти зібрані саме за період російського вторгнення на територію України були значно більшими , ніж тексти за 2021 рік. Тому процес збору довоєнного матеріалу був важчим та більш кропітким. Також на сторінках деяких ЗМІ в Facebook була відсутня можливість фільтрації дописів за датою публікації , що теж ускладнило нам роботу над корпусом.

	ВОЄННІ ТЕКСТИ					
1	"18 травня – День пам'яті жертв геноциду кримськотатарського народу У травні 1944 року радянська влада депортувала з Криму 183 155 кримчан. Більшість з них — це жінки та діти. Протягом 12 років кримські татари мали статус «спецпереселенців». Переважна частина депортованих була направлена на спецпоселення до Узбекистану, частина – до ГУЛАГу, а ще частина – для поповнення спецконтингенту для московського вугільного басейну. За офіційними даними, майже половина кримських татар загинула від голоду та хвороб у дорозі чи на засланні. Вічна пам'ять кожному, хто став жертвою злочинних дій кремля. 18 травня – День пам'яті жертв геноциду кримськотатарського народу У травні 1944 року радянська влада депортувала з Криму 183 155 кримчан. Більшість з них — це жінки та діти. Протягом 12 років кримські татари мали статус «спецпереселенців». Переважна частина депортованих була направлена на спецпоселення до Узбекистану, частина – до ГУЛАГу, а ще частина – для поповнення спецконтингенту для московського вугільного басейну. За офіційними даними, майже половина кримських татар загинула від голоду та хвороб у дорозі чи на засланні. Вічна пам'ять кожному, хто став жертвою злочинних дій кремля."	https://www.facebook.com/1plus1.ua/posts/pfbid02rAeCofa7wR7x1EHoH3KRGVgMGAZErnPTj6ur3cBSnjCxEEyMHCsHyf88jVh8TZLWl?__cft__[0]=AZV2axVt4MJhiqCvTkeWB-_OHOTfOFwOetub09dtQIwV7TRudZbKUH_4kbeUjhFC4sNNcDCGPg7sDkS4ovOAUS0kIjK7-uL1JccgzIdf8chFqxOSokQuQxT1wieTrsaL-GX5plsLUVVpfKLIAwWLbdxGNqn7zDxD5SX4gYfGN0jdwWvOp_h5QYtw4zxHFhMsIRkz3LIRMB17Lsrork2PFAnzDZ7HuuCWyxUr6tsjYUd0w&__tn__=%2C0%2CP-R	Дата публікації-18.05.2023	83 слова	Джерело-фейсбук сторінка телеканалу "1+1"	
2	"Сьогодні в Україні відзначають День вишиванки Українська вишиванка – не просто святковий одяг. Споконвіку сорочка, вишита дружиною або мамою, вважалася потужним оберегом від нещасть: кожен стібок зберігав у собі любов і тепло душі. У кожному регіоні України є свої традиції вишивання і символізм орнаментів, але незмінно лише одне: вишиванка — це частина генетичного коду українця. Сьогодні День вишиванки має особливий сенс — українці змушені відстоювати свою національну автентичність і боротися за неї.	https://www.facebook.com/1plus1.ua/posts/pfbid079JSa8Xq48RFc8MNtRHWju79PegiSdiicWMvFaZV2D2TBkeE71H7bopGisyPJbKNl?__cft__[0]=AZWkmoFgRWnB9DX0f9ysvVlsGXJr9gc3F9RvilOMc-mB-6dEfNt6cqjS_nAUDH896MZ	Дата публікації-18.05.2023	100 слів	Джерело-фейсбук сторінка	

Фрагмент корпусу текстів, розміщених після початку повномасштабного вторгнення

Повністю корпус зібраних текстів наводиться у Додатку 1 до цієї роботи.

2.5. Частотний словник. Алгоритм роботи з частотним словником.

У цьому розділі слід згадати працю "Частотний словник сучасної української художньої прози: у 2-х т." видається в 1981 році під редакцією Перебийніса В.І. і присвячена аналізу словникового матеріалу сучасної української художньої прози. Головна тема роботи полягає у створенні частотного словника, який відображає вживання слів у сучасній українській художній прозі. Частотний словник складається зі списків слів, які були зібрані та класифіковані залежно від частотності їх використання у текстах прози. Два томи роботи містять ці списки слів разом з їх статистикою використання.Результати цього дослідження мають велике значення для мовознавців, лінгвістів та дослідників, які цікавляться українською літературою та словниковим матеріалом. Створення частотного словника дозволяє виявляти популярність та тенденції використання слів у сучасній українській художній прозі.

За загальноприйнятою класифікацією словники на лінгвістичні та нелінгвістичні. Лінгвістичні словники, у свою чергу, поділяються на текстоорієнтовані та системоорієнтовані. Одним із видів текстоорієнтованих словників є частотні словники, які відображають частоту вживання мовних одиниць у тексті.

Процес створення таких словників є важливим завданням комп'ютерної лексикографії. Він полягає у складанні списків мовних одиниць, що зустрічаються у певному наборі текстів, разом із відповідними частотами їх вживання. Частотні підрахунки часто проводяться для словоформ і лексем.

Частотні словники є інструментом аналізу мовлення, який дозволяє визначити частоту вживання слів у тексті чи колекції текстів. Вони базуються на підрахунку кількості входжень кожного слова у тексті і визначенні його рангу за частотою вживання. Ці словники можуть бути складені для конкретних мов, жанрів, періодів чи тематик.

Частотні словники дозволяють виявити ключові слова, що найчастіше зустрічаються у тексті, і тим самим ідентифікувати тематичні акценти. Вони також можуть допомогти виявити тенденції та зміни у вживанні слів, а також виявити слова, що ймовірно використовуються у специфічних контекстах, наприклад, воєнному періоді. Частотні словники можуть бути використані для порівняння текстів, виявлення стилістичних особливостей та дослідження лінгвістичних характеристик.

Наприклад, частотний словник може показати, що певні терміни або сленгові вирази стають більш поширеними у мовленні під час воєнного періоду. Також, використання частотного словника може допомогти ідентифікувати особливості мовлення у текстах з різних історичних або соціокультурних контекстів.

Створення частотного словника включає наступні етапи:

1. Збір даних: На цьому етапі необхідно мати доступ до текстового матеріалу, на основі якого буде будуватися словник. Тексти можуть бути зібрані з різних джерел, таких як книги, статті, інтернет-ресурси тощо.
2. Токенізація: Тексти необхідно розбити на окремі слова або словоформи. Цей процес називається токенізацією. В результаті отримується список мовних одиниць, з яких складатиметься словник.
3. Лематизація (опціонально): Для зменшення кількості дублікатів може застосовуватися лематизація, тобто приведення всіх словоформ до базової лексеми. Наприклад, слова "біжити", "біжить", "біжимо" будуть зведені до однієї лексеми "бігти".
4. Розрахунок частот: Для кожної мовної одиниці в словнику обчислюється частота її вживання у тексті. кожна виявлена словоформа додається до словника, і перевіряється, чи вона вже присутня в ньому. Якщо так, то збільшується абсолютна частота цієї словоформи в словнику. Частота може виражатися у відносних або абсолютних числах. Відносна частота відображає відношення вживання даної мовної одиниці до загальної кількості слів у тексті.
5. Упорядкування: Після обчислення частот, словник може бути упорядкованим за алфавітом або за зростанням абсолютної частоти. Це дозволяє легко здійснювати пошук і отримувати більш значущу інформацію про словникові елементи. Якщо словник будується для повного тексту, то процес перевірки і додавання повторюваних елементів відбувається до тих пір, поки не закінчиться вхідний текст.
6. Редагування і форматування: На останньому етапі слід перевірити, чи не містить словник помилок і непотрібної інформації. Також може знадобитися форматування словника для його зручного використання.

Портал mova.info (Інститут філології Київського університету імені Тараса Шевченка) надає доступ до 6 частотних словників, які покривають різні напрями, такі як науковий, публіцистичний та художній стиль. Ці частотні словники є дуже корисним інструментом для вивчення різних аспектів

лексики. Вони дозволяють досліджувати подібність та розподіл слів у різних сферах вживання. Вивчення цих словників надає цікаві відомості про якісну розмежування слів залежно від їхньої сфери вживання.

Отже, створення частотного словника включає підготовчі етапи (збір даних, токенізація, лематизація) і основні етапи (розрахунок частот, упорядкування, редагування і форматування). Такий алгоритм дозволяє створити частотний словник, в якому зберігається інформація про вживання різних мовних одиниць у тексті. Це може бути корисно для вивчення мови, аналізу стилістичних особливостей тексту, а також для інших лексикографічних досліджень.

2.6. Створення частотних словників словоформ.(програмне забезпечення)

Частотний словник словоформ зібраних нами текстів було створено автоматично, за допомогою написаного нами програмного забезпечення. Використовувалися мова програмування Python та інтегроване середовище розробки PyCharm. Програма реалізовує створення частотного словника на основі декількох текстових файлів формату docx, за допомогою бібліотеки Spacy для обробки мовних даних.

Для розробки коду та обробки тексту були використані різні програмні функціонали, такі як бібліотеки, модулі та функції:

1. **Python** – основна мова програмування, на якій написаний цей код.
2. **Spacy** – це високоефективна та швидка бібліотека для обробки природної мови, написана на Python і Cython. Вона використовується для різних завдань NLP, таких як розбиття тексту на слова (токенізація), визначення частин мови (POS-тегування), визначення іменованих сутностей (NER), тощо. В даному коді Spacy використовується для попередньої обробки тексту та отримання лем слів.

3. **sqlite3** – модуль Python для роботи з базами даних SQLite. SQLite - це вбудована в додаток СУБД, яка не вимагає окремого сервера. В даному коді sqlite3 використовується для створення бази даних, таблиць в базі даних, і вставки даних в таблиці.

4. **docx** – модуль Python для читання, запису і створення документів Word (.docx). В даному коді він використовується для читання тексту з файлів Word.

5. **collections.defaultdict** – це клас вбудованої бібліотеки Python **collections**, який використовується для створення словника зі значеннями за замовчуванням. В даному коді він використовується для створення частотного словника слів.

6. **os** – це вбудований модуль Python, який використовується для взаємодії з операційною системою.

7. **enum** – це вбудований модуль Python для створення перелічуваних типів. В даному коді він використовується для створення типу **TableType**, який представляє імена таблиць в базі даних.

Код та опис процесів:

2. Код для main.py:

```
import os
import spacy

from typing import List, Tuple, Dict
from collections import defaultdict
from docx import Document

from database_utils import create_db, fill_db, TableType

# Завантаження моделі мови spacy для української мови
nlp = spacy.load("uk_core_news_lg")

# Список шляхів до файлів текстів та типу таблиць, в які потрібно занести дані
TEXT_PATH: List[Tuple[str, TableType]] = [("war_texts_part1.docx",
TableType.DURING_WAR),
                                          ("war_texts_part2.docx",
TableType.BEFORE_WAR)]


# Функція для читання даних з файлу формату docx
def read_docx_file(file_path: str) -> List[str]:
    print(f"Reading document: {file_path}")
```

```
    document: Document = Document(file_path)
    text: List[str] = [paragraph.text for paragraph in document.paragraphs]
    return text


# Функція для попередньої обробки текстових даних з використанням моделі мови spacy
def preprocess_text_data(text_data: List[str]):
    print("Preprocessing document")
    return nlp.pipe(text_data)


# Функція для створення частотних словників
def get_word_frequency_dict(preprocessed_data) -> Dict[str, int]:
    print("Creating frequency dict...")
    word_freq = defaultdict(int)

    for doc in preprocessed_data:
        for token in doc:
            if token.is_alpha and not token.is_oov:
                word_freq[token.text] += 1

    return word_freq


def main():

    # Створення бази даних
    create_db()

    # Обробка кожного тексту
    for text_path, table_name in TEXT_PATH:

        # Формування шляху до файлу
        text_path: str = os.path.join(os.getcwd(), "texts", text_path)

        # Читання даних з файлу
        doc_data: List[str] = read_docx_file(text_path)

        # Попередня обробка текстових даних
        pp_data = preprocess_text_data(doc_data)

        # Отримання частотного словника слів
        word_freq: Dict[str, int] = get_word_frequency_dict(pp_data)

        # Заповнення бази даних
        fill_db(
            words_data=word_freq,
            table_name=table_name,
            tp=nlp
        )

    print("Created word frequency dicts")


if __name__ == '__main__':
    main()
```

3. Код для database_utils.py

```
import os
import enum
import sqlite3
```

```python
class TableType(enum.Enum):
    BEFORE_WAR = "posts_before_war"
    DURING_WAR = "posts_during_war"


# Шлях до бази даних
DB_PATH = os.path.join(os.getcwd(), "texts.db")


def create_db() -> None:

    # Створення бази даних
    print("Creating database...")

    # Видалення бази даних, якщо вона вже існує
    if os.path.exists(DB_PATH):
        os.remove(DB_PATH)

    # Підключення до бази даних та отримання курсора
    conn = sqlite3.connect(DB_PATH)
    cursor = conn.cursor()

    # Створення таблиці для постів періоду під час війни
    cursor.execute(
        """
        CREATE TABLE posts_during_war (
            id INTEGER PRIMARY KEY,
            word TEXT NOT NULL,
            lemma TEXT NOT NULL,
            freq INT NOT NULL
        )
        """
    )

    # Створення таблиці для постів періоду до війни
    cursor.execute(
        """
        CREATE TABLE posts_before_war(
            id INTEGER PRIMARY KEY,
            word TEXT NOT NULL,
            lemma TEXT NOT NULL,
            freq INT NOT NULL
        );
        """
    )

    # Збереження змін у базі даних та закриття з'єднання
    conn.commit()
    conn.close()


def fill_db(words_data: dict, table_name: TableType, tp) -> None:
    # Заповнення бази даних словами та їх характеристиками для відповідного
періоду
    print(f"Filling database table: {table_name.value}\n")

    try:
        # Підключення до бази даних та отримання курсора
        conn = sqlite3.connect(DB_PATH)
        cursor = conn.cursor()

        # Проходження крізь частотний словник слів та вставка значень в таблицю
        for word, freq in words_data.items():
```

```
                # Використання моделі мови для отримання леми слова
                token = tp(word)[0]
                lemma = token.lemma_

                cursor.execute(
                    f"""
                    INSERT INTO {table_name.value} ('word', 'lemma', 'freq')
                    VALUES ("{word}", "{lemma}", "{freq}")
                    """
                )

            # Збереження змін у базі даних та закриття з'єднання
            conn.commit()
            conn.close()

    except ConnectionError as error:
        print(f"Error during connection to database: {error}")
```

Повний код програми наводиться у Додатку 2 до цієї роботи.

Загальний алгоритм роботи програми:

1. Створення бази даних.
2. Читання текстових файлів та отримання списку рядків тексту.
3. Обробка текстових даних з використанням моделі мови Spacy.
4. Створення частотного словника на основі оброблених даних.
5. Заповнення бази даних створеним частотним словником.(Програма використовує TableType.DURING_WAR та TableType.BEFORE_WAR для вказівки типу таблиці, в яку потрібно занести дані).
6. Повторення кроків 2-5 для кожного заданого текстового файлу.
7. Виведення повідомлення про успішне створення частотних словників.

Основні функції програми:

- **read_docx_file(file_path: str) -> List[str]:** Ця функція приймає шлях до файлу формату docx та повертає список рядків, які містяться у цьому файлі.
- **preprocess_text_data(text_data: List[str]):** Ця функція приймає список рядків тексту та застосовує попередню обробку до цих даних з використанням моделі мови Spacy. Результатом є оброблені дані у форматі Doc.

- **get_word_frequency_dict(preprocessed_data) -> Dict[str, int]:** Ця функція приймає оброблені дані та створює частотний словник. Вона обходить кожне слово в оброблених даних, перевіряє, чи воно є алфавітним та не є невідомим словом (OOV - Out of Vocabulary), і збільшує частоту цього слова в словнику.
- **main():** Головна функція програми. Вона виконує всі необхідні кроки для створення частотного словника на основі вказаних текстових файлів. Зокрема, вона створює базу даних, читає дані з файлів, проводить попередню обробку та отримує частотний словник. Потім вона заповнює базу даних цими частотними словниками.
- **create_db():** Ця функція створює нову базу даних SQLite шляхом виконання операторів SQL. Спочатку він перевіряє, чи файл бази даних уже існує, і видаляє його, якщо існує. Потім він встановлює з'єднання з базою даних, створює дві таблиці (posts_before_war і posts_during_war) із певними стовпцями та вносить зміни до бази даних.
- **fill_db(words_data: dict, table_name: TableType, tp):**функція встановлює з'єднання з базою даних, створює об'єкт курсора та виконує ітерацію по словнику words_data. Для кожного слова він отримує свою лему (ймовірно, використовуючи функцію tp) і вставляє слово, лему та частотні значення у відповідну таблицю за допомогою оператора SQL INSERT.Нарешті, функція фіксує зміни в базі даних і закриває з'єднання.

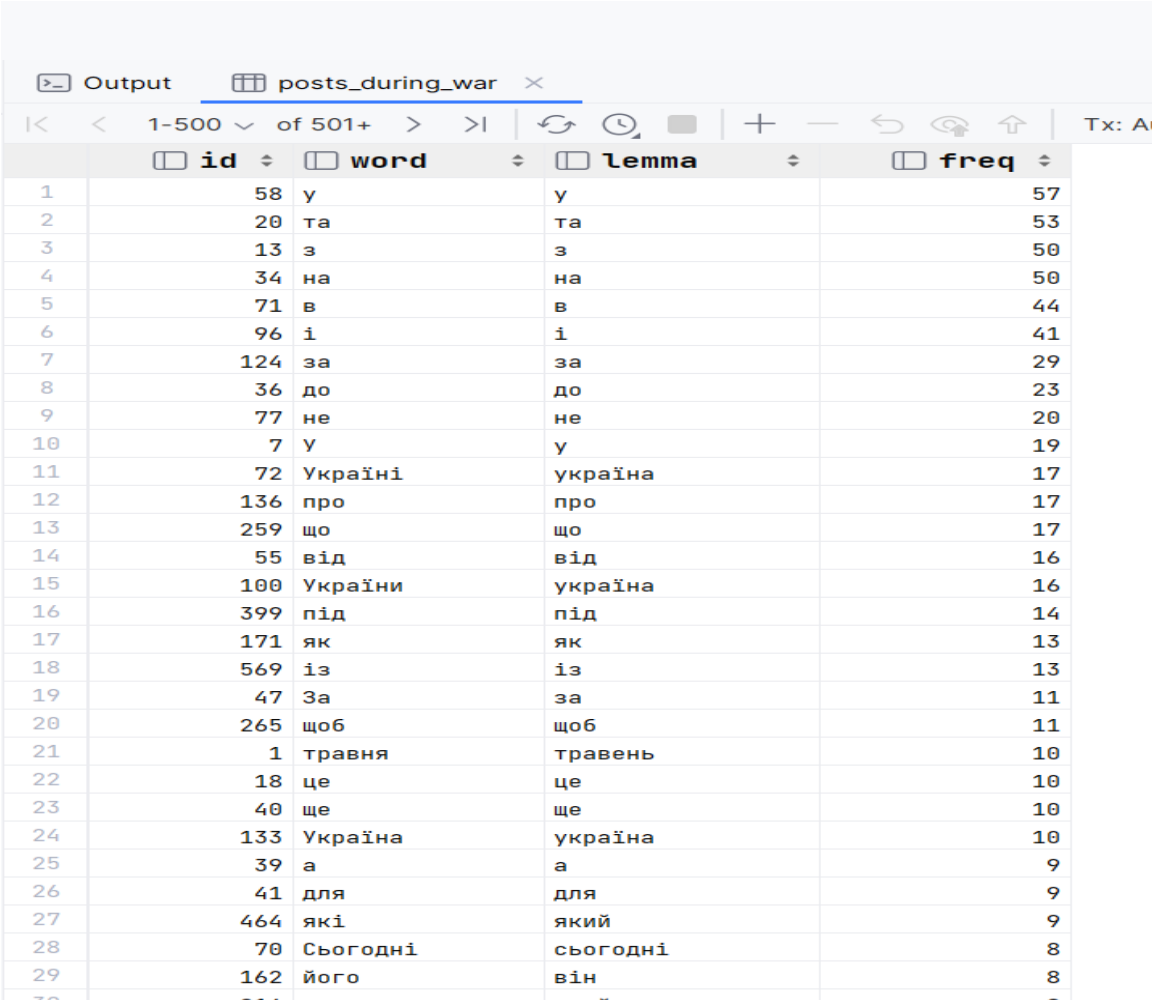
Результат роботи програмного забезпечення: Результатом роботи програми було створення двох частотних словників. Один з яких створений на базі корпусу текстів до повномасштабного вторгнення, а інший базується на новинних текстах в період повномасштабного вторгнення.

Output posts_before_war

1-500 of 501+ Tx: Auto

	id	word	lemma	freq
1	17	на	на	52
2	42	і	і	42
3	126	у	у	40
4	86	та	та	33
5	10	що	що	30
6	23	з	з	30
7	27	в	в	26
8	8	про	про	24
9	327	до	до	24
10	66	не	не	22
11	30	за	за	20
12	369	України	україна	15
13	1	У	у	14
14	103	для	для	14
15	72	як	як	12
16	238	із	із	12
17	101	це	це	10
18	190	від	від	10
19	200	а	а	10
20	277	й	й	10
21	317	які	який	8
22	380	Україні	україна	8
23	184	Україна	україна	7
24	477	За	за	7
25	1069	народний	народний	7
26	1070	депутат	депутат	7

На зображені представлено фрагмент корпусу довоєнних текстів.



Output posts_during_war

1-500 of 501+

	id	word	lemma	freq
1	58	у	у	57
2	20	та	та	53
3	13	з	з	50
4	34	на	на	50
5	71	в	в	44
6	96	і	і	41
7	124	за	за	29
8	36	до	до	23
9	77	не	не	20
10	7	У	у	19
11	72	Україні	україна	17
12	136	про	про	17
13	259	що	що	17
14	55	від	від	16
15	100	України	україна	16
16	399	під	під	14
17	171	як	як	13
18	569	із	із	13
19	47	За	за	11
20	265	щоб	щоб	11
21	1	травня	травень	10
22	18	це	це	10
23	40	ще	ще	10
24	133	Україна	україна	10
25	39	а	а	9
26	41	для	для	9
27	464	які	який	9
28	70	Сьогодні	сьогодні	8
29	162	його	він	8

На зображені представлено фрагмент корпусу текстів за період війни.

Повністю переглянути частотні словники словоформ можна у Додатку 3 до цієї роботи.

Структура частотного словника :

Id- значення , що вказує скільки разів дане слово зустрічалось у тестах.

Word- саме слово ,представлене у граматичні формі, в якій зустрічається в тексті .

Lemma- початкова форма слова, позиційний атрибут, який програма приписує кожну словоформу в корпусі за словником.

Freq – частота досліджуваного слова в масиві текстів.

Висновки до розділу 2

Отже, статистична параметризація відіграє важливу роль у лінгвістичному аналізі українськомовних текстів. Застосування цього підходу дозволяє отримувати кількісні показники, виявляти регулярності та закономірності, аналізувати стилістичні та семантичні особливості текстів, а також проводити порівняльні дослідження. Це сприяє кращому розумінню та опису мовних явищ у контексті воєнного періоду та загального розвитку української мови.

Створення корпусу текстів є важливим етапом лінгвістичних досліджень, оскільки він надає вченим можливість аналізувати та вивчати мову на підставі реальних мовних висловлювань. Корпуси текстів збираються і організовуються для подальшого використання в дослідженнях різних аспектів мови, таких як лексика, граматика, стилістика та багато інших.

Перш за все створення корпусу текстів дозволяє отримати репрезентативний зразок мовлення, що дозволяє виявити тенденції, закономірності та варіацію в мовному вживанні. Крім того, створення корпусу текстів дає змогу проводити різноманітні мовні дослідження, розробляти нові методи та інструменти для аналізу текстів, робити висновки про мовленнєві спільноти, культурні аспекти та комунікативну динаміку.

Корпуси текстів є важливим ресурсом для розвитку лінгвістики та інших суміжних галузей, таких як машинний переклад, комп'ютерна лінгвістика та інші області, де мовна аналітика є важливою складовою.

Частотні словники грають важливу роль у вивченні мови, оскільки надають систематичну та об'єктивну інформацію про вживання слів та лексичних одиниць в текстах. Вони допомагають лінгвістам та дослідникам мови краще розуміти лексичну структуру мовлення, його особливості в різних сферах комунікації та стилістичних контекстах.

Підсумовуючи , створення частотних словників сприяють систематизації та розумінню мови як комплексного системного явища, допомагають вивчати

мову з більш об'єктивної та емпіричної перспективи. Вони виступають як основний інструмент для аналізу текстів, аналітичних досліджень та розробки лексикографічних ресурсів, що забезпечують розвиток теоретичних і практичних аспектів мовознавства. Аналізуючи частотність вживання слів, дослідники можуть виявити популярні лексичні одиниці, а також рідкісні та специфічні слова, що дають унікальну інформацію про мовленнєві спільноти, жанри текстів та культурні особливості.

Додатково, частотні словники допомагають виявити семантичні зв'язки та синонімічність слів, дозволяючи лінгвістам вивчати семантичні поля та фразеологію. Вони також служать основою для розробки лексикографічних ресурсів, словників та тезаурусів, що покращують розуміння та використання конкретних слів.

Розділ 3. Граматична та лексична параметризація новинних текстів за двома вибірками

Отже, дослідження включало в себе такі процеси : збір матеріалів, ручне внесення даних в корпус, створення програмного забезпечення частотного словника словоформ .

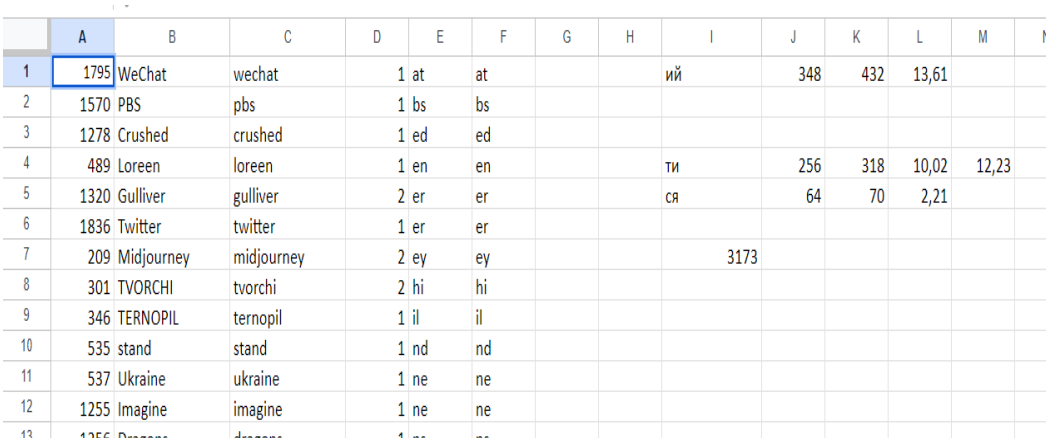
В результаті роботи програми ми отримали два частотних словника словоформ. Один з яких це частотний словник словоформ на базі новинних текстів 2021 року , інший словник укладено на базі медіатекстів ,що були опубліковані в період повномасштабного вторгнення.

На базі отриманих даних з частотних словників словоформ , ми вирішили перевірити чи тексти загалом статичні чи динамічні та порівняти процентні дані між собою. Загалом перед нами постало питання: “Чи змінились показники статичності та динамічності новинних текстів з початком російського вторгнення на територію України?”

Ми розглянули співвідношення прикметників та дієслів, щоб перевірити, чи тексти загалом статичні , чи динамічні ,якщо у тексті великий відсоток прикметників, ми маємо повне право вважати його статичним . Наприклад, якщо в тексті переважають дієслова , то текст буде динамічним.

Процес роботи :

Слід зазначити , що частини мови в отриманих частотних словниках слоформ не вказані. Тому в окрему колонку ми відібрали та внесли дві останні літери кожної леми, використовуючи функцію RIGHT в Екселі. Наступним етапом дослідження було відсортування внесених даних за алфавітом. Зображення з отриманим результатом:



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	1795	WeChat	wechat	1	at	at			ий	348	432	13,61	
2	1570	PBS	pbs	1	bs	bs							
3	1278	Crushed	crushed	1	ed	ed							
4	489	Loreen	loreen	1	en	en			ти	256	318	10,02	12,23
5	1320	Gulliver	gulliver	2	er	er			ся	64	70	2,21	
6	1836	Twitter	twitter	1	er	er							
7	209	Midjourney	midjourney	2	ey	ey			3173				
8	301	TVORCHI	tvorchi	2	hi	hi							
9	346	TERNOPIL	ternopil	1	il	il							
10	535	stand	stand	1	nd	nd							
11	537	Ukraine	ukraine	1	ne	ne							
12	1255	Imagine	imagine	1	ne	ne							

На зображенні представлено фрагмент з результатом підрахунків за період квітень-травень 2023 року.

З отриманих обрахунків ми побачити сумарні частоти лем, що закінчуються на –ий (прикметники) та на –ти і –ся (дієслова). Результати будуть не точними, адже до прикметників додані іменники, прізвища прикметникової форми.

Тексти до повномасштабного вторгнення

	кількість лем	сума частот		% від усіх слововживань
ий (прикметники)	278	334		**12,70**
ти (дієслова)	225	261	9,92	**12,17**
ся (дієслова)	56	59	2,24	

загальна кількість слововживань у всіх текстах
2630

Тексти в час повномасштабного вторгнення

	кількість лем	сума частот		% від усіх слововживань
ий (прикметники)	348	432		**13,61**
ти (дієслова)	256	318	10,02	**12,23**
ся (дієслова)	64	70	2,21	

загальна кількість слововживань у всіх текстах
3173

Повний процес обрахунку зібраних текстів наводиться у Додатку 4 до цієї роботи.

Отже , отримавши значення у відсотковому коефіцієнті , ми маємо змогу порівняти обидва корпуси на статичність та динамічність. Частотність вживання прикметників у новинних текстах за період 2023 року зросла на 0.91%. Слід зазначити , що відсоток досить незначний , а значить стверджувати ,що тексти стали більш статичними ми не можемо. Тобто тексти новинних дописів зберегли баланс статичності. Розглянемо отримані дані частотності дієслів. У цьому співвідношенні практично нічого не змінилося- медіатексти з берегли свою динамічність та повномасштабне вторгнення ніяк не вплинуло на показники.

Тобто тексти дописів зберегли баланс статичність – динамічність.

Наступним ми провели дослідження емоційної забарвленості зібраних новинних текстів.

Процес роботи : цього разу задля визначення емоційного забарвлення ми помаркували лексику певними маркерами. Слід уточнити, що маркували не в текстах, а саме у словнику. Тому контекст не враховувався, і якщо певна лексема

в публіцистичних текстах може бути полісемантичною або омонімічною і не завжди позначати слова з певної семантичної групи, то її не маркувала.

Були позначені лексеми з таких семантичних груп:

1. Дотичні до воєнної та військової теми (в таблиці маркер 1).
   Приклади: “бойові” “Генштаб” “армія” “передова” “ракетний” “окупація” “полк” “військо” “сапер” “бліндаж” та інш.
2. Лексика з позитивним забарвленням (в таблиці маркер 2).
   Приклади: “перемагати” “дарувати” “допомагати” “честь” “рівноправ’я” “вдячність” “творець” “незалежність” “відновити” та інш.
3. Лексика з негативним забарвленням (в таблиці маркер 3)
   Приклади: “страждати” “гинути” “вибухати” “скалічити” “потирпати” “втрачати” “гетто” “жорстокість” та інш.
4. Лексика, дотична до теми пандемії (в таблиці маркер 4).
   Приклади: “лікарі” “хвороб” “хворіє”.

Важливий момент, якщо одна лексема відносилася до двох груп, вона позначалася двома маркерами.Три маркери для однієї лексеми в нашому дослідженні не зустрічалися. Також під час збору та обробки воєнної лексики нами були відкинуті слова, які є полісемічним або мають інше значення,але в нинішніх умовах сприймаються як маркери воєнної теми. Наприклад: “Гостомель”, “Бахмут”, “оборона”, “приліт” та інші.

Принципи, за яким маркувалися позитивні та негативні лексеми:

У процесі маркування ми базувались на емоційних відтінках, позитивні лексеми маркувались як вирази задоволення, радості, позитивного ставлення, тоді як негативні лексеми маркувались як смерть, незадоволення, неприємність, негативну оцінку. Деякі позитивні лексеми позначали активність, дію, силу, успіх, тоді як негативні лексеми вказували на пасивність, втрату ,неконструктивність, провал.

Слід пам'ятати, що маркування позитивних та негативних лексем може бути суб'єктивним, оскільки оцінка певних слів може залежати від особистого досвіду, культурних контекстів і мовних варіацій.

Результати наведені в таблицях.

Тексти до повномасштабного вторгнення

воєнна		
1	17	18
1, 2	1	1
1, 3	3	3
разом	***21***	***22***
%		***0,84***
позитивна		
2	63	76
1, 2	2	2
2, 3	1	1
разом	***66***	***79***
%		***3,00***
негативна		
3	69	71
1, 3	2	2
2, 3	3	3
3, 4	1	1
разом	***75***	***77***
%		***2,93***
дотична теми пандемії		
4	11	15
3, 4	1	1
разом	***12***	***16***
%		***0,61***

Тексти в час повномасштабного вторгнення

воєнна		
1	47	59
1, 2	1	1
1, 3	21	43
разом	***69***	***103***
%		***3,25***
позитивна		
2	99	131
1, 2	27	29
2, 3	1	1
разом	***127***	***161***
%		***5,07***
негативна		
3	86	105

1, 3	27	29
2, 3	21	43
3, 4	1	1
разом	***135***	***178***
%		***5,61***

дотична теми пандемії		
4	3	3
3, 4	1	1
разом	***4***	***4***
%		***0,13***

Повний процес обрахунку зібраних текстів наводиться у Додатку 4 до цієї роботи.

Підсумок дослідження:

Отже, промакрування частотного словника словоформ та отримання відсоткових значень дає нам змогу прослідкувати тенденції в новинних текстах у 2021 та 2023 році. Також хочу додати, що у даному дослідженні ми не спирались на словники тональності та не шукали токсичної лексики. Слід зазначити, що в обрахунках ми не опускали слова , що мають різне значення , заради точності підрахунків.

Отримавши дані, ми маємо змогу зробити наступні висновки:

1. **Воєнна лексика.**

   У медіатекстах з початком повномасштабного вторгнення значно зросла. До прикладу в 2021 році на офіційних сторінках українськомовних ЗМІ процентний коефіцієнт воєнної лексики становив - 0.84% , а з початком воєнних дій коефіцієнт становить - 3.25%, що свідчить про зростання воєнної тематики в медіапросторі на - 2.41%. Отже, в даному контексті воєнна тематика почала переважати в новинних текстах у 2023 році.

2. **Позитивна лексика.**

   У 2021 році позитивна лексика у медіатекстах становила - 3%. З початком війни на території України новинні тексти почали містити в собі -5%. Отже, коефіцієнт використання позитивно забарвленої лексики зріс на -2%. Ми можемо зробити висновок , що у 2023 році пости стали більш емоційно забарвленими.

3. **Негативна лексика.**

   З цих даних ми отримали такий самий результат як і в позитивній лексиці.Частотність вживання негативної лексики зросла на -2.68%(2021 рік - 2.93% , 2023 рік-5.07 %). Отже, в період повномасштабного вторгнення новинні тексти стали більш емоційно забарвленими.

4. **Лексика дотична до пандемії.**

   Останнім маркером у нашому досліджені була лексика дотична до пандемії ковіду. У 2021 році процент вживання становить 0.61 % , що обумовлено саме піком захворюваності та важливістю даної теми. У 2023 році процентний коефіцієнт - 0.13 %. Отже, робимо висновок, що використання даної лексики значно зменшилась.

Висновки до розділу 3

Підсумовуючи проведені дослідження , ми можемо стверджувати , що статичність та динамічність новинних текстів з початком повномасштабного вторгнення росії на територію України залишилась незмінною. Аналіз емоційного забарвлення текстів за період 2021 та 2023 років, дало нам можливість прослідкувати збільшення воєнної тематики в медіатекстах за 2023 рік .Також завдяки процентним даним позитивної та негативної лексики побачити зростання емоційного забарвлення в постах українськомовних ЗМІ у воєнний період.

Висновки

Отже, у даній роботі ми ставили перед собою мету-дослідити зміни, що відбулись та надалі відбуваються в українськомовному медіапросторі ,а саме змін параметрів медіатекстів, що були опубліковані в соціальній мережі Facebook до початку повномасштабного вторгнення Російської Федерації в Україну та в час активних воєнних дій.

Першим етапом нашої роботи було створення двох корпусів текстів. Для отримання корпусу нами було опрацьовано офіційні сторінки українськомовних ЗМІ в мережі Facebook , а саме “1+1”, “СТБ” та “Радіо Свобода”. Обравши джерела , ми приступили до збору та фільтрації матеріалів. Загалом , нами було зібрано 100 новинних текстів з вищеперерахованих сторінок, а саме 50 постів за період 2021 року та 50 постів за період воєнних дій на території України. Зібрані тексти ми відфільтрували від непотрібних нам символів та внесли в таблицю Excel. В двох корпусах ми вказували покликання на пост , дату створення , кількість слів у тексті та джерело з якого було взято новинний текст.

Наступним етапом було створення програмного забезпечення для частотного словника словоформ зібраних нами текстів. Для написання коду програми ми використовували мову програмування Python та інтегроване середовище розробки PyCharm. Наша програма реалізує створення частотного словника на основі декількох текстових файлів формату docx, за допомогою бібліотеки Spacy для обробки мовних даних. Також нами було використано такі модулі ,як **sqlite3, docx ,enum, os** . На виході ми отримали два частотних словника словоформ з такою структурою :**Id-** значення , що вказує скільки разів дане слово зустрічалось у тестах,**word-** саме слово ,представлене у граматичні формі, в якій зустрічається в тексті, **lemma-** початкова форма слова, позиційний атрибут, який програма приписує кожну словоформу в корпусі за словником, **freq** – частота досліджуваного слова в масиві текстів. Також слід додати ,що один з них створений на базі корпусу новинних текстів до

повномасштабного вторгнення, а інший базується на медіатекстах в період повномасштабного вторгнення.

Отримавши частотний словник словоформ , ми перейшли до заключного етапу нашого дослідження , а саме аналіз граматичних та лексичних особливостей новинних текстів за період повномасштабного вторгнення Російської Федерації в Україну та довоєнний час. Спочатку ми вирішили перевірити чи змінилась статичність та динамічність в зібраних текстах. Для цього ми підрахували відсотковий коефіцієнт прикметників та дієслів у текстах, адже якщо в тексті переважають дієслова, то ми маємо повне право назвати тексти динамічними ,а якщо в текстах великий відсоток прикметників, то текст буде статичним.Щоб підрахувати кількість прикметників та дієслів ,ми внесли закінчення (-ти- -ся- для дієслів та -ий-для прикметників) в таблицю Excel та за допомогою функції отримали відсоткове значення.Проаналізувавши співвідношення , ми прийшли до висновку ,що новинні тексти за період 2021 року та за період воєнних дій зберегли свою статичність та динамічність.

Наступним етапом нашого аналізу було визначення емоційної забарвленості в зібраних нами текстах. Хочу зазначити ,що в даній роботі ми не спирались на словники тональності та не шукали токсичну лексику. Нами було обрано 4 маркери : 1-воєнна лексика , 2 -позитивна ,3-негативна, 4- дотична до пандемії. У процесі маркування ми базувались лише на емоційних відтінках слова та відкидали лексику з багатьма значеннями,для більш точного результату. Закінчивши маркування , ми внесли дані в таблицю Excel та за допомогою функції вивели відсоткові співвідношення для двох вибірок. Порівнявши отримані дані за 2021 та 2023 роки ,ми прослідкували збільшення воєнної тематики та емоційного забарвлення, спад теми пандемії в новинних текстах у час повномаштного вторгнення.

У даній роботі ми провели дослідження лексико-граматичних особливостей новинних медійних текстів на сучасному етапі. Проте, варто зазначити ,що для забезпечення повноцінного дослідження в майбутньому варто залучити більший обсяг матеріалів та провести маркування не лише

однозначних слів, а й слів з полісемією . Це забезпечить- повне охоплення воєнної лексики й сприятиме формуванню більш об’єктивних висновків.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. В. Блінцова Процес медіатизації політики та функції мас-медіа у сучасному суспільстві-2013.-С.140
2. В. Туркель і А. Краймбл, «Підрахунок частот слів за допомогою Python», Пр.ogramminghistorian.org.2018.URL: https://programminghistorian.org/en/lessons/counting-частот.
3. В.А.Широков, О.В.Бугаков, Т.О.Грязнухіна та ін. Корпусна лінгвістика . – К.: Довіра, 2005.-С.142.
4. Галина Наєнко, Ярина Ходаківська. До питання про початки наукової мови Куліша // Записки Наукового Товариства імені Шевченка в Америці. Нова серія. Т.2-3. Нью-Йорк, 2021. С.185 – 200 (США).
5. Георгій Почепцов. Токсичний інфопростір. Як зберегти ясність мислення і свободу дії - Харків : Віват, 2022. - 381, [1] с.
6. Д.Бодненко А.Біленко Використання Facebook в роботі філолога//2016-С.127
7. Дарчук Н. П. Статистичні характеристики лексики як відображення структури тексту // Мовознавчі студії.— К.: Наукова думка, 1976.— С. 102.
8. Дарчук Н. П.. Комп’ютерне анотування українського тексту: результати і перспективи. Видавничий дім “Освіта України”, 2013-С. 543.
9. Дарчук Н.П. Корпус української мови – джерело лексикографічних досліджень/2018 р. Міжнародна наукова школа-семінар «Бази даних і системи для потреб сучасного мовознавства» Доповідь виставлена в репозиторії Інституту славістики ПАН.
10. Дейвид Патрикаракос. «Війна у 140 знаках. Як соціальні медіа змінюють військові конфлікти XXI століття». – Київ: Yakaboo Publishing, 2019.

11. Демська-Кульчицька О. Основи національного корпусу української мови.— К.: Інститут української мови національної академії наук України, 2005.— С.140.
12. Дж.Томпсон: Мас медіа та сучасне суспільство //-С.440
13. Дзюбіна, Оксана Ігорівна. Структура, семантика та прагматика сленгових неологізмів соціальних мереж Twitter та Facebook (на матеріалі англійської мови) [Текст] : автореф. дис. ... канд. філол. наук : 10.02.04 / Дзюбіна Оксана Ігорівна ; Львів. нац. ун-т ім. Івана Франка. - Львів, 2016. - 18 с.
14. Добросклонская Т. Г. Медиалингвистика. Системный подход к изучению языка СМИ / Т. Г. Добросклонская. – М. : Флинта – Наука, 2008. – С.264 . Єрмолова А. "Частотні словники та їх використання". -URL: http://dspace.univer.kharkov.ua/bitstream/123456789/5991/2/Ermolova.pdf . -С.137
15. Досенко, Анжеліка. Інтернет-журналістика: комунікативні маркери : навчально-методичний посібник / А. Досенко, І. Погребняк ; Київський ун-т ім. Б. Грінченка. – К. : Центр учбової літератури, 2020. – 183,[1] с
16. Є. Бистрицький, Р. Зимовець, С. Пролеєв. Комунікація і культура в ґлобальному світі. — К.: Дух і Літера, 2020. — 416 с. — Бібліогр.: 395–409 с. — ISBN 978-966-378-799-2.
17. Жуковська В . В. Вступ до корпусної лінгвістики: навчальний посібник / В.В. В. Жуковська – Житомир : Вид-во ЖДУ ім. І. Франка, 2013. – 140 с.
18. Зіґфрід Вайшенберґ, Ганс Й. Кляйнштойбер, Бернгард Пьорксен. Журналістика та медіа : Довідник / Перекл. з нім. П.Демешко та К.Макєєв; за загал. ред. В.Ф. Іванова, О.В. Волошенюк. — К. : Центр Вільної Преси, Академія Української Преси, 2011. — 529 с.
19. Карпіловська Є. Тенденції розвитку сучасного українського лексикону: чинники стабілізації інновацій // Українська мова. – 2007. – № 4. – С. 15.
20. Коломієць, Неля Василівна. Лінгвістичні особливості організації гіпертексту інтернет-новин (на матеріалі англійської мови) [Текст] : дис... канд. філол. наук: 10.02.04 / Коломієць Неля Василівна ; Київський

національний ун-т ім. Тараса Шевченка. - К., 2004. - 214 арк. - арк. 178-206.

21. Комп'ютерно орієнтована освіта майбутніх філологів:[навчально-методичний посібник для студентів ВНЗ]./В.І.Бобрицька, С.М.Процька–Полтава : Скайтек, 2016–С. 136
22. Маршалл Маклюэн. С появлением Спутника планета стала глобальным театром, в котором нет зрителей, а есть только актеры // Кентавр / пер. В. П. Терин. — М., 1994. — № 1. — С. 20—31.
23. Математична лінгвістика: Навчальний посібник. — К. : Вид. центр КНЛУ, 2014. — С.125
24. Медіатекст у сучасному комунікативному дискурсі ЗБІРНИК ТЕЗ. URL:https://mku.edu.ua/wp-content/uploads/2021/04/Zbirnyk-tez-2020-2021-Mediatekst-ostannij.pdf.
25. Макнамі, Роджер. Зафейсбучені: як соціальна мережа штовхає світ до катастрофи [Текст] / Роджер Макнамі ; пер. з англ. Ірина Серебрякова та Оксана Макарова. - Київ : Книголав, 2021. - 375 с. - (Полиця нон-фікшн). - Назва на корінці : Зафейсбучені. - Пер. изд. : Zucked: Waking Up to the Facebook Catastrophe / Roger McNamee. - New York, 2019. - 2000 прим. - ISBN 978-617-7820-72-6
26. Недбай В. В. Інноваційні медіа-технології в системі політичних комунікацій-Одеса, 2012-С.32
27. Ніна Зражевська Теорія медій та суспільства//Навчальний посібник з дисципліни "Теорія медій та суспільства"-Київ 2021-С.192
28. О. Заславська. Віртуалізація простору політичної комунікації: особливості та тенденції-2016-С.99.
29. Перебийніс В. С. Статистична стилістика // Українська мова: Енциклопедія / редкол.: В. М. Русанівський та інші.— 2-ге вид., випр. і доп.— К.: в-во Українська енциклопедія ім.. М. П. Бажана, 2004.— С. 644.
30. Перебийніс В. С., Муравицька М. П., Дарчук М. П. Частотні словники та їх використання.— К.: Наукова думка, 1985.— С. 78.

31. Полюга Л. Статистичний аналіз лексики поетичних творів І. Франка // Іван Франко і національне відродження.— Львів: ЛДУ ім. І Франка; Ін-т франкознавства, 1991.— С. 164–166.
32. Рєзіна О. В. Методичні аспекти навчання студентів створення цифрових частотних словників..URL: https://www.researchgate.net/publication/336647784_METODICNI_ASPEKTI_NAVCANNA_STUDENTIV_STVORENNU_CIFROVIH_CASTOTNIH_SLOVNIKIV
33. Різун В. В. Лінгвістика впливу / В. В. Різун, Н. Ф. Непийвода, В. М. Корнєєв. – К. : ВПЦ «Київський університет», 2005. – С.148 .
34. Роль мас-медіа у формуванні ціннісної системи суспільства / О. Дубас // Сучасна українська політика. Політики і політологи про неї. — К., 2009. — Вип. 16. — С. 242-248.
35. С.Рум'янцева Медіа-технології у виборчому процесі: світові тенденції та українська практика-2017-С.20.
36. Традиційна та комп'ютерна лексикографія: Навч. посібник. — К. : Вид. центр КНЛУ, 2009. — С. 218 (у співавторстві)
37. Фергюсон, Ніл. Площі та вежі. Соціальні зв'язки від масонів до фейсбуку / пер. Катерина Диса. К.: Наш Формат, 2018. — 376 с. ISBN 978-617-7552-77-1
38. Хлівнюк. Т.Сучасна система мас-медіа як умови досягнення соціального консенсусу-2013-С.126.
39. Частотний словник сучасної української художньої прози: у 2-х т. / за ред. Перебийніс В. І. — К. : Наукова думка, 1981. — Т. 1. — С.863, Т. 2. — С.856 . (у співавторстві)
40. Biber D. Representativeness in corpus design // Literary and Linguistic Computing.- 1993.Vol. 8. - N4.- C. 243 - 257.