УДК 811.111'01/01
DOI https://doi.org/10.24919/2663-6042.22.2024.2

ПОТЕНЦІАЛ КОРПУСНОГО ІНСТРУМЕНТАРІЮ ДЛЯ ВИВЧЕННЯ БАГАТОЯРУСНИХ СТРУКТУР МУЛЬТИЛІНГВАЛЬНОГО КОРПУСУ

Андрущенко О. Ю.
Київський національний лінгвістичний університет, Університет Аугсбург

У статті запропоновано новий інструментарій для аналізу мультилінгвального корпусу текстів, розробку якого здійснено в контексті модуля Jean Monnet (Erasmus+) «Мультилінгвальний корпус та його ресурси для дослідження Європеїстики» у Київському національному лінгвістичному університеті. Здійснено огляд синхронних (моно- та мультилінгвальних) й діахронних корпусів текстів, які з'явилися і над якими працюють останніми десятиліттями. Запропонований підхід з використанням програмного забезпечення #LancsBox, випрацьованого в Ланкастерському університеті, та ANNIS, який допомагає реалізовувати багатошарову розмітку текстів, уможливлює аналіз структур текстів з урахуванням фонетичного, лексичного, граматичного та інформаційно-структурного рівнів. Програми дають змогу працювати з власними даними дослідника, а також з наявними корпусами, візуалізуючи мовні факти та автоматично анотуючи тексти, що робить їх інтуїтивно доступними. Пошук та візуалізація даних за допомогою #LancsBox сприяє виявленню колокацій і колігацій досліджуваних одиниць у паралельних корпусах текстів, що забезпечено такими інструментами програми, як KWIC, Words, GraphColl та Ngrams. У результаті дослідження принципу роботи ANNIS з'ясовано, що програма містить розмітку для анотації таких мовних рівнів, як фонологія, морфологія, синтаксис, семантика й інформаційна структура (ІС). Саме спроможність аналізу останньої в мультилінгвальних корпусах текстів викликає особливу зацікавленість в науковому плані, оскільки сприяє дослідженню змін у засобах репрезентації елементів ІС у різних мовах. Зроблено висновок, що лінгвістичний корпус текстів є потужним інструментом, який відкриває нові перспективи для мовознавчої науки та сприяє об'єктивному аналізу мови. Використання обох окреслених програм при аналізі паралельних корпусів хоча і становить виклики (особливо пов'язані з неоднозначністю анотації ІС), проте має й певну перевагу – уможливлює простежити, як взаємодіють усі мовні рівні у процесі мовлення.

Ключові слова: *мультилінгвальний корпус, #LancsBox, KWIC, Words, GraphColl, Ngrams, ANNIS.*

Andrushenko O. Ju. The potential of corpus toolkit for investigating multilevel structures of the multilingual corpus. *The article proposes a new corpus toolkit for analysing a Multilingual Texts Corpus elaborated within the framework of the Jean Monnet (Erasmus+) module "Multilingual corpus and its resources for the European studies" at the Kyiv National Linguistic University. The article reviews synchronic (mono- and multilingual) and diachronic corpora of texts developed in recent decades. The approach described suggests implementing #LancsBox software, developed at Lancaster University, as well as ANNIS, which allow multilayer text marking, enable the analysis of text structures taking into account phonetic, lexical, grammatical and information-structural levels. The programs allow working with the researcher's own data, as well as with the existing corpora, visualizing linguistic facts and automatically annotating texts and making them intuitively accessible. The data search and visualization by means of #LancsBox helps identify collocations and collations of the units under analysis in parallel text corpora with the assistance of such program tools kits as KWIC, Words, GraphColl and Ngrams. After studying the working principles of ANNIS, it has been found that the program contains the markup for annotating such language levels as phonology, morphology, syntax, semantics and information structure (IS). It is the ability to analyse the latter in multilingual corpora of texts that arises special interest in the scientific plane, as it contributes to the study of changes in the means of representation of IS elements in different languages. To summarize, text corpora are a powerful tool that opens new perspectives for linguistic research facilitating the objective language analysis. The usage of both programs mentioned above in the analysis of parallel corpora, although presents challenges specifically related to the ambiguity of IS annotation, but nevertheless allows tracing the process of language level interaction in the speech process.*

Key words: *multilingual corpus, #LancsBox, KWIC, Words, GraphColl, Ngrams, ANNIS.*

Постановка проблеми та обґрунтування актуальності її розгляду. Корпусні дослідження стали незамінним інструментом у вивченні мови, оскільки вони забезпечують можливість автоматизованого пошуку лінгвістичних даних. Це дає змогу отримати різнобічний та вичерпний аналіз мовних явищ, використовуючи великі обсяги матеріалів, що містяться у структурованих та анотованих колекціях текстів природних мов. Корпусна система даних істотно спрощує пошук матеріалу, проте вимагає глибокого знання різних підходів і методик лінгвістичних досліджень [14]. Хоча корпус забезпечує статистичну верифікацію матеріалу для підтвердження або спростування гіпотез, він також є безцінним джерелом якісної інформації, що містить приклади природної комунікації, оскільки корпуси текстів дають змогу простежувати, як реальна мова використовується людьми в повсякденному спілкуванні [19; 25]. Доцільність створення та використання корпусів можна окреслити так: великий за обсягом

та збалансований матеріал корпусу забезпечує типовість даних та гарантує повноту подання усього спектру явищ у мові; факти мовлення, що містяться у корпусі, представлені у своїй природній та контекстуальній формі, а отже, це уможливлює здійснити всебічний та об'єктивний аналіз; багаторазовість використання створеного масиву тексту з різною метою; швидкість отриманих даних [6].

Методика роботи з великим масивом даних на основі історичних корпусів знайшла відбиток у дослідженнях інформаційної структури речення в англійській мові на основі аналізу фокусувальних адвербів [2; 5; 6].

Аналіз можливості використання різного програмного забезпечення для дослідження багатоярусних структур речення (фонетика, лексика, граматика, інформаційна структура) у текстах мультилінгвального корпусу є сьогодні на часі, що і зумовлює *актуальність* пропонованого дослідження.

Аналіз останніх досліджень і публікацій. Науковий внесок у сучасну лінгвістику постійно зростає завдяки новим колективним монографіям [див.: 24; 26; 31], навчально-науковим посібникам та статтям [див.: 4; 28; 29], які детально окреслюють теоретико-методологічну базу дослідження та технологічні можливості програмного забезпечення для роботи з корпусами текстів. Результати різноаспектних корпусних досліджень сприяють перегляду багатьох лінгвістичних постулатів та демонструють якісно нові характеристики конкретних одиниць як однієї мови [24], так і багатьох мов. Зокрема, важливим є проєкт зі створення мультимовного корпусу в Україні «Романо-германо-слов'янський корпус наукових аутентичних текстів з лінгвоантропогенезу: розробка технологій нового покоління» (керівник – проф. Корольова А. В.) [1].

Перший комп'ютеризований корпус текстів *The Brown Corpus*, створений у 1960-х рр., налічував 500 текстових уривків загальним обсягом 1 млн. слів, виокремлених із американських газет, журналів та книг. Принципи добору текстів та можливі завдання на той час викликали жваві дискусії, оскільки корпус супроводжувався значною кількістю матеріалів його первинної статистичної обробки, яка спиралася на професійну інтуїцію укладачів корпусу [22]. А отже, для досягнення максимальної об'єктивності з'явилась необхідність в побудові формалізованих, прозорих для перевірки і контролю процедур аналізу, до яких належать корпуси другого покоління, що почали розроблятися у 80-х роках XX ст. Зокрема, монолінгвальні *The Intelligent Web-based Corpus, British National Corpus, American National Corpus, CoRola, TS Corpus* та ін. мають обсяг близько 14 мрд. слововживань [15], представляючи різножанрові тексти і демонструючи лематизацію, колігацію, семантичне тегування тощо.

Корпусні дослідження не обмежуються вивченням мов у синхронії. Уже в 70-х роках XX ст. започатковано розробку діахронійних корпусів [30]. У 1990-х роках укладання Гельсинського [13] та ARCHER [9] корпусів уможливило отримати безпрецедентний доступ до вивчення різних етапів діахронійного розвитку англійської мови [8]. Такі корпуси текстів, як *EBOO, COHA, Hansard Corpus* (обсягом від 755 млн. до 1.6 млрд. слів), дають змогу одержати дані про частоту вживання окремих лексичних одиниць, фраз та граматичних конструкцій у динаміці протягом кожного десятиліття, побудувати конкорданс для слів, зіставивши їх у різні періоди розвитку мови [3, 16]. Також варто згадати корпуси, що містять синтаксичну анотацію та використовуються для аналізу історичних стадій розвитку англійської мови: *The York Toronto-Helsinki Parsed Corpus of Old English Prose*, *The York-Helsinki Parsed Corpus of Old English Poetry*, *The Penn-Helsinki Parsed Corpus of Middle English*, *The Parsed Corpus of Early English Correspondence* та *The Penn-Helsinki Parsed Corpus of Early Modern English*.

Завдяки сучасному програмному забезпеченню з 1990-х рр. з'явилась можливість розвивати «мультилінгвальні корпуси текстів» [18], термін, який найчастіше використовують для позначення паралельних корпусів, тобто письмові тексти і їх переклад однією або кількома мовами. Найвідомішими корпусами є *English-Norwegin Parallel Corpus (ENPC), Englis-Swedish Parallel Corpus (ESPC), Europarl Corpus, OPUS* та ін.

Формулювання мети і завдань дослідженняю Метою статті є описати сучасні програми, використовувані для роботи з електронними текстами в межах проєкту «Мультилінгвальний корпус та його ресурси для дослідження Європеїстики» (КНЛУ) (програма Erasmus+). Серед **завдань** дослідження – 1) схарактеризувати основні інструменти програмних продуктів *#LancsBox* та *ANNIS*; 2) проаналізувати переваги та виклики при використанні кожної програми для роботи з великими масивами текстів.

Виклад основного матеріалу дослідження. Для автоматичного аналізу мультилінгвальних корпусів та інтерпретації отриманих даних використовують програмне забезпечення *#LancsBox*, уперше розроблене в Ланкастерському університеті 2015 року. Програма може працювати як з власними даними дослідника, так і з наявними корпусами, візуалізуючи мовні факти та автоматично анотуючи тексти, що робить її зручною для користувачів [10]. Програмне забезпечення включає: 1) обробку даних користувача або вже наявних корпусів, які можна імпортувати у форматах txt, .xml, .doc, .docx, .pdf, .odt, .xls, .xlsx та ін.; 2) візуалізацію мовних фактів; 3) аналіз даних будь-якою мовою; 4) автоматичну анотацію даних та тегування частин мови (POS); 5) сумісність з основними операційними системами (*Windows, Mac, Linux*) [11, 38–40].

Основною перевагою *#LancsBox* є можливість автоматизованого вивчення асоціацій слів та ідентифікації співрозмовників за трьома традиційними критеріями – такими як: відстань (визначення діапазону навколо ключового слова – «колокаційне вікно»), частота (важливий показник типовості асоціації слів) і ексклюзивність [12]. Інші критерії, відповідно до S. Gries, передбачають спрямованість (сила зв'язку між словами), дисперсію (розподіл вузлів та колокацій у корпусі) і дистрибуцію токенів серед колокацій (оцінка сили колокаційного зв'язку та рівня конкуренції слотів навколо ключового слова

з іншими типами колокацій) [21, 139]. Розробники *#LancsBox* також враховують взаємозв'язки між окремими колокаціями.

Для спрощення пошуку даних і візуалізації отриманих результатів при автоматичному статистичному аналізі в програмі використовуються такі інструменти з пакета *#LancsBox*: *KWIC* (забезпечує контекстуальною інформацією про досліджуваний токен, створюючи список усіх прикладів використання пошукового терміна в корпусі; подвійне клацання на вузлі відкриває спливаюче вікно з розширеним текстом для більш детального вивчення слова в контексті), *Words* (уможливлює шукати слова в одному класі), *GraphColl* (надає інформацію про колокаційні патерни, візуалізуючи праві та ліві колокації в мережі відповідно до трьох параметрів: сили, частоти та позиції). Інструмент *Words* також дає змогу детально аналізувати частоту типів, лем і частиномовних категорій, а також порівнювати корпуси за допомогою пошуку ключових слів. Інструмент *Ngrams* допомагає глибше аналізувати частоту типів нграмів, лем і частиномовних категорій, а також порівнювати корпуси за допомогою техніки ключових нграмів [10, 12]. На рис. 1 показано фрагмент роботи програми.

Останніми десятиріччями спостерігається тенденція до розробки багатоярусних корпусів, які успішно використовуються для аналізу паралельних текстів: *ANNIS* [27], *PROIEL* [20], позаяк одноярусні обмежуються лише лематизацією або тегуванням частин мови. Ці багатошарові архітектури пропонують більш глибоку й різнопланову анотацію мовних явищ із різних перспектив їхніх функцій у реченні (синтаксичної, семантичної, інформаційно-структурної) [17; 23]. Так, корпус *ANNIS* уможливлює анотацію таких мовних рівнів, як:

1. Фонологія та інтонація, що містить яруси для фонетичної транскрипції та загальної орфографії. Інші яруси, представлені у корпусі, включають інформацію, що стосується фонетики, фонології та просодії висловлення. Останні метадані є необхідними для аналізу інформаційно-структурних характеристик.

2. Морфологія, що охоплює дані про морфемну сегментацію (morph), переклад морфів (gloss), а також тегування кожного слова як частиномовної категорії (pos).

3. Синтаксис, який, ґрунтуючись на морфологічній інформації, репрезентує структуру конституентів у реченні, включаючи синтаксичні функції та семантичні ролі.

4. Семантика, що передбачає інформацію про семантичні та прагматичні риси: означеність (DefP), злічуваність (C) та істоту (A).

5. Інформаційна структура, що містить такі рівні, як інформаційний статус, топік, фокус.

Загальний вигляд даних, вміщених у корпусі, представлено на рис. 2.

Можливість анотації інформаційної структури речення в корпусах текстів викликає особливу зацікавленість в мультилінгвальному плані, оскільки сприяє дослідженню змін щодо засобів репрезентації її елементів у різних мовах. Принципи анотації інформаційної структури речення в окресленому корпусі застосовані для розмітки інформаційно-структурних компонентів речення в мультилінгвальному корпусі, що дає можливість порівняти не лише лексичний та граматичний інвентар, а й проаналізувати відмінності мов в текстовому плані. Проте запропонована схема має окремі недоліки, що насамперед стосується писемних текстів: складність анотації фонетичного рівня, а отже, неоднозначність анотування топіка та фокусу речення, що підтверджено розрахунками *F-scores* та *Kappa* при аналізі даних двох анотацій двох укладачів корпусу. Тому для досягнення об'єктивного результату аналізу даних корпусів, використовуваних у дослідженні, необхідна розробка більш формальної і однозначної методики, яка б сприяла визначенню компонентів інформаційної структури об'єктивно. Вбачаємо за доцільне використання програмного забезпечення *#LancsBox* та *ANNIS* з метою досягнення об'єктивного результату.

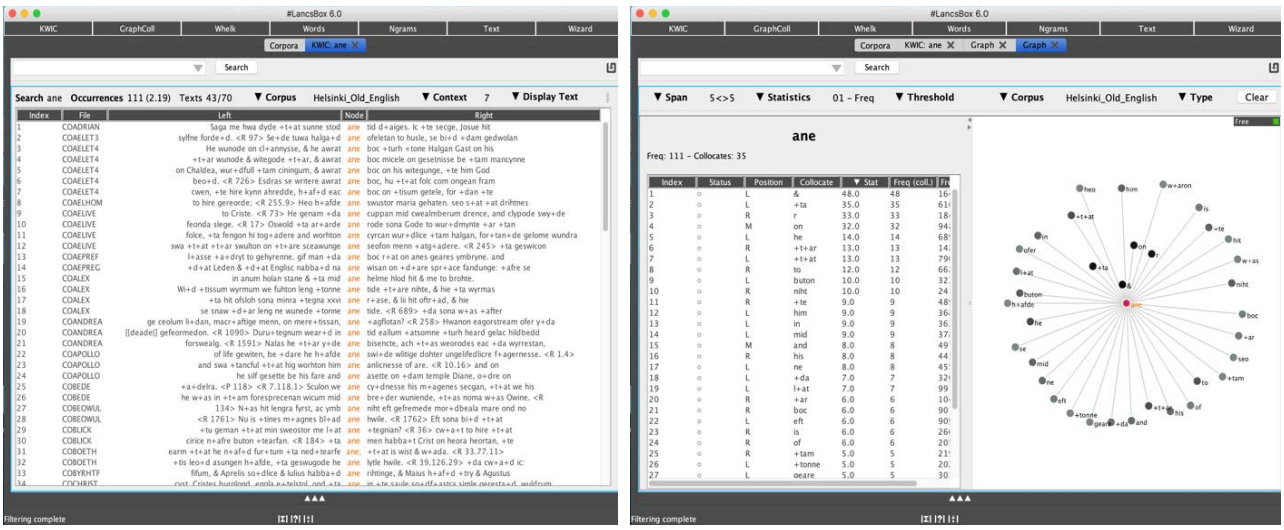

Рис. 1. Візуалізація пошуку за допомогою *KWIC* та *GraphColl* у програмному забезпеченні *#LancsBox*

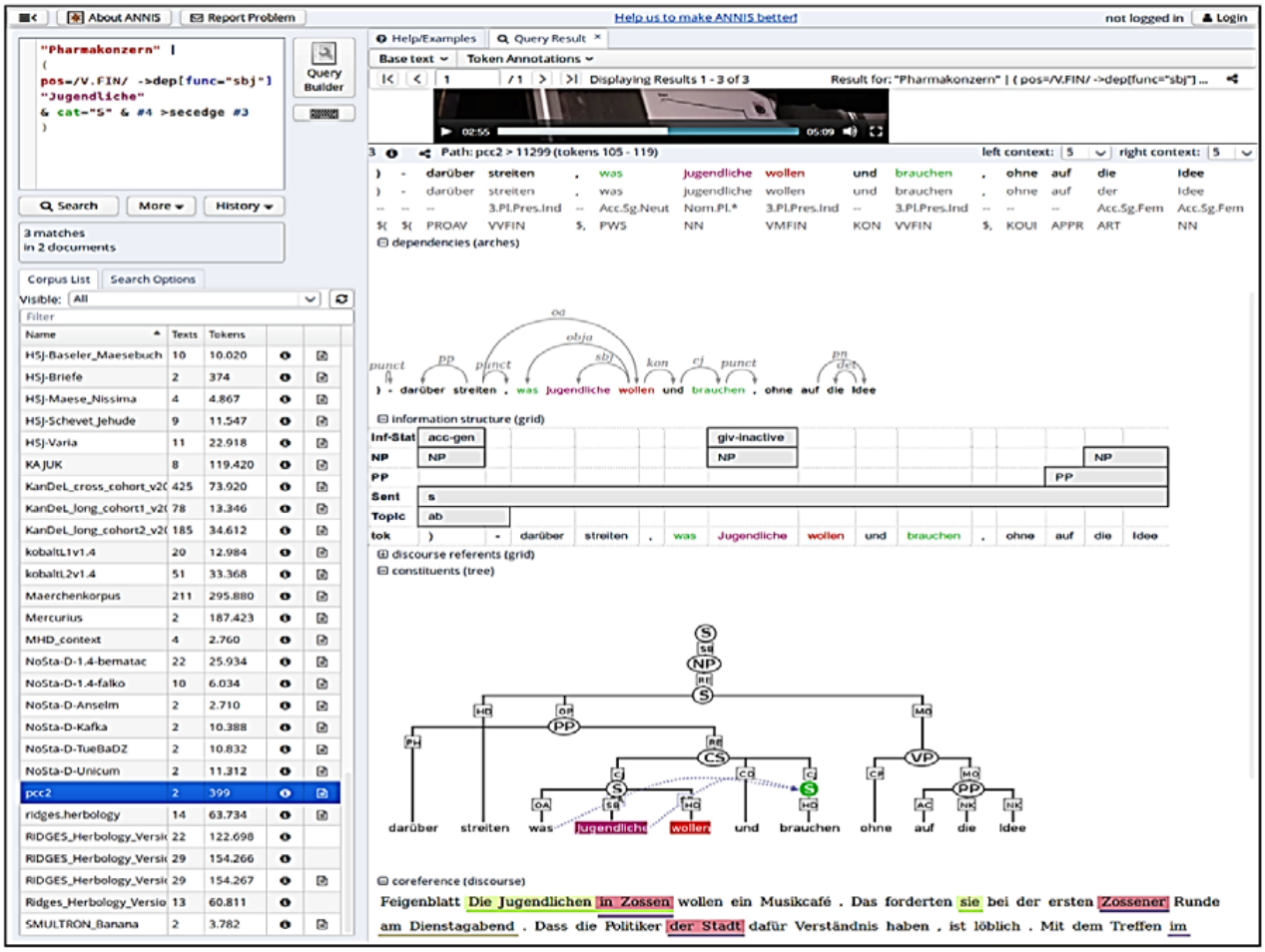

Рис. 2. Графічна репрезентація анотацій різних рівнів у корпусі текстів *ANNIS*

Висновки та перспективи дослідження. Лінгвістичний корпус текстів є потужним інструментом для проведення мовознавчих досліджень, який відкриває нові перспективи для мовознавчої науки, що сприяє об'єктивному аналізу мови. Використання програми *#LancsBox* найбільш доцільне для візуалізації мовних фактів мультилінгвальних корпусів та порівняння даних, зокрема оприявнення лексичного оточення одиниці з оптимальним набором кількісних даних. Вбудоване програмне забезпечення *KWIC*, *GraphColl*, *Wizard* уможливлює повною мірою не лише отримати нові кількісні дані, а й здійснити аналіз якісних характеристик лексем у реченні. Сучасні корпуси текстів на прикладі *ANNIS* сприяють багатомовному аналізу всіх мовних рівнів, серед іншого й інформаційно-структурного. Розробка багатоярусних корпусів, здійснювана останніми десятиліттями, дає змогу простежити, як взаємодіють усі мовні рівні у процесі мовлення. Проте існує низка проблем, пов'язаних із неоднозначністю анотації компонентів інформаційної структури речення в письмових текстах, оскільки в них важко визначити інтонацію, яка є вирішальною при розмежуванні топіка і фокусу речення в окремих європейських мовах. Тому існує необхідність випрацювання формальної методики аналізу ІС речення в мультилінгвальних корпусах текстів.

Подяка. *Співфінансується Європейським Союзом. Проте висловлені думки належать лише авторці і не обов'язково збігаються з поглядами Європейського Союзу чи Європейського виконавчого агентства з питань освіти та культури. Ні Європейський Союз, ні орган, що надає гранти, не відповідають за наведені в статті погляди.*

Acknowledgements. *Co-funded by the European Union. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.*

ЛІТЕРАТУРА

1. Корольова А. В. Принципи розробки бази даних «Лінгвоантропогенез» : концепція, структура і зміст. *Науковий часопис НПУ імені М. П. Драгоманова. Серія 9. Сучасні тенденції розвитку мов*. 2014. № 15. С. 119–125.

2. Андрущенко О. Ю. Інформаційно-структурні перетворення адитивного адверба *even* (на матеріалі пам'яток і текстів корпусів англійської мови XII–XVII ст.). *Вісник Київського національного лінгвістичного університету. Серія : Філологія*. 2021. № 24 (1). С 16–32.

3. Alexander M., Davies M., Dallachy F. The Hansard Corpus 1803–2005. 2015. URL: https://www.english-corpora.org/hansard/.

4. Andrushenko O. Integrated methods for studying focusing adverbs in modern and historical English corpora. *Innovative pathway for the development of modern philological sciences in Ukraine and EU countries.* Riga : Baltija Publishing, 2022. P. 26–54.

5. Andrushenko O. Iu. Lancsbox software options for the prospective investigation of the multilingual corpus for European studies. *Вісник КНЛУ. Серія : Філологія*. 2023. № 26 (1). С. 16–32.

6. Andrushenko O. The landscape of Middle English focusing adverb even. *Litera : Journal of Language, Literature and Culture Studies*. 2022. No. 32 (2). P. 1–24.

7. Baker P. Corpus methods in linguistics. *Research methods in linguistics* / L. Litosseliti (Ed.). London, New York : Continuum, 2010. P. 93–113.

8. Bennett P., Durrell M., Scheible S., Whitt R. New methods in historical corpora. Tübingen : Narr, 2013. 282 p.

9. Biber D., Reppen R. Introduction. *The Cambridge handbook of English corpus linguistics* / D. Biber, R. Reppen (eds.). Cambridge : CUP, 2015. P. 1–8.

10. Brezina V., Weill-Tessier P., & McEnery T. #LancsBox 5.x and 6.x [software]. 2018. URL: http://corpora.lancs.ac.uk/lancsbox.

11. Brezina V. Statistics in corpus linguistics : A practical guide. Cambridge : Cambridge University Press, 2018. 316 p.

12. Brezina V., Timperley M., & McEnery T. #LancsBox 4.x [software]. 2018. URL: http://corpora.lancs.ac.uk/lancsbox.

13. Brinton L., Bergs A. The history of English. Early Modern English. Berlin : Walter de Gruyter, 2017. 344 p.

14. Conrad S. Register in corpus linguistics : the role and legacy of Douglas Biber. *Corpus Linguistics and Linguistic Theory*. 2023. No. 19 (1). P. 7–21.

15. Davies M. The 14 billion word iWeb corpus [Volume 22. Provo] : [Brigham Young University]. 2018. URL: https://www.english-corpora.org/ iweb/. Retrieved October 8, 2022.

16. Davies M. The best of both worlds : Multi-billion word "dynamic" corpora. Proceeding of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019 / P. Banski et al. (eds.). Manhein : Leibniz Institute für Deutsche Sprache, 2019. P. 23–28.

17. Domínguez M., Farrús M., Wanner L. The Information structure–prosody interface in text-to-speech technologies. An empirical perspective. *Corpus Linguistics and Linguistic Theory*. 2022. No. 18 (2). P. 419–445.

18. Doval I., Sanchez Nieto M. T. Parallel corpora for contrastive and translations studies : New resources and applications. Amsterdam, Philadelphia : John Benjamins Publ., 2019. 301 p.

19. Egbert J., Baker P. Introduction. *Triangulating methodological approaches in corpus-linguistic research* / P. Baker, J. Egbert (eds.). New York, London : Routledge, 2016. P. 1–19.

20. Geyken A., Gloning T. A living text archive of 15th–19th-century German. Corpus strategies, technology, organization. *Historical corpora. Challenges and perspectives* / J. Gippert, R. Gehrke (eds.). Tübingen : Narr, 2015. P. 165–179.

21. Gries S. 50-something years of work on collocations : What is or should be next... *International journal of corpus linguistics*. 2013. No. 18 (1). P. 137–166.

22. Lange C., Leuckert S. Corpus linguistics for world Englishes : A guide for research. New-York : Routledge, 2020. 236 p.

23. Lavidas N., Truslew H. D. T. Postclassical Greek and treebanks for a diachronic analysis. Postclassical Greek : contemporary approaches to philology and linguistics / D. Rafiyenko, I. Seržant (eds.). Berlin : Walter de Gruyter, 2020. P. 163–202.

24. López-Couso M. J., Méndez-Naya A., Núñez-Pertejo B. P., Palacios-Martínez I. M. Corpus linguistics on the move. Exploring and understanding English through corpora. Leiden, Boston : Brill, Rodopi, 2016. 368 p.

25. McEnery T. Corpus linguistics : An introduction. Edinburgh : Edinburgh University Press, 2001. 256 p.

26. McEnery T., Hardie A. Corpus linguistics. Cambridge : CUP, 2011. 311 p.

27. Petrova S. Information structure and word order variation in the Old High German Tatian. Information structure and language change / R. Hinterhölzl, S. Petrova (eds.). Berlin : Mouton de Gruyter, 2009. P. 251–279.

28. Rühlemann Ch. Corpus linguistics for pragmatics : A guide for research. New-York : Routledge, Taylor & Francis Group, 2019. 210 p.

29. Stefanowitsch A. Corpus linguistics : a guide to the methodology. Berlin : Language Science, 2020. 510 p.

30. Whitt R. Using diachronic corpora to understand the connection between genre and language change. *Diachronic corpora, genre, and language change* / R. Whitt (ed.). Amsterdam, Philadelphia : John Benjamins Publ., 2018. P. 1–18.

31. Yang D., Li W. (Eds). Corpus-based approaches to grammar, media and health discourses. Systemic Functional and other perspectives. Singapore : Springer, 2020. 396 p.

REFERENCES

1. Korolyova, A. V. (2017). Pryntsypy rozrobky bazy dannyh "Lingvoantropogenes": kontseptsiya, struktura i zmist [Principles of the development of the database "Linguoanthropogenesis": concept, structure and content]. *Naukovyy chasopys NPU imeni M. P. Drahomanova. Seria 9. Suchasni tendentsii rozvytky mov*, 15, 119–125 [in Ukrainian].

2. Andrushenko, O. Iu. (2021). Informatsiyno-strukturni harakterystyky adytyvnogo adverba even (na materiali pamyatok I tekstiv korpusiv angliyskoi movy XII–XVII st.) [Information-structural transformations of additive adverb even (based on English language corpora of XII–XVII cen.)]. *Visnyk Kyivskogo natsionalnogy lingvistychnogo universytetu. Seria: Philologiya***,** 24 (1), 16–32 [in Ukrainian].

3. Alexander, M., Davies, M & Dallachy, F. (2015). *The Hansard Corpus 1803–2005*. URL: https://www.english-corpora.org/hansard/.

4. Andrushenko, O. (2022). Integrated methods for studying focusing adverbs in modern and historical English corpora. *Innovative pathway for the development of modern philological sciences in Ukraine and EU countries*. Riga: Baltija Publishing, 26**–**54.

5. Andrushenko, O. (2023). Lancsbox software options for the prospective investigation of the multilingual corpus for European studies. *Вісник КНЛУ. Серія: Філологія*, 26 (1), 16**–**32.

6. Andrushenko, O. (2022). The landscape of Middle English focusing adverb even. *Litera: Journal of Language, Literature and Culture Studies*, 32 (2), 1**–**24.

7. Baker, P. (2010). Corpus methods in linguistics. *Research methods in linguistics / L. Litosseliti* (ed.). London, New York: Continuum, 93–113.

8. Bennett, P., Durrell, M., Scheible, S., Whitt, R. (2013). *New methods in historical corpora*. Narr.

9. Biber, D., Reppen, R. (2015). Introduction. *The Cambridge handbook of English corpus linguistics* / *D. Biber & R. Reppen* (eds.). Cambridge: CUP, 1–8.

10. Brezina, V., Weill-Tessier, P., & McEnery, T. (2020). #LancsBox 5.x and 6.x [software]. URL: http://corpora.lancs.ac.uk/lancsbox.

11. Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge University Press.

12. Brezina, V., Timperley, M., & McEnery, T. (2018). #LancsBox 4.x [software]. URL: http://corpora.lancs.ac.uk/lancsbox.

13. Brinton, L., Bergs, A. (2017). *The history of English. Early Modern English.* Walter de Gruyter.

14. Conrad, S. (2023). Register in corpus linguistics: the role and legacy of Douglas Biber. *Corpus Linguistics and Linguistic Theory*, 19 (1), 7–21.

15. Davies, M. (2018). The 14 billion word iWeb corpus. Retrieved October 8, 2022, [Vol. 22. Provo]: [Brigham Young University]. URL: https://www.english-corpora.org/ iweb/.

16. Davies, M. (2019). The best of both worlds: Multi-billion word "dynamic" corpora. *Proceeding of the Workshop on Challenges in the Management of Large Corpora (CMLC-7)* / P. Banski et al. (eds.). Manhein: Leibniz Institute für Deutsche Sprache, 23–28.

17. Domínguez, M., Farrús, M. & Wanner, L. (2022). The Information Structure–prosody interface in text-to-speech technologies. An empirical perspective. *Corpus Linguistics and Linguistic Theory*, 18 (2), 419–445.

18. Doval, I. & Sanchez Nieto, M. T. (2019). Parallel corpora for contrastive and translations studies: New resources and applications. Amsterdam, Philadelphia: John Benjamins Publ.

19. Egbert, J., Baker, P. (2016). Introduction. *Triangulating methodological approaches in corpus-linguistic research* / P. Baker & J. Egbert (eds.). New York, London: Routledge, 1–19.

20. Geyken, A. & Gloning, T. (2015). A living text archive of 15th-19th-century German: Corpus strategies, technology, organization. *Historical corpora. Challenges and perspectives* / J. Gippert & R. Gehrke (eds.). Tübingen: Narr, 165–179.

21. Gries, S. (2013). 50-something years of work on collocations: What is or should be next... *International journal of corpus linguistics*, 18 (1), 137–166.

22. Lange, C., Leuckert, S. (2020). Corpus linguistics for world Englishes: A guide for research. New-York: Routledge.

23. Lavidas, N. & Truslew, H. D. T. (2020). Postclassical Greek and treebanks for a diachronic analysis. *Postclassical Greek: contemporary approaches to philology and linguistics* / D. Rafiyenko & I. Seržant (eds.). Berlin: Walter de Gruyter, 163–202.

24. López-Couso, M. J., Méndez-Naya, A., Núñez-Pertejo, B. P. & Palacios-Martínez, I. M. (2016). Corpus linguistics on the move. Exploring and understanding English through corpora. Leiden, Boston: Brill, Rodopi.

25. McEnery, T. (2001). Corpus linguistics: An introduction. Edinburgh: Edinburgh University Press.

26. McEnery, T., Hardie, A. (2011). Corpus linguistics. Cambridge: CUP.

27. Petrova, S. (2009). Information structure and word order variation in the Old High German Tatian. *Information structure and language change* / R. Hinterhölzl & S. Petrova (eds.). Berlin: Mouton de Gruyter, 251–279.

28. Rühlemann, Ch. (2019). Corpus linguistics for pragmatics: A guide for research. New-York: Routledge, Taylor & Francis Group.

29. Stefanowitsch, A. (2020). Corpus linguistics: a guide to the methodology. Berlin: Language Science.

30. Whitt, R. (2018). Using diachronic corpora to understand the connection between genre and language change. *Diachronic corpora, genre, and language change* / R. Whitt (ed.). Amsterdam, Philadelphia: John Benjamins Publ., 1–18.

31. Yang, D. & Li, W. (Eds). (2020). Corpus-based approaches to grammar, media and health discourses. Systemic Functional and other perspectives. Singapore: Springer.