ISSN 2307-8332. Вісник ОНУ. Сер.: Філологія. 2024. Т. 29, вип. 2(30)

УДК 004.912:811.161.2
DOI: https://doi.org/10.18524/2307-8332.2024.2(30).320406

Марія МАЛИШЕВА
Одеський національний університет імені І. І. Мечникова
PhD (доктор філософії з філології)
старша викладачка кафедри прикладної лінгвістики
м. Одеса
mariiamalysheva@onu.edu.ua
ORCID iD: https://orcid.org/0000-0002-1910-4833

КОРПУС ТЕКСТІВ ПУБЛІЧНИХ КАНАЛІВ МЕСЕНДЖЕРА TELEGRAM: СТВОРЕННЯ ТА ПРАКТИЧНЕ ЗАСТОСУВАННЯ

Анотація. *У статті представлено процес створення та конкретні приклади практичного застосування корпусу текстів публічних каналів месенджера Telegram. Метою розвідки обрано розроблення корпусу текстів публічних українськомовних каналів месенджера Telegram, що передбачало розв'язання таких завдань: розробити методику формування корпусу текстів із публічних Telegram-каналів, створити та апробувати скрипти для автоматизованого збирання, очищення і аналізу текстових даних, завантажити опрацьовані тексти в корпусний менеджер, визначити перспективи подальшого використання корпусу та його удосконалення. Розроблення корпусу текстів публічних українськомовних каналів месенджера Telegram виконано в три етапи: на першому етапі обрано в месенджері Telegram публічний канал, який став джерелом текстових даних, і завантажено історію публікацій; на другому етапі переведено отримані дані у формат, який можна використовувати в спеціалізованому програмному забезпеченні для створення корпусів текстів та керування ними; на третьому етапі завантажено попередньо-опрацьовані тексти в обраний корпус-менеджер. Для підготовки файлу написано два скрипти на мові програмування Python із використанням бібліотек SpaCy, pandas тощо (один скрипт для вилучення текстів дописів та збереження їх в окремий файл, і другий скрипт для очищення текстів та статистичного аналізу). Для ілюстрації можливостей корпусу в контексті дослідження мережевого дискурсу зроблено запити: пошук дієслів довжиною понад 15 літер, пошук хештегів, пошук власних назв, пошук атрибутивних словосполучень. Запити сформульовано за допомогою мови корпусних запитів CQL та регулярних виразів. Перспективи дослідження передбачають розширення корпусу текстами з Telegram каналів інших блогерів, вдосконалення етапу підготовки та фільтрування текстів, залучення іншого програмного забезпечення для створення та керування корпусами текстів.*

Ключові слова*: корпусна лінгвістика, корпус текстів, Telegram, українська мова, мережевий дискурс, Sketch Engine.*

Постановка проблеми в загальному вигляді. У сучасних умовах стрімкого розвитку інформаційно-комунікаційних технологій дослідження мереже-

вої комунікації набуває особливого значення. Месенджери стають важливим джерелом текстових даних, що відображають актуальні суспільні, політичні та культурні тенденції. Проте фіксуємо недостатню кількість спеціалізованих корпусів, що ґрунтуються виключно на текстах з Telegram, зокрема, українськомовних. Це обмежує можливості для системного аналізу специфіки мережевого дискурсу, виявлення мовних трендів і дослідження впливу цифрового середовища на мову, та спонукає до створення нового корпусу, який дасть змогу більш детально досліджувати мережеву комунікацію. Розробленню такого корпусу також сприяло проходження програми підвищення кваліфікації науково-педагогічних працівників у Єнському університеті за темою «Використання лінгвістичних корпусів у викладанні мовних дисциплін і дослідженнях мови» [3].

Ступінь розроблення проблеми в мовознавстві. Дослідження мережевого дискурсу та застосування корпусних методів у мовознавстві активно розвиваються. Відомі українськомовні корпуси, такі як ГРАК [4], Лабораторія Української [2], ПАВУК [6], UberText 2.0 [5] та інші, включають тексти з різних джерел, зокрема інтернету та месенджерів, також зафіксовано спробу створити українськомовний корпус повідомлень із Twitter для подальшого автоматизованого виявлення токсичних текстів [1], але корпусів, побудованих виключно на українськомовних текстах з Telegram, у відкритому доступі немає.

Мета дослідження полягає в створенні корпусу текстів публічних українськомовних каналів месенджера Telegram та передбачає розв'язання таких **завдань**: розробити методику формування корпусу текстів із публічних Telegram-каналів, створити та апробувати скрипти для автоматизованого збирання, очищення і аналізу текстових даних, завантажити опрацьовані тексти в корпусний менеджер, визначити перспективи подальшого використання корпусу та його удосконалення. Під час реалізації мети дослідження застосовано **методи** автоматизованого збирання та очищення текстових даних із залученням скриптів на мові програмування Python [9], використано бібліотеки для обробки текстів (pandas [8], SpaCy [11] тощо). Для створення корпусу використано платформу Sketch Engine [10]. У процесі аналізу даних застосовано пошукові запити на мові CQL, регулярні вирази, а також методи базового статистичного аналізу. **Джерельну базу** становлять тексти публічного українськомовного каналу Telegram, обраного для створення корпусу (канал Сергія Стерненка [12]).

Викладення основного матеріалу дослідження. Розроблення корпусу текстів публічних українськомовних каналів месенджера Telegram передбачає три етапи:

1) обрання в месенджері Telegram публічного каналу, який стане джерелом текстових даних, і завантаження історії публікацій в каналі;

2) переведення отриманих дані у формат, який можна використовувати у спеціалізованому програмному забезпеченні для створення корпусів текстів та керування ними;

3) завантаження попередньо-опрацьованих текстів в обраний корпус-менеджер.

Розглянемо кожний етап більш детально.

1. Обрання каналу, який стане джерелом текстових даних і завантаження історії публікацій. На першому етапі для створення корпусу обрано канал відомого блогера та волонтера Сергія Стерненка [12]. Для того, щоб завантажити історію дописів, необхідно натиснути на символ, що позначає три крапки, у верхньому правому кутку вікна (див. Рис. 1), відкривається спадне меню. У спадному меню необхідно обрати «Export chat history». З'являється вікно з переліком даних, які можна завантажити. Для створення корпусу текстів публічних українськомовних каналів месенджера Telegram потрібні лише тексти, тому необхідно прибрати прапорці з усіх пунктів. Telegram надає технічну можливість завантажити дані у двох форматах: HTML або JSON. З огляду на те, що текстові дані потребують опрацювання перед завантаженням у корпусний менеджер, для завантаження обрано формат JSON. У відкритому вікні необхідно натиснути на кнопку «SAVE», після цього історію каналу буде завантажено на пристрій (комп'ютер або ноутбук; для мобільної версії зазначеного функціоналу немає). Після завершення завантаження отримано файл result.json.

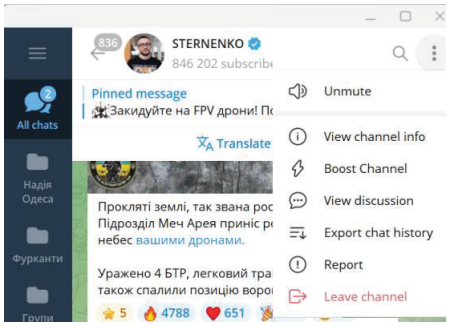

Рис. 1. Завантаження історії публікацій

2. Переведення отриманих дані у формат, який можна використовувати в спеціалізованому програмному забезпеченні для створення корпусів текстів та керування ними. Інформація у файлі result.json зберігається у вигляді таблиці, що містить велику кількість технічних даних (наприклад, айді автора повідомлення, дати, розміри зображення в дописі тощо). Авторкою написано скрипт на мові програмування Python [9] (із залученням вбудованої бібліотеки json), з метою дістати з файлу лише тексти дописів і зберегти їх у новому файлі з розширенням txt для подальшого опрацювання. Отримані тексти все ще не є придатними для завантаження в корпусний менеджер, оскільки вони містять зайві символи. З огляду на це, за допомогою скрипту на Python із застосуванням регулярних виразів та бібліотеки pandas [8] прибрано емодзі, посилання та зайві пробіли. Також за допомогою зокрема бібліотеки SpaCy [11] додатково проведено базовий статистичний аналіз текстів. Відповідно, набір даних налічує 840432 токени, 706954 слів та 65469 речень. Очищені тексти, готові до завантаження в корпусний менеджер, збережено у файлі cleaned_text.txt.

3. Завантаження попередньо-опрацьованих текстів в обраний корпус-менеджер. Отриманий файл cleaned_text.txt є готовим для завантаження в про-

грамне забезпечення для створення та керування корпусами текстів. Авторкою обрано платформу Sketch Engine [10], а саме її пробну версію, черезії поширеність та зручність у використанні. Процес завантаження текстів є інтуїтивно зрозумілим. Спочатку необхідно ввести назву корпусу, обрати формат – одномовний чи багатомовний (обрано одномовний), потім обрати мову корпусу – українську, за бажанням можна додати короткий опис корпусу. У наступному вікні обрано варіант «I have my own texts», оскільки підготовлено текст для формування корпусу. На наступному етапі відбувається завантаження тексту з його подальшим опрацюванням платформою Sketch Engine та компіляцією. У результаті отримано корпус текстів публічного каналу месенджера Telegram з такими основними статистичними характеристиками: кількість токенів складає 835287, кількість слів – 675618, кількість речень – 57781. Якщо порівняти з даними, отриманими в результаті статистичного аналізу датасету, виявляємо розбіжність у 5–10% відсотків, що, ймовірно, зумовлено різницею в методах токенізації та обробки тексту, а також у правилах розбиття тексту на речення. Отриманий корпус має автоматичну морфологічну розмітку, з переліком тегів можна ознайомитися на сторінці «General Info». Для ілюстрації можливостей корпусу зроблено запити.

Пошук довгих дієслів (дієслів, що містять понад 15 літер): [word=». {15,}»&tag=»V.*»]. Для написання запиту використано мову CQL та регулярні вирази. Сформовано конкорданс (шукані слова в контексті, див. Рис. 2) та частотні списки.

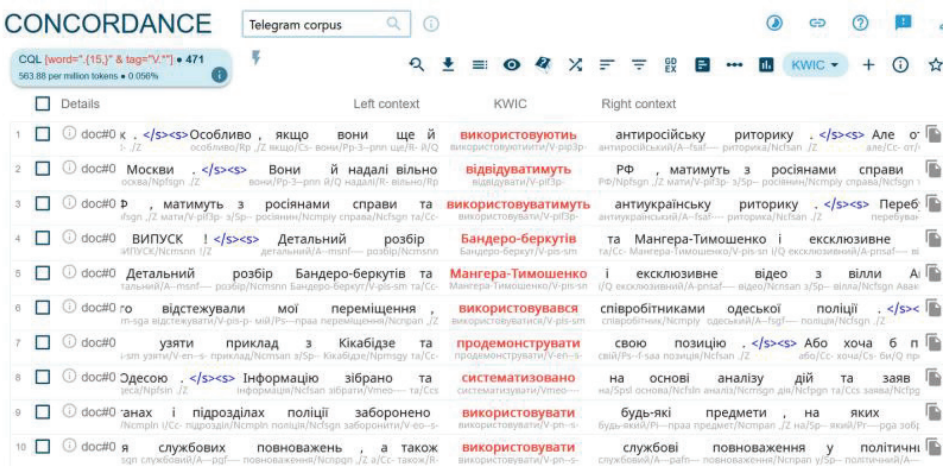

Рис. 2. Пошук слів, що містять понад 15 літер.

У частотних списках для зручності тут і далі подано лематизовані лексеми. Також тут і далі збережено авторську орфографію та пунктуацію. Загалом знайдено 471 випадок вживання таких дієслів. Найпоширенішими лемами є:

використовувати (85 випадків) та *використовуватися* (29 випадків). Також частотними є леми: *госпіталізувати* (16), *демілітаризувати* (16), *законтрактувати* (14), *спостерігатися* (13), *продемонструвати* (12), *застосовуватися* (11), *підтверджуватися* (11), *держав-союзниця* (6). Поміж поодиноких лем, зокрема, виявлено: *тренуватися, продовжувати, відвертатися, вдосконалюватися, оголошувати, текла-червоніла, віце-президентка, Експерментувати, завантажуватися, електропередачі, поп-фсбшник*.

У конкордансі та частотних списках виявлено похибки розмітки, наприклад, такі іменники, як *віце-президентка, поп-фсбшник, держава-союзниця* та інші теговано як дієслова. Також виявлено, що не всі лексеми приведено до правильної леми, наприклад, *текла-червоніла, використовуваютись* та інші, що свідчить про недосконаліть алгоритмів морфологічного аналізу, які використовуються для розмітки текстів у корпусі. Такі помилки можуть виникати через складність української мови, зокрема, через нетипові або нові лексеми, які не завжди правильно обробляються автоматизованими системами.

Пошук хештегів: [word=»#.*»] (див. конкорданс на Рис. 3). Такий пошуковий запит є актуальним в контексті дослідження мережевого дискурсу, зокрема для вивчення трендів та тематичного розподілу текстів. Загалом у корпусі знайдено 521 випадок використання хештегів. Найпоширенішим хештегом є *#звітую* (335 випадків). Розуміння цього вбачаємо в волонтерській діяльності автора каналу. Також частотними є хештеги, пов'язані з обговоренням соціальних чи політичних питань: *#Гандзюк* (21 випадок), *#літакМедведчука* (20 випадків), *#Гандзюк.* (8), *#колисядеМедвудчук* (4). Здебільшого вживаються українськомовні хештеги, але присутні й англомовні (наприклад, *#stoprussia* (8), *#SendNatoToUkraine* (4), *#CloseTheSky* (4), *#BanRussafronSwift* (4)), що свідчить про орієнтацію на міжнародну аудиторію.

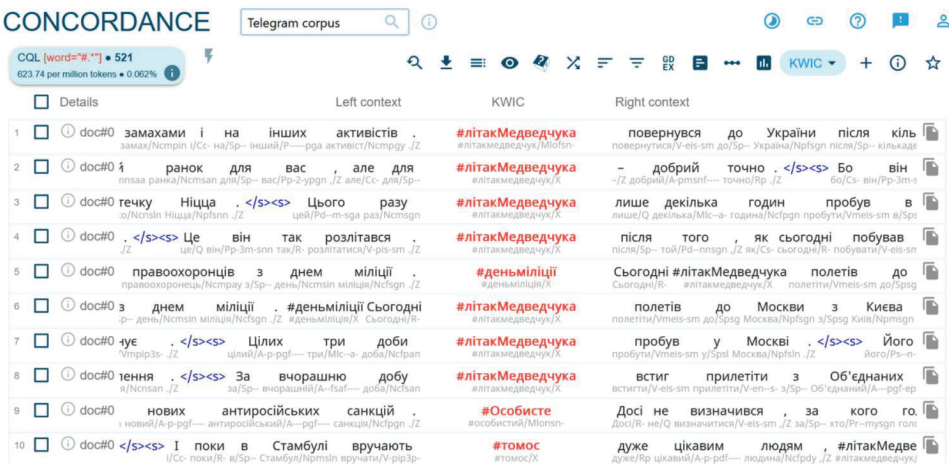

Рис. 3. Пошук хештегів

Пошук власних назв: [tag=»NP.*»] (див. конкорданс на Рис. 4). Такий пошук є цікавим у контексті вивчення трендів у суспільному дискурсі.

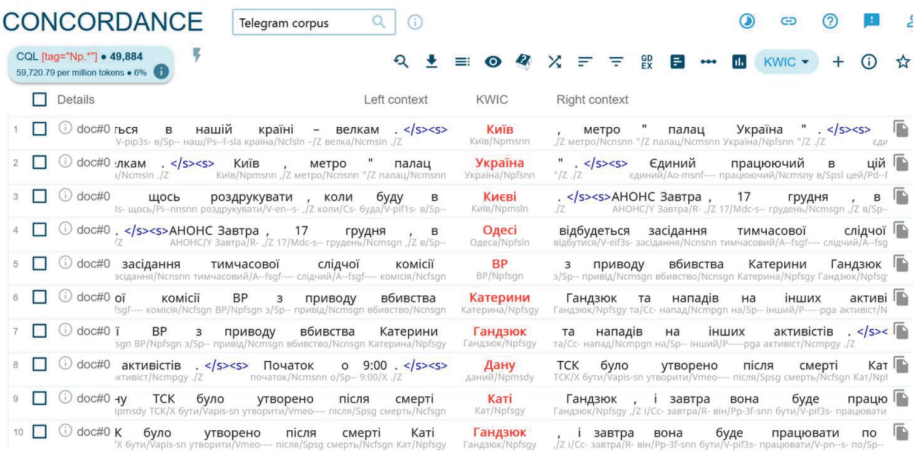

Рис. 4. Пошук власних назв

Загалом ідентифіковано 49884 випадки використання власних назв. Найчастотнішими є *Україна* (3894 вживання), *росія* (1480) та *Київ* (1255). Також частотними є *Крим* (823), *Одеса* (723), *Зеленський* (644), *ЗСУ* (644), *СБУ* (639), *путін* (622), *РФ* (596), *США* (489). До поодиноких вживань (із частотою 1) належать, зокрема, *крейсер, НЕМА, КРАСА, Старовєрів, Михоєв, Маукєнов, Апреленка, республіка, Лємєновий, Зелінський, ярмак*. Отже, у корпусі домінують географічні назви Україна, Київ, Одеса, Крим, що свідчить про часте обговорення регіонів і міст у текстах. Частотними є згадки персоналій Зеленського та путіна, що відображає суспільний інтерес. Також виявлено хибно анотовані лексеми *НЕМА*, *КРАСА*. Морфологічні аналізатори часто покладаються на формальні ознаки (великі літери тощо), і якщо загальне слово написане з великої літери, його можуть помилково визначити як власну назву.

Пошук атрибутивних словосполучень є доречним у контексті дослідження суспільних настроїв. Зроблено два запити. Перший запит передбачав пошук сполучень прикметника з іменником, що позначає власну назву, істоту: [tag=»A.*»][tag=»NP...Y»] (див. конкорданс на Рис. 5).

У корпусі ідентифіковано 254 випадки конструкцій, що відповідають заданому шаблону. Найчастотнішим результатом є словосполучення *обов'язковий Публікувати* (27 вживань). Менш частотними є словосполучення *проросійський Олег* (4), *проросійський Мураєва* (4), *п'ятирічний Кирил* (3), *УН Льовочкін* (3), *клятий Степаненко* (3), *запобіжний Антоненко* (2), *засуджений Сергій* (2), *судимий Кива* (2), *розвідувальний БпЛА* (2). Поодиноко вживаються такі словосполучення: *Шановний Паня, загиблий Владислав, запропонований Джонсон,*

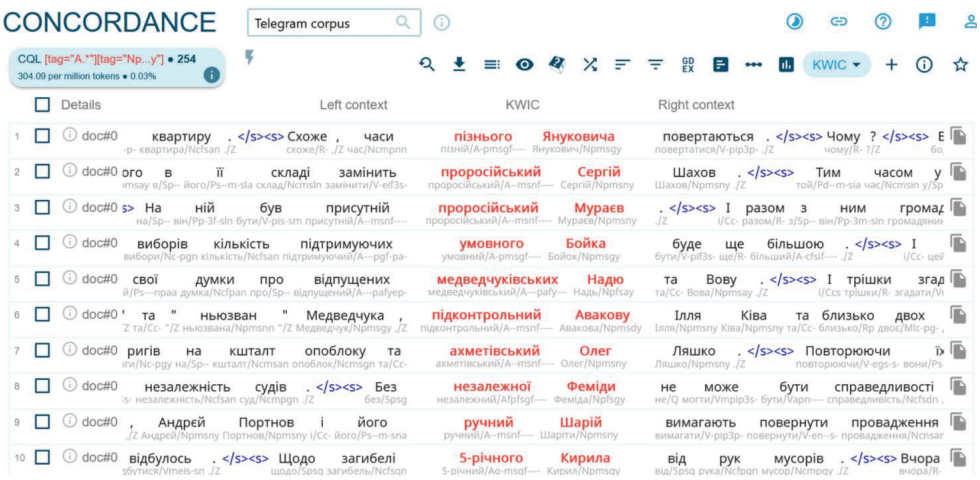

Рис. 5. Пошук сполучень прикметника з іменником, що позначає власну назву, істоту

базований Іван, харківський обло, чудовий дроном-камікадзе, заснований Олена, вечірній Байден, державний корда, російський дрон-камікадзе. Переважно атрибутивні словосполучення вживаються в корпусі поодиноко. Найчастотніша конструкція *обов'язковий Публікувати* є результатом помилки тегування, коли дієслово неправильно ідентифікується як власна назва. Вважаємо, алгоритм на основі формальних ознак (великої літери та контексту) помилково позначив дієслово як власну назву. Другий запит передбачав пошук атрибутивних словосполучень за шаблоном «так званий + іменник» та або «так званий + прикметник + іменник»: [word=»так»][lemma=»званий»]([tag=»N.*»]|[tag=»A.*»][tag=»N.*»]) (див. конкорданс на Рис. 6).

У корпусі знайдено 372 випадки конструкцій, що відповідають заданому шаблону. Найбільш частотним результатом є *так званий росія* (337 випадків вживань). Менш частотними є *так званий білорусь* (5), *так званий спецоперація* (5), *так званий російський федерація* (2), так *званий російський культура* (2). Решта є поодинокими: *так званий губернатор, так званий ПМР, так званий російський опозиція, так званий Росія, так званий госдум, так званий Нагірний, так званий готель, так званий Нагірний Карабах, так званий парламент, так званий РФ, так званий асвабадітєль, так званий віцепрем'єр, так званий СВО, так званий білоруський суд*. Вираз «так званий» часто вживається для створення іронії, критики або сумніву щодо легітимності чи автентичності означуваного об'єкта, що підтверджено даними з корпусу.

Висновки й перспективи дослідження. Отже, здійснено спробу створити корпус текстів публічних каналів месенджера Telegram та продемонструвати можливі сфери його застосування. Корпус демонструє високий рівень частотності географічних назв, персоналій і тематично маркованих конструкцій, що

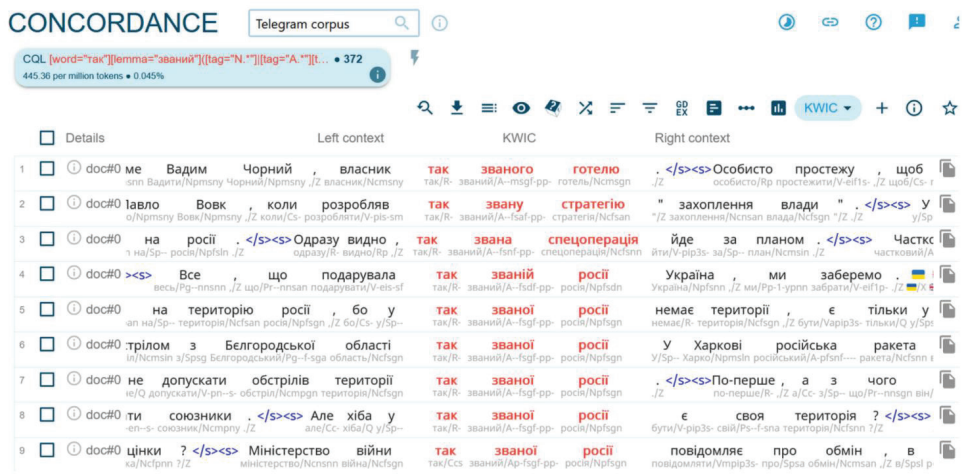

Рис. 6. Пошук атрибутивних словосполучень за шаблоном «так званий + іменник» та або «так званий + прикметник + іменник»

свідчить про домінування суспільно-політичної тематики. Виявлені атрибутивні словосполучення вказують на іронічно-критичний характер деяких текстів, що є характерним для мережевого дискурсу. Аналіз хештегів підкреслює увагу до волонтерської діяльності, міжнародних подій і соціальних трендів. Водночас, аналіз показав низку похибок морфологічного тегування, пов'язаних із формальними обмеженнями алгоритмів, що вказує на потребу в удосконаленні методів автоматичної обробки текстів. Розроблений корпус та скрипти потребують вдосконалення, тому перспективи дослідження передбачають розширення корпусу текстами з Telegram каналів інших блогерів, вдосконалення етапу підготовки та фільтрування текстів, залучення іншого програмного забезпечення для створення та керування корпусами текстів. Розроблені інструменти для підготовки текстів є відкритими для широкого загалу на GitHub [7].

Література

1. Bobrovnyk, K. (2019). Automated Building and Analysis of Ukrainian Twitter Corpus for Toxic Text Detection. *Proceedings of the* 3*d International Conference Computational Linguistics And Intelligent Systems*, II, 55–56. Режим доступу: https://goo.su/smSyTB8

Використані джерела

2. *Лабораторія Української.* (n.d.). Режим доступу: https://mova.institute/
3. *Програма підвищення кваліфікації науково-педагогічних працівників* (2024). Єнський університет. Режим доступу: https://cutt.ly/ze0YrDwt (дата доступу 24.12.2024).
4. Шведова, М., фон Вальденфельс, Р., Яригін, С., Рисін, А., Старко, В., Ніколаєнко, Т., та ін. (2017–2024). *Генеральний регіонально анотований корпус української мови* (ГРАК). Київ, Львів, Єна. Режим доступу: https://uacorpus.org/Kyiv/ (дата доступу 24.12.2024).

5. Chaplynskyi, D. (2023). *Introducing UberText* 2.0*: A Corpus of Modern Ukrainian at Scale. In Proceedings of the Second Ukrainian Natural Language Processing Workshop* (UNLP) (pp. 1–10). Dubrovnik, Croatia: Association for Computational Linguistics.
6. Kieraś, W., Kobyliński, Ł., Komosińska, D., Nitoń, B., Rudolf, M., Shvedova, M., & Zwierzchowska, A. (2023). *PAWUK: Polish Automatic Web Corpus of Ukrainian Language*. Instytut Podstaw Informatyki PAN, Warszawa. Режим доступу: https://pawuk.ipipan.waw.pl (дата доступу 24.12.2024).
7. Malysheva, M. (2024). Telegram Public Channel Corpus. *GitHub*. Режим доступу: https://github.com/mariiamalysheva/Telegram-Public-Channel-Corpus (дата доступу 24.12.2024).
8. *pandas Development Team.* (2025). pandas. Режим доступу: https://pandas.pydata.org/ (дата доступу 24.12.2024).
9. *Python Software Foundation.* (2025). Python. Режим доступу: https://www.python.org/ (дата доступу 24.12.2024).
10. *Sketch Engine.* (n.d.). Sketch Engine. Режим доступу: https://www.sketchengine.eu/ (дата доступу 24.12.2024).
11. *spaCy.* (2025). spaCy. Режим доступу: https://spacy.io/ (дата доступу 24.12.2024).
12. Sternenko, S. (2025). *STERNENKO*. Режим доступу: https://t.me/ssternenko (дата доступу 24.12.2024).

References

1. Bobrovnyk, K. (2019). Automated Building and Analysis of Ukrainian Twitter Corpus for Toxic Text Detection. Proceedings of the 3d International Conference Computational Linguistics And Intelligent Systems, II, 55–56. Retrieved from https://goo.su/smSyTB8 (access date 24.12.2024).
2. Laboratoriia Ukrainskoi. (n.d.). [Laboratory of Ukrainian]. Retrieved from https://mova.institute/ (access date 24.12.2024). [In Ukrainian]
3. Prohrama pidvyshchennia kvalifikatsii naukovo-pedahohichnykh pratsivnykiv (2024). [Professional development programme for academic staff]. University of Jena. Retrieved from https://cutt.ly/ze0YrDwt (access date 24.12.2024). [In Ukrainian]
4. Shvedova, M., fon Valdenfels, R., Yaryhin, S., Rysin, A., Starko, V., Nikolaienko, T., ta in. (2017–2024). General Regionally Annotated Corpus of the Ukrainian Language (GRAC) [Heneralnyi rehionalno anotovanyi korpus ukrainskoi movy (HRAK)]. Kyiv, Lviv, Jena. Retrieved from https://uacorpus.org/Kyiv/ (access date 24.12.2024). [In Ukrainian]
5. Chaplynskyi, D. (2023). Introducing UberText 2.0: A Corpus of Modern Ukrainian at Scale. *In Proceedings of the Second Ukrainian Natural Language Processing Workshop* (UNLP) (pp. 1–10). Dubrovnik, Croatia: Association for Computational Linguistics.
6. Kieraś, W., Kobyliński, Ł., Komosińska, D., Nitoń, B., Rudolf, M., Shvedova, M., & Zwierzchowska, A. (2023). *PAWUK: Polish Automatic Web Corpus of Ukrainian Language*. Instytut Podstaw Informatyki PAN, Warszawa. Retrieved from https://pawuk.ipipan.waw.pl (access date 24.12.2024). [In Ukrainian]
7. Malysheva, M. (2024). Telegram Public Channel Corpus. *GitHub.* Retrieved from https://github.com/mariiamalysheva/Telegram-Public-Channel-Corpus (access date 24.12.2024). [In Ukrainian]
8. *pandas Development Team.* (2025). pandas. Retrieved from https://pandas.pydata.org/ (access date 24.12.2024). [In Ukrainian]
9. *Python Software Foundation*. (2025). Python. Retrieved from https://www.python.org/ (access date 24.12.2024).
10. *Sketch Engine.* (n.d.). Sketch Engine. Retrieved from https://www.sketchengine.eu/ (access date 24.12.2024).
11. *spaCy.* (2025). spaCy. Retrieved from https://spacy.io/ (access date 24.12.2024).
12. Sternenko, S. (2025). *STERNENKO*. Retrieved from https://t.me/ssternenko (access date 24.12.2024). [In Ukrainian]

ISSN 2307-8332. Вісник ОНУ. Сер.: Філологія. 2024. Т. 29, вип. 2(30)

Mariia MALYSHEVA

CORPUS OF TEXTS FROM PUBLIC CHANNELS OF THE TELEGRAM MESSENGER: CREATION AND PRACTICAL APPLICATION

Abstract. *The article presents the process of creating and practical applications of a text corpus based on public channels of the Telegram messenger.*

The aim of this study is to develop a corpus of texts from public Ukrainian-language Telegram channels. The research objectives are as follows: to develop a methodology for building a text corpus from public Telegram channels, to create and test scripts for the automated collection, cleaning, and analysis of textual data, to upload the processed texts into a corpus manager, and to explore the potential for further applications and improvements of the corpus. To achieve these objectives, automated methods for text collection and cleaning were employed using Python scripts and text processing libraries (e.g., json, pandas, SpaCy). The Sketch Engine platform was utilized for the creation and management of the corpus. Data analysis involved the use of CQL (Corpus Query Language) search queries, regular expressions, and basic statistical analysis. The source data consisted of texts from a selected public Ukrainian-language Telegram channel (Serhiy Sternenko's channel).

The corpus development process was conducted in three stages: (1) selecting a public Telegram channel as the source of textual data and downloading its publication history; (2) converting the downloaded data into a format compatible with specialized corpus creation and management software; and (3) uploading the pre-processed texts into the chosen corpus manager. Two Python scripts were developed to prepare the data. The first script extracted the text of posts and saved them to a separate file, while the second script performed text cleaning and statistical analysis. The cleaning process involved the removal of unnecessary symbols, emojis, and links.

To illustrate the potential applications of the corpus in online discourse research, several queries were performed: searching for verbs exceeding 15 characters in length, identifying hashtags, extracting proper names, and analyzing attributive phrases. The queries for attributive phrases focused on two patterns: adjective-noun combinations denoting proper names or entities, and phrases matching the structure 'so-called + noun' or 'so-called + adjective + noun.' These queries were formulated using CQL and regular expressions.

The findings indicate that the created corpus is a valuable resource for studying online discourse. Future research directions include expanding the corpus by incorporating texts from additional Telegram channels, refining the text preparation and filtering processes, and evaluating the applicability of alternative software solutions for corpus creation and management.

This research contributes to the field of computational linguistics and offers a novel resource for the analysis of Ukrainian-language online discourse.

Keywords: *corpus linguistics, text corpus, Telegram, Ukrainian language, network discourse, Sketch Engine.*