МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
УКРАЇНСЬКИЙ КАТОЛИЦЬКИЙ УНІВЕРСИТЕТ

Гуманітарний факультет

Кафедра філології

# КОРПУСНО-СТАТИСТИЧНЕ ДОСЛІДЖЕННЯ ЛЕКСИЧНИХ ОСОБЛИВОСТЕЙ ПУБЛІЦИСТИЧНОГО ТЕКСТУ (НА МАТЕРІАЛІ ГАЗЕТИ «ДЕНЬ»)

Студентки IV курсу

групи ГФІ17/Б

**Христини Борецької**

Науковий керівник:

кандидат філ. наук

**Старко Василь**

Львів 2021

**ЗМІСТ**

# ВСТУП

В сучасну епоху постійного напливу інформації відчувається потреба автоматичного опрацювання природномовного матеріалу. Завдяки розвитку комп’ютерних технологій, які прорвалися в лінгвістику в 60-х роках ХХ століття, стали можливими зберігання, впорядкування та редагування мовного матеріалу. Таким чином у сфері мовознавства й інформаційних технологій постала нова галузь – комп'ютерна лінгвістика, яка стала самостійним та самодостатнім напрямом наприкінці 80-х – початку 90-х років ХХ століття. Ця галузь дозволяє глибоко і вичерпно вивчити мовний матеріал, адже вона послуговується як і кількісними, так і статистичними методами дослідження. Актуальність корпусної лінгвістики підтверджується великими кількостями наукових розвідок, посібників і передусім створенням корпусів текстів, які слугують ключовим інструментарієм для дослідження в цій галузі науки. В Україні корпусна лінгвістика перебуває на етапі її становлення і розвитку. Серед невеликої кількості українських корпусів одним з найкращих є ГРАК (Генеральний регіонально анотований корпус української мови), який перебуває в процесі розвитку, тому поповнення корпусу новими текстами, зокрема публіцистичними, буде цінним внеском в розвиток корпусної лінгвістики в України. Корпусний менеджер ГРАКу дозволяє проводити різні статистичні дослідження, які здійснюються на матеріалі корпусу. Ця робота виконана в межах напрямку корпусної лінгвістики.

**Актуальність теми** та проведеного корпусно-статистичного дослідження лексичних особливостей публіцистичного стилю зумовлена тим, що попри розвиток корпусних досліджень в Україні все ще відчувається брак досліджень публіцистики з застосуванням статистичних метрик, а також існує потреба в дослідженні згаданих текстів в порівнянні з текстами інших жанрів, яке базується на корпусних даних.

**Новизна дослідження** дипломної роботи полягає в тому, що вперше в галузі української корпусної лінгвістики на матеріалах корпусу ГРАК, до якого додано попередньо опрацьовані тексти, проаналізовано лексичні особливості

публіцистичного стилю з застосуванням кількісних, статистичних методів й оцінкою статистичної значущості.

**Метою цієї роботи** є підготувати публіцистичні тексти української газети «День» 2000-2019 років для подальшого їх додавання в корпус ГРАК, а також дослідити характерну для публіцистичного стилю лексику. Відповідно до мети в роботі передбачається розв'язання таких **завдань**:

- сформулювати загальні положення корпусної лінгвістики та її значення, схарактеризувавши стан корпусної лінгвістики в Україні;
- пояснити та обґрунтувати важливість корпусів текстів у дослідженні мови;
- вивчити та описати структуру корпусу ГРАК;
- з'ясувати важливість публіцистики в корпусі та дослідити представленість публіцистичних текстів в українських корпусах;
- проаналізувати особливості наявних досліджень мови газети «День»;
- опрацювати тексти «Дня» за 2010-2019 рр., підготувати їх до внесення в корпус ГРАК;
- створити фокусний і референтний підкорпуси в межах ГРАКу;
- вивчити метод ключових слів, який передбачає проведення тестів на статистичну значущість, проаналізувати застосування цього методу до корпусних даних;
- встановити ключові слова для текстів газети «День»;
- проаналізувати виявлені ключові слова й одиниці.

**Об'єктом дослідження** в роботі виступає мовлення сучасного публіцистичного видання, а **предметом** є характерна для публіцистичного стилю лексика у протиставленні з текстами інших стилів.

В процесі роботи було застосовано такі **методи**:

- метод ключових слів, який охоплює низку процедур;

- аналітичний — для визначення сутності та головних завдань корпусної лінгвістики, було проаналізовано її об’єкт та предмет вивчення;
- метод попереднього опрацювання тексту — послідовність операцій із текстами для приготування їх до додання в корпус;
- метод контекстного аналізу — з’ясування особливостей вживання слів у конкретних контекстах.

**Матеріалом цієї роботи** є публіцистичні тексти української газети «День», опубліковані у 2000-2019 роках.

Загальний обсяг роботи становить 84 сторінок та має таку **структуру**:

**У першому розділі** визначено основні теоретичні засади корпусної лінгвістики, а також проаналізовано розвиток цієї галузі в Україні. В підрозділі 1.1 детально розглянуто історію становлення корпусної лінгвістики як науки, окреслено її сутність, предмет і завдання, а також досліджено становище цієї галузі в Україні. В підрозділі 1.2 описано особливості структури українського корпусу ГРАК. Підрозділ 1.3 присвячено описанню важливості публіцистики в корпусі, а також проаналізовано наповненість українських корпусів текстами цього стилю.

**Другий розділ** детальніше розглядає особливості текстів українського видання «День». У підрозділі 2.1 проаналізовано попередні дослідження мови газети «День», зокрема розглянуто вживання іншомовних слів, простежено особливості написання заголовків тощо. У підрозділі 2.2 описано процес попереднього опрацювання текстів газети «День» 2000-2019 років для їх подальшого додання в корпус ГРАК.

**Третій розділ** пропонує розгляд методу ключових слів та сам аналіз ключових слів для газети «День». У підрозділах 3.1 і 3.2 докладно описано принципи побудови фокусного і референтного підкорпусів відповідно. Підрозділ 3.3. розглядає поняття ключових слів і ключовості. В підрозділі 3.4 здійснено опис метрик статистичної значущості, зокрема детально розглянуто тест хі-квадрат, метрику логаритмічної правдоподібності, а також Баєсів інформаційний критерій.

Підрозділ 3.5 описує процес відбору ключових слів зі згенерованих списків потенційних ключових слів за допомогою метрик статистичної значущості. Тоді як в підрозділі 3.6. здійснено аналіз виявлених ключових слів газети «День».

В кінці вміщено висновки до цілої роботи й список використаної літератури. В додатку 1 подано ілюстрації здійснених операцій в програмі Notepad++ для очищення текстів газети «День» від позатекстових елементів, в додатку 2 — наведено список ключових слів для згаданого видання.

# РОЗДІЛ І
# КОРПУСНА ЛІНГВІСТИКА

У цьому розділі буде описано сутність корпусної лінгвістики, розглянуто її поняття та засоби, за допомогою яких спеціалісти можуть досліджувати мову та мовлення, а також буде проведений аналіз стану корпусної лінгвістики в Україні з описом наявних українських корпусів, зокрема велику увагу приділено ГРАКу та його публіцистичному стилю.

## 1.1. Корпусна лінгвістика і корпус. Розвиток корпусної лінгвістики в Україні

З початку 60-х років минулого століття і станом на сьогодні розвиток лінгвістичної науки відзначається помітним розмахом такої галузі, як корпусна лінгвістика. Вона є надзвичайно перспективною у сфері лінгвістики, оскільки дає можливість науковцям та дослідникам, які мають справу не тільки з лінгвістикою, але й з іншими царинами науки, працювати з дуже корисним засобом-інструментарієм – корпусом, який забезпечує швидке, повне та глибоке вивчення певних теоретичних та практичних аспектів, сприяє вирішенню найрізноманітніших завдань, дає змогу оптимізувати та вдосконалити роботу з різних напрямків.

Поняття «корпусна лінгвістика» міцно становилося в науковому вжитку в останні десятиліття XX століття з виходом в 1983 році збірника наукових праць «Corpus Linguisitcs: Recent Developments in the Use of Computer Corpora in English Language Research» за матеріалами конференції ICAME «Conference on the Use of Computer Corpora in English Language Research»[1]. Анатоль Стефанович у книжці «Corpus linguistics. A guide to the methodology» вводить досить загальне, але цілком влучне визначення для поняття «корпусна лінгвістика», твердячи, що це будь-яка форма мовного дослідження на основі даних, отриманих із корпусу[2]. Тобто

[1] Світлана Голощук. Історичні передумови розвитку корпусної лінгвістики / С. Л. Голощук // *Національний університет «Львівська Політехніка»*. Львів 2017, ст. 83-84.
[2] Anatol Stefanowitsch. *«Corpus linguistics. A guide to the methodology»* // Language Science Press 2020, p.22.

корпусна лінгвістика тісно пов’язана з емпіричним дослідженням використання живої мови за допомогою комп’ютерів та електронних корпусів.

Значення корпусної лінгвістики для дослідження мови лінгвістами та й, загалом, для науковців важко переоцінити, проте місце цієї галузі серед інших наук було достатньо нетривке і повністю неокреслене. Е. МакЕнері, Е. Вілсон, Р. Ксіао та інші мовознавці розцінювали корпусну лінгвістику як особливу «методологію», спосіб вивчення і дослідження мовного матеріалу, напроти таких «звичайних» царин лінгвістики, як орфографія, фонетика, синтаксис. Згадані лінгвістичні галузі вивчають лише певний аспект мовної системи, тоді як методи корпусної лінгвістики можна застосувати до теоретичних та практичних аспектів різних наук. Тим більше корпусні дані можна використати, вивчаючи мовні одиниці будь-якого мовного рівня.

Методи та принципи використання корпусів для дослідження мови та для навчання мові має теоретичне підґрунтя, проте одне теоретичне підґрунтя ще не є науковою теорією. Тому, за Е. МакЕнері, Е. Вілсон, Р. Ксіао, корпусну лінгвістику розглядають як набір методів, за якими вивчають мову, з багатьма опціями використання в мовознавчих напрямах[3].

Натомість більшість сучасних лінгвістів, які працюють напряму з корпусом, не погоджуються з твердженням, що корпусна лінгвістика має лише методологічну роль, а розглядають цю галузь, як окрему повноцінну науку прикладної лінгвістики зі своїм предметом, об’єктом, метою, методами дослідження, а теж теоретичним і практичним підґрунтям. Наприклад, Тоґніні-Боннелі, цитуючи Хеллідей, зазначає, що корпусна лінгвістика наново об'єднує діяльність зі збору та теоретизації даних, і стверджує, що це призводить до якісних змін у нашому розумінні мови[4]. Інші лінгвісти вказують на зв'язок між використанням обчислювальних, а отже, алгоритмічних та статистичних методів, з одного боку, та якісною зміною

[3] Anthony McEnery, Richard Xiao, Yukio Tono. *Corpus-based Language Studies: an Advanced Resource Book*. London 2006, p. 6-7.
[4] Elena Tognini-Bonelli. *Corpus Linguistics at Work*. Amsterdam 2001, p. 1.

спостережень, що випливають із цього підходу, з іншого. Тобто корпусна лінгвістика заслуговує вийти зі статусу методологічної бази опрацювання мови та стати окремою дисципліною, самостійним науковим напрямком.

Отож чим займається корпусна лінгвістика? Ця галузь прикладної лінгвістики 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»[5]. Об'єктом корпусної лінгвістики є корпус текстів, який являє собою структуроване зібрання машиночитних текстів найрепрезентативнішого реального писемного або усного мовного матеріалу, представлений в електронному вигляді, уніфікований, структурований, розмічений, не модифікуючи при цьому мовленнєвої дійсності, що уможливлює розгляд фактичного корпусного матеріалу як емпіричну базу лінгвістичного дослідження.

Корпусний аналіз характеризується певним переліком ознак:

1) емпіричний підхід до аналізу мовних даних;

2) використання репрезентативних машиночитних вибіркових колекцій тексту, як об'єкт дослідження;

3) широкомасштабне залучення комп'ютерних технологій для дослідження мовного матеріалу;

4) застосування кваліативних і квантитативних аналітичних методик (наприклад, вивчення частоти вживання лінгвістичних одиниць, статистичні дослідження сполучуваності й т.ін.).

Дані, які лінгвісту вдалося отримати за допомогою корпусного аналізу, дозволяють сформулювати якісно нові висновки про мову й окреслити абсолютно нові напрями досліджень. З тої причини, що корпус – це колекція «живих» текстів, які реально функціонують чи функціонували в мові, дослідник має змогу

[5] Вікторія Жуковська. *Вступ до корпусної лінгвістики: навчальний посібник*. Житомир 2013, с. 9.

абстрагуватися від попередніх напрацювань, висновків, суб'єктивності, й абсолютно об'єктивно оцінити мовний матеріал та зробити актуальні висновки.

Як твердять Е. МакЕнері й Е. Харді, «корпусні розвідки переорієнтовують традиційний підхід до вивчення мови, а результати аналізу даних корпусу сприяють переоцінці низки лінгвістичних теорій»[6]. Тому важливість корпусу важко переоцінити.

Велике число праць у сфері корпусного дослідження мови можна поділити на два напрямки. Один з яких зосереджений на дослідженні проблем, які стосуються теорії та практики створення корпусів: розмір корпусу, збалансованість, репрезентативність, вибірковість, структурування і т. д. А інший напрям спрямований вже на дослідження цих корпусів, тобто підведення висновків за допомогою корпусних методів на основі того мовного матеріалу, який показаний в корпусі. Проте цей поділ дуже умовний, оскільки укладачі корпусу одночасно створюють мовознавчі дослідження на їх основі. Таким чином, об'єкт корпусної лінгвістики – корпус – має двосторонній характер: він одночасно виступає і вихідним мовленнєвим матеріалом для корпусної лінгвістики, і є результатом діяльності цього мовознавчого напряму.

Предметом корпусної лінгвістики виступають теоретичні й практичні засади створення та використання мовних корпусів.

Головним завданням корпусної лінгвістики є системне зображення структурованого спілкування мовою. Ця галузь прикладної лінгвістики відзначається тим, що підходить до прикладних проблем в мовознавстві опосередковано в комунікативному процесі й при цьому важливо тримати в полі зору процес змістовного мовлення та об'єктивно описувати цей процес.

Метою корпусної лінгвістики є об'єктивний мовознавчий опис та дослідження мовної системи. Важливо зазначити, що для цього опису корпусна

[6] Anthony MacEnery, Andrew Hardie. *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press 2012, p. 1.

лінгвістика бере під детальне вивчення конкретну людську комунікацію. Менш важливою ціллю корпусної лінгвістики вважається «вироблення особливого способу відображення мовного матеріалу в корпусі текстів»[7].

Питання виникнення та розвитку корпусної лінгвістики є таким же важливим, як і питання теоретичного підґрунтя зазначеної галузі. Аналізуючи теоретичні засади корпусної лінгвістики, висновується, що структуралізм чи структурна лінгвістика вважається теоретичним підґрунтям вищезгаданої галузі прикладної лінгвістики. Маючи в основі структурний підхід, за допомогою методів корпусної лінгвістики, аналізується реальний текст, при тому виділяючи узагальнені інваріантні одиниці (схеми речень, морфем, фонем) і співвідносячи їх з конкретними мовленнєвими сегментами, виходячи «з детермінованих правил реалізації, які визначають межі варіювання мовних одиниць у мовленні, у такий спосіб декларуючи примат реального тексту в лінгвістичному дослідженні, що власне і є ідеєю корпусної лінгвістики»[8]. Таким чином, лінгвіст працює з об'єктивними кількісними даними, за допомогою яких досягається більш ґрунтовних та передусім актуальних висновків. Корпусна лінгвістика сприяє підтвердженню або спростуванню гіпотези про функціонування мови, окреслює нові напрями вивчення мови, які до застосування корпусних методів не потрапляли в коло питань дослідників.

Як вже було зазначено, центральним для корпусної лінгвістики є корпус текстів, оскільки він являється основною метою корпусної лінгвістики, а з іншого — є об'єктом дослідження цієї лінгвістичної царини. Сучасне корпусне мовознавство має у своєму складі значну кількість корпусів різних типів і розмірів. Досить велике число корпусів свідчить про їхню потрібність поміж науковцями, оскільки корпуси створюють з певною метою та для розв'язання конкретних питань. З цієї причини багато країн створюють корпуси текстів для якісного дослідження мовного матеріалу. Наприклад, збалансовані та репрезентативні

[7] Вікторія Жуковська. *Вступ до корпусної лінгвістики: навчальний посібник*, с. 11.
[8] Орися Демська-Кульчицька. Деякі аспекти корпусної лінгвістики // *Українська мова, № 1, 2005,* с. 48.

корпуси-мільйонники Brown Corpus, Ланкастерський корпус англійської мови (LancasterOslo-Bergen Corpus, LOB), Мангеймський корпус німецької мови, Чеський національний корпус, COCA можуть використовуватися для різноманітних лінгвістичних цілей. Важко переоцінити корисність одного із видів корпусу — національного корпусу, який є великий за обсягом та об'єднує у своїй структурі тексти найрізноманітніших жанрів і типів (завдяки сучасним технологіям можливе вбудовування в національні корпуси аудіо- і відеоматеріалів). Серед відомих національних корпусів наступні: the British National Corpus (100 млн. слововживань), the American National Corpus (22 млн.), the PELCRA Reference Corpus of Polish Corpus (100 млн.), the Czech National Corpus (більше 100 млн.), the Hungarian National Corpus (187,6 млн.), the Hellenic National Corpus (корпус сучасної грецької мови, 47 млн. слововживань), the Slovak National Corpus (339 млн.), the Modern Chinese Language Corpus (100 млн. знаків) та ін.[9]

Варто зауважити, що розвиток корпусної лінгвістики в Україні відрізняється від розвитку світової традиції тим, що у вітчизняній лінгвістиці відсутній період першого покоління («стадія ранньої корпусної лінгвістики (1910–1960-ті рр.), коли відбувається формування теоретичного підґрунтя та прагматичних передумов виникнення напряму й створення текстових зібрань для лінгвістичного дослідження переважно на паперових носіях»[10]). Цей період в українській лінгвістиці представлений численними лінгвостатистичними дослідженнями на матеріалі доелектронних корпусів.

Таким чином, серед особливостей розвитку корпусної лінгвістики першого покоління в Україні, як і в інших країнах СРСР, можна назвати відсутність мотивації побудови електронних корпусів у радянських лінгвістів.

В період, коли в багатьох країнах рух в корпусній лінгвістиці спрямований на створення універсальних національних корпусів текстів, українські лінгвісти надають перевагу укладанню частотних словників. Прорив української корпусної

[9] Вікторія Жуковська. *Вступ до корпусної лінгвістики: навчальний посібник*. Житомир 2013, с. 66.
[10] Орися Демська-Кульчицька. Деякі аспекти корпусної лінгвістики, с. 45.

лінгвістики відбувся в першому десятиріччі XXI ст. Результатом цього стали теоретичні дослідження, які вивчали засади побудови та використання електронних машиночитних корпусів текстів[11].

Отже, українському мовознавству притаманний довготривалий період ранньої корпусної лінгвістики, на відміну від світової традиції, який переростає одразу в період корпусної лінгвістики другого покоління (починається з 1960 рр., безпосередньо пов'язаний зі значним розвитком комп'ютерних технологій). Подібні риси розвитку корпусної лінгвістики притаманні іншим країнам пострадянського простору, а «це сприяє усвідомленню досвіду побудови сучасних електронних корпусів, зокрема, Web-корпусів»[12].

Сьогодні українська комп'ютерна лінгвістика представлена досить вагомим практичним доробком – це, зокрема, БрУК, лінгвістичні портали (mova.info), словникові сайти (r2u, e2u), комплекси засобів (brown-uk, lang-uk) тощо. Детальніше про українські корпуси нижче.

1. Лінгвістичний портал mova.info. Корпус розроблений співробітниками лабораторії комп'ютерної лінгвістики Інституту філології Київського національного університету імені Тараса Шевченка Н. Дарчук (керівниця проєкту), В. Сорокіним (програміст), О. Сіруком, Я. Ходаківською, Н. Чейлитко, М. Лангенбахом[13]. Дослідницький корпус сучасної української мови нараховує 100 мільйонів токенів, включаючи юридичні тексти, академічні, поетичні тексти, публіцистику та художні тексти. Розмір публіцистичних текстів за 2008-2010 роки складає 44 655 032 слововживання. Цей корпус текстів — це тексти в електронній формі, призначені не для читання, а для з'ясування різних питань, пов'язаних з українською мовою.

---

[11] Тетяна Бобкова. Проблеми періодизації корпусної лінгвістики у світовому та українському мовознавстві / Т. В. Бобкова // *Науковий вісник кафедри ЮНЕСКО Київського національного лінгвістичного університету. Філологія, педагогіка, психологія.* Вип. 28, 2014, с. 50.

[12] Там само.

[13] Лінгвістичний портал MOVA.info
Київ: 2003-2021. [Електронний ресурс], режим доступу:
http://www.mova.info/carticle.aspx?l1=210&DID=5347

2. Ukrainisches Gemischt-Korpus – це корпус українських різножанрових інтернет-текстів, зібраних в Лейпцизькому університеті, за матеріалами з 2014 року[14]. Він містить 102 429 857 речень та 1 546 330 404 лексеми. За результатами пошуку можна дізнатися про приклади вживання слова в текстах, сполучуваність і тривимірний граф сполучуваності. Однак недоліком є відсутність морфологічного маркування, що дозволяє шукати лише словоформи.

3. Корпус інтернет-текстів Лабораторія української (бл. 140 тисяч токенів). Проєкт створений Наталією Коцибою, Богданом Москалевським та Михайлом Романенком, який включає чотири паралельні корпуси, що розмічені автоматичним аналізатором зі знятою омонімією, та загальний корпус Звідусіль, який налічує 2 848 203 658 токенів. Підкорпус публіцистики становить 17 млн слововживань. Кожен текст в корпусі розмічений за жанровими особливостями й типом тексту.[15]

4. «Корпус української мови». Обсяг корпусу складає 6,6 Гб україномовних текстів з електронної бібліотеки «Чтиво», тому тексти можуть бути недостатньо якісними і відповідно не відповідають вимогам корпусної лінгвістики. З цієї причини результати можуть бути нерелевантними й вимагають уважного розгляду та критичного мислення. Публіцистика в цьому корпусі складається з 57839 творів 19446 авторів.

5. БрУК[16]

Творцями корпусу є Василь Старко та Андрій Рисін. Корпус отримав таку назву на прикладі добре відомого Браунського корпусу англійської мови. Це анотований невеликий корпус сучасної української мови, який налічує 1 млн токенів. Тексти корпусу поділяються на два види – інформативні й художні. Публіцистичний стиль складає 25% усіх текстів. Наразі він у стадії розроблювання.

[14] Ukrainisches Gemischt-Korpus / – Leipzig: 1998-2021.
[Електронний ресурс], режим доступу: https://corpora.unileipzig.de/de?corpusId=ukr_mixed_2014
[15]Лабораторія української
[Електронний ресурс], режим доступу: https://mova.institute/
[16] Корпус сучасної української мови (БрУК) / В. Старко, А. Рисін
[Електронний ресурс], режим доступу: https://github.com/brown-uk

6. Польсько-український паралельний корпус обсягом понад 3 мільйони словоформ. Корпус містить оригінальні та перекладені тексти польською та українською мовами, створені здебільшого в 20 столітті, які належать до різних жанрів: художньої літератури, публіцистики, документів, підручників, пресрелізів. Тексти лематизовані, набори тегів для української мови є стандартизовані[17].

## 1.2. Генеральний анотований корпус української мови (ГРАК): його особливості та опис

Гостроактуальною проблемою в українській корпусній лінгвістиці залишається потреба національного репрезентативного збалансованого корпусу, який стане добрим інструментарієм для лінгвістів, які досліджують мову та підбивають підсумки щодо розвитку мови, які відповідають рівню дослідження таких мов, як-от англійська чи німецька, а також науковцям з областей інших наук. Наразі корпусна лінгвістика має у своєму арсеналі добрий «універсальний» корпус – Генеральний регіональний анотований корпус української мови, який все оновлюється та збільшується. Корпус знаходиться у вільному доступі в Інтернеті, налічує 822 871 459 млн токенів (GRAC v.12) та зображує більшість жанрів написаних текстів.

Особливість ГРАКу в тому, що це єдиний корпус української мови, де кожен текст має своє метадані, а також регіональну та жанрову анотації (тобто кожен текст анотується за жанрами та регіоном). Наступною перевагою ГРАКу є те, що приблизно 50% текстів припадають на різні регіони України або країни діаспори.

Спираючись на те, що корпус охоплює період з 1816 по 2020 роки, підводимо до висновку, що він буде добрим засобом для дослідження науковцями – соціологами, лінгвістами, лексикографами – оскільки вони можуть вивчати мову в синхронному та діахронічному зрізах відповідно до стилю та жанру. Основними виконавцями цього корпусу є Марія Шведова з Василем Старком (Україна), технічну реалізацію здійснюють у співпраці з Рупрехтом фон Вальденфельсом

[17] Polsko ukraiński korpus równoległy
[Електронний ресурс], режим доступу: domeczek.pl/~polukr/index.php

(Німеччина), Сергієм Ярихіним (Україна), Андрієм Рисіном (США), Тимофієм Ніколаєнком (Україна), Михайлем Круком (США), Міхалом Возняком (Польща), наукові консультанти проєкту: Рупрехт фон Вальденфельс (Німеччина), Дмитро Січінава (Україна), Михайло Назаренко (Україна). Наведені нижче дані про ГРАК спираються на інформацію, наведену на сайті корпусу[18] та в публікації М. Шведової[19].

Остання версія ГРАКу (GRAC v. 12) була створена 5 травня 2020 року. Вона містить понад 70 тисяч текстів різних жанрів і близько 22 тисяч авторів. Основні стилі ГРАКу: ACA – науковий, FIC – художній, JOU – публіцистичний. У порівнянні з одинадцятою версією стало більше публіцистики: додалися окремі номери газети «Літературна Україна» (1991-1998, обсяг 1,2 млн токенів); окремі номери 1992-1995 років газети «Кримська світлиця» (570 тис токенів); журнал «Всесвіт» (1991-1998, окремі номери, >2 млн токенів); радянські газети різних регіонів 1945-1991 рр. обсягом 4 млн токенів. Важливо, що в корпусі, починаючи з десятої версії, є семантична розмітка. Кожен текст в корпусі має вказану дату створення, також багато текстів мають дані, які стосуються їх публікації. Особливість та цінність ГРАКу в тому, що він має створений підкорпус текстів української діаспори, оскільки раніше в галузі корпусної лінгвістики такі тексти були в занедбаному стані, а тепер вони є полем для дослідження та вивчення нормативних варіацій.

Щодо стилів, жанрів та тематики ГРАКу: приблизно половину корпусу складають художні прозові тексти, хоча в перспективі включення поетичних текстів (станом на початок 2020 року «Енеїда» Котляревського є єдиним поетичним текстом). З нехудожніх текстів великий підкорпус становлять публіцистичні тексти і тексти з інформаційних сайтів в Інтернеті (до корпусу взяті тексти тільки з сайтів,

[18] Марія Шведова. 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: 2017–2021. [Електронний ресурс], режим доступу: http://uacorpus.org/

[19] Maria Shvedova. The General Regionally Annotated Corpus of Ukrainian (GRAC, uacorpus.org): Architecture and Functionality. Kyiv 2020.

які мають одну, україномовну, версію). Інший досить об'ємний підкорпус складається з наукових та навчальних текстів: монографії, дисертації, наукові статті, підручники. Також заслуговує уваги окремий підкорпус релігійних текстів, який включає два українські переклади Біблії. Цікавим для дослідників є підкорпус его-текстів, який складається зі спогадів, щоденників, листів та дописів людей у соціальній мережі Facebook, серед яких є блоги людей з усіх областей України та навіть з діаспори. Хоч і невеликий, але, без перебільшення, дуже важливий є підкорпус розмовних жанрів, а саме інтерв'ю та промови. Останні два згадані підкорпуси слугують чудовим інструментарієм для науковців: лексикографів, соціологів, лінгвістів, оскільки на основі цих корпусів дослідники можуть робити висновки про «живу мову», а саме вивчати вживання варіантних форм в мовленні людей, висновувати про форми, які «прижилися» у мовленні людей, а які зникли, а відтак шукати причин функціонування тих чи інших форм у мовленні. Корпус також включає словники з фразеологізмами, ідіомами, Словник української мови Б. Грінченка і «Російсько-український словник сталих виразів» І. Виргана та М. Пилинської. Інструменти корпусу дозволяють працювати зі вказаними словниками та здійснювати пошук не тільки лексеми, але й лексико-граматичного зразка, які прослідковуються в прикладах та цитованих ідіомах.

ГРАК єдиний корпус української мови, тексти якого проанотовані за регіональною розміткою, яка є взагалі унікальною для слов'янських корпусів. Кожен автор тексту приписаний до одного або кількох регіонів, залежно від того, де він народився, вчився або жив тривалий час. Приблизно третя частина текстів в корпусі – переклади з 69 мов (GRAC v.7). Тому, якщо текст є перекладним, то йому відповідає регіон, приписаний перекладачеві. Важливий момент, що деколи українські переклади здійснені не з мови оригіналу, а з російського перекладу. Якщо це зазначено у виданні, тоді вноситься в корпус, що текст перекладено з російського перекладу. Проте не всі тексти проаналізовані за можливістю перекладу з російського перекладу. Суть регіональної розмітки текстів полягає в адміністративному поділі. Корпус включає тексти з усіх областей України та з

Криму. Усі регіони об’єднані в шість макрорегіонів. Приблизно половина текстів проанотовані за регіональною розміткою. Тексту присвоюється тег того регіону, де автор жив з народження, вчився і жив понад 10 років, а якщо після цього проміжку часу автор переїхав в інше місто чи країну, то текст буде належати до кількох регіональних підкорпусів. Якщо невідомо місце народження чи проживання автора, тоді тексту присвоюється регіональний тег відповідно до місця публікації тексту. Найбільший обсяг текстів становлять тексти з регіональним тегом «Київ», оскільки більшість авторів в певний етап свого життя переїжджали до Києва. Така ж ситуація, коли автор емігрував або жив закордоном понад 10 років, йому присвоювали тег тої країни, де він знаходиться. Таким чином ГРАК має підкорпус текстів авторів діаспори. Їхні твори розмічені за країнами (Німеччина, Канада, США, Франція, Велика Британія, Польща). Отже, регіонально анотований корпус – це цінний інструмент для дослідження регіональних відмінностей в українській мові, а в перспективі – база для систематизованого опису регіональних українських стилів.

Тексти в корпусі анотуються за роком написання чи за останнім роком, коли текст міг бути написаний. Коли точної дати невідомо, тексти позначаються періодом. У випадку якщо текст перекладний, він отримує рік створення перекладу. Крім дати самого твору, додатково може зазначатися рік видання. Важливо зазначити, що в корпусі міститься інформація про авторів: рік народження, стать, місце народження, регіон/регіони (де він навчався або проживав понад 10 років).

Медіа в ГРАКу представлені трьома типами періодичних видань: газети, журнали та інформаційні сайти. Корпусні інструменти дозволяють вибрати певну назву видання: DOC.MEDIANAME або тип видання: DOC.MEDIATYPE. Як вже зазначалося, кожне періодичне видання має інформацію про регіон, де воно публікується або публікувалося. У випадку якщо автор тексту у виданні невідомий, тоді тег регіону цього тексту відповідатиме місцеві видання.

В основі морфологічної розмітки корпусу лежить ВЕСУМ, який розроблений спеціалістами гурту r2u – Василем Старком та Андрієм Рисіним.

Тобто тексти корпусу тегуються на морфологічному рівні на основі ВЕСУМу, відповідно леми можна визначити лише для тих слів, які є в словнику. Кожне слово отримує певні граматичні теги. Наприклад, словоформа *вона* матиме такі теги – /|вона|/|noun:f:v_naz:&pron:pers:3|, де |вона| – це лема, /|noun: – іменник, f – жіночий рід, v_naz: – називний відмінок, &pron: – займенник, pers:3 – третя особа однини.

ГРАК представлений не тільки сучасними текстами, але й старими текстами XX-XIX століть. Відповідно це ставить перед збирачами матеріалу, а надалі перед науковцями, які формуватимуть висновки на основі своїх досліджень, певний перелік проблем, серед яких основні – проблема сучасних видань старих текстів і наявність кількох правописів у старих виданнях. Корпус поповнюється переважно текстами за сучасними чи радянськими виданнями і якщо відомо, то про це зазначається в паспорті тексту. Користувач має бути пильним при дослідженні текстів та інтерпретації даних, тому що текст може містити не тільки орфографічні, граматичні, але й лексичні виправлення. У випадку, коли відомо про редакторське втручання в текст, то тоді зазначається автора тексту, назву, рік редагування і тільки після цього первинний рік створення. ГРАК містить невелику кількість текстів за первинним, старим виданням зі збереженням правописом. Тексти в корпусі створені за трьома орфографічними системами: скрипниківкою, желехівкою, сучасним правописом. Морфологічна анотація текстів за желехівкою відбувається лише частково. Таким чином програма лематизує такі слова, леми яких є у ВЕСУМі. До прикладу, словоформа *мякий* отримає лему *м'який*, а словоформі *цїлком* буде присвоєна лема *цілком* тощо.

Програма NoSketch Engine – це корпусний менеджер, доступний для вільного використання, за допомогою якого користувач здійснює пошук за лемою (lemma), словоформою (word) і граматичними тегами. Якщо дослідник хоче побудувати складний пошуковий запит, наприклад, здійснити пошук тексту, час написання якого 1920 рік, місце написання — Київ, стать автора — чоловіча, то це він може зробити, використавши CQL-код. Також можливою опцією є формування власного підкорпусу, у якому користувач може побудувати різні частотні

словники, шукати в текстах словоформу, лексему, граматичну ознаку, будувати складні запити для сполучень кількох слів, обробляти результати пошуку.

Користувач бачить результати пошуку у вигляді конкордансу. Його вигляд можна змінити, вибравши в меню зліва «Вигляд». Важливо те, що при зміні інтерфейсу, він змінюється для всіх. Користуючись цієї опцією, дослідник може вибрати інформацію про текст, яка буде зображатися в конкордансі. Таким чином, можна вибрати, щоб було видно автора, назву, рік написання тексту, жанр, регіон тощо. У випадку, коли потрібна повна інформація про текст, потрібно натиснути на синій рядок з назвою.

Наступною опцією ГРАКу є будування частотного списку. Для того, щоб скористатися цією можливістю, користувачеві потрібно на сторінці конкордансу натиснути «Частота» (у меню зліва). На наступній відкритій сторінці потрібно вибрати атрибут, залежно від мети пошуку: якщо потрібний частотний список словоформ, то треба вибрати атрибут word, якщо потрібний частотний список лексем, то вибираємо lemma. Після цього користувачеві треба натиснути «Побудувати частотний список».

Не менш цікавою та важливою особливістю ГРАКу є побудова частотних графіків, яка слугує чудовим інструментом у вивченні мовних явищ. Графік показує співвідношення кількох мовних явищ щодо років, що дозволяє прослідкувати, наприклад, функціонування фемінітивів в текстах.

Новою можливістю ГРАКу (GRAC v. 10) є анотування найчастотнішої лексики корпусу за семантичною розміткою. В основі цієї розмітки лежить присвоєння семантичних ознак слову. Наприклад, прикметнику *малий* присвоюється три значення, й кожне позначено іншим тегом: **size** (розмір), **age** (вік) та **degree** (ступінь).

Таким чином, ГРАК є добрим інструментарієм в дослідженні української мови завдяки його детальній розмітці, можливості укладати різноманітні підкорпуси та здійснювати пошук, будувати графіки. Науковці можуть вивчати

регіональні відмінності в українській мові, щоб системно описати українські регіональні стилі. ГРАК дозволяє дослідникам порахувати й порівняти частотність слів, а також граматичних одиниць в регіонах, що дає можливість глибоко проаналізувати функціонування мови, а також знайти нові підходи до її лексикографічного та граматичного опису.

### 1.3. Публіцистика в українських корпусах. Публіцистичний підкорпус ГРАКу

Зі зростанням масової комунікації та інформаційних технологій в другій половині XX - на початку XXI ст. розвивається великий інформаційний простір, який залишив свій слід на мові та мовленні. Тому публіцистика займає важливе місце в системі функційних стилів мови й мовлення. Оскільки вона є стилем громадського життя і нею послуговуються ЗМІ, то, відповідно, відбиває актуальні проблеми, і таким чином відзначається своєю експресивністю, а також оцінним характером. С. Єрмоленко, Л. Мацько, П. Дудик, О. Стишов, М. Навальна, О. Пономарів та ін. досліджували природу та характерні риси публіцистичного стилю. Варто зазначити, що цей стиль охоплює досить велику кількість площин: суспільно-культурну, засоби масової інформації, публічну площину і, звісно, громадсько-політичну. М. Навальна у своїй роботі вказала, що основне призначення публіцистичного стилю «не тільки інформування читачів про суспільно-політичне життя, а й формування громадської думки. Ефективність соціально-політичного впливу на читача пов’язана з посиленням логічного аспекту висловлення і водночас з емоційним напруженням викладу»[20].

Цінність та потрібність публіцистичного стилю полягає в його реалізації стандарту літературної мови. Оскільки функціонування стилю досить широке, українська мова відчуває свою силу та престижність, входячи в користування широких кіл громадськості. Т. Коць згадує такі важливі ознаки цього стилю: стисліcть, логічність викладу з одночасним інформативним навантаженням, вживання слів та висловів із прозорою семантикою, використання суспільно-

[20] Марина Навальна. *Лексика української газетної періодики початку XXI ст.: джерела поповнення та стилістичне використання:* монографія. Переяслав-Хмельницький 2018, с. 16.

політичної термінології, мовних кліше, штампів, переосмислення лексики й трансформація фразеологізмів[21], тобто це свідчить, що публіцистичному стилю притаманна дуже розмаїта лексика – від застарілих і діалектних слів, разом із мовними штампами до сленгізмів.

Як зазначається, мова – це як живий організм, вона росте, змінюється, реагує на трансформації суспільства, світу. Відтак, на думку Л. Архипенко, на фоні появи нових явищ, понять та постійної зміни та руху саме публіцистичний стиль першим «опиняється перед потребою словесного оформлення нових реалій»[22]. Тому публіцистика – це добре поле для досліджень, оскільки дає змогу простежити зміни в мовленні: занепад одних слів, появи інших, а також дозволяє дослідити розмовну мову, адже промови, інтерв'ю є одними із жанрів публіцистики.

Публіцистика реалізується через певні мовностильові засоби. Тобто їй приманний певний стиль: чітка логіка викладу, ґрунтовний аналіз та орудування реальними фактами разом з образністю, яка найвиразніше помітна в журналістських жанрах. Суб'єктивна оцінка, ставлення автора до висловленого, яке передбачає використання емоційно забарвленої лексики, використання водночас книжної та розмовної мови, змішування стандартизованої й експресивної лексики є притаманними публіцистиці. Синоніми, антоніми, пароніми, неологізми, історизми, екзотизми забезпечують значну варіативність, багатство лексики, в чому полягає й особливість публіцистичного тексту. Наступною рисою публіцистики є використання риторичних засобів, тобто риторичних питань, окличних речень, вставних конструкцій, звертань. Важливим фактом є те, що з публіцистичних творів в українській мові виокремилися стійкі словосполучення. Висновуючи з вищесказаного, мовними особливостями публіцистичних текстів є принципова «неоднорідність стилістичних засобів, використання спеціальної

[21] Тетяна Коць. *Публіцистичний стиль в українській літературній мові кінця XIX – початку XXI ст.: нормативно-аксіологічний аспект*: Дисертація на здобуття наукового ступеня доктора філологічних наук / Інститут української мови НАН України. Київ 2019, с. 2.
[22] Людмила Архипенко. Мова ЗМІ як об'єкт лінгвістичних досліджень: історія становлення, специфіка функціонального стилю // *Культура народов Причерноморья. № 101*, 2007, с. 74.

термінології та емоційно забарвленої лексики, поєднання стандартних і експресивних засобів мови, використання й абстрактної, і конкретної лексики»[23].

Як зазначалося, публіцистичні тексти мають високу цінність, тому їхнє значення в корпусі важко переоцінити. Наприклад, проєкт «Лінгвістичний портал mova.info» містить Корпус української мови, в якому підкорпус публіцистики становить 17 млн слововживань; у БрУК категорія текстів Преса (репортажі, огляди, редакційні статті, листи до редакції; національні й регіональні видання; тематично – політика, спорт, суспільство, економіка й фінанси, короткі новини, культура – театр, література, музика, танці) становить 25% всіх текстів в корпусі. А в проєкті «Лабораторія української» розподіл за жанрами має такий вигляд: репортажі, промови, оголошення, твіти складають 6% всіх текстів, а новини 5,6%, коли наукові праці 2,9 %, енциклопедичні статті – 4,6 %. Наразі підкорпус публіцистичних текстів ГРАКу нараховує 88 254 161 токен, 68 660 464 словоформи. Таким чином, публіцистичні тексти є вагомою частиною корпусу, оскільки допоможуть дослідникам розв'язати певні питання, а то й поставити абсолютно нові, які раніше в науці не виникали. Тому потрібність публіцистики в корпусах, зокрема в ГРАКу, є актуальною.

## Висновки до І-го розділу

Отже, корпусна лінгвістика – це дуже динамічна галузь науки, яка стрімко розвивається у світі. Вона дозволяє вичерпно і глибоко дослідити мову і мовлення за допомогою інформаційних технологій. Центральним в корпусній лінгвістиці є корпус текстів, який слугує добрим інструментом для дослідження. Свідченням розвитку корпусної лінгвістики в Україні є кілька корпусів, серед яких найкращим є ГРАК, який стрімко оновлюється. В цьому розділі здійснено детальний опис особливостей структури та функціонування українського корпусу ГРАКу й окрему увагу приділено аналізу публіцистичним підкорпусам українських корпусів. Також проаналізовано особливості публіцистичного стилю й обґрунтовано потребу включати тексти цього стилю в українські корпуси, зокрема в ГРАК.

[23] Марина Нетреба. Стилістичні особливості публіцистичних текстів // М. М. Нетреба. *Інформаційне суспільство* 2015, с. 7.

# РОЗДІЛ ІІ

# ОСОБЛИВОСТІ МОВИ ГАЗЕТИ «ДЕНЬ»

Цей розділ присвячено огляду наявних досліджень мови газети «День», тексти якої буде досліджено у практичній частині. Розглянемо особливості стилю газети й окреслимо роль газети в українських ЗМІ. Також буде здійснено опис процесу редагування текстів газети «День» для додання їх до корпусу ГРАК.

## 2.1. Попередні дослідження мови газети «День»

Мова газетних видань як жанрового різновиду публіцистичного стилю зображає актуальний стан української літературної мови та загальний стан національної культури, оскільки журналістський текст повністю живе в мові та реалізується в ній. Як вже зазначалося, лексика ЗМІ дуже чутлива як і до внутрішніх, так і до зовнішніх впливів, тому вона постійно еволюціонує. Дослідження мови українських періодичних видань стало популярною та важливою цариною в сфері лінгвістичних досліджень, оскільки, аналізуючи публіцистику, можна отримати дані про актуальний стан мови: наприклад, визначити частотність вживання запозичених слів, новоутворів, окреслити лексику, яка характеризує тексти тих чи інших років.

Всеукраїнська газета «День» викликає неабиякий інтерес в дослідників. Варто виділити роботи Олени Ткаченко, Ірини Лебедь, Христини Білограць, Наталі Прокопенко, які досліджували особливості мови та змісту вже згаданої газети з різних перспектив.

За результатами дослідження Христини Білограць та Ірини Лебедь[24], газета «День» вирізняється своїми цікавими та захопливими назвами статей. Варто згадати, що заголовки в ЗМІ дуже важливі, оскільки це елемент привернення уваги. Саме від того, наскільки влучно і виразно підібраний заголовок, залежить, чи

[24] Ірина Лебедь. Заголовки публікацій у ЗМІ на прикладі газети «День» і журналу «Країна» / І. Лебедь, Х. Білограць. *Вісник Національного університету «Львівська політехніка».* Серія: Журналістські науки. Вип. 4, 2020, с. 63-67.

потенційний читач зверне увагу на журналістський текст. Невдалий заголовок часто стає причиною того, що цей текст залишається поза увагою та прочитанням читача. Дослідниці, вивчаючи заголовки суспільно-політичної газети «День», дійшли висновку, що в цьому виданні «активно використовуються інформаційні, спонукально-наказові, проблемні, констатуючо-описові та рекламно-інтригуючі заголовки»[25]. Аби зацікавити читача, журналісти послуговуються стилістично забарвленою лексикою, застосовують мовну гру. В заголовках «Дня» можна знайти метонімію – *«Прочитати Донбас»*, тавтологію – *«Чи буде українським Українське море»*, метафори – *«У пеклі протікає кран»*, риторичні запитання – *«Воно нам треба?»*, фразеологізми – *«Як із козла молока»*, антитези – *«Маленькі речі – великі теми»* та низку інших художніх засобів. Також журналісти використовують різноманітні лексичні засоби та графічні зображення, наприклад, хештеги – *#ЧужихДітейНеБуває*, *#СтопБулінг*; плюси, знак рівності – «*підтримка + критика = критична підтримка»*; виокремлюють текст різними кольорами, щоб привернути увагу потенційного читача. До того ж заголовки пишуться лексемами з професійної та запозиченої лексики, наприклад, *«Плюралізм гри в монополію»*. Звісно, вживання іншомовних слів є небезпекою, оскільки є ймовірність, що незнайоме слово відштовхне читача і відповідно, стаття залишиться непрочитаною. Але важливо, що автори статей передбачають це і, у випадку, якщо іншомовне слово справді недостатню відоме читацькій авдиторіїї й обтяжує сприйняття тексту, залишають короткий коментар під заголовком. Прикладом може бути такий заголовок: *««Політичний цугцванг», де «цугцванг» – шаховий термін, який означає примушення до здійснення ходу»* (Газета «День», 2018).

Отже, газета «День» відзначається вживанням забарвленої лексики у заголовках: професійна лексика, іншомовні слова, художні засоби та риторичні фігури. Це, безперечно, збільшує шанси тексту на прочитання.

Важливим питанням для дослідження є вживання запозичених слів в текстах газетних видань. До цієї теми звернулися Наталія Прокопенко та Анастасія

[25] Ірина Лебедь. Заголовки публікацій у ЗМІ на прикладі газети «День» і журналу «Країна», с. 63.

Дєдова у праці «Використання іншомовної лексики в текстах газети «День»: проблемний підхід»[26] і дослідниця Солодовник Дарія в роботі «Використання іншомовних слів у газеті «День»»[27]. Небезпека вживань запозичених слів полягає в тому, що через перенасичення тексту цими словами читач відмовиться від його прочитання, оскільки відчуватиме себе загубленим у понаднормовій кількості незнайомих слів. Провівши аналіз публікацій «Дня», Солодовник Дарія дійшла висновку, що автори текстів вживають велику частину запозиченої лексики, серед якої англіцизми (наприклад, *імідж*, *мас-медіа*, *саміт*, *тендер*, *спічрайтер*, *шейпінг*, *лізинг*, *хіт*) становлять найбільшу групу. Також спостерігається велика частина запозичень з латини, яка вживається на позначення понять політичної, фінансово-економічної й юридичної царин та інші: *адвокат*, *резолюція*, *консорціум*, *санкція*, *субсидія*. Невелику частину складають запозичення з французької мови: *макіяж*, *кутюр'є*, *фуршет*, *фритюр*, *бутик*. Цікавим є дослідження Наталії Прокопенко й Анастасії Дєдової, які аналізували статті газети «День», виокремлюючи елементи, де вживання запозичених слів було виправдане, а де іншомовна лексика переобтяжувала текст. Таким чином авторки наводять матеріал журналістки Людмили Засєди як ілюстрацію доречного вживання іншомовного слова: *Не впевнена, чи варто писати про щось релаксне, якщо протягом останніх місяців посварились найстійкіші* (Газета «День», 2019). Однак також спостерігається невдале тавтологічне вживання в одному реченні слова *експерт* та похідних від нього слів, що обтяжує сприйняття інформації і спотворює звучання. Для підтвердження цієї тези авторки праці наводять низку прикладів вживання згаданого слова у журналістських текстах, серед яких є такий: *Окрім того, жінки не так часто стають суб'єктами новин, а щодо експертної думки, практично в усіх галузях кількість жінок-експертів значно нижча, ніж їхніх колег-чоловіків* (Газета «День», 2019). Проаналізувавши функціонування слова в наведених

[26] *Журналістська освіта в Україні: світові професійні стандарти* / уклад. О. Г. Ткаченко. Суми 2019, с. 7-11.
[27] Дарія Солодовник. Використання іншомовних слів у газеті «День» / Д. Солодовник, Н. І. Голубінка. *ЗМІ та демократичний розвиток України: збірник матеріалів ІІІ Всеукраїнської конференції студентів та молодих дослідників*. Львів 2018, с. 178–184.

реченнях, дослідниці висновують, що авторам/редакторам текстів потрібно чергувати вживання іншомовного слова та його українського відповідника або підбирати слова із синонімічного ряду. З метою перевірити тексти, в яких спостерігається активне вживання іншомовної лексики, на адекватність сприйняття читачами, дослідниці провели анкетування, опитавши 15 людей. Ідея дослідження полягала в тому, що респонденти, прослухавши уривки тексту з іншомовними словами, повинні були пояснити їх значення, дібрати українські відповідники або ж відповісти, чи вживання того чи іншого запозичення в наданому реченні доречне. Навівши перелік прикладів уривків з іншомовною лексикою з текстів газети «День», які брали участь в дослідженні, а також відповіді реципієнтів, Наталія Прокопенко й Анастасія Дєдова дійшли висновку, що іноді запозичені слова ускладнюють сприйняття тексту, тому доречно підбирати українські відповідники до цих слів, проте все ж таки «деякі визначення доцільно вживати у вихідному варіанті»[28], адже вони точно доносять сенс повідомлення і допоможуть уникнути двозначності.

Серед переліку досліджень газети «День» важливе місце посідає робота Наталі Прокопенко у співпраці з Каріною Хачатар'ян «Діалогічність як фундаментальна ознака мови газети «День»»[29]. Авторки аналізують стилістичні прийоми взаємодії та мету використання мовних одиниць у діалогічному дискурсивному газетному середовищі. Діалогічність – невіддільна особливість газети «День», оскільки в кожному п'ятничному випуску є інтерв'ю з певним фахівцем. Розмови між журналістом та респондентом побудовані переважно у формі неформальної бесіди або репортажу, що спрощує процес сприйняття тексту читачем і не перевантажує читача інформацією. Слушно відзначити успішну традицію видання – запрошувати на розмови фахівців в тих чи інших сферах, тому гостями редакції часто є представники різних країн світу, консолідацій, що

[28] Журналістська освіта в Україні: світові професійні стандарти / уклад. О. Г. Ткаченко. Суми 2019, с. 10.
[29] Наталія Прокопенко. Діалогічність як фундаментальна ознака мови газети «ДЕНЬ» // Н. М. Прокопенко, К. А. Хачатар'ян. *Філологічні трактати. Т. 10, № 1*, 2018, с. 58-65.

пояснює популярність газети серед читачів. Як зазначають дослідниці, діалогічне мовлення видання характеризується неповнотою висловлювань, еліптичністю, складною структурою речень, розлогими думками, використанням окличних, іноді спонукальних речень. На думку авторок, журналісти найчастіше вживають такі типи діалогів (за комунікативною метою): інформативні, конфліктні, оцінні, імперативні діалоги у формі повідомлення, констатації фактів, суперечок, звинувачень, обурень, аргументацій, критики, похвали, переконування. Серед основних особливостей діалогічного мовлення газети «День» є низький ступінь імпровізованості, поміркована експресія і високий прагматизм. Аргументація, переконання відбуваються при дотриманні норм етикету. Важливим висновком, до якого дійшли авторки праці, є твердження, що структурній організації діалогів притаманна гендерна розрізненість. Проте, як твердять дослідниці, оброблювання мовного матеріалу стирає ці відмінності між жіночим та чоловічим спілкуванням. Завдяки цьому прагматизмові не відбувається проникнення у внутрішній світ партнера, що уможливлює неупереджене ставлення та формулювання якісних та фахових коментарів.

В широкому переліку праць, які стосуються дослідження газети «День», особливе місце посідає робота Олени Григорівни Ткаченко «Державна мовна політика України у висвітленні газети «День»»[30]. Це дослідження є важливим, оскільки тема державної мови є гарячою й актуальною в українському суспільстві, а ЗМІ є одними із засобів, які беруть участь у формуванні мовної політики. І газета «День» не залишається осторонь цієї важливої проблеми. Редакція тижневика вважає, що державна мовна політика – це невіддільна складова національної політики та національної безпеки держави, тому весь зміст, особливість якого в тематичному та жанровому розмаїтті матеріалів, пронизаний єдиною ідеєю: мова – це об'єднавчий чинник і мірило не тільки державності, але й рівня демократії, соціально-політичного устрою.

[30] Олена Ткаченко. Державна мовна політика України у висвітленні газети «День» // *OPERA SLAVICA, XXIV, 4*. Брно 2014, с. 44-49.

Отже, газета «День» слугує хорошим джерелом для дослідження та вивчення проблем, пов’язаних не тільки з синтаксично-стилістичними особливостями текстів, але й з питаннями суспільно-державного рівня.

## 2.2. Опрацювання текстів газети «День» для корпусу ГРАК

Предметом опрацювання цієї роботи стали тексти всеукраїнського видання «День». Вони були стягнуті автоматичним способом із сайту газети, очищені від html-кодів та інших позатекстових елементів та передані нам для подальшого їх опрацювання. В результаті було упорядковано понад 51 000 текстів загальним обсягом понад 47 221 118 млн токенів (словоформ, пунктуаційних знаків та спецсимволів). Нашим завданням було вилучити всі зайві елементи у файлах, залишивши лише назву та сам текст публікації. Нижче подано список елементів, які належало вилучити (ілюстрації в Додатку 1 на Рис. 1, 2, 3):

- ім’я та прізвище автора;
- розриви рядків;
- вставки в тексті на кшталт: *«День» №211, 1999 рік, Фото Олександра БУРКОВСЬКОГО*;
- рядок в кінці публікації такого типу: «*Підготувала Мар’яна ОЛІЙНИК, «День»»*;
- вилучення зайвого пробілу після дефісу, наприклад, *екс- міністр, віце- прем’єр, по- друге*;
- видалення інтернет-посилань тощо

Всі операції над текстами було здійснено через програму Notepad++, що давало змогу використовувати регулярні вирази та здійснювати масові заміни в тисячах файлах.

Опрацювання текстів газети «День» відбувалося в кілька етапів: упорядкування (1) назв статей, (2) підзаголовків, (3) розривів рядків і (4) вилучення позатекстових елементів, приклади яких наведено вище. Детально розглянемо перший блок, який полягав в опрацюванні назв статей і складався з кількох кроків. Завданням цього етапу було проставити два розриви рядка після назви статті для

того, щоб вже у відредагованому тексті після заголовку публікації з'явився порожній рядок. Спершу, скориставшись опцією *Find in Files* (комбінація клавіш CTRL+F), за допомогою регулярного виразу – *\A^( *)*[31], введеного в графу *Find what*, відбулося охоплення зайвих пробілів перед назвою статті, які вдалося вилучити за допомогою заміни на ніщо у графі *Replace with* (див. Додаток 1, Рис. 4). Опісля, ввівши регулярний вираз *(?-s)\A^.*$*[32] у графу *Find what* із заміною на *$0%%%*[33] у полі *Replace with*, в кінці першого рядка було проставлено символ *%%%* і розрив рядка (див. Додаток 1, Рис. 5), вилучення якого було наступним і заключним етапом в цьому блоці (див. Додаток 1, Рис. 6). Отже, здійснивши всі перелічені операції, в кінець першого рядка додано технічний символ і вставлено наступний порожній рядок.

Також публікації 2011-2019 років рясніли підзаголовками всередині текстів, які теж потребували опрацювання. Варто зазначити, що цей етап упорядкування текстів був дещо ускладнений, оскільки деякі підзаголовки розпочиналися не зі слів, а з цифр або ж з лапок і відповідно кожен випадок потребував застосування окремих регулярних виразів.

Інструментарій програми Notepad++, а саме опцію здійснення масового пошуку та заміни елементів в тисячах файлів – *Find in Files* – було застосовано для опрацювання текстів на решти етапах (див. Додаток 1, Рис. 7-12).

Після очищення текстів від зайвих елементів наступним завдання було їх розділення на дві теки: інтерв'ю та неінтерв'ю. Суть цього поділу полягає в тому, що при внесенні текстів в корпус ГРАК тексти-інтерв'ю будуть введені окремо з розміткою як стиль SPO, тому їх потрібно відділити від інших. Важливо, щоб інтерв'ю були в окремій теці, оскільки вони максимально наближені до усного розмовного мовлення, а це дуже важливий тип матеріалу для корпусу.

---

[31] *\A^* – відповідає початку лише першого рядка у файлі, а *( *)* – охоплює всі пробіли до першої літери в рядку

[32] *(?-s)* – вмикає опцію новий рядок для решти регулярного виразу, *\A^* – відповідає лише першому рядку у файлі, а *.*$* – охоплює цілий рядок (за винятком закінчень рядків)

[33] *$0* – це вміст знайденого рядка, а *%%%* – технічний символ

Отож, вилучивши всі позатекстові елементи та розділивши публікації на дві теки: інтерв'ю і неінтерв'ю, тексти були готові для додавання в корпус. Опрацьовані нами статті газети «День» 2000-2019 років були внесені до ГРАКу-11 15 лютого 2021 року.

## Висновки до ІІ-го розділу

Отже, газета «День» слугує добрим джерелом для дослідження публіцистичного стилю української мови, про що свідчать вже наявні публікації українських дослідників, в яких здійснений різнобічний огляд статей газети. Для того, щоб окреслити особливості зазначеного видання та довести необхідність його подальшого дослідження, здійснено огляд наявних праць, в яких вивчається певний мовний або стилістичний аспект, наприклад, вживання іншомовних слів, особливості написання заголовків або ж діалогічний дискурс. Також було описано процес підготовки текстів газети «День» 2000-2019 років для додання їх в корпус ГРАК з детальним аналізом етапів, які були здійснені за допомогою програми Notepad++.

Таким чином, в цьому розділі було проведено аналіз попередніх досліджень газети «День» та окреслено етапи опрацювання текстів видання «День» 2000-2019 років.

# РОЗДІЛ ІІІ

# КОРПУСНО-СТАТИСТИЧНИЙ АНАЛІЗ І ПОРІВНЯННЯ ЛЕКСИКИ ПУБЛІЦИСТИЧНОГО СТИЛЮ З ТЕКСТАМИ ІНШИХ СТИЛІВ

У цьому розділі буде описана практична частина дослідження лексичних особливостей публіцистичного стилю. Для того, щоб розуміти, як відбувався аналіз лексики, буде здійснено опис статистичних метрик, а також пояснено основні принципи методу ключових слів, на якому базується дослідження.

Аналіз здійснюється на матеріалі корпусу ГРАК, корпусний менеджер якого дозволяє створювати підкорпуси з необхідними текстами для порівняння текстів публіцистичного стилю з текстами інших стилів.

## 3.1. Побудова фокусного підкорпусу

Фокусний підкорпус включає тексти газети «День», оскільки дослідження лексичних особливостей тексту здійснюється саме на прикладі цього видання. Оскільки ми опрацьовували публікації «Дня» 2000-2019 років, то й до фокусного підкорпусу входять тексти саме цих років. Назва створеного підкорпусу — «DAY_foc_2000_2019». Багатоаспектна розмітка текстів ГРАКу дала можливість створити згаданий підкорпус за необхідними параметрами. Користуючись атрибутом *DOC.MEDIANAME,* ми вказали назву видання, тексти якого повинні складати фокусний підкорпус. А за допомогою менеджера ГРАКу обрали публікації тих років, які повинні бути в корпусі, а саме — на сторінці пошуку у *DOC.DATE* ввели регулярний вираз «*200.|201.»*, де крапка замінює один символ (в цьому випадку — цифру).

Таким чином, фокусний підкорпус текстів газети «День» 2000-2019 років «DAY_foc_2000_2019» має обсяг 47 221 118 токенів, ~ 36 883 847 словоформ.

## 3.2. Побудова референтного підкорпусу

Позаяк мета дослідження полягає в тому, щоб визначити особливості публіцистичного стилю на прикладі газети «День» 2000-2019 років, то фокусний

підкорпус, який складається з текстів публіцистичного стилю, доцільно порівнювати з референтним підкорпусом, наповненим текстами, власне, непубліцистичного стилю. Беручи до уваги графік наповнення ГРАКу-11 за кількістю токенів, типами текстів та за роками, робимо висновок, що публіцистика займає дуже велике місце в корпусі, збільшуючи свій обсяг та значно відриваючись від інших типів з 2000 року і до сьогодні, що дає підставу виокремити публіцистичні тексти для референтного підкорпусу.

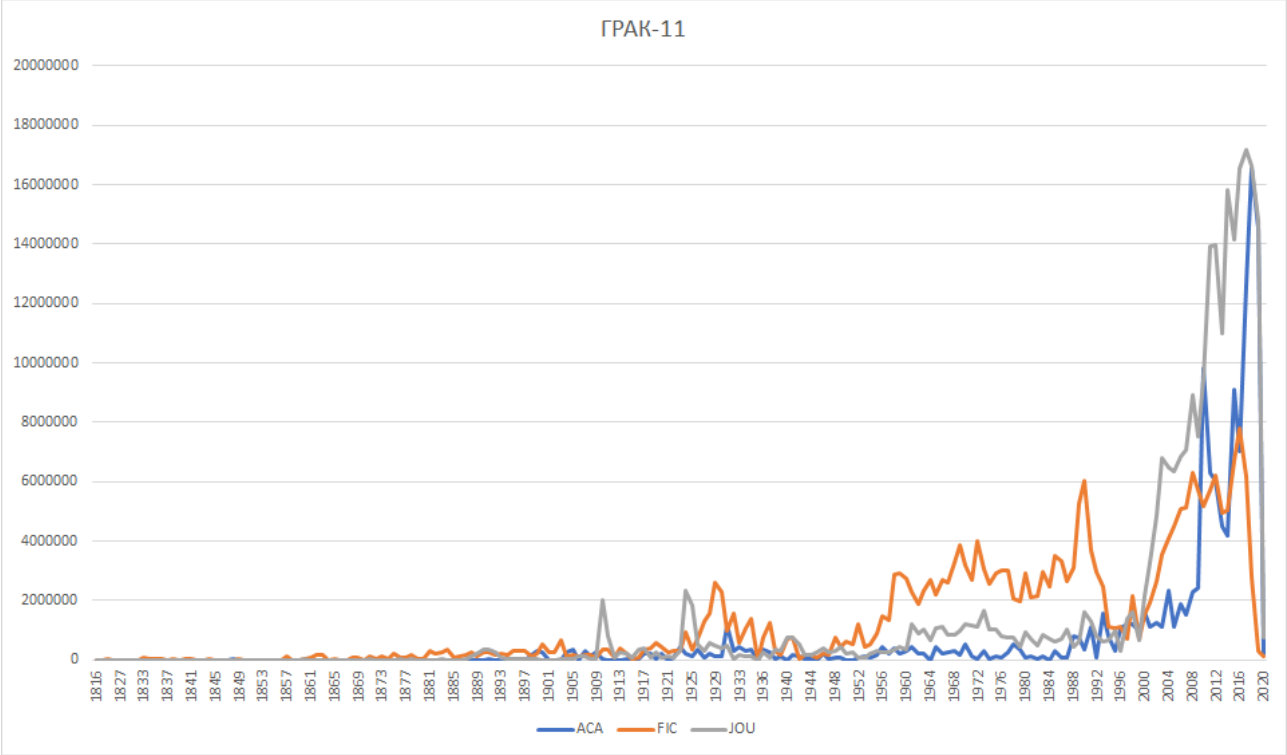


Рис. 3.1. Графік наповнення ГРАК-11 за типами текстів.

Тому, створюючи референтний підкорпус, назва якого «NON_JOUR_2000_2019», за допомогою зручного та багатофункціонального інтерфейсу, в атрибуті *DOC.STYLE* вказуємо всі стилі, окрім JOU — журналістика та SPO — усний, оскільки в останній стиль входять інтерв'ю, яким теж притаманні ознаки публіцистичного стилю.

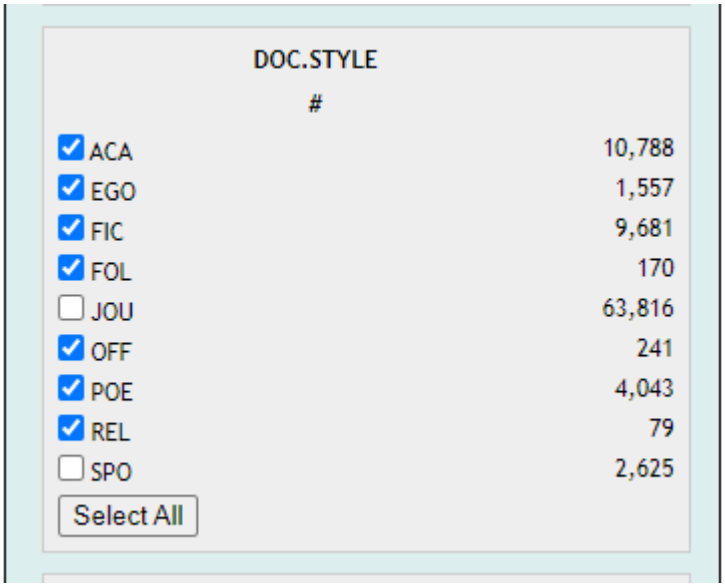


Рис. 3.2. Вибір стилів для референтного підкорпусу за допомогою атрибуту *DOC.STYLE*.

З огляду на те, що фокусний підкорпус містить тексти 2000-2019 років, відповідно для усунення чинника часу, який може негативного вплинути на дослідження, викрививши результати, для референтного підкорпусу теж потрібно вибрати тексти 2000-2019 років. Для цього в атрибуті *DOC.DATE* вводимо регулярний вираз «*200.|201.*».

Таким чином отримуємо референтний підкорпус «NON_JOUR_2000_2019» непубліцистичних текстів обсягом 213 675 507 токенів, ~ 166 899 368 словоформ.

### 3.3. Поняття ключовості

Дедалі частіше науковці з галузі корпусної лінгвістики цікавляться новим методом дослідження текстів — аналіз ключовості (англ. *keyness*) тексту та ключових слів. Їхні спостереження є дуже важливими для корпусного мовознавства, оскільки сприяють визначенню особливостей, ознак того чи іншого типу тексту, дозволяють визначити певні слова, які відрізняють певний корпус від іншого. Поняття ключовості, як воно розуміється в корпусній лінгвістиці, ввели в середині 90-х років минулого століття, а процедура аналізу ключовості вперше

була включена в Wordsmith Tools[34]. Майк Скотт ввів термін «ключове слово», що визначається як «слово, яке трапляється з незвичною частотою у певному тексті у порівнянні з якимсь референтним корпусом»[35]. Тобто ключовими можна вважати ті слова, частота використання яких перевищує певні норми. Дослідження ключових слів передбачає певну процедуру. Спершу ми підраховуємо випадки вживання певного слова в кожному корпусі — у фокусному й референтному, потім ділимо кожне число (кількість вживання певного слова у фокусному підкорпусі; кількість вживань того ж слова у референтному) на кількість слів у цьому корпусі, за бажанням множимо на 1 000 або 1 000 000, щоб отримати частоти на тисячу або мільйон, і ділимо одне число на інше, щоб отримати співвідношення (оскільки тисячі або мільйони скасовуються, коли ми робимо ділення, немає різниці, використовуємо ми тисячі чи мільйони). Для корпусних лінгвістів корисно знайти співвідношення для всіх слів і відсортувати їх за співвідношенням, щоб знайти слова, які найбільше пов'язані з кожним корпусом на відміну від іншого[36]. Тобто для визначення і дослідження ключових слів необхідно порівняти частотність слів.

Як вже зазначалося, нашим завданням є виокремити та дослідити особливості саме публіцистичного стилю, що й ми виконуватимемо, вдаючись до аналізу ключових слів й інтерпретації отриманих даних. Ця розвідка передбачає кілька етапів:

- генерування списку потенційних ключових слів за допомогою корпусного менеджера ГРАКу — платформи NoSketch Engine;
- визначення ключових слів, які є статистично значущими (процес відбору зі списку потенційних ключових);

[34] Wordsmith Tools – інтегрований набір програм для вивчення поведінки слів у текстах. Користувач має змогу використовувати інструменти, щоб дізнатись, як слова використовуються у його текстах або в текстах інших.
[35] Costas Gabrielatos. Keyness analysis: Nature, metrics and techniques / ред. Taylor, C. & Marchi, A. / *Corpus Approaches To Discourse: A critical review*. Milton 2018, с. 225.
[36] Adam Kilgarriff. Simple maths for keywords // *Proceedings of the Corpus Linguistics Conference*. Liverpool 2009, p.1.

- інтерпретація даних: групування слів за подібністю, пояснення відмінної частотності тих чи інших слів у фокусному й референтному підкорпусах.

ГРАК має зручну функцію генерування списків потенційних слів, яку ми й використовуємо для цього дослідження. Варто зауважити, що ці списки генеруються з використанням метрики величини ефекту під назвою *відношення зі згладжуванням*, яка показує, наскільки певне слово у фокусному підкорпусі є більш частотне напротивагу частотності того ж слова у референтному підкорпусі. Важливість цієї метрики в тому, що вона включає зміну, яка дозволяє користувачеві сфокусувати увагу на словах вищої або нижчої частоти. Ідея цього полягає в тому, що користувач, вказуючи коефіцієнт згладжування – 100, 1000, 1000…, отримує на вершині списку слова з більшою частотою, тобто більш вживані слова, тоді як при нижчому значенні – 0,1, 1 – увага фокусується на низькочастотних (більш рідкісних слова).

Отож оцінка ключовості слова обчислюється за такою формулою:

$$\frac{fpm_{rmfocus}+N}{fpm_{rmref}+N},$$

Рис. 3.3. Формула для визначення ключовості слова.

де $fpm_{rmfocus}$ — відносна частотність слова (на мільйон) у фокусному підкорпусі, а $fpm_{rmref}$ — відносна частотність слова (на мільйон) у референтному підкорпусі. Тоді як N — це коефіцієнт згладжування, який за замовчуванням дорівнює 1. Спершу для обраного слова слід отримати абсолютну частотність, підрахувавши випадки вживання цього слова спершу у фокусному підкорпусі. Після цього, для визначення відносної частотності, потрібно поділити отримане число (значення абсолютної частотності вибраного слова) на загальну кількість слів у цьому корпусі, після цього додавши коефіцієнт згладжування N (це може бути 1 за замовчуванням, або, як зазначалося, можна додавати й 0,1, 100, 1000 чи 10000, щоб

подивитися на різні частини списку) ($fpm_{rmfocus}$ + N). Опісля необхідно поділити це число на відносну частотність того ж слова в референтному корпусі і додати коефіцієнт згладжування ($fpm_{rmref}$ + N). Отриманий показник (score) показує, у скільки разів частіше це слово трапляється у фокусному підкорпусі.

Вищезгадана метрика під назвою «відношення зі згладжуванням» належить до метрик, які вимірюють величину ефекту. Ці метрики вказують на «величину» спостережуваної знахідки, тобто показують, наскільки різниця, яку ми знайшли між частотою одного й того ж слова в двох корпусах, є сильною чи слабкою. Тоді як метрики для визначення статистичної значущості не виявляють величини різниці частот, але вони дають змогу встановити, з якою ймовірністю виявлені в корпусі дані могли трапитися за умови правдивості нульової гіпотези і, відповідно, з якою ймовірністю вона продиктована реальними відмінностями в досліджуваній ділянці. З цієї причини для аналізу ключових слів варто застосовувати метрики величину ефекту, і, вже маючи списки потенційних ключових слів, перевіряти їх на статистичну значущість, таким чином визначаючи справді ключові слова для поглибленого аналізу.

Користуючись інтерфейсом ГРАКу (рис.3.3), застосовуючи метрику для вимірювання величини ефекту «відношення зі згладжуванням», ми згенерували списки потенційних ключових слів спочатку за лемами і, зважаючи на омонімію лем, також згенерували списки потенційних ключових слів за словоформами.

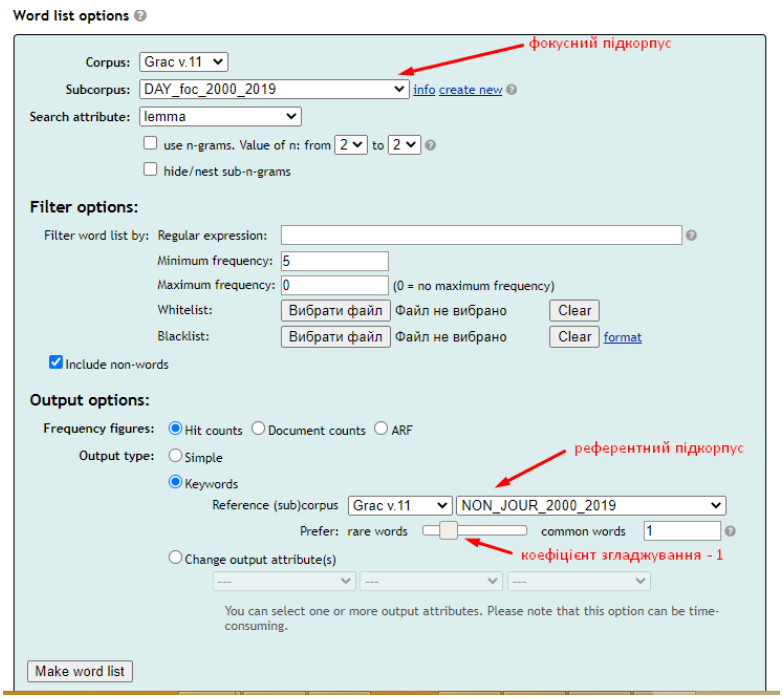


Рис. 3.4. Генерування списку потенційних слів за лемами з коефіцієнтом згладжування 1.

Також, щоб подивитися на різні частини списку слів за частотностями, ми змінювали коефіцієнт згладжування на 10, 100, 1000, 10000, кожного разу генеруючи новий список.

| | | | | | | |
|---|---|---|---|---|---|---|
| 2 | # Word list | | | | | |
| 3 | | | | | | |
| 4 | # Corpus: grac11 | | | | | |
| 5 | # Subcorpus: DAY_foc_2000_2019 | | | | | |
| 6 | # Reference corpus: grac11 | | | | | |
| 7 | # Reference subcorpus: NON_JOUR_2000_2019 | | | | | |
| 8 | | | | | | |
| 9 | | | | | | |
| 10 | | | | | | |
| 11 | lemma | Freq | Freq/mill | Freq_ref | Freq_ref/mill | Score |
| 12 | ---------- | | | | | |
| 13 | \|Інтерфакс-Україї | 2193 | 46.4 | 6 | 0.0 | 46.1 |
| 14 | \|прес-служба\| | 3765 | 79.7 | 185 | 0.9 | 43.3 |
| 15 | \|Івшина\| | 2181 | 46.2 | 46 | 0.2 | 38.8 |
| 16 | \|фотовиставка\| | 2284 | 48.4 | 64 | 0.3 | 38.0 |
| 17 | \|Ющенко\| | 10945 | 231.8 | 1218 | 5.7 | 34.7 |
| 18 | \|кабмін\| | 3403 | 72.1 | 261 | 1.2 | 32.9 |
| 19 | \|БЮТ\| | 2235 | 47.3 | 107 | 0.5 | 32.2 |
| 20 | \|Путін\| | 11687 | 247.5 | 1472 | 6.9 | 31.5 |
| 21 | \|Гонгадзе\| | 1976 | 41.8 | 94 | 0.4 | 29.8 |
| 22 | \|віце-прем'єр\| | 2097 | 44.4 | 137 | 0.6 | 27.7 |
| 23 | \|Тимошенко\| | 7798 | 165.1 | 1304 | 6.1 | 23.4 |

Рис. 3.5. Вершина списку потенційних ключових слів за лемами з коефіцієнтом згладжування 1.

### 3.4. Метрики статистичної значущості

Наступним етапом цього дослідження є відбір ключових слів зі списків потенційних ключових слів, які вдалося згенерувати за допомогою інтерфейсу ГРАКу. Цей відбір здійснюється, користуючись метриками статистичної значущості (оскільки вони допомагають з'ясувати, чи є отримані дані результатом справжньої різниці між двома слова, а не випадковим явищем).

Ідея статистичної значущості випливає з наступного: отриманий результат, вважається статистично значущим, якщо ймовірність спостереження такого явища менша або дорівнює вибраному порогу значущості. У нашому випадку на Рис. 3.5 вище видно, що слово *кабмін* більш притаманне текстам публіцистичного стилю, оскільки у фокусному підкорпусі текстів газети «День» воно трапляється набагато частіше, ніж у референтному підкорпусі всіх типів текстів, крім публіцистичного, але, щоб бути впевненим, що це відбувається не лише через випадковість, потрібно виконати певні калькуляції – тести на статистичну значущість.

Важливим і ключовим в статистиці є поняття нульової гіпотези ($H_0$). Це термін, що використовується для позначення припущення за замовчуванням в більшості статистичних тестів, а саме, що всі варіації у певній вибірці зумовлені випадковими причинами[37]. Потім нульова гіпотеза перевіряється на основі альтернативної гіпотези або $H_1$, яка, як правило, говорить про те, що спостережувана різниця не зумовлена випадковістю.

Для своїх досліджень корпусні лінгвісти можуть користуватися кількома типами тестів на статистичну значущість: хі-квадрат тест (chi-square або $\chi^2$), точний тест Фішера (Fisher exact test), логаритмічна правдоподібність (log-likelihood), Баєсів інформаційний критерій (Bayesian Information Criterion), t-критерій Стьюдента, t-критерій Вілкоксона та інші.

Хі-квадрат часто використовуваний в корпусній лінгвістиці тест на визначення статистичної значущості. Цей непараметричний тест порівнює різницю

[37] Gard B. Jenset. Basic statistics for corpus linguistics: *Handout for methods seminar in English linguistics* // Bergen 2008, p. 5.

між спостережуваними частотами в даних з тими, які можна було б очікувати, якби не діяв жоден інший фактор, крім випадковості (очікувані частоти). Чим ближче ці два результати один до одного, тим більша ймовірність того, що на спостережувані частоти впливає лише випадковість. Варто зазначити, що хі-квадрат тест є ненадійним, коли йдеться про обчислення дуже малих цифр. Наступне обмеження тесту хі-квадрат – це неможливість використання пропорційних даних (відсотків тощо)[38].

$$x^2 = \sum \frac{(O-E)^2}{E},$$

Рис. 3.6. Формула для обчислення значення тесту хі-квадрат.

де O – це спостережувані частоти, E – це очікувані частоти, а ∑ – означає, що потрібно виконати обчислення для кожного окремого елемента у досліджуваному наборі даних і підсумувати отримані результати.

Розрахувавши значення хі-квадрата за формулою вище, потрібно заглянути в набір статистичних таблиць, щоб побачити, наскільки значущим є значення хі-квадрат. Наступне необхідне значення – число ступенів вільності (degrees of freedom), яке теж вираховується дуже просто: *(кількість стовпців у табличці частот - 1) * (кількість рядків у таблиці частот - 1).* Після цього, користуючись таблицею значень хі-квадрат, у рядку для відповідної кількості ступенів вільності шукаємо найближче значення хі-квадрат до значення, яке ми обчислили, і зчитуємо значення ймовірності для цього стовпця. Чим воно ближче до 0, тим більшою є різниця, що означає, що це не обумовлено випадковістю. Значення, близьке до 1, означає, що різниця швидше за все обумовлена випадковістю. На практиці в корпусній лінгвістиці граничною точкою, яка приймається як різниця між

[38] Significance Testing
[Електронний ресурс], режим доступу:
https://www.lancaster.ac.uk/fss/courses/ling/corpus/Corpus3/3SIG.HTM

значущим і незначущим результатом, зазвичай є 0,01 або нижче, тобто значення ймовірності менше/дорівнює 0,01 записується як p ≤0,01 і вважається значущим.

Наступна метрика – точний тест Фішера, який є більш доречним при використанні невеликих наборів даних. Він належить до класу точних тестів, оскільки значущість відхилення від нульової гіпотези (наприклад, значення ймовірності) може бути обчислена точно, а не покладатися на приблизне число, яке стає точним у межах, коли обсяг вибірки зростає до нескінченності, як і в багатьох статистичних тестах. Серед його переваг є те, що він менш консервативний, ніж метрика хі-квадрат, тобто він може точніше виявити справжній зв'язок між даними. А обмеження застосування цього критерія стосується складності його розрахунку, який полягає в обчисленні факторіалів[39]. В цій роботі використання цього тесту виявилося недоцільним, оскільки значення частотностей досліджуваних слів високі (див. Додаток 2).

*Метрика логаритмічної правдоподібності (log-likelihood, LL)*, створена Полом Рейсоном, дозволяє здійснювати тести на значущу різницю в частоті між двома корпусами.[40]

Вирахувати значення LL можна за такою формулою[41]:

$$LL = 2*((a*\ln(a/E1)) + (b*\ln(b/E2))),$$

де a – це спостережувана частотність досліджуваного явища у фокусному підкорпусі; b – спостережувана частотність цього ж досліджуваного явища у референтному підкорпусі; E1 та E2 – це очікувані частотності досліджуваного явища у фокусному й референтному підкорпусах відповідно.

В основі цього тесту лежать чотири значення. Припустімо, що ми перевіряємо різницю між корпусом 1 та корпусом 2 у частоті якогось мовного

[39] Fisher's exact test
[Електронний ресурс], режим доступу: https://en.wikipedia.org/wiki/Fisher%27s_exact_test
[40] Statistics in corpus linguistics
[Електронний ресурс], режим доступу: http://corpora.lancs.ac.uk/clmtp/2-stat.php
[41] Log-likelihood and effect size calculator
[Електронний ресурс], режим доступу: http://ucrel.lancs.ac.uk/llwizard.html

явища, яке умовно позначимо X. Для такого розрахунку потрібні наступні числа: частота вживання X в корпусі 1, загальна кількість можливих вживань X в Корпусі 1 (кількість випадків, скільки воно було вжите, плюс кількість потенційних випадків, тобто коли це могло статися, але не відбулося), частота X у корпусі 2; загальна кількість можливих вживань X у Корпусі 2. У випадку, коли треба перевірити різницю у частоті слова, тоді замість «числа можливих вживань» потрібна загальна кількість слів в корпусі. Тоді коли необхідно розглянути частоту певного типу речень (наприклад, стверджувального на відміну від питального), кількістю можливих вживань була б загальна кількість речень у корпусі. Важливо розуміти, що всі показники повинні бути абсолютними частотами. Оскільки тест на статистичну значущість враховує розмір корпусу, тому не слід використовувати нормалізовані частоти як вхідні дані. Маючи 4 значення, про які сказано вище, їх можна вставити у формулу (Рис.3.7), яка доступна на сайті[42], та вирахувати значення за допомогою онлайн-калькулятора.

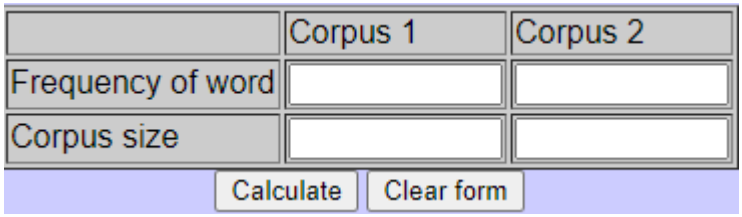


Рис. 3.7. Онлайн-калькулятор для обчислення значення логаритмічної правдоподібності.

Розрахуємо значення LL для слова *кабмін*, яке наведене вище на Рис. 3.5. Ввівши показники частотності з двох корпусів та обсяги цих корпусів, ми отримуємо результат у такому вигляді:

[42] Log-likelihood and effect size calculator
[Електронний ресурс], режим доступу: http://ucrel.lancs.ac.uk/llwizard.html

| Item | O1 | %1 | O2 | %2 | LL |
|---|---|---|---|---|---|
| Word | 3403 | 0.01 | 261 | 0.00 + | 9855.64 |

Рис. 3.8. Результати тесту на логаритмічну правдоподібність, отримані за допомогою онлайн-калькулятора.

де O1 і O2 – це спостережувані частоти, власне, цифри, які ми ввели; %1 і %2 – це спостережувані частоти у нормалізованій (відсотковій) формі; знак «+» свідчить про те, що слово частіше зустрічається в Корпусі 1 (знак «-» вказуватиме на те, що воно, навпаки, частіше в Корпусі 2); LL – це є результат тесту на логаритмічну правдоподібність, який вказує на те, чи справді різниця у частотах є статистично значущою. Інтерпретуючи дані, потрібно керуватися пороговими значеннями LL, які відповідають різним рівням значущості, які для своїх робіт встановлюють самі дослідники. Нижче наведені відповідності між значенням ймовірності і LL:

- $p \leq 0.0001$ відповідає LL ≥15,13 (дуже вагомий доказ);

- $p \leq 0{,}001$ відповідає LL≥10,83 (вагомий доказ);

- $p \leq 0{,}01$ відповідає LL≥6,63 (дані достатньо переконливі);

- $p \leq 0{,}05$ відповідає LL≥3,84 (дані переконливі)[43].

### 3.5. Відбір ключових слів із текстів газети «День»

Маючи списки потенційних ключових слів, потрібно відсіяти з них власне ключові слова, тобто перевірити отримані частоти за допомогою метрик статистичної значущості для подальшого поглибленого аналізу.

Першим завданням на етапі відбору ключових слів зі списків є встановлення порогу статистичної значущості. У соціальних розвідках науковці встановлюють його на рівні 0,5, проте в серйозних корпусних дослідженнях цей поріг становить

[43] Log-likelihood and effect size calculator
[Електронний ресурс], режим доступу: http://ucrel.lancs.ac.uk/llwizard.html

0,01 або й нижче[44]. Однією з найпопулярніших метрик в корпусних дослідженнях є метрика логаритмічної правдоподібності. Коли значення log-likelihood (LL) дорівнює або вище за 6,63, це відповідає рівню значущості *p*=0,01, що є мінімальним пороговим значенням для корпусних досліджень. Проте недоліком метрики LL є те, що вона залежить від обсягу корпусу, тобто чим більший корпус, тим більше значення LL. Тому ця метрика не є цілком об'єктивною, адже в дуже великих корпусах всі або більшість різниць частот будуть значущими.

З цієї причини для відбору ключових слів слід застосовувати критерій BIC (Bayesian Information Criterion, Баєсів інформаційний критерій), який обчислюється за такою формулою (Рис. 3.9.) (для тесту на логаритмічну правдоподібність з одним ступенем вільності).

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

,

Рис. 3.9. Формула для обчислення критерія BIC (ступінь вільності дорівнює 1).

де LL – це значення логаритмічної правдоподібності, N – розмір двох корпусів разом узятих (кількість токенів). У випадку, коли ступенів вільності (df) 2 або більше, Баєсів приблизний критерій обчислюється за такою формулою:

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

Рис. 3.10. Формула для обчислення критерія BIC (ступенів вільності 2 або більше).

Цей критерій можна швидко обчислити, користуючись таблицею[45], яка зображена на Рис. 3.11. В стовпець A потрібно ввести досліджувані слова, в стовпець B – частотності цих слів у фокусному підкорпусі, в стовпець C – частотності в референтному підкорпусі. Всі потрібні значення беремо з раніше

[44] Costas Gabrielatos. Keyness analysis: Nature, metrics and techniques, с. 239.
[45] Downloadable spreadsheet incorporating the log-likelihood calculation and the set of effect size measures [Електронний ресурс], режим доступу: http://ucrel.lancs.ac.uk/people/paul/SigEff.xlsx

згенерованих списків потенційних ключових слів. Таким чином, в стовпці K отримано значення BIC.

Керуючись значеннями BIC, потрібно відібрати слова для ручного аналізу. Відповідно до даних в таблиці, ключовими словами будуть ті, значення BIC яких дорівнює числу 2 і більше, що свідчить про переконливість даних проти нульової гіпотези. У випадку BIC=0-2 — спостережувані дані, найімовірніше, є результатом простої випадковості. Тобто чим вище значення BIC, тим краще. Згідно з даними на Рис. 3.12, значення BIC=2 відповідає $p$=0.00018 і LL=13.98. Ці вимоги до статистичної значущості є достатньо суворими. Але вони добре працюють у випадку наших корпусів, оскільки розмір референтного підкорпусу є досить великим і різниці в частотах є разючими.

Таким чином, внісши дані в таблицю Excel, а саме частотності вибраних слів в обох корпусах – фокусному й референтному, а також обсяги підкорпусів у словах (важливо, що не в токенах, оскільки до токенів входять не тільки словоформи, але й пунктуаційні знаки і спецсимволи), ми отримали автоматично обчислені значення критерія BIC, що зображено на Рис. 3.11, а відтак, опираючись на ці дані, вибрали слова для поглибленого аналізу.

| | observed frequencies | | expected frequencies | | Over/under-use | | normalised frequencies | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | corpus1 | corpus2 | corpus1 | corpus2 | | Log Likelihoo | corpus1 | corpus2 | %DIFF | Bayes Factor BI |
| \|УЄФА\| | 680 | 19 | 126,52 | 572,48 | + | 2157,74 | 0,00002 | 0,00000 | 16094,73 | 2138,60 |
| \|Марчук\| | 1433 | 265 | 307,33 | 1390,67 | + | 3533,86 | 0,00004 | 0,00000 | 2346,91 | 3514,73 |
| \|коаліція\| | 6422 | 1883 | 1503,17 | 6801,83 | + | 13814,71 | 0,00017 | 0,00001 | 1443,26 | 13795,57 |
| \|Тігіпко\| | 827 | 72 | 162,71 | 736,29 | + | 2354,29 | 0,00002 | 0,00000 | 5097,46 | 2335,16 |
| \|РНБО\| | 883 | 91 | 176,29 | 797,71 | + | 2450,27 | 0,00002 | 0,00000 | 4290,74 | 2431,14 |
| \|ракський | 1522 | 301 | 329,95 | 1493,05 | + | 3689,66 | 0,00004 | 0,00000 | 2188,06 | 3670,53 |
| TOTAL | 36883847 | 166899368 | | | | | | | | |

Рис. 3.11. Таблиця для автоматичного вираховування значення BIC.

| BIC | Переконливість даних проти нульової гіпотези | Значення ймовірності | LL |
|---|---|---|---|
| 2-6 | Достатньо переконливі дані проти нульової гіпотези | 0,00018 | 13,98 |
| 6-10 | Вагомі докази проти нульової гіпотези | 0,000014 | 18,81 |
| >10 | Дуже вагомі докази проти нульової гіпотези | 0,0000024 | 22,22 |

Рис. 3.12. Таблиця відповідностей між значеннями BIC і *p*-value[46].

### 3.6. Поглиблений аналіз ключових слів

Відбираючи ключові слова, ми опиралися на критерій BIC, застосувавши поріг BIC >10, оскільки розмір референтного підкорпусу досить великий і спостережувані різниці в частотах вагомі. Отриманий список ключових слів (див. Додаток 2), який вдалося отримати, можна поділити на кілька груп:

- лексика, зумовлена особливостями публіцистичного стилю;
- лексика, пов’язана з додатковими напрямами діяльності газети «День»;
- лексика, яка формує дискурсивний портрет видання;
- лексика, вживання якої можна пояснити мовними преференціями авторів/редакторів статей.

Отож першу групу, а це — лексика, вживання якої зумовлене особливостями публіцистичного стилю, — представляють власні назви: прізвища або імена політиків, культурних діячів. Яскравими представниками цієї групи є такі слова: *Ющенко*, *Путін, Тимошенко, Янукович, Кучма, Яценюк, Азаров, Кінах, Лукашенко, Марчук, Тігіпко, Медведєв, Кличко, Медведчук, Черновецький, Саркозі, Саакашвілі, Клінтон, Трамп, Порошенко, Клінтон, Турчинов, Гройсман, Маккейн*. Значення BIC цих слів дуже високі (див. Додаток 2), але такі показники мають

[46] Costas Gabrielatos. Keyness analysis: Nature, metrics and techniques, с. 240.

цілком логічне пояснення, оскільки цей перелік ключових слів є підтвердженням такої особливості публіцистичного стилю: часте вживання власних назв, зокрема прізвищ актуальних політиків. Адже публіцистичні тексти тісно пов'язані з подіями на політичній арені й висвітлюють останні події, які відбулися в світі, а політики і політикині є, по суті, представниками цих процесів, тому вживання їхніх прізвищ цілком очевидне явище в публіцистиці. Виразними прикладами, які підкріплюють твердження про те, як швидко публіцистика реагує на події в суспільно-політичному житті, є прізвища *Гонгадзе, Пукач* (BIC=6008,53 / BIC=1335,70), які рясніють в статтях досліджуваної газети «День», найбільше в 2000-2006 роках. Це можна пояснити розслідуванням резонансного вбивства Григорія Гонгадзе в 2000 році і притягненням до кримінальної відповідальності винного в цій справі Олексія Пукача, тому така частотність цих прізвищ є резонна й обґрунтована. Наступне ключове слово *Марчук* (BIC=3514,73), високу частотність якого теж слід пояснити. Євген Марчук – це український державний діяч, екс-прем'єр-міністр України і чоловік Лариси Івшиної, яка є головною редакторкою газети «День», що, власне, пояснює часте фігурування Євгена Марчука в статтях та його участь в інтерв'ю, оскільки він був частим гостем газети. Наступні ключові слова — власні назви *Кофман, Шопен, Джамала, Бетховен, Вагнер, Поклітару, Пономарьов, Пуччіні* (див. Додаток 2) є прізвищами композиторів, співачок, співаків, хореографів, які свідчать про наступну особливість публіцистичного стилю: широке вживання культурно-освітньої лексики. Варто зазначити, що Джамала, відома українська співачка, виборола перемогу для України на Євробаченні 2016 року та й відома своєю активною громадянською позицією. Звісно, що після перемоги цієї співачки на Євробаченні вона була бажаною і частою гостею в студіях всеукраїнських газет, каналів. Газета «День» реагує на суспільно-культурні події в соціумі, як бачимо на прикладі високої частотності вживання слова *Джамала*. Важливо згадати також про слова *Лариса* (BIC=5461,99), *Івшина* (BIC=7007,15), ключовість яких можна пояснити зважаючи на те, що Лариса Івшина головний редактор газети «День». Велике значення BIC, а саме 3745,66, отримало слово *Мейс*, яке є ключовим для «Дня»,

оскільки Джеймс Мейс був консультантом цієї газети і працював в редакції. Широкий перелік ключових слів-власних назв *Дональд, Трамп, Качиньський, Обама, Маккейн, Байден, Клінтон, Ердоган, Саркозі* (див. Додаток 2), які є прізвищами закордонних політиків і політикинь. Насиченість статей такою лексикою вказує на те, що газета «День» висвітлює не лише українські політичні процеси, а реагує на зміни та події на світовій політичній арені, що притаманно текстам публіцистичного стилю. Також групу ключових слів-власних назв доповнюють слова-назви політичних партій, компаній, організацій: *БЮТ, Газпром, Нафтогаз, НАТО, Бі-Бі-Сі, Нацрада, МВФ, УЄФА, РНБО, ОДА, Укрнафта, Євросоюз, ВР, АТО, МОЗ, Укртелеком* (див. Додаток 2). Точні найменування є характерними для текстів газетного стилю, оскільки однією з головних функцій публіцистичного стилю є інформаційна. Звісно, інформаційна функція притаманна усім стилям мови, проте ця функція у публіцистичному стилі є особлива своїм характером інформації. Автори вважають за необхідне інформувати якомога ширшу авдиторію якомога точною інформацією про найбільш актуальні для суспільства проблеми, а власні назви сприяють цьому, оскільки прив'язують ту чи іншу подію до конкретної людини чи організації, таким чином поглиблюючи сприйняття.

Наступні ключові слова, ВІС значення яких лежить в межах від 200 до понад 15000, стосуються безпосередньо суспільно-політичної лексики, якій притаманні семи, що стосуються понять «суспільство», «політика», «держава» або дотичних до них. Суспільно-політичну лексику можна умовно поділити на кілька підгруп.

Перша з них включає лексеми, які відображають політичне життя суспільства. Тобто це поняття, які називають систему політичних поглядів, течій, ідей, які сформовані конкретними організаціями, політичними партіями, громадськими рухами; номени суб'єктів політики, політичних груп і явищ. Серед переліку ключових слів газети «День» до цієї підгрупи можна віднести такі слова: *путінський, демократія, сепаратист, соціал-демократ, олігархат,*

*авторитаризм*, *європейськість*, *сталінський*, *регіонал*, *коаліція*, *шовінізм*, *нацизм*, *опозиціонер*, *олігарх*, *коаліційний*.

Друга підгрупа із групи ключових слів, що належать до суспільно-політичної лексики, налічує лексеми, які позначають назви носіїв державної влади: *урядовець*, *депутат*, *депутатка*, *депутатство*, *нардеп*, *чиновник*, *віце-прем'єр*, *президент*, *прес-секретар*, *екс-міністр*, *прем'єр*, *спікер*, *держсекретар*, *гендиректор*, *віце-президент*; назви політичної діяльності (акції, проєкти): *мораторій*, *політреформа*, *законопроект*, *популізм*, *прес-конференція*, *тендер*, *саміт*, *інавгурація*, *приватизація*, *реприватизація*, *транш*, *синдикат*, *вибори*, *відставка*, *люстрація*, *екзит-пол*, *імпічмент*; лексика на позначення державного устрою і законодавчого апарату: *кабмін*, *облдержадміністрація*, *мінфін*, *облрада*, *міноборони*, *генпрокуратура*, *обленерго*, *Єврокомісія*, *міськрада*.

Третю підгрупу суспільно-політичної лексики складають лексеми на позначення ідейно-моральних понять: *незалежність*, *патріотизм*, *соборність*, *солідарність*, *проукраїнський*, *стурбованість*, *протест*, *співвітчизник*, *перемога*, *пострадянський*, *геноцид*, *голокост*, *миротворчий*, *патріотичний*, *постгеноцидний*, *українець*, *тероризм*, *агресія*.

Серед переліку ключових слів яскраво виділяються лексеми на позначення понять з мистецтва, тобто музичні терміни, наприклад: *сопрано*, *філармонія*, *диригент*, *композитор*, *мюзикл*, *квартет*, *піаніст*, *концерт*, *оркестр*, *дует*, *вокаліст*, *консерваторія*, *ансамбль*, *вокал*, *естрадний*, *оперета*, *піаніст*, *симфонія*, *хормейстер*, *джаз*, *капела*, *тенор*; театральна лексика: *вистава*, *опера*, *сценографія*, *спектакль*, *сценічний*, *театральний*, *сценограф*, *оперета*, *режисер-постановник*, *театрал*; слова з мистецтва танцю: *балет*, *хореографічний*, *балетний*, *хореографія*, *балетний*, *балетмейстер*. Наявність такої спеціальної термінології підтверджує тезу, що публіцистика відіграє важливу роль в системі стилів мовлення, оскільки її завданням є задовольнити як й інтелектуальні, так й естетичні потреби читача. Тому й публіцистичному жанру притаманна принципова неоднорідність мовних засобів і, як показано вище, вживання професійної лексики.

До того ж специфіка газети «День» полягає, зокрема, в широкому висвітленні культурного життя країни, що пояснює високу частотність цієї групи лексики.

Спираючись на проведений вище аналіз ключових слів, бачимо домінантну рису в лексиці сучасних ЗМІ – детермінілогізація спеціальної лексики, тобто терміни втрачають свою стилістичну нейтральність, системність, однозначність. Таким чином слова-терміни зі сфери економіки (*бартер*, *субсидія*), зі спорту (*аутсайдер*, *раунд*) вийшли зі своїх терміносистем і стали широковживаними, що можна побачити, досліджуючи ключові слова газети «День». Наприклад, ключове слово *аутсайдер* (ВІС=136) цікаво звучить в наступному контексті: *Уже не типовими, радше поодинокими, стають моральні аутсайдери, які не можуть і не хочуть змиритися з основною нашою проблемою і нещастям — відсутністю масової соціальної ініціативи* (Газета «День», 2003), що свідчить про вживання цієї лексеми не тільки в прямому значенні, але й переносному. Подібна ситуація зі ключовим словом *вакуум* (ВІС=42,34). Тут важливо згадати про контексти вживання цього слова для того, щоб показати, що воно може вживатися не тільки в науці, зокрема в галузі фізики. Добре ілюструє приклад вживання цієї лексеми в значенні «відсутність інформації», «інформаційної порожнечі»: *Сільські люди опинилися в інформаційному вакуумі, і я впевнений, що більшість із них не знають, хто керує областю і що в ній відбувається* (Газета «День», 2002). Тоді як на противагу цьому значенню бачимо приклад, де термін *вакуум* вживається в науковому контексті: *Пустий простір в атомі, а також в іншій речовині і є фізичним вакуумом* (Газета «День», 2003). Отож через розвиток переносного значення, розширення змістового обсягу поняття, накопичення нових сем відбувається активне вживання іншомовних слів-термінів за межами їх терміносистем.

Принагідно тут можна згадати про економічну лексику, якою рясніють публіцистичні тексти. Зважаючи на високі значення ВІС (від 332 (показник слова *здешевлення*) до 33 942 (показник слова *гривня*), лексеми з економічної галузі можна вважати ключовими словами. Виразними представниками цієї групи слів є:

*дефолт*, *транзит*, *гривня*, *кубометр*, *тонна*, *мільярд*, *подорожчання*, *долар*, *гектар*, *девальвація*, *концерн*, *видобуток*, *експортер*, *дотація*, *мільйон*, *квота*, *інфляція*, *дефіцит*, *профіцит*, *імпортер*, *товарообіг*, *кредитувати*, *тисяча*, *гривневий*, *акциз*, *здешевлення*, *інвестувати*, *барель*, *ціна*, *збитковий*. Як бачимо, серед переліку ключових слів є лише загальнозрозумілі терміни, які розраховані на сприйняття широкою авдиторією, тоді як вужчі слова-терміни, наприклад, *демпінг* (BIC=-1,53), *кліринг* (BIC=-16,21) не є ключовими словами. Ключовість вище перелічених слів може бути зумовлена тим, що ЗМІ, зокрема «День», у своїх публікаціях значну увагу приділяють економічним проблемам, що і пояснює частотне вживання економічних термінів.

Досить важливим є дослідження ключових слів, які є запозиченими з іноземних мов, оскільки з вживанням іншомовних лексем існує певна небезпека, зокрема в тому, чи вжиті вони в правильному контексті, чи доречно і грамотно. Пропонуємо спершу поглянути на перелік ключових-іншомовних слів: *гран-прі*, *лауреат*, *прес-реліз*, *призер*, *шоу*, *приз*, *хіт*, *де-факто*, *істеблішмент*, *експозиція*, *консорціум*, *альянс*, *бієнале*, *аншлаг*, *фест*, *холдинг*, *кулуари*, *пікет*, *реверсний*, *експонувати*, *ажіотаж*, *стенд*, *контингент*, *блокбастер*, *есдек*, *дебют*, *форум*, *реванш*, *ескалація*, *інсталяція*, *вето*, *ротація*, *ювілей*, *гала-концерт*. Для прикладу, проаналізуємо детальніше слово *контингент* (BIC=773). Газета «День» подає його в такому контексті: *Нині дитячий контингент змінюється в кращий бік* (Газета «День», 2000). В цьому фрагменті лексема *контингент* звучить цілком природно, зрозуміло і невимушено, тому немає підстав, щоб протестувати проти вживання цього слова. Лексема *контингент* позначає сукупність людей, що становлять подібну з певного погляду групу; склад осіб певної категорії. На Рис. 3.13 бачимо, що згадана лексема спостерігається в публікаціях газети ще в 2000 роках, тому можна зробити висновок, що українському читачеві це слово не є чужим і незрозумілим.

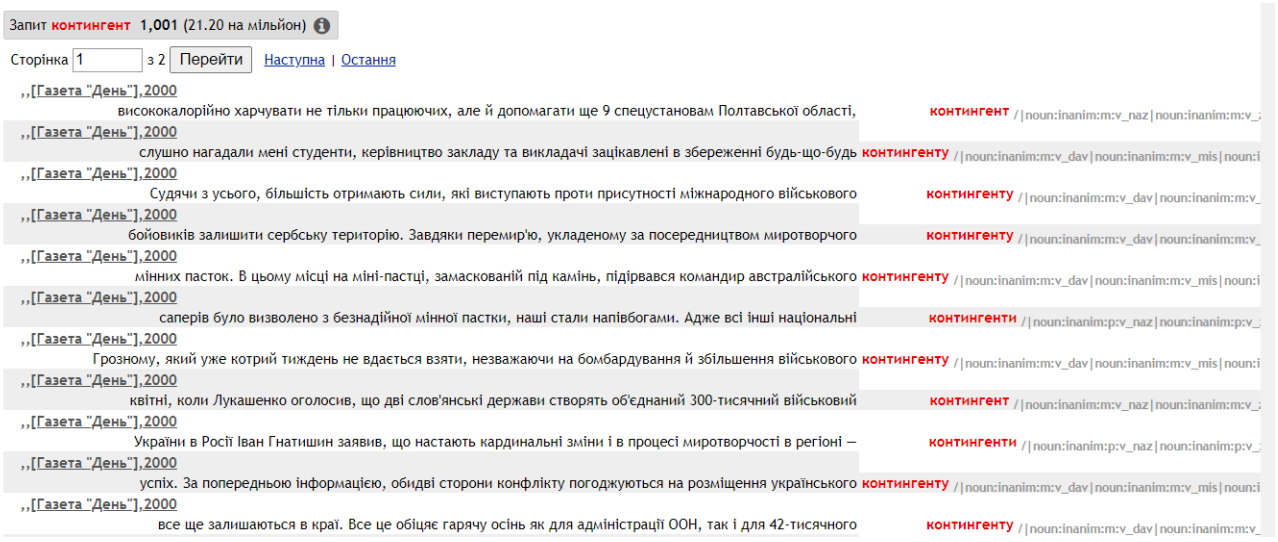


Рис. 3.13. Приклади вживання леми *контингент* в публікаціях газети «День».

Слово *консорціум* (BIC=1923) означає «тимчасові статутні об'єднання промислового і банківського капіталу для досягнення спільної мети»[47]. Походить воно від латинського *consortium*, що означає «спільнота, співучасть». У газеті «День» слово вжите у такому реченні: *На його думку, консорціум не працюватиме повноцінно, якщо він не зможе розпоряджатися газотранспортною системою* (Газета «День», 2003). Слово *консорціум*, так само, як і слово *контингент* вживалося в публікаціях ще з 2000 року, як видно на Рис. 3.14.

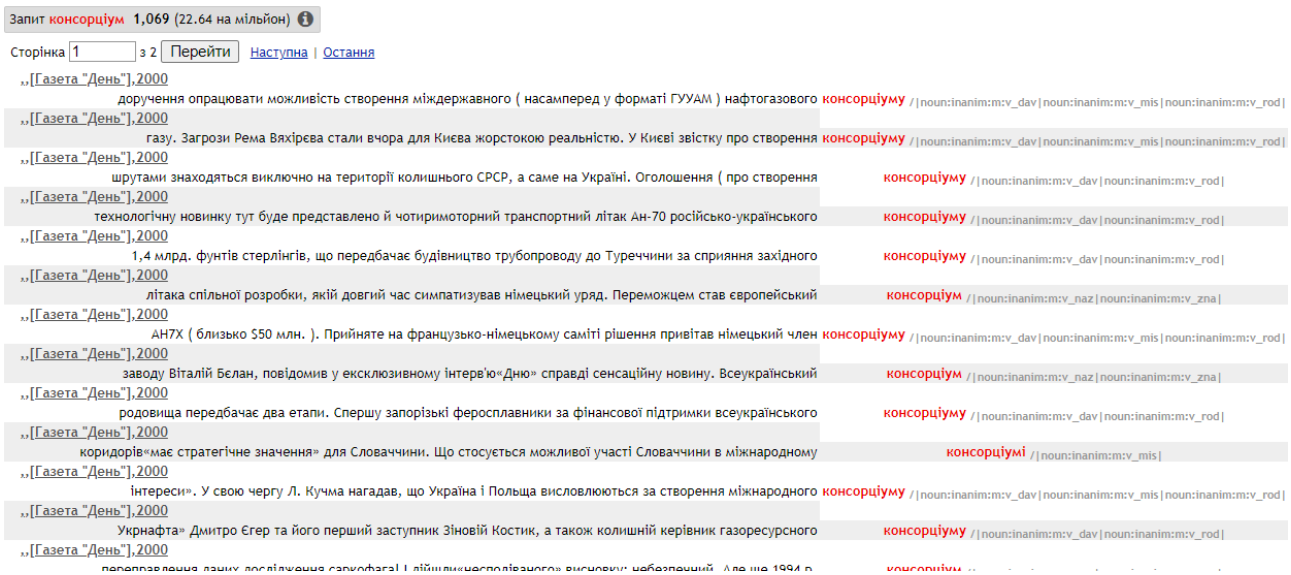


Рис. 3.14. Приклади вживання леми *консорціум* в публікаціях газети «День» .

[47] *Словник української мови: в 11 томах* // Том 4, 1973, с. 265.

Вживання терміну *консорціум* у наведених реченнях на Рис. 3.14 є виправдане і доцільне, оскільки воно є назвою явища, взятого з мови-джерела разом з позначуваним поняттям. Переглянувши інші ключові-іншомовні слова газети «День», які наведені вище, можна зробити висновок, що деякі з них прийшли в нашу мову кілька століть тому, зокрема *ювілей*, *приз*, *форум*, *концерт* та вже пройшли процес адаптації, тоді як деякі увійшли до складу української мови порівняно нещодавно (*блокбастер*, *контингент*, *ескалація*, *бієнале*) і набувають поширення.

Отож високочастотне вживання і ключовість іншомовної лексики в публіцистичному стилі, зокрема в статтях «Дня», є логічна і резонна, оскільки публіцистика надзвичайно швидко реагує на зміни, йде нога в ногу з актуальними світовими подіями і, відповідно, відображає їх, запозичуючи разом із тим лексику.

Цікавими прикладами, які ілюструють, наскільки публіцистика є актуальною і чутливою до змін в суспільстві, є ключові слова *помаранчевий* (BIC=579) і *касетний* (BIC=579). В газеті «День» лексема *помаранчевий* вживається в такому контексті: *Разом із тим ці кольори я б не стільки пов'язувала з політичними силами, тому що за тими й іншими стоять люди, які надали б їм певного значення: помаранчевий — це колір змін, колір повороту до нової держави, народ якої усвідомлюватиме свою силу і значимість* (Газета «День», 2004)*; На думку співачки, «Помаранчева революція» свідчить про те, що Україна — уже Європа, вона готова до вступу в Євросоюз* (Газета «День», 2004). Що не менш важливо, згадана лексема починає активно вживатися в публікаціях саме з 2004 року, до цього часу наявні 2 випадки вживання в 2000 року і 2 випадки в статтях 2002 року. Це можна пояснити тим, що 22 листопада 2004 року розпочалася Помаранчева революція, події якої, відповідно, відобразилися в публіцистиці, яка миттєво реагує на зміни в світі та інформує читача про них. Така ж історія зі словом *касетний*. Восени 2000 року було оприлюднено касетні записи із кабінету Президента України Леоніда Кучми, які свідчили про причетність тодішнього президента та інших високопосадовців і політиків до вбивства журналіста Г.

Гонгадзе. В газеті «День» це слово вживається саме в такому контексті практично майже у всіх знайденнях, що проілюстровано на Рис. 3.15.

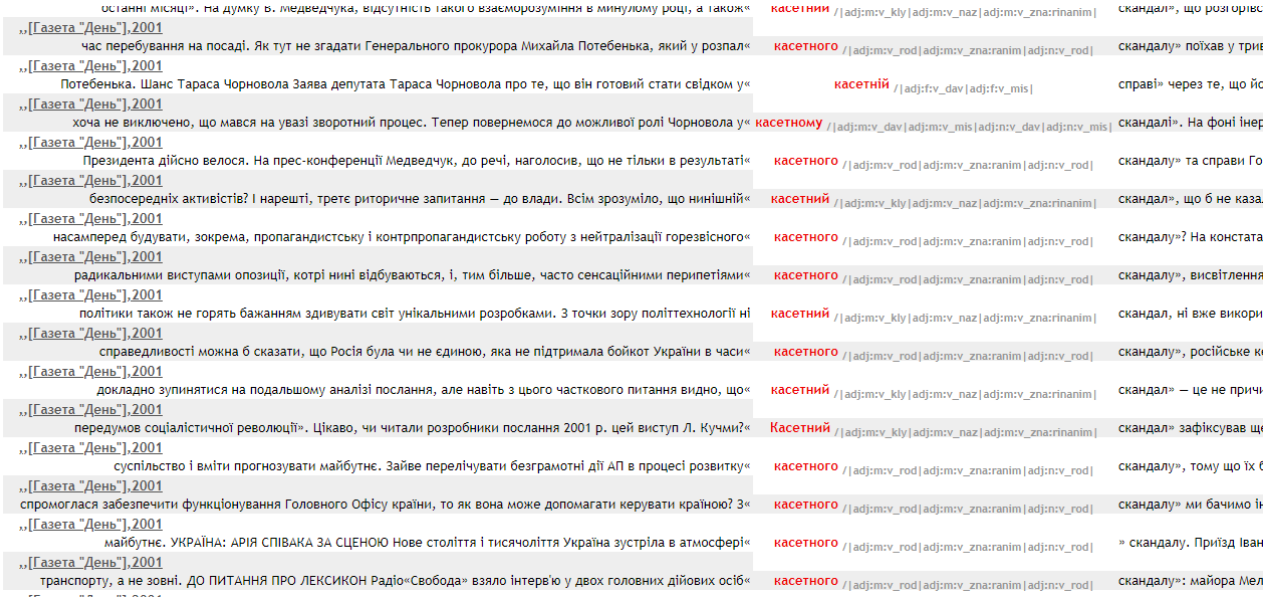


Рис. 3.15. Приклади вживань леми *касетний* в публікаціях газети «День».

Таким чином, ці слова підтверджують тезу, що публіцистика надзвичайно швидко реагує на суспільно-політичні події, оскільки однією з її головних ознак є інформаційність.

Також варто зазначити про категорію ключових слів, до якої входить загальновживана лексика (див. Додаток 2): *поспіль*, *абсолютно*, *нещодавно*, *минулий*, *щороку*, *сьогоднішній*, *безумовно*. Серед цих слів можна виокремити лексеми на позначення часу: *нещодавно*, *минулий*, *щороку*, *сьогоднішній*. Їхню ключовість можна пояснити тим, що ЗМІ часто пишуть про повторювані події, що бачимо в такому рядку, взятого з конкордансу: *Я відвідую фотовиставку «Дня» щороку з моменту її заснування* (Газета «День», 2006), і недавні події (*нещодавно*, *сьогоднішній*). До цих лексем можна віднести і *поспіль*, яке теж вказує на повторюваність подій, що проілюстровано в цьому рядку: *Тож уже традиційно сьомий рік поспіль ця організація оголошує список «Ворогів Інтернету» — країни, в яких зафіксували випадки перешкоджання свободі слова в світовій павутині* (Газета «День», 2014). Але також це слово вживається в значенні безперевної повторюваності певної події протягом певного проміжку часу: *Ми могли вчитися в*

*день по 16 годин, або два дні поспіль без жодної зупинки, а потім вже відпочивати* (Газета «День», 2008). Отож ці досліджувані слова є ключовими з тієї причини, що газети «День» повідомляє свою читацьку авдиторію про події, які повторюються певний проміжок часу або відбулися недавно, використовуючи лексеми на позначення часу.

Наступними словами, які також цікаво проаналізувати є *абсолютно* й *безумовно*. Розглянемо глибше слово *абсолютно* (ВІС=8336). Абсолютна частота у фокусному підкорпусі – 9 480, у референтному підкорпусі – 11 200. Для того, щоб зрозуміти, в яких контекстах вживається досліджуване слово, було здійснено аналіз конкордансу, в результаті якого можна зробити висновок, що ця лексема у фокусному підкорпусі вживається для підсилення певної думки, тези, наприклад, наведемо рядок з конкордансу: *Абсолютно очевидно, що в умовах так званої адміністративної реформи «по-українськи», коли спостерігається зростання бюрократичних та контрольних структур, навіть прийняття позитивного рішення про створення соціополісів спричинить наступне* (Газета «День», 2000). Такий стиль висловлювання думки характерний для публіцистики, якій притаманне вживання емоційно-оцінної лексики для того, щоб вплинути на читача та переконати його в певній думці, що якоюсь мірою досягається за допомогою таких слів як *абсолютно*, *безумовно*, що, власне, і пояснює їхню ключовість. Варто зазначити, що синонімами до слова *абсолютно* можуть бути *цілковито*, *повністю*, *цілком*. Розглянемо абсолютні частоти цих слів у фокусному підкорпусі: *цілковито* – 587, *цілком* – 9 930, *повністю* – 6 195. Таким чином, можна сказати, що автори «Дня», надають перевагу *цілком* й *абсолютно* над *цілковито* і *повністю*. Такі різниці в частотах можуть бути зумовлені мовними преференціями авторів/редакторів.

Вирізняються з-поміж переліку потенційних ключових слів лексеми, які можна об’єднати в одну категорію під назвою – логічні слова-зв’язки. Представляють цю групу такі слова: *утім*, *попри*, *однак*, *по-друге*, *по-перше*, *завдяки*, *щоправда*, *хоча*, *причому* та інші. Для прикладу, розглянемо значення ВІС

слова *утім* та його функціонування в текстах. Абсолютна частота цього слова у фокусному підкорпусі 3951, а в референтному – 5321. У такому випадку ВІС цього слова дорівнює 2961,91, а це дуже високий показник, який свідчить про статистичну значущість цього слова. Для того, щоб пояснити ключовість слова *утім,* було здійснено аналіз знайдень в конкордансі на пошуковий запит за цією лемою. Ілюстрація контексту, в якому вживається *утім*: *Він також додав, що це стосується і справи екс-міністра Юрія Луценка. Утім, єврокомісар наголосив, що від України «чекають прогресу»* (Газета «День», 2012). Як бачимо на прикладі цього уривку, вживання *утім* й інших слів з цієї групи забезпечує логічний виклад думки та добру аргументацію, яка дуже важлива в публіцистичному стилі, оскільки є потужним засобом впливу на реципієнта. Чітка аргументація є сильним інструментом публіцистичного тексту, адже змушує читача задуматися над порушеною проблемою, звернути увагу на її актуальність, а слова-зв'язки сприяють логічному викладу таких аргументів, тому ключовість таких слів в газеті «День» є обґрунтованою.

Варте уваги слово *підтримка*, яке опинилося на вершині аналізованого списку. На перший погляд, це просте загальновживане слово, вживання якого не зумовлене ані діяльністю газети «День», ані її сферами зацікавлення. Відповідно, щоб пояснити ключовість цього слова, потрібно проаналізувати конкорданси обох підкорпусів. Лема *підтримка* у фокусному підкорпусі трапляється 17 402 рази, а у референтному – 25 130 разів. При таких даних значення ВІС цього слова 11956,08, що є дуже вагомим доказом проти нульової гіпотези. Перед тим, як здійснювати пошук за цим словом в ГРАКу, було зроблено припущення, що ключовість слова *підтримка* може полягати в тому, що в публіцистичному стилі, в цьому випадку в текстах газети «День», воно часто вживається в сполуці *за підтримки*, яка є публіцистичним штампом. Щоб перевірити цю гіпотезу, було здійснено пошукові запити за фразою *за підтримки* в обох підкорпусах. В результаті отримано такі значення: 1173 рази ця фраза трапляється у фокусному підкорпусі, і 718 – в референтному підкорпусі. Такі дані мають показник ВІС 1766,64, що свідчить про

ключовість фрази *за підтримки* для фокусного підкорпусу. Варто зазначити, що слово *підтримувати* (BIC=5782,81) також є ключовим для текстів «Дня», для того, щоб пояснити його ключовість потрібно проаналізувати контексти, в яких воно вживається і, можливо, в них також трапляється *підтримка*. Для цього було згенеровано 200 випадкових контекстів із фокусного підкорпусу, скориставшись функцією в інтерфейсі ГРАКу, і на основі цього проаналізовано, стосовно чого найчастіше вживають лексеми *підтримувати* й *підтримка*. Отож у фокусному підкорпусі досліджувані слова найбільш частотно вживаються в значенні підтримки законопроектів, демократичних сил, України, влади, військових, бійців, партій, президента, курсу включення України в ЄС, НАТО, також газета подає результати певних голосувань, анкетувань, використовуючи колокації таких типів: *підтримують 70 % опитуваних*; *підтримують лише 3 % виборців*. Цікаво, що в цих випадково згенеровано контекстах з фокусного підкорпусу вирізнялося вживання *підтримувати* зі словами *театр*, *актор*, *драматургія*. Проаналізувавши також згенерованих 200 випадкових контекстів з референтного підкорпусу, можна зробити висновок, що слова *підтримувати, підтримка* частотно вживають (1) стосовно *вогнища*, *горіння*, *життя* та інші, (2) з іменами, наприклад, *підтримує Григорій*, (3) із займенниками: *мене*, *нас, її, його*, (4) рідше трапляються вживання в значенні *підтримувати Україну*, *демократичні сили*. Тоді як в порівнянні з випадковими контекстами у фокусному підкорпусі, у двісті випадкових контекстах з референтного підкорпусу немає жодної колокації прикладу *підтримувати президента*, *підтримувати театр*, *підтримувати законопроект*, *підтримувати курс включення України в ЄС*. Щоб перевірити, чи колокації такого типу є ключовими для текстів «Дня» було використано функцію в інтерфейсі корпусного менеджера ГРАКу – фільтр лем, за допомогою якої можна задати контекстні обмеження пошуку в вигляді лем. Таким чином, ми знайшли, що слова *підтримувати* та *підтримка*, абсолютна частота яких у фокусному підкорпусі 26 569 сумарно, вживаються 5 963 рази в контексті з лемами *вибори*, *проект*, *законопроект*, *опитаний*, *НАТО*, *ЄС*, *військовий*, *президент*, *політика*, *Україна*, *влада*, *демократичний*, *виборець*, *театр*, *актор*, *політика*, *курс*. А в референтному

підкорпусі слова *підтримувати* і *підтримка* в контекстах з цими ж лемами трапляються 3 895 разів. Для того, щоб перевірити, чи ця різниця є статистично значущою, ми знайшли значення BIC, яке дорівнює – 1598,02, що дає підставу вважати досліджувані дані статистично значущими.

Отже, слова *підтримка* і *підтримувати* є ключовими для газети «День» з тієї причини, що статті видання – це публіцистичний стиль мовлення, якому притаманні як і штампи, серед яких і є фраза *за підтримки*, так і вживання суспільно-політичної лексики, саме тому колокації типу *підтримувати президента*, *підтримувати законопроект*, *підтримувати політику* та інші є ключовими для текстів «Дня». А також, зважаючи на зацікавленість газети культурним життям, можна пояснити частотне вживання *підтримувати* і *підтримка* стосовно театру, драматургії, акторів.

Досліджуючи згенерований список потенційних ключових слів за словоформами з коефіцієнтом згладжування 1000, вирізняється помітна група таких ключових слів: *заявив*, *повідомляє*, *вважає*, *повідомив*, *говорить*, *зазначив*, *розповідає*, *вважаю*, *наголосив*, *підкреслив*. Детальніше проаналізуємо слово *заявив*. Абсолютна частота у фокусному підкорпусі – 11 802, у референтному – 5 175, BIC – 21 515,13. Для виявлення контексту, в якому функціонує досліджуване слово, було здійснено аналіз конкордансу. Ось приклад використання цього слова в тексті: *«Я переконаний, що МВФ не дасть Україні чергового траншу, оскільки це повне ігнорування Меморандуму», — заявив Ющенко* (Газета «День», 2009). Тобто вживання цього слова відбувається тоді, коли необхідно передати певну думку, ідею, тезу і вказати джерело, що є поширеним явищем в публіцистиці, головною особливістю якої є швидка реакція на події й інформативність. Тому ключовість слова *зазначив* й інших слів цієї групи можна пояснити особливостями текстів публіцистичного стилю, а саме важливістю та необхідністю якомога точно повідомити читача.

Аналізуючи список ключових слів, виділяється друга група лексем, вживання яких зумовлене діяльністю газети «День»: проведення проєктів, шкіл, фотовиставок.

Одним з найпомітніших представників цієї групи є високочастотне (абсолютна частота – 11 375) слово – *конкурс* (BIC=22299,80). Газета «День» відома тим, що вона проводить багато конкурсів для своїх читачів. Заглянувши в конкорданс, знаходимо інформацію, що в 2000 році газета «День» започаткувала новий конкурс «Громадський форум»: *Сьогодні «День» оголошує про новий масштабний конкурс для читачів. У рамках уже нашого фірмового проекту «Експерт «Дня» ми спільно з фондом «Відродження» розпочинаємо акцію «Громадський форум»* (Газета «День», 2000). Наступний рядок теж ілюструє вживання цієї словоформи в контексті згаданого проєкту: *«День» повинен активніше залучати молодих людей до участі в конкурсах у рамках «Громадського форуму», стимулювати їхню політичну активність, застосовуючи для цього різноманітні форми заохочення* (Газета «День», 2000). Чи не найчастіше слово *конкурс* вживається в контексті Міжнародного фотоконкурсу газети «День», який в 2020 році був проведений в двадцять друге. Подаємо рядок з конкордансу з вживанням цього слова в контексті проведення Міжнародного фотоконкурсу «Дня»: *Продовжувати проведення традиційних міжнародних фотоконкурсів нашу газету спонукають і численні відгуки відвідувачів виставки «День нового тисячоліття»* (Газета «День», 2000). Також одним з найпопулярніших конкурсів, які проводить газета «День», є «Експерт «Дня». Тому словоформа *конкурс* часто вживається в контексті його проведення: *Крім того, найактивніші учасники конкурсу «Експерт «Дня», люди, яких ми по праву називаємо професійними читачами, отримали персональні анкети* (Газета «День», 2000). Серед інших конкурсів, які проводить газета «День», і які трапляються в конкордансі, — «День кохання» і «Сімейний альбом»: *Приїхали представники обласного радіо й телебачення. Цього дня співробітники газети «День» вітали переможця конкурсу «День кохання» Василя Марковича Кучеренка, керівника місцевих ветеранів* (Газета

«День», 2000)*; 9 жовтня минулого року в нашій газеті у рамках конкурсу «Сімейний альбоми України» було вміщено публікацію «Пригоди навколо старої фотографії»* (Газета «День», 2000). Тому ключовість слова *конкурс* пояснюється тим, що газета «День» описує і «зовнішні» конкурси, які проводяться на державному, суспільному рівнях, і ще й пише про «внутрішні» конкурси (деякі назви яких наведені вище), які проводить сама.

Ще одним словом, яке входить до цієї категорії, є *академія* (BIC=6149). Однією із гіпотез, яка може пояснити високочастотність вживання цієї лексеми (7371 – абсолютна частота), є те, що газета «День» активно співпрацює з Острозькою академією, тому у своїх публікаціях журналісти активно пишуть про нові події, зумовлені цим зв'язком. Тому, проаналізувавши конкорданс, ілюструємо приклад, який підтверджує наше твердження: *Більше того, як сказав ректор академії Ігор Пасічник, «територія Острога тепер оголошується не лише територією читання, а й викладання «Дня». Тут буде запроваджено спецкурс для студентів «День» як конвергентне медіа»* (Газета «День», 2015). Поруч з Острозькою академією спостерігаються вживання досліджуваної леми в сполуці з Києво-Могилянською академією. По-перше, це можна пояснити тим, що певна кількість експертів, гостей газети «День» є саме викладачами, професорами Києво-Могилянської академії, наприклад, Володимир Моренець, Лариса Масенко, Тетяна Ярошенко, Михайло Мінаков, тому у публікаціях зазначаються їхні посади та навчальний заклад, в якому викладають. А по-друге, газета «День» вже понад 18 років організовує Літню школу журналістики, серед учасників якої є багато студентів Києво-Могилянської, Острозької академій. Пишучи про цю школу в статтях, редактори згадують про її учасників і відповідно про навчальні заклади, де вони навчаються. Ілюстрацією цього слугує речення: *Марта ФРАНЧУК, Національний університет «Києво-Могилянська академія»: — У кожному періоді мого життя є фрази, які закарбовуються в душі і характеризують певний момент. … Гадаю, ця цитата підходить і для Літньої школи, адже вона показує, що всі ці лектори, які тут зібрались, може не є такими відомими, але всі мали цю*

*мету і йшли крізь терни, через побиття, як от Олександр Єльяшкевич* (Газета «День», 2015). Отож ключовість слова *академія* зумовлена позамовними реаліями.

Третім словом в цій групі є *бібліотека* (абсолютна частота – 5160), ключовість якого варто проаналізувати та пояснити. Спершу варто зазначити, що «День» багато пише про культурне й освітнє життя країни, що було показано вище, а словоформа *бібліотека* і належить до такого типу лексики, тому це перша причина, яка пояснює високу частотність цієї лексеми, а відтак її ключовість. Ось приклад функціонування словоформи *бібліотека* в статті: *Як повідомили в управлінні освіти та науки Дніпропетровської міськради, вже у п'ятницю третину книжок було відправлено в шкільні бібліотеки, решта надійде до шкіл упродовж тижня* (Газета «День», 2016). Але також існує і позамовна причина, яка пояснює частотне вживання цього слова, яка полягає в тому, що «День» з 2002 року створив свою Бібліотеку «Україна Incognita», яка щороку представляє певні новинки, про що й, власне, журналісти пишуть в статтях. Підтвердженням цього можуть слугувати такі рядки, взяті з конкордансу: *За словами автора та ведучого програми Миколи Івановича Єдомахи ( який працює на радіо понад тридцять років ), одним із приводів для цієї зустрічі стали книги з «Бібліотеки «Дня» — «Україна Incognita» та «Дві Русі» ( під загальною редакцією Лариси Івшиної. — М. М. )* (Газета «День», 2003)*; Книгозбірні військових частин-переможців отримають книги «Україна Incognita» та «Дві Русі» з «Бібліотеки «Дня», а особисто автори листів-переможців — мобільні телефони від «Епосу»* (Газета «День», 2003). Отже, по-перше, ключовість цього слова можна пояснити тим, що газета «День» активно висвітлює в своїх статтях культурно-освітнє життя країни, а, по-друге, високочастотне вживання лексеми *бібліотека*, як і лексем *конкурс,* й *академія,* зумовлене діяльністю видання, тобто організацією проєктів, проведенням шкіл, фотоконкурсів, створенням своєї бібліотеки.

З-поміж великої кількості потенційних ключових слів можна виокремити третю групу слів, що складають дискурсивний портрет «Дня», який висвітлює

ділянки, які цікаві газеті, на які вона звертає увагу в своїх текстах. До цієї категорії слів входять: *заповідник*, *театр*, *школа*, *музика*, *голодомор*, *церква* та інші.

Проаналізуймо детальніше слово *заповідник*. Його частотність у фокусному підкорпусі – 2309, у референтному – 1391, відповідно значення ВІС цього слова – 3528,20, що свідчить про його ключовість. Для того, щоб дослідити та зрозуміти причини високої частотності *заповідника* в «Дні», ми здійснили пошуковий запит за цим словом у фокусному підкорпусі. Простеживши знайдення у конкордансі, висновується, що ця словоформа часто трапляється у поєднанні з такими назвами охоронних зон:

- «Софія Київська»: *Тоді ж архітектор Лариса Скорик казала про те, що будівельні роботи на вулицях поблизу Національного заповідника «Софія Київська» ставлять під загрозу збереження старовинних будівель, унікальних мозаїк і фресок цього музею* (Газета «День», 2004).

- «Хортиця»*: Із метою збереження пам'ятних місць, пов'язаних з історією запорозького козацтва, постановою Кабінету Міністрів України створено Національний заповідник «Хортиця», до якого включені території острова Хортиця та прилеглих до нього островів* (Газета «День», 2005)*;*

- «Асканія-Нова»: *Дорогою з Криму ми зупинилися в природному заповіднику Асканія-Нова* (Газета «День», 2007)*;*

- *«*Горгани»: *Від Івано-Франківська через місто Надвірну до села Зелена, де розташована садиба Горганського природоохоронного науково-дослідного відділення природного заповідника «Горгани», всього 65 кілометрів* (Газета «День», 2007)*;*

- «Бабин Яр»: *Під час створення інституту за часів Ющенка йому були підпорядковані заповідник «Бабин Яр», меморіали жертв Голодомору та «Биківнянські могили»* (Газета «День», 2017).

З цього можна зробити висновок, що журналісти газети «День» не лише активно пишуть про актуальні суспільно-політичні події на світовій арені, але й разом з тим проявляють небайдужу позицію до питань екології, зокрема, порушуючи теми українських охоронних зон, а також цікавляться культурою та історією.

Наступну групу ключових слів представляють слова, вживання яких зумовлене мовними преференціями авторів/редакторів статей або ж спричинене недопрацюванням коректорів на етапі вичитування тексту.

Однією зі знахідок цього дослідження є вживання субстантивованого активного дієприкметника теперішнього часу *бажаючий* і його правильного відповідника *охочий*. Обидва слова є ключовими для газети «День». Але, відповідно до правил чинного правопису, потрібно уникати вживання субстантивованих активних дієприкметників теперішнього часу, оскільки вони не притаманні українській мові. Попри це необхідно зазначити, що вживання лексеми *бажаючий* (1049 – абсолютна частота) є частотніше в 2000-2010 роках, тоді як з 2010 до 2019 року набирає обертів вживання слова *охочий* (1194 – абсолютна частота) і витісняє *бажаючий*, що є позитивною зміною. Наприклад, вже в 2019 році спостерігається значна перевага вживання леми *охочий* – 56 слововживань, над лемою *бажаючий* – лише 9 знайдень. Тобто ключовість слова *бажаючий* можна пояснити тим, що коректори на певному етапі роботи не доопрацювали тексти статей, залишивши помилкове вживання.

Однією з особливостей української фонетичної системи, яка відрізняє українську мову, наприклад, від російської, є вживання літери Ґ, що має довгу і складну історію. Хоч її використовують не так часто, але є певні слова, які її вимагають. Одне з них – *ґрунт* і похідні від нього: *ґрунтовий, ґрунтовно, ґрунтувати*. Попри це газета «День» часто вживає форму *грунт* (945 – абсолютна частота), і це слово є серед переліку ключових слів «Дня» (BIC=148,66). Але важливо, що з кожним роком його частотність зменшується, наприклад, з 2016 по 2019 рік спостерігаємо вже лише 35 знайдень. Вживання цього слова в текстах та

його ключовість можна пояснити таким чином. Першою причиною може бути те, що автори/редактори при перекладі публікацій користуються автоматичними засобами для перекладу, наприклад, сервісом Google Translate, який «не знає» літери Ґ і в результаті перекладу видає літеру Г. Щоб перевірити цю тезу, з конкордансу було обрано рядок зі статті «Дня» 2000 року, де вжито слово *грунт*: *Але ландшафтним чином організована природа — структура багатоскладова, яка сама по собі є і тлом, і грунтом для окремого, більш ретельного розгляду: з урахуванням міждисциплінарних досліджень* (Газета «День», 2018). Тоді за допомогою Google Translate ми отримали переклад цього рядка російською: *Но ландшафтно организованная природа — структура многосложная, которая сама по себе есть и фон, и почва для отдельного, более тщательного рассмотрения: с учетом междисциплинарных исследований*. Ввівши цей перекладений рядок в пошукову стрічку в Google, ми знайшли цю ж статтю, але вже російською, тому можемо припустити, що переклад текстів з російської на українську міг відбуватися із застосуванням онлайн-перекладачів. Але разом з тим тут очевидне недопрацювання коректорів статей, які недогледіли вживання літери *г* в текстах, що може бути зумовлене тим, що (1) редагування могло відбуватися за допомогою правописника, якому теж незнайома літера *Ґ* (наприклад, правописник, який вмонтований в MS Word), або ж (2) через великий обсяг роботи, покладений на коректора, що є причиною не завжди якісного вичитування. А наступна причина може стосуватися історії вживання літери *Ґ*. Відомо, що правопис 1933 року, затверджений Н. Кагановичем і А. Хвилею (Олінтером), з ідеологічних міркувань вилучив українську літеру Ґ, опираючись на те, що це націоналістичний елемент штучності, використання якого відтягало розвиток української мови від завдань соціалістичного будівництва. Було потрібно аж 56 років для того, щоб її реабілітували: Ґ знову займає своє законне місце в українській абетці 14 листопада 1989 р., коли відбулося затвердження п'ятої редакції «Українського правопису». Але, як бачимо, навіть ще в 2000 роках з тих чи інших причин трапляється вживання літери Г замість Ґ, що може бути відлунням скасування цієї літери правописом 1933 року.

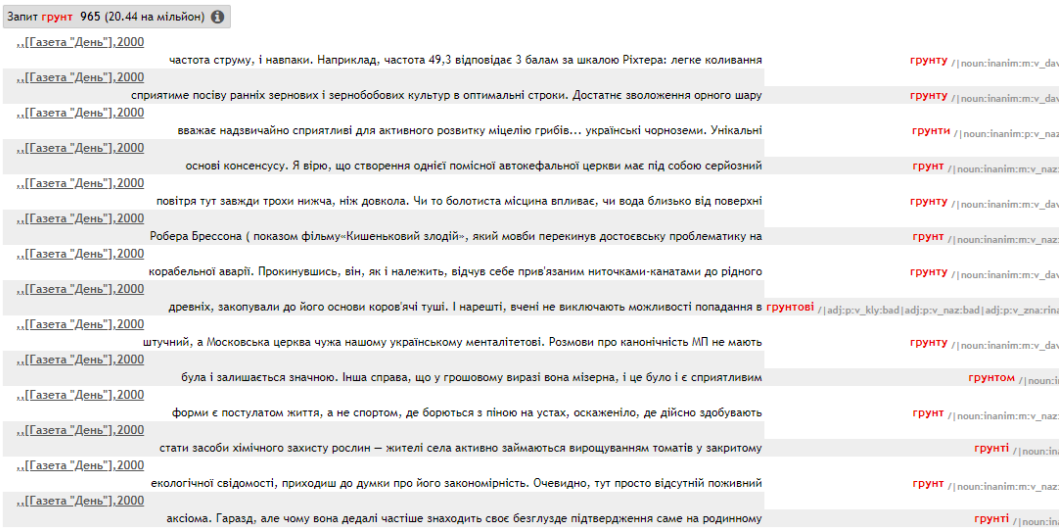


Рис. 3.16. Приклади вживання леми *ґрунт* в статтях «Дня» 2000 року.

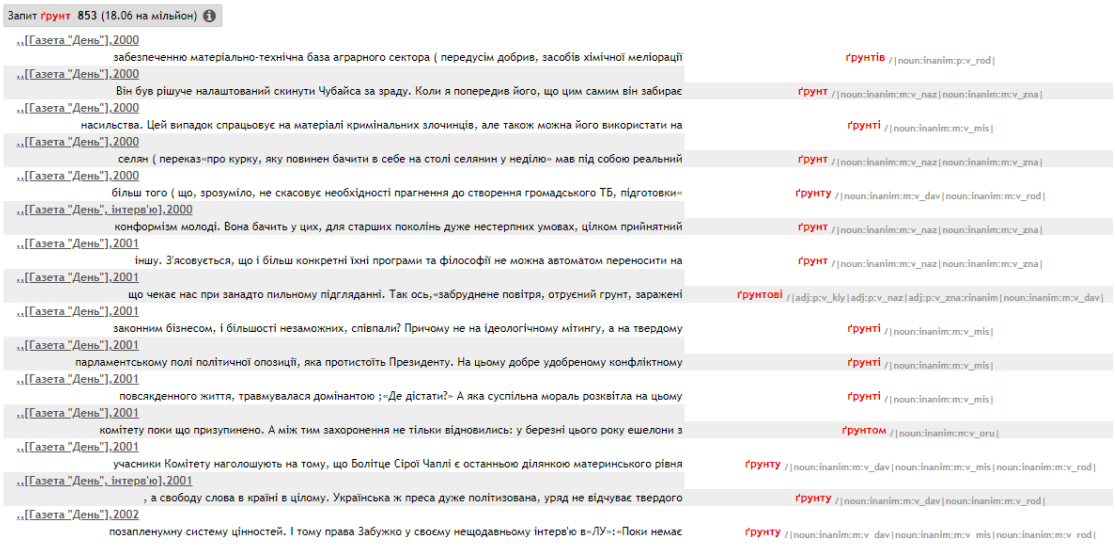


Рис. 3.17. Приклади вживання леми *ґрунт* в статтях «Дня» в 2000-2002 роках.

Аналізуючи список ключових слів, ми вирішили ґрунтовно дослідити словосполуку *згідно з* (BIC=1873), яка вживається з іменником в орудному відмінку. За правилами української мови ця сполука взаємозамінна з *відповідно до*. Цікаво, що частотність *згідно з* у фокусному підкорпусі – 15 340 – набагато переважає над частотністю *відповідно до* – 2796. Тоді як в референтному підкорпусі маємо такі значення: частотність *згідно з* – 45 550, *відповідно до* – 28 017. З цих показників виглядає так, що газета «День» використовує форму *згідно з* пропорційно частіше, ніж це відбувається в непубліцистичних текстах, і частіше, ніж форму *відповідно до*. Для того, щоб переконатися, що це відбувається не через випадковість, нам потрібно виконати наступний розрахунок – тест хі-квадрат на

статистичну значущість, що можна зробити швидко і легко за допомогою онлайн-калькулятора[48]. Ввівши потрібні дані, результат тесту на статистичну значущість такий: значення хі-квадрат = 3349,48, значення ймовірності < 0, 0000000 (тобто занадто мале для вимірювання). Опираючись на те, що обраний поріг статистичної значущості в цьому дослідженні p < 0,01, отриманий результат можна розцінювати як статистично значущий. Тому слід сказати, що досліджувані значення не є результатом простої випадковості, а зумовлені певними причинами, які ми спробуємо визначити. Отож відомо, що *згідно з* і *відповідно до* часто вживаються у поєднанні з лемами *закон, Конституція* в публіцистичному стилі, коли йде мова про певні рішення на політичному рівні. Тому у нас виникла гіпотеза, що «День» цитує певні закони, нормативно-правові акти, витяги, в яких, можливо, сполука *згідно з* використовується частіше, ніж *відповідно до* і саме через це цитування частотність *згідно з* в газеті «День» більша. Для того, щоб перевірити цю гіпотезу, було створено підкорпус текстів офіційно-ділового стилю в корпусі ГРАК, в якому ми за допомогою пошуку знайшли, що в текстах офіційно-ділового стилю сполука *відповідно до* трапляється 2039 разів, а *згідно з* – 1007, тобто в підкорпусі текстів офіційно-ділового стилю *відповідно до* вживається частіше на противагу підкорпусу фокусному. Тому наша гіпотеза про причину більшої частотності *згідно з* над *відповідно до* в статтях «Дня», пов'язану з цитуванням офіційно-ділового стилю, не підтверджується, оскільки в статтях офіційно-ділового стилю *відповідно до* вдвічі превалює над *згідно з*. Отже, в цьому випадку можна припустити, що статистично значуща різниця у вживаннях *згідно з* і *відповідно до* пов'язана з мовними преференціями авторів/редакторів, які надають перевагу у вживанні першому варіанту.

Ще одне цікаве для дослідження слово – *прийняти.* Абсолютна частота у фокусному підкорпусі – 9579, у референтному – 27 408, значення BIC – 1359,77. В українській мові часто це дієслово вживається неправильно, наприклад, замість сполуки *брати/взяти участь*, можуть казати *приймати/прийняти участь*. Тому ми

[48] UCREL Significance Test System
[Електронний ресурс], режим доступу: http://corpora.lancs.ac.uk/sigtest/

вирішили перевірити, чи трапляються в газеті «День» помилкові вживання слова *прийняти*, зокрема в наведеній сполуці. Для цього спершу було здійснено пошуковий запит у фокусному підкорпусі, використавши CQL[49], в результаті якого отримано 138 знайдень, а після цього за цим же пошуковим запитом в референтному підкорпусі було отримано 830 знайдень. Як виявилося, в обох підкорпусах трапляються хибні використання слова *прийняти*. Це може свідчити про те, що публіцистику редагують не настільки ретельно, що може бути спричинене великою навантаженістю на коректорів, які працюють в дуже швидкому темпі і недостатньо добре вичитують тексти.

## Висновки до ІІІ-го розділу

Отже, в цьому дослідженні було застосовано метрики статистичної значущості, а саме критерій BIC і тест хі-квадрат, в результаті чого було виявлено і проаналізовано 290 ключових слів та зроблено такі висновки. Велику частину ключових слів займають лексеми, які притаманні публіцистичному жанру, а саме суспільно-політична, економічна лексика. З великої кількості суспільно-політичних слів вирізняються власні назви, зокрема прізвища політиків, політикинь, громадських діячів/діячок, які причетні до творення історії української держави та до актуальних подій, що підкреслює роль і завдання публіцистики – інформувати читача про останні події у світі та країні зокрема. Також серед когорти ключових слів виділяється низка словоформ, які не пов’язані з позамовними реаліями і які не продиктовані особливостями стилю. Наявність таких лексем в більшості випадків пов’язана з мовними вподобанням редакторів чи авторів статей, які активно вживають той чи інший варіант, формуючи певний стиль газети. Іншою причиною вживання ключових слів може бути активна діяльність «Дня», а саме проведення різноманітних шкіл, проєктів, конкурсів, про які й згадують журналісти у своїх публікаціях.

[49]CQL-код:
[lemma="приймати"][tag="adj"][lemma="участь"]|[lemma="прийняти"][tag="adj"][lemma="участь"]|[lemma="приймати"][lemma="участь"]|[lemma="прийняти"][lemma="участь"]

# ВИСНОВКИ

Отже, розвиток корпусної лінгвістики в Україні за останні десятиліття набуває все нових обертів: створюється все більше статей, наукових праць, посібників в цьому напрямку. А найголовніше – розбудовуються українські корпуси, серед яких один з найкращим є ГРАК. Отож в цій роботі було окреслено досягнення корпусної лінгвістики в Україні з описом кількох наявних українських корпусів, де найбільшу увагу приділено аналізу особливостей структури та функціонування українського корпусу ГРАК, на матеріалах якого було здійснене практичне дослідження. Таким чином, серед результатів роботи є дослідження основних засад корпусної лінгвістики, обґрунтування необхідності корпусів для вивчення мови та мовлення, аналіз представленості публіцистичних текстів в українських корпусах, а також огляд й опис наявних досліджень мови газети «День», тексти якої є матеріалом цього дослідження. Також важливою частиною цієї праці є редагування текстів 2000-2019 років вищезазначеного видання, які було додано в ГРАК. В результаті ми підготували понад 51 000 текстів загальним обсягом понад 47 221 118 млн токенів (словоформ, пунктуаційних знаків та спецсимволів), що становить 5,73 % від загального обсягу ГРАКу-12.

Ключовим результатом цього дослідження є визначення характерної лексики публіцистичного стилю на матеріалі газети «День» 2000-2019 років. Для цього було використано метод ключових слів, який передбачає застосування метрик статистичної значущості. Спершу було збудовано фокусний підкорпус, який включав тексти газети «День» 2000-2019 років, і референтний підкорпус з текстами непубліцистичного стилю того ж часового проміжку. За допомогою корпусного менеджера ГРАКу ми згенерували списки потенційних ключових слів, з яких було обрано справді ключові слова шляхом обчислення статистичної значущості із застосуванням таких статистичних метрик: Баєсового інформаційного критерію й тесту хі-квадрат. Загалом виявлено 290 ключових слів, про статистичну значущість яких свідчать отримані високі показники BIC (від 14,43 до 102 923). Такі високі значення зумовлені тим, що порівняння

здійснювалося між двома абсолютно протилежними стилями: публіцистика і непубліцистика. Особливо сильно значущими словами були *президент*, *гривня*, *депутат*, *конкурс*, *мільйон* та інші (показник ВІС перелічених слів перевищує 20 000). В результаті дослідження було виявлено такі характерні лексичні особливості газети «День»: вживання лексики, яка зумовлена рисами публіцистики (власні назви, економічна лексика, суспільно-політична лексика, логічні слова-зв'язки та інше); використання лексем, які пов'язані з діяльністю газети «День»; вживання слів, які формують дискурсивний портрет видання; лексика, використання якої зумовлене мовними вподобаннями авторів/редакторів «Дня» або ж недопрацюванням редакторів на етапі вичитування тексту.

## СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ І ЛІТЕРАТУРИ

### Наукова література:


1. Вікторія Жуковська. *Вступ до корпусної лінгвістики: навчальний посібник*. Житомир 2013
2. Дарія Солодовник. Використання іншомовних слів у газеті «День» / Д. Солодовник, Н. І. Голубінка. *ЗМІ та демократичний розвиток України: збірник матеріалів ІІІ Всеукраїнської конференції студентів та молодих дослідників*. Львів 2018.
3. *Журналістська освіта в Україні: світові професійні стандарти* / уклад. О. Г. Ткаченко. Суми 2019.
4. Ірина Лебедь. Заголовки публікацій у ЗМІ на прикладі газети «День» і журналу «Країна» / І. Лебедь, Х. Білограць. *Вісник Національного університету «Львівська політехніка».* Серія : Журналістські науки. Вип. 4, 2020.
5. Людмила Архипенко. Мова ЗМІ як об'єкт лінгвістичних досліджень: історія становлення, специфіка функціонального стилю // *Культура народов Причерноморья*, № 101, 2007.
6. Марина Навальна. *Лексика української газетної періодики початку XXI ст.: джерела поповнення та стилістичне використання*: монографія. Переяслав-Хмельницький 2018.
7. Марина Нетреба. Стилістичні особливості публіцистичних текстів // М. Нетреба. *Інформаційне суспільство*. 2015.
8. Наталія Прокопенко. Діалогічність як фундаментальна ознака мови газети «ДЕНЬ» // Н. Прокопенко, К. Хачатар'ян. *Філологічні трактати. Т. 10, № 1*, 2018.
9. Олена Ткаченко. Державна мовна політика України у висвітленні газети «День» // *OPERA SLAVICA, XXIV, 4*. Брно 2014.
10. Орися Демська-Кульчицька. Деякі аспекти корпусної лінгвістики // *Українська мова*, № 1, 2005.

11. Світлана Голощук. Історичні передумови розвитку корпусної лінгвістики / С. Голощук // *Національний університет «Львівська Політехніка»*. Львів 2017.
12. Тетяна Бобкова. Проблеми періодизації корпусної лінгвістики у світовому та українському мовознавстві / Т. Бобкова // *Науковий вісник кафедри ЮНЕСКО Київського національного лінгвістичного університету. Філологія, педагогіка, психологія*. Вип. 28, 2014, с. 46-52.
13. Тетяна Коць. *Публіцистичний стиль в українській літературній мові кінця XIX – початку XXI ст.: нормативно-аксіологічний аспект*: Дисертація на здобуття наукового ступеня доктора філологічних наук / Інститут української мови НАН України. Київ 2019.
14. Adam Kilgarriff. Simple maths for keywords // *Proceedings of the Corpus Linguistics Conference.* Liverpool 2009.
15. Anatol Stefanowitsch. *«Corpus linguistics. A guide to the methodology».* Language Science Press 2020.
16. Costas Gabrielatos. Keyness analysis: Nature, metrics and techniques / ред. Taylor, C. & Marchi, A. / *Corpus Approaches To Discourse: A critical review*, Milton 2018.
17. Elena Tognini-Bonelli. *Corpus Linguistics at Work.* Amsterdam 2001.
18. Gard B. Jenset. Basic statistics for corpus linguistics: *Handout for methods seminar in English linguistics*. Bergen 2008.
19. Maria Shvedova. The General Regionally Annotated Corpus of Ukrainian (GRAC, uacorpus.org): Architecture and Functionality. Kyiv 2020.
20. Anthony MacEnery, Andrew Hardie. *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press, 2012.
21. Anthony McEnery, Richard Xiao, Yukio Tono. *Corpus-based Language Studies: an Advanced Resource Book*. London, 2006.

**Електронні ресурси:**

22. Корпус сучасної української мови (БрУК) / В. Старко, А. Рисін [Електронний ресурс], режим доступу: https://github.com/brown-uk

23. Лабораторія української
[Електронний ресурс], режим доступу: https://mova.institute/
24. Лінгвістичний портал MOVA.info
Київ: 2003-2021. [Електронний ресурс], режим доступу:
http://www.mova.info/carticle.aspx?l1=210&DID=5347
25. Шведова М. 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: 2017–2021. [Електронний ресурс], режим доступу: http://uacorpus.org/
26. Downloadable spreadsheet incorporating the log-likelihood calculation and the set of effect size measures
[Електронний ресурс], режим доступу:
http://ucrel.lancs.ac.uk/people/paul/SigEff.xlsx
27. Fisher's exact test
[Електронний ресурс], режим доступу:
https://en.wikipedia.org/wiki/Fisher%27s_exact_test
28. Log-likelihood and effect size calculator
[Електронний ресурс], режим доступу: http://ucrel.lancs.ac.uk/llwizard.html
29.Polsko ukraiński korpus równoległy
[Електронний ресурс], режим доступу: domeczek.pl/~polukr/index.php
30. Significance Testing
[Електронний ресурс], режим доступу:
https://www.lancaster.ac.uk/fss/courses/ling/corpus/Corpus3/3SIG.HTM
31. SLOVNYK.ME // *«Уроки державної мови» з газети «Хрещатик»*
[Електронний ресурс], режим доступу:
https://slovnyk.me/dict/khreshchatyk/%D0%BF%D0%BE%D0%BA%D1%80%D0%B0%D1%89%D0%B5%D0%BD%D0%BD%D1%8F
32. Statistics in corpus linguistics
[Електронний ресурс], режим доступу: http://corpora.lancs.ac.uk/clmtp/2-stat.php

33. UCREL Significance Test System
[Електронний ресурс], режим доступу: http://corpora.lancs.ac.uk/sigtest/
Ukrainisches Gemischt-Korpus / – Leipzig: 1998-2021.
[Електронний ресурс], режим доступу:
https://corpora.unileipzig.de/de?corpusId=ukr_mixed_2014

**Довідкова література**

34. *Словник української мови: в 11 томах* // Том 4, 1973.

# ДОДАТОК 1

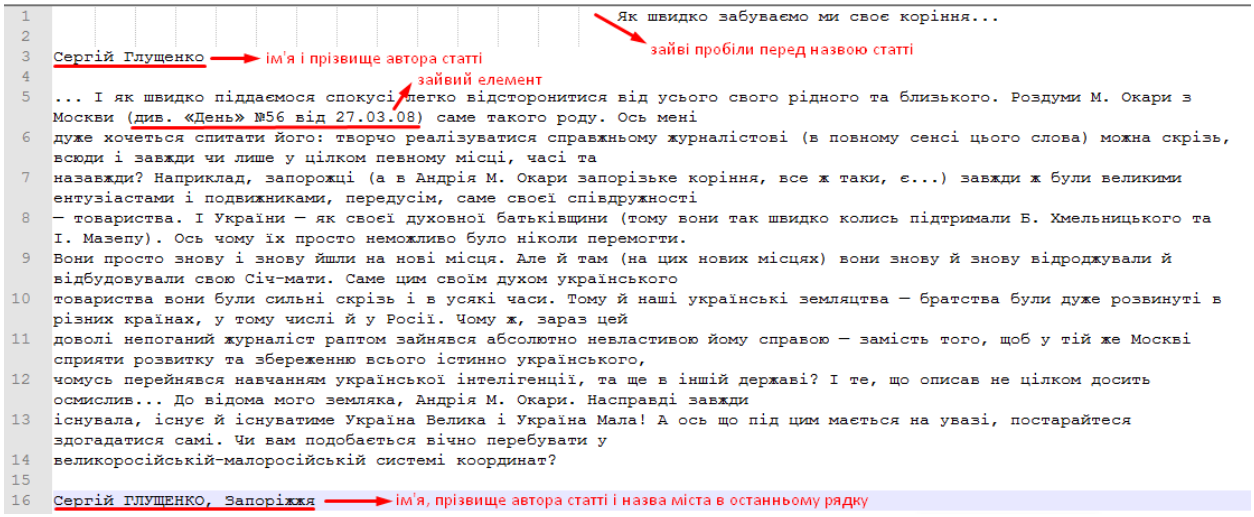


Рис. 1. Вигляд тексту до редагування (елементи до вилучення позначені червоним кольором).

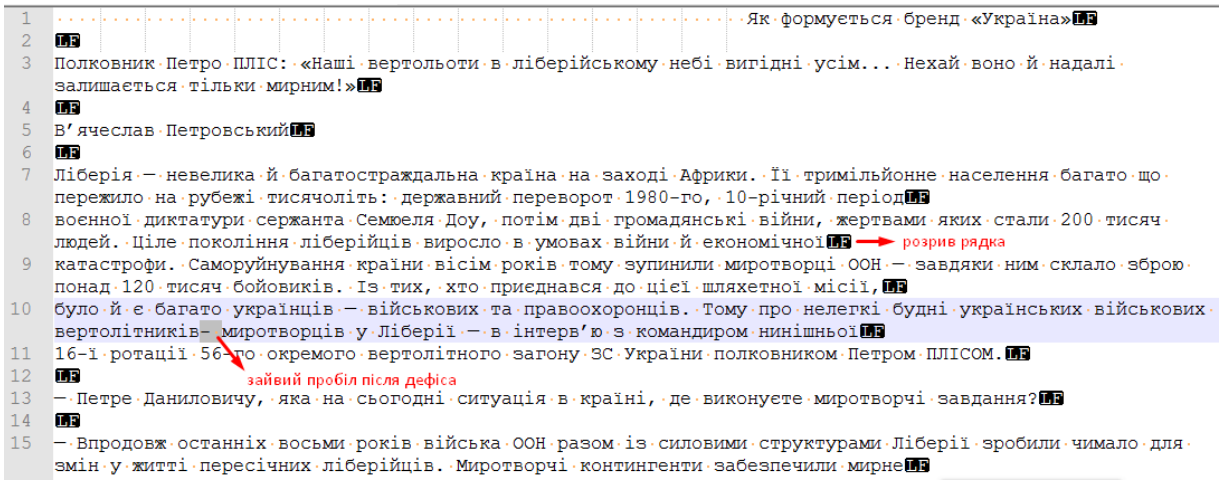


Рис. 2. Вигляд тексту до редагування (елементи до вилучення позначені червоним кольором).

сьогодні. По-друге,  питання результатів. «Коли я жертвую ресурси та НКО
(некомерційна організація. — О. Г.) їх отримує, що саме вона з ними зробить, щоб вплинути на долі
людей?» — щодня ставлять собі питання жертводавці. Саме тому неприбуткові
організації мають загострити фокус на аналізі своїх результатів у сфері зміни якості життя людей.
Організації, які це зрозуміють, матимуть найвищу якість взаємовідносин із
донорами, залучатимуть більше грошей та більше партнерів.

— На які дані ви посилаєтесь у даному випадку?

інтернет-посилання

— «Уряд» меценатів знаходиться у Школі благодійності Університету у штаті Індіана
(http://www.philanthropy. iupui.edu/), яку вважають найавторитетнішим джерелом ресурсів про
благодійність. По-перше, кожні два роки разом із Bank of America вони роблять спеціальний огляд, що
містить емпіричні дані, на які я покладаюся. По-друге, наша команда (тобто
співробітники компанії Westfall Group. — О. Г.) повсякчас шукає у Інтернеті будь-які дослідження щодо
благодійності. Наприклад, одного разу ми наштовхнулись на огляд, який нас
насправді зацікавив: то було дослідження про третій сектор у Азії. З нього слідувало, що місто №1, що
«дає» гроші, — це Пекін. Саме тому, що ми шукаємо актуальні новини та дані
про благодійність, агенція може претендувати на ефективність.

Рис. 3. Вигляд тексту до редагування (елемент до вилучення – інтернет-посилання).

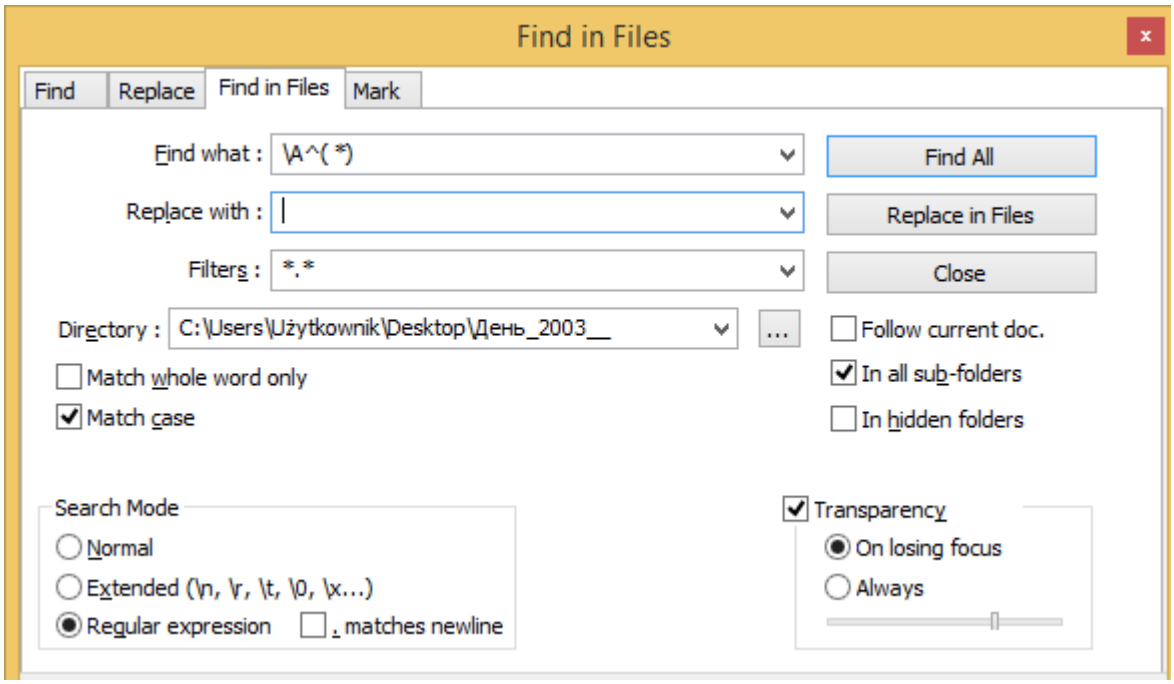


Рис. 4. Ілюстрація видалення зайвих пробілів перед назвою статті за допомогою опції *Find in Files*.

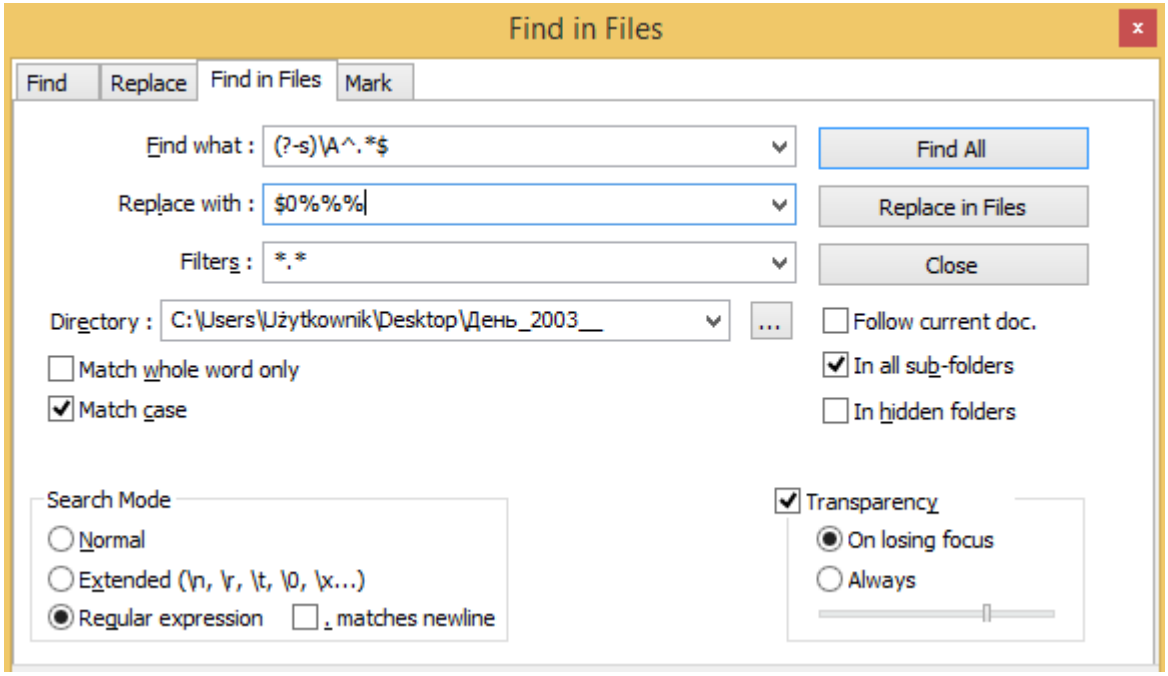


Рис. 5. Проставлення технічного символу %%% і розриву рядка в кінця першого рядка за допомогою опції *Find in Files*.

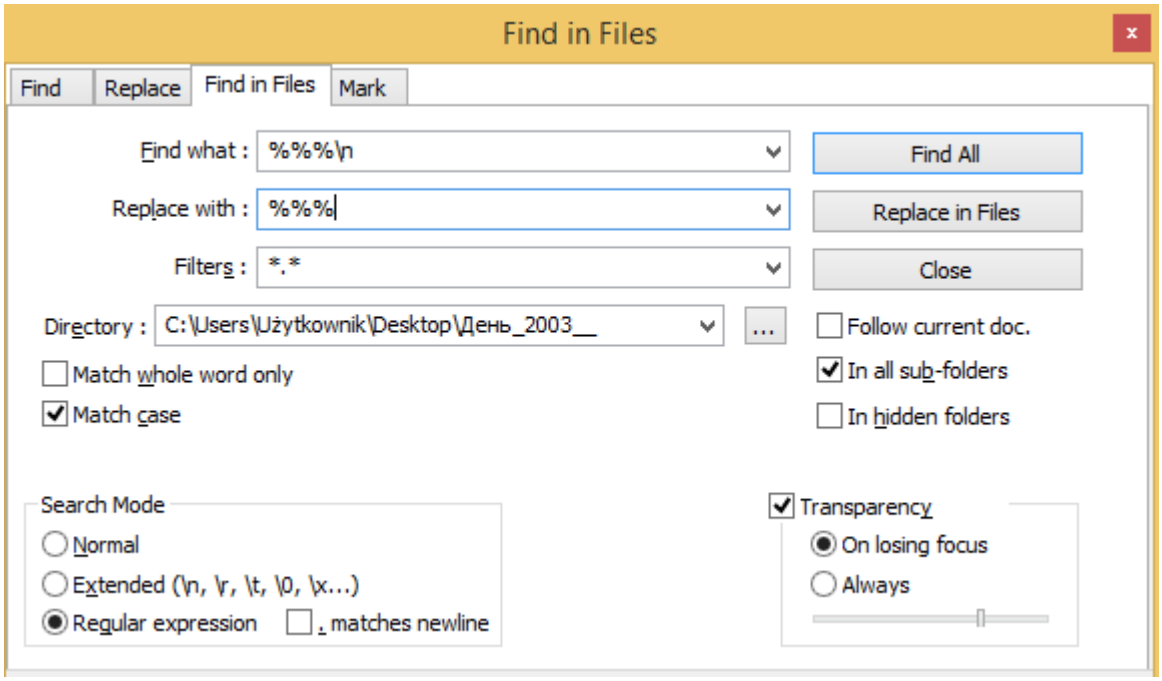


Рис. 6. Вилучення розриву рядка в кінці першого рядка файлу, користуючись опцією *Find in Files*.

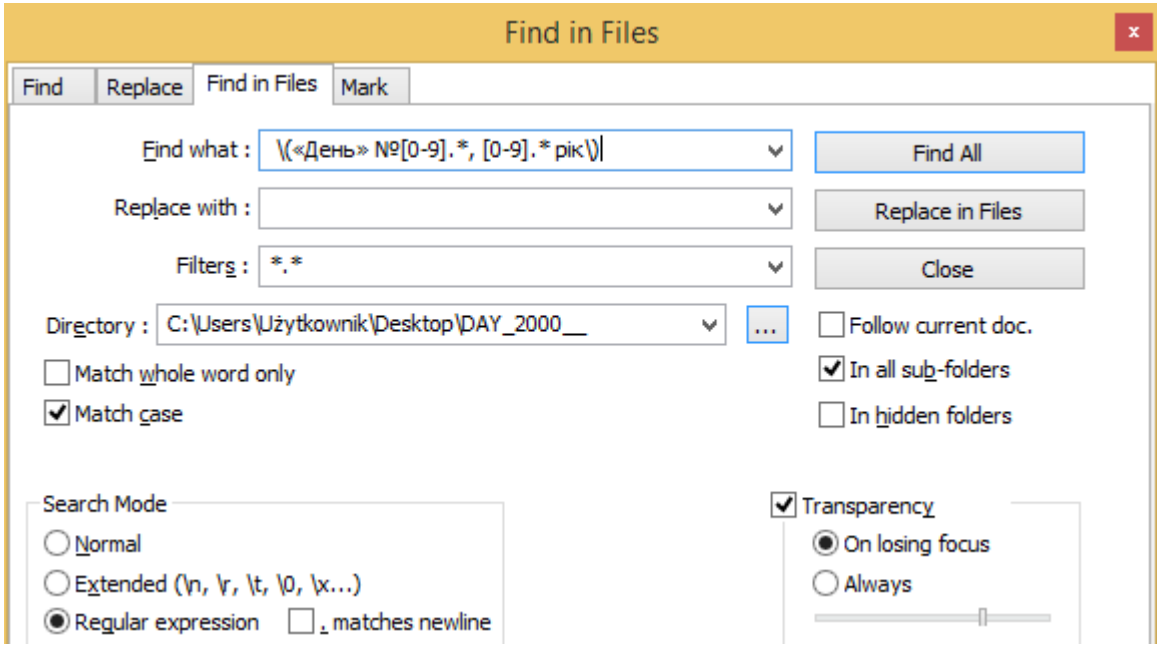


Рис. 7. Видалення вставки *«День» №211, 1999 рік*» за допомогою регулярного виразу із заміною на ніщо.

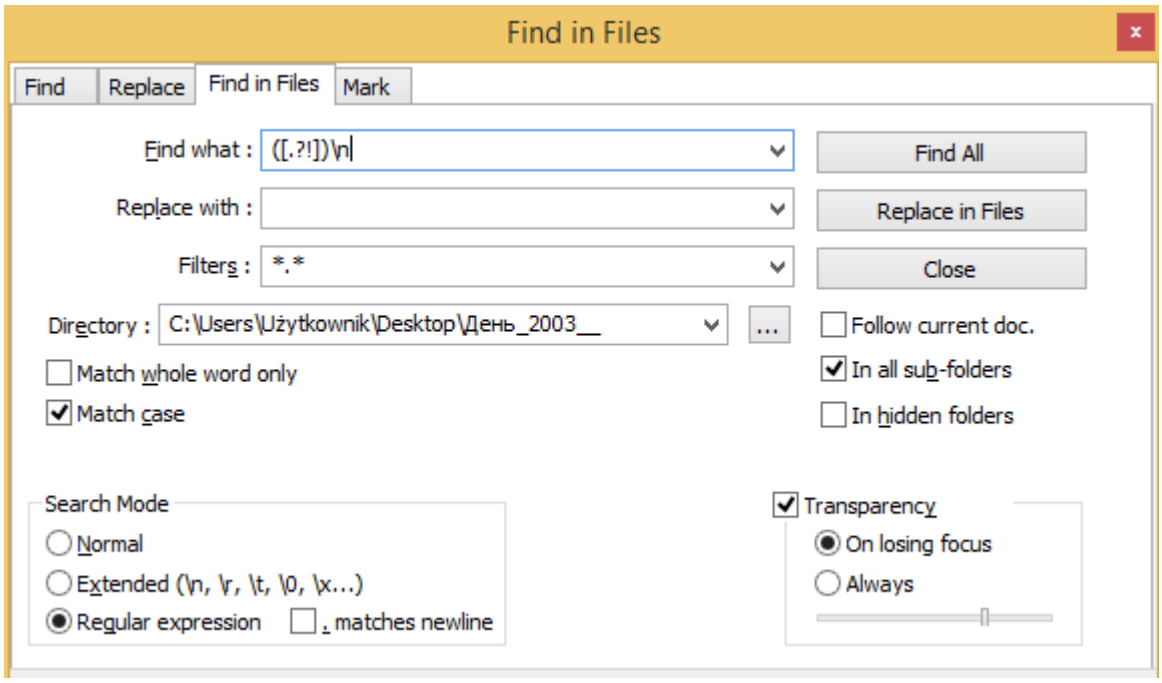


Рис. 8. Пошук правдивих розривів рядків в текстах «Дня» 2003 року.

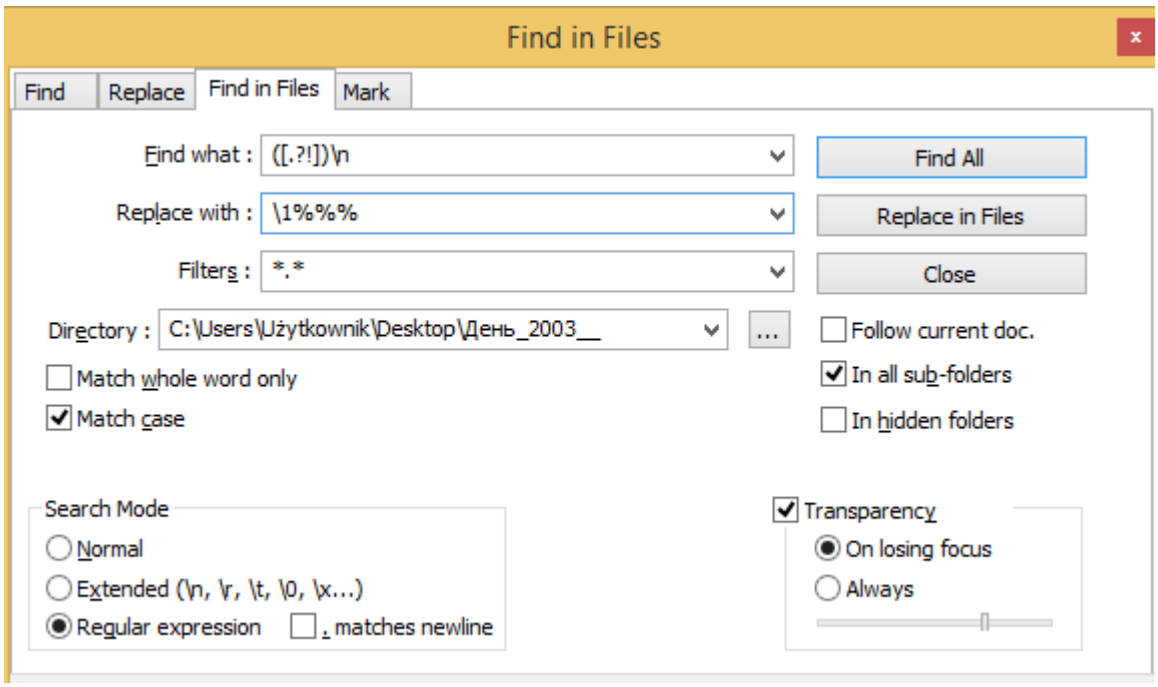


Рис. 9. Заміна правдивих розривів рядків на допоміжний символ %%% у текстах 2003 року.

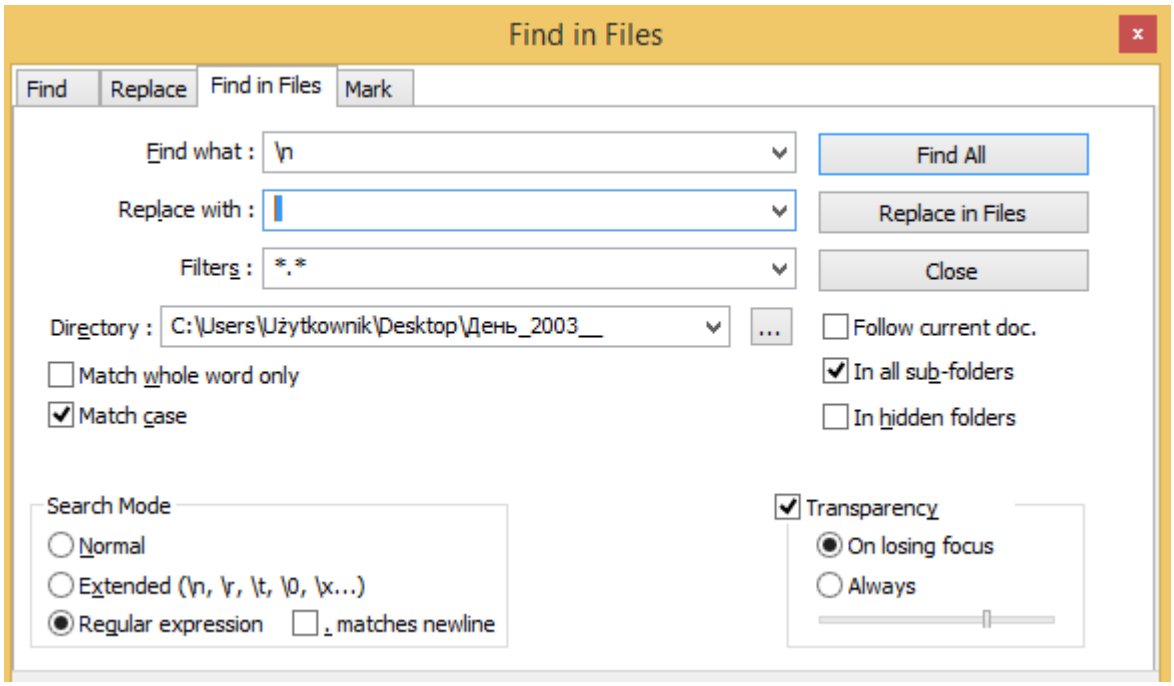


Рис. 10. Заміна зайвих розривів рядків на проміжок (пробіл).

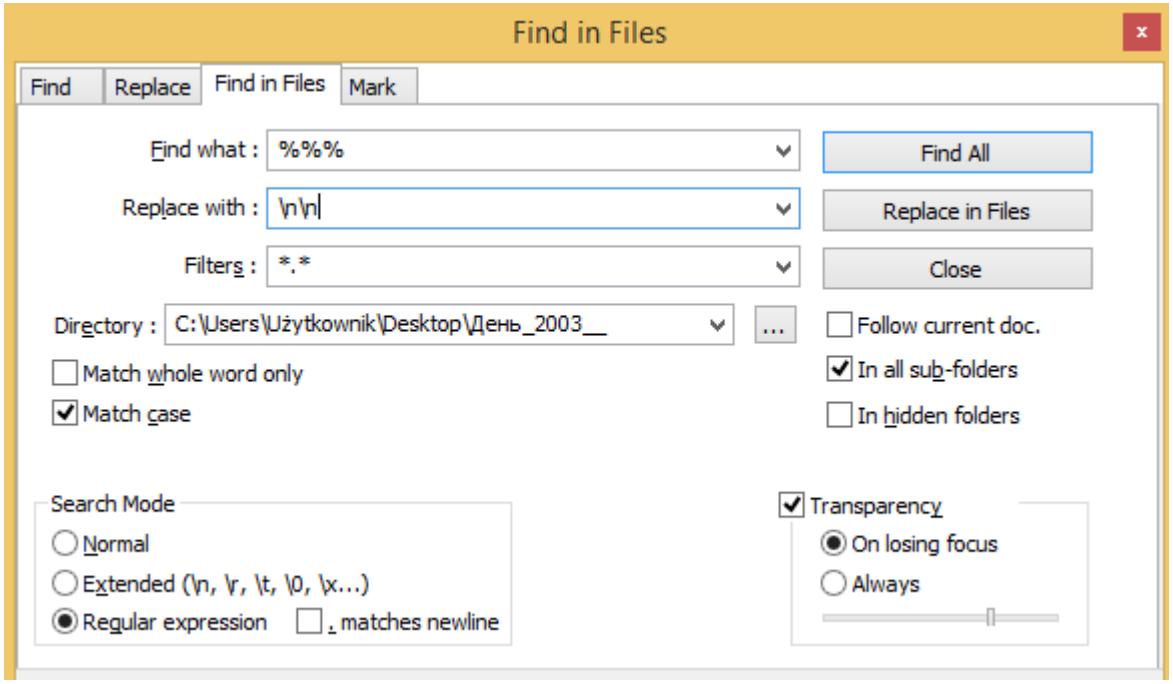


Рис. 11. Відновлення абзаців в текстах.

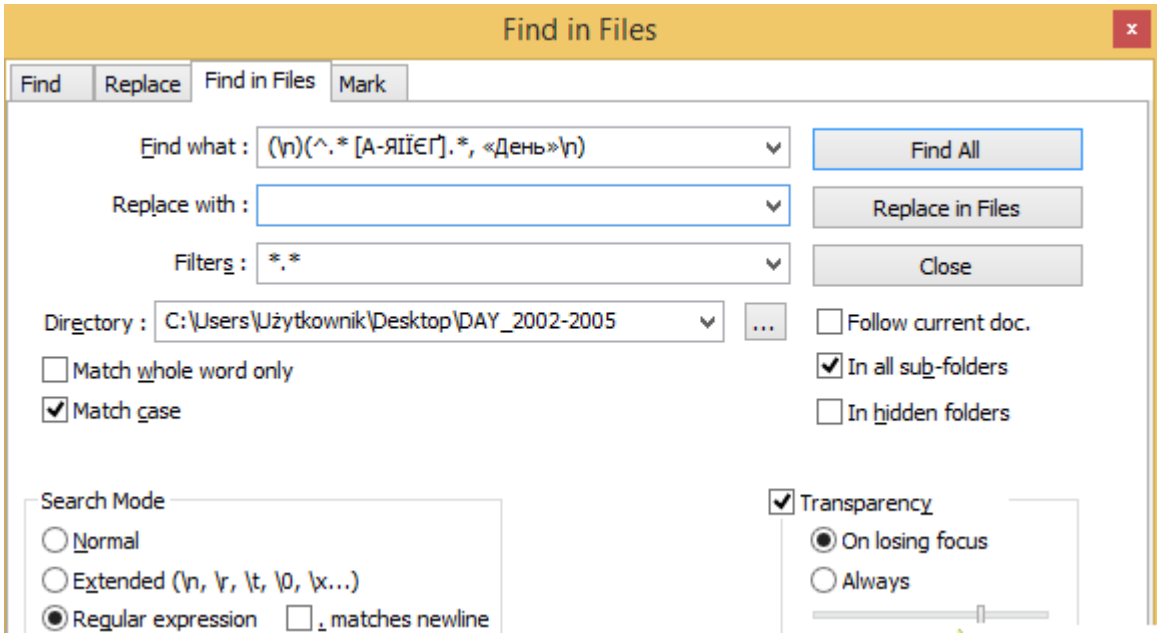


Рис. 12. Видалення вставки в кінці тексту такого типу «*Підготувала Клара Гудзик, «День»*».

# ДОДАТОК 2

## Список ключових слів газети «День»

| Слово | Абс. частота у фок. підкорпусі | Абс. частота у реф. підкорпусі | BIC |
|---|---|---|---|
| абсолютно | 9480 | 11200 | 8336,04 |
| авторитаризм | 417 | 492 | 348,93 |
| агресія | 3375 | 3343 | 3540,52 |
| ажіотаж | 414 | 191 | 717,89 |
| Азаров | 1074 | 69 | 3158,8 |
| академія | 7371 | 9115 | 6149,41 |
| акциз | 497 | 441 | 559 |
| акція | 12180 | 10339 | 14680,4 |
| альянс | 3738 | 2124 | 5931,37 |
| ансамбль | 1786 | 1152 | 2611,42 |
| аншлаг | 404 | 107 | 880,26 |
| АТО | 3066 | 1040 | 6230,17 |
| аутсайдер | 174 | 204 | 135,53 |
| бажаючий | 1049 | 1093 | 1034,87 |
| Байден | 267 | 78 | 555,96 |
| балет | 2423 | 597 | 5499,53 |
| балетмейстер | 348 | 49 | 893,39 |
| балетний | 581 | 161 | 1255,13 |
| барель | 307 | 147 | 517,31 |
| бартер | 92 | 152 | 32,72 |
| безумовно | 4095 | 5251 | 3263,83 |
| Бетховен | 297 | 307 | 281,62 |
| Бі-Бі-Сі | 1338 | 78 | 3982,21 |
| бібліотека | 5160 | 12589 | 1249,81 |
| бієнале | 366 | 78 | 850,49 |
| блокбастер | 157 | 83 | 241,21 |
| БЮТ | 2235 | 107 | 6794,67 |
| Вагнер | 341 | 299 | 381,53 |
| вакуум | 401 | 1129 | 42,34 |
| вважаю | 5776 | 3264 | 9205,12 |
| вважає | 12538 | 15257 | 10669,8 |
| вето | 700 | 400 | 1091,54 |
| видобуток | 1968 | 1630 | 2403,43 |
| визначитися | 1757 | 1722 | 1852,37 |
| вистава | 9575 | 3891 | 18075,68 |
| відставка | 3665 | 2327 | 5433,79 |
| віце-президент | 1503 | 479 | 3118,15 |
| віце-прем'єр | 2097 | 137 | 6174 |
| вокал | 396 | 131 | 795,88 |
| вокаліст | 404 | 105 | 885,74 |
| воркшоп | 71 | 62 | 65,58 |
| ВР | 2357 | 716 | 4987,85 |
| Газпром | 2432 | 270 | 6646,77 |
| гала-концерт | 214 | 22 | 574,94 |
| гектар | 1657 | 732 | 2993,57 |
| гендиректор | 432 | 21 | 1296,07 |
| геноцид | 2348 | 1378 | 3648,06 |
| генпрокуратура | 1013 | 289 | 2180,76 |
| говорить | 9754 | 15021 | 6106,8 |
| голодомор | 4728 | 1843 | 9081,34 |
| голокост | 895 | 471 | 1468,7 |
| Гонгадзе | 1976 | 94 | 6008,53 |
| гран-прі | 711 | 40 | 2115,01 |
| гривневий | 211 | 107 | 338,72 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | ескалація | 367 | 376 | 355,72 |
| гривня | 15044 | 3814 | 33942,68 | естрадний | 566 | 277 | 958,86 |
| Гройсман | 553 | 193 | 1095,42 | Єврокомісія | 868 | 251 | 1857,11 |
| грунт | 945 | 2556 | 148,66 | європейськість | 180 | 133 | 222,49 |
| дебют | 694 | 267 | 1324,27 | Євросоюз | 4727 | 1587 | 9654,38 |
| девальвація | 634 | 401 | 926,46 | завдяки | 9578 | 31010 | 752,36 |
| демократія | 8186 | 9188 | 7606,65 | зазначив | 6772 | 3345 | 11626,16 |
| депутат | 19757 | 10230 | 33115,69 | законопроект | 6936 | 3240 | 12252,25 |
| депутатка | 82 | 52 | 102,97 | заповідник | 2308 | 1391 | 3528,2 |
| депутатство | 78 | 49 | 97,7 | заявив | 11802 | 5175 | 21515,13 |
| держсекретар | 1184 | 220 | 2897,21 | збитковий | 321 | 238 | 410,7 |
| де-факто | 703 | 512 | 934,38 | згідно з | 15340 | 45550 | 1873,08 |
| дефіцит | 2654 | 3050 | 2391,79 | здешевлення | 243 | 161 | 332,57 |
| дефолт | 447 | 162 | 868,13 | Івшина | 2181 | 46 | 7007,15 |
| джаз | 694 | 669 | 731,44 | імпічмент | 370 | 343 | 395,3 |
| Джамала | 150 | 76 | 235,38 | імпортер | 434 | 350 | 526,45 |
| диригент | 1830 | 597 | 3767,3 | інавгурація | 407 | 122 | 849,59 |
| дистриб’ютор | 153 | 100 | 204,29 | інвестувати | 1097 | 582 | 1796,4 |
| долар | 9173 | 6863 | 12183,29 | інсталяція | 423 | 160 | 805,64 |
| Дональд | 624 | 361 | 963,79 | інфляція | 2393 | 2469 | 2408,47 |
| досконально | 49 | 71 | 14,43 | істеблішмент | 335 | 257 | 418,34 |
| дотація | 662 | 456 | 914,35 | кабмін | 3403 | 261 | 9836,51 |
| дует | 814 | 373 | 1434,84 | капела | 463 | 392 | 540,82 |
| екзит-пол | 297 | 27 | 821,09 | касетний | 375 | 212 | 579,6 |
| екс-міністр | 728 | 46 | 2139,03 | Качиньський | 336 | 2 | 1105,8 |
| експозиція | 2850 | 1428 | 4845,24 | квартет | 447 | 252 | 695,71 |
| експонувати | 91 | 73 | 95,74 | квота | 1202 | 1016 | 1436,52 |
| експортер | 718 | 493 | 995,51 | Кінах | 858 | 19 | 2738,38 |
| Ердоган | 377 | 266 | 503,76 | Кличко | 1294 | 230 | 3203,07 |
| есдек | 159 | 87 | 239,52 | Клінтон | 1222 | 605 | 2079,71 |

| | | | |
|---|---|---|---|
| коаліційний | 1106 | 278 | 2484,38 |
| коаліція | 6422 | 1883 | 13795,57 |
| композитор | 4174 | 1691 | 7879,71 |
| конкурс | 11375 | 4239 | 22299,8 |
| консерваторія | 771 | 367 | 1332,13 |
| консорціум | 1069 | 473 | 1923 |
| контингент | 1001 | 1298 | 772,57 |
| концерн | 677 | 439 | 974,59 |
| концерт | 7043 | 3775 | 11571,19 |
| Кофман | 187 | 50 | 395,88 |
| кредитувати | 196 | 74 | 363,33 |
| кубометр | 1458 | 154 | 4010,57 |
| кулуари | 401 | 127 | 819,84 |
| Кучма | 3449 | 1292 | 6733,4 |
| Лариса | 4041 | 2943 | 5461,99 |
| лауреат | 2677 | 1015 | 5195,23 |
| Лукашенко | 1542 | 277 | 3810,77 |
| люстрація | 415 | 298 | 549,43 |
| Маккейн | 298 | 38 | 777,6 |
| Марчук | 1433 | 265 | 3514,73 |
| МВФ | 2913 | 725 | 6594,96 |
| Медведєв | 1082 | 159 | 2793,14 |
| Медведчук | 1057 | 156 | 2725,66 |
| Мейс | 1510 | 267 | 3745,66 |
| месидж | 292 | 74 | 640,14 |
| минулий | 14661 | 10598 | 19972,57 |
| миротворчий | 1107 | 768 | 1534,23 |
| мільйон | 16338 | 9353 | 25876,29 |
| мільярд | 7674 | 2072 | 16957,56 |
| міноборони | 690 | 107 | 1753,74 |
| мінфін | 957 | 105 | 2609,18 |
| міськрада | 1635 | 713 | 2971,9 |
| МОЗ | 1272 | 313 | 2879,15 |
| мораторій | 977 | 212 | 2290,64 |
| музика | 8991 | 14661 | 5155,73 |
| мюзикл | 586 | 107 | 1430,52 |
| наголосив | 3026 | 1281 | 5593,95 |
| наголосити | 4318 | 4950 | 3913,88 |
| нардеп | 838 | 255 | 1759,94 |
| НАТО | 13598 | 2856 | 32419,57 |
| Нафтогаз | 2320 | 268 | 6296,26 |
| нацизм | 576 | 459 | 711,69 |
| Нацрада | 866 | 55 | 2546,67 |
| незалежність | 9569 | 14293 | 6262,44 |
| непросто | 917 | 1112 | 765,73 |
| нещодавно | 7910 | 4613 | 12381,61 |
| Обама | 2660 | 1373 | 4449,59 |
| облдержадміністрація | 2014 | 224 | 5499,36 |
| обленерго | 427 | 18 | 1297,04 |
| облрада | 921 | 106 | 2489,59 |
| однак | 20110 | 74635 | 579,7 |
| ОДА | 1128 | 187 | 2836,16 |
| олігарх | 2917 | 1117 | 5638,77 |
| олігархат | 152 | 31 | 346,36 |
| опера | 150 | 48 | 293,49 |
| оперета | 518 | 286 | 819,22 |
| опозиціонер | 1134 | 241 | 2677,35 |
| оркестр | 3518 | 1950 | 5661,83 |
| охочий | 1194 | 1504 | 958,71 |
| патріотизм | 1602 | 2212 | 1151,39 |
| патріотичний | 1660 | 2126 | 1313,66 |
| перемога | 7925 | 11023 | 5716,13 |
| по-друге | 4646 | 11815 | 991,09 |
| по-перше | 5398 | 13146 | 1315,16 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| піаніст | 1032 | 343 | 2101,03 | профіцит | 181 | 75 | 319,93 |
| підтримати | 6325 | 6623 | 6305,19 | Пукач | 454 | 27 | 1335,7 |
| | | | | Путін | 11687 | 1472 | 31299,85 |
| підтримка | 17402 | 25130 | 11956,08 | путінський | 906 | 141 | 2306,91 |
| пікет | 432 | 196 | 756,27 | Пуччіні | 151 | 17 | 393,76 |
| повідомив | 7575 | 5199 | 10688,63 | радувати | 775 | 664 | 909,11 |
| | | | | раунд | 670 | 425 | 978,31 |
| повідомляє | 10202 | 2238 | 24026,15 | реально | 2935 | 3631 | 2435,84 |
| | | | | реванш | 368 | 370 | 363,57 |
| подорожчання | 453 | 150 | 912,85 | реверсний | 230 | 60 | 495,4 |
| Поклітару | 149 | 11 | 414,5 | регіонал | 1372 | 105 | 3955,52 |
| покращання | 591 | 290 | 1000,66 | режисер-постановник | 181 | 33 | 428,79 |
| політреформа | 503 | 18 | 1551,07 | реприватизація | 352 | 11 | 1090,01 |
| помаранчевий | 2603 | 1499 | 4092,24 | РНБО | 883 | 91 | 2431,14 |
| | | | | розповідає | | | |
| Пономарьов | 193 | 96 | 311,55 | ротація | 555 | 375 | 773,73 |
| попри | 5126 | 15244 | 609 | Саакашвілі | 1310 | 447 | 2544,8 |
| популізм | 706 | 204 | 1507,34 | | | | |
| Порошенко | 3001 | 1181 | 5733,29 | саміт | 3278 | 1218 | 6420,55 |
| | | | | Саркозі | 687 | 104 | 1755,26 |
| поспіль | 1325 | 1747 | 1007,56 | сепаратист | 1005 | 594 | 1543,88 |
| постгеноцидний | 175 | 11 | 499,96 | симфонія | 744 | 633 | 877,09 |
| | | | | синдикат | 533 | 130 | 1198,62 |
| пострадянський | 2627 | 3533 | 1966,44 | соборність | 356 | 582 | 184,94 |
| президент | 51694 | 18587 | 102923,1 | солідарність | 1631 | 2196 | 1211,85 |
| прем'єр | 879 | 166 | 2137,16 | | | | |
| прес-конференція | 3110 | 587 | 7611,13 | сопрано | 287 | 151 | 458,03 |
| | | | | соціал-демократ | 775 | 706 | 862,3 |
| прес-реліз | 648 | 165 | 1441,72 | спектакль | 1937 | 974 | 3280,73 |
| прес-секретар | 979 | 160 | 2467,07 | співвітчизник | 1270 | 1218 | 1360,82 |
| приватизація | 3672 | 1393 | 7131,72 | спікер | 1928 | 377 | 4668,55 |
| приз | 2150 | 782 | 4242,17 | сталінський | 1611 | 2323 | 1091,72 |
| призер | 470 | 69 | 1202,71 | | | | |
| прийняти | 9579 | 27408 | 1359,77 | стенд | 586 | 695 | 495,13 |
| примітно | 344 | 310 | 375,78 | стурбованість | 556 | 623 | 499,74 |
| причому | 6426 | 12047 | 2888,09 | | | | |
| протест | 4618 | 4643 | 4783,5 | субсидія | 989 | 911 | 1094,86 |
| проукраїнський | 515 | 237 | 898,83 | сценічний | 1399 | 681 | 2404,91 |

| | | | |
|---|---|---|---|
| сценограф | 250 | 20 | 700,91 |
| сценографія | 443 | 71 | 1110,84 |
| сьогоднішній | 5594 | 6167 | 5290,72 |
| театр | 19229 | 9565 | 32926,42 |
| театрал | 195 | 95 | 318,6 |
| театральний | 4424 | 3180 | 6037,53 |
| тендер | 1365 | 301 | 3193,33 |
| тенор | 343 | 289 | 395,13 |
| тероризм | 2108 | 2505 | 1826,76 |
| Тимошенко | 7798 | 1304 | 19680,52 |
| тисяча | 24272 | 29985 | 20316,75 |
| Тігіпко | 827 | 72 | 2335,16 |
| товарообіг | 420 | 334 | 514,61 |
| тонна | 4453 | 1512 | 9053,66 |
| Трамп | 1370 | 1051 | 1769,94 |
| транзит | 1407 | 899 | 2065,83 |
| транш | 473 | 95 | 1122,88 |
| Турчинов | 386 | 64 | 957,91 |
| УЄФА | 680 | 19 | 2138,6 |
| українець | 18794 | 24021 | 15107,27 |
| Укрнафта | 602 | 24 | 1844,82 |
| Укртелеком | 534 | 42 | 1522,34 |
| урядовець | 1252 | 1905 | 781,16 |
| фест | 503 | 179 | 986,74 |
| філармонія | 1331 | 304 | 3081,89 |
| форум | 5137 | 3459 | 7336,41 |
| хіт | 537 | 416 | 677,03 |
| холдинг | 612 | 257 | 1120,32 |
| хореографічний | 465 | 84 | 1134,22 |
| хореографія | 440 | 80 | 1070,49 |
| хормейстер | 166 | 22 | 421,42 |
| хоча | 22348 | 79647 | 931,15 |
| церква | 10491 | 38685 | 313,12 |
| ціна | 16819 | 21681 | 13378,95 |
| Черновецький | 522 | 22 | 1589,9 |
| чиновник | 6867 | 4775 | 9601,7 |
| школа | 17606 | 46417 | 3391,88 |
| шовінізм | 221 | 497 | 48,35 |
| Шопен | 321 | 198 | 467,24 |
| шоу | 1754 | 1203 | 2461,43 |
| щоправда | 5555 | 11407 | 2072,92 |
| щороку | 2365 | 2346 | 2471,85 |
| ювілей | 2348 | 1284 | 3801,7 |
| Ющенко | 10945 | 1218 | 29968,02 |
| Янукович | 9308 | 1610 | 23310,18 |
| Яценюк | 1945 | 275 | 5076,69 |
| | | | |