УДК 81'374.8-022.218

ЧЕРНИШ Оксана – *кандидат філологічних наук, доцент, декан факультету педагогічних технологій та освіти впродовж життя, Державний університет «Житомирська політехніка», вул. Чуднівська, 103, Житомир, 10005, Україна (Chernyshoxana@gmail.com)*

ORCID: https://orcid.org/0000-0002-2010-200X

ПАНЧЕНКО Наталія *– старший викладач кафедри педагогічних технологій та мовної підготовки, Державний університет «Житомирська політехніка», вул. Чуднівська, 103, Житомир, 10005, Україна (panchenko.nataliia@ztu.edu.ua)*

ORCID: https//orcid.org/0000-0002-8171-3939

DOI: https://doi.org/10.24919/2522-4565.2023.53.11

Бібліографічний опис статті: Черниш, О. Панченко, Н. (2023). Сучасні корпусні технології у створенні електронного словника. *Проблеми гуманітарних наук: збірник наукових праць Дрогобицького державного педагогічного університету імені Івана Франка. Серія «Філологія»*, 53, 94–98, doi: https://doi.org/10.24919/2522-4565.2023.53.11

СУЧАСНІ КОРПУСНІ ТЕХНОЛОГІЇ У СТВОРЕННІ ЕЛЕКТРОННОГО СЛОВНИКА

Анотація. *У статті розглянуто роль корпусної лінгвістики як самостійної галузі, що розробляє та вдосконалює методики збору реальних мовних явищ, писемних та усних текстів, а також способів їх збереження та аналізу. Корпусна лінгвістика має вагоме значення, оскільки сприяє оптимізації епістемічної функції, пов'язаної зі збереженням і передачею знань та відображенням національної самосвідомості. Автори зазначають, що дослідження корпусної лінгвістики спрямовані переважно на вивчення питань теорії та практики створення корпусів, типологію корпусу, структурування та принципи відбору базових одиниць, а також вивчення мови за допомогою корпусних методів. У статті розкрито об'єкт, предмет, особливості та мету корпусної лінгвістики, що полягає у здійсненні об'єктивного лінгвістичного опису мовної системи на основі вивчення людської комунікації; з'ясовано етапи та особливості її розвитку. Установлено, що базовим поняттям корпусної лінгвістики є корпус тексту, під яким розуміють (у широкому сенсі) низку письмових чи усних текстів, використовуваних із метою дослідження мови. Значну увагу приділено створенню мовних корпусів. Наголошено, що корпусні технології можуть бути застосовані для створення електронних словників. Вони особливо корисні в процесах лематизації та стемінгу. Послуговуючись базовою формою слова, вони скорочують обсяг словника, зменшуючи тим самим кількість записів і полегшуючи пошук необхідного слова. Звернено увагу на корпусну організацію лінгвальних даних, яка потребує врахування типології текстових корпусів, оскільки це сприяє виробленню стратегій та принципів створення. У дослідженні встановлено, що корпусні технології ефективні в міжмовному зіставленні лексичних одиниць, покликаному встановити спільні та відмінні ознаки різних мов, що допоможе окреслити особливості їх уживання.*

Ключові слова: *електронний словник, зіставлення слів, корпусна лінгвістика, корпус тексту, корпусні технології, створення корпусів.*

CHERNYSH Oksana – *Candidate of Philological Sciences, Associate Professor, Dean of the Faculty of Pedagogical Technologies and Lifelong Learning, Zhytomyr Polytechnic State University, 103 Chudnivska, str., Zhytomyr, 10005, Ukraine (Chernyshoxana@gmail.com)*

ORCID: https://orcid.org/0000-0002-2010-200X

PANCHENKO Natalia *– Senior Lecturer at the Department of Pedagogical Technologies and Language Studies, Zhytomyr Polytechnic State University, 103 Chudnivska, str., Zhytomyr, 10005, Ukraine (panchenko.nataliia@ztu.edu.ua)*

ORCID: https//orcid.org/0000-0002-8171-3939

DOI: https://doi.org/10.24919/2522-4565.2023.53.11

To cite this article: Chernysh, O., Panchenko, N. (2023). Suchasni korpusni tekhnolohii u stvorenni elektronnoho slovnyka [Modern corpus technologies in an electronic dictionary compilation]. *Problemy humanitarnych nauk: zbirnyk naukovych prats Drohobytskoho derzhavnoho pedahohichnoho universytetu imeni Ivana Franka. Seriia «Filolohiia» – Problems of the humanities: a collection of scientific works of Ivan Franko Drohobytsk State Pedagogical University. Series "Philology",* 53, 94–98, doi: https://doi.org/10.24919/2522-4565.2023.53.11 [in Ukrainian].

MODERN CORPUS TECHNOLOGIES DICTIONARY COMPILATION

Summary. *The article examines the role of corpus linguistics as an independent field that develops and improves methods of collecting natural language phenomena, written and spoken texts, and methods of their preservation and analysis. Corpus linguistics is essential because it contributes to optimizing the epistemic function related to the preservation and transmission of knowledge and the reflection of national self-awareness. The authors note that corpus linguistics research mainly studies issues of the theory and practice of creating corpora, corpus typology, structuring and principles of selection of basic units, and language learning using corpus methods. The article reveals the object, subject, features, and purpose of corpus linguistics, which consists of implementing an objective linguistic description of the language system based on the study of human communication and the stages and features of its development. It was found that the basic concept of corpus linguistics is a corpus of text, which in a broad sense is understood as some written or spoken texts used for language research. Therefore, considerable attention is paid to creating and using language corpora. The authors emphasize that corpus technologies can be used to create electronic dictionaries, in particular in the aspect of lemmatization and stemming, because corpus technologies reduce the volume of the dictionary by using the primary form of the word, which reduces the number of entries and facilitates the search for the required word. It is also essential to pay attention to the corpus organization of linguistic data, which needs to consider the typology of text corpora, as this contributes to the strategy and principles of its creation. The research found that corpus technologies are effective in comparing words, as they contribute to establishing standard and distinctive features of different languages, which allows outlining the peculiarities of their use.*

Key words: *electronic dictionary, word matching, corpus linguistics, text corpus, corpus technologies, corpus creation.*

Постановка проблеми. Наше сьогодення тісно пов'язане з розвитком комп'ютерних технологій, які сприяють з'яві нових галузей лінгвістики. Корпусна лінгвістика як самостійна галузь має вагоме значення, тому що зберігає й передає знання, відображаючи національну самосвідомість. Оскільки перед лінгвістами постає завдання системного аналізу мови, змістовного спілкування мовою, то на сьогодні актуальним є здійснення її точного опису за допомогою корпусних методів, які можуть бути використані у створенні електронних словників.

Аналіз останніх досліджень та публікацій. Вітчизняні та зарубіжні мовознавці (Ф. Бацевич, В. Жуковська, О. Демська-Кульчицька, Т. Бобкова, О. Тоньїні-Бонеллі та інші) у своїх публікаціях значну увагу приділяють дослідженням лінгвістичних корпусів та вивчають словесний матеріял на основі спостережень мовної діяльності, тобто тексту. **Актуальність нашої розвідки** визначається опрацюванням даних корпусу та низки письмових і усних текстів, використовуваних із метою дослідження мови. Їх вивчення сприяє встановленню спільних та відмінних ознак різних мов, що допоможе окреслити особливості їх уживання. **Мета статті** – розглянути корпусні технології та проаналізувати їхню роль у створенні електронного словника. **Завдання –** з'ясувати роль корпусних методів у створенні електронних словників та розглянути типологію текстових корпусів.

Виклад матеріалу. Як самостійна галузь мовознавства корпусна лінгвістика виникла у 60-х роках XX століття. Вона вивчає загальні принципи побудови, обробки та використання даних лінгвістичних корпусів із залученням сучасних комп'ютерних технологій, а також розроблення методики збору реальних мовних явищ, писемних та усних текстів, а також способів їх збереження та аналізу (Жуковська, 2013, с. 9). Корпусна лінгвістика має вагоме значення, оскільки сприяє оптимізації епістемічної функції, пов'язаної з правдивістю інформації, збереженням і передачею знань, а також із відображенням національної самосвідомості.

Розвиток цієї науки безпосередньо пов'язаний з розвитком комп'ютерних технологій; у ньому виділяють такі етапи (Tognini-Bonelli, 2001, с. 16–17):

– середина 60-х – початок 80-х років XX століття – період набуття знань про організацію та підтримку корпусів кількістю до 1 млн. слів. Цей період характеризується відсутністю матеріалів в електронному форматі та нагальною потребою набору текстів вручну;

– 1980–2000 рр. – період а) з'яви сканерів і б) розширення можливостей комп'ютерного набору, що суттєво полегшило доступ до великих за обсягом текстових матеріалів в електронному форматі, а отже, сприяло значному збільшенню розмірів корпусів – до 20 млн. слововживань;

– з початку 2000-го року і до сьогодні – період електронних (віртуальних) текстів, що уможливило створення корпусів необмеженого розміру.

Дослідження корпусної лінгвістики спрямовані в основному на вивчення питань теорії та практики створення корпусів, типології корпусів, їх призначення, обсягу, параметризації предметної галузі, репрезентативності, структурування та принципів відбору базових одиниць, зберігання тощо. Крім того, корпусна лінгвістика включає дослідження самих лінгвістичних корпусів і вивчення мови за допомогою корпусних методів.

Основне завдання корпусної лінгвістики полягає в системному відображенні процесів використання мови, її точний опис та вичерпний аналіз. Відтак мета корпусної лінгвістики полягає у здійсненні об'єктивного лінгвістичного опису мовної системи на основі вивчення людської комунікації.

Корпусна лінгвістика має низку особливостей (Жуковська, 2013, с. 17), а саме:

– використовує дані корпусу у своїх дослідженнях;

– послуговується квантитативними методами для об'єктивізації своїх досліджень;

– розглядає текст як певну фізичну сутність;

– займається укладанням граматики конкретних мов;

– основну увагу приділяє формі та досліджує текст у глобальній перспективі;

– робить висновки на основі спостережень мовленнєвої діяльності, репрезентованої у вигляді текстів;

– використовує ймовірнісні методи й статистику для первинної обробки мовленнєвого матеріалу;

– працює з лінгвальними даними в їхньому природному контекстуальному оточенні;

– надає перевагу індуктивним методам обробки емпіричного мовного матеріалу.

Отже, об'єкт корпусної лінгвістики – текстовий корпус як вихідний її матеріал і як результат дослідження. Предмет корпусної лінгвістики – теоретичні основи й практичні механізми створення та використання мовних корпусів. Відповідно сучасна корпусна лінгвістика вивчає (McEnery, Hardie, 2012, с. 3–21):

– формат представлення текстів у корпусі;

– режим відбору та врахування даних і накопичення їх у корпусі;

– механізми використання анотованих / неанотованих корпусів;

– багатомовні та одномовні корпуси

й здійснює корпуснобазовані та корпуснокеровані дослідження.

Базовим поняттям корпусної лінгвістики є корпус тексту, під яким розуміють (у широкому сенсі) низку письмових чи усних текстів, використовуваних із метою дослідження мови (Лендау, с. 270; Kennedy, p. 67; Meyer, p. 1–13). У вузькому сенсі корпус тексту – це тексти в електронній формі, що відображають особливості певної мови та сприяють її вивченню (Демська-Кульчицька, 2005, с. 20; Francis;

Leech, p. 29–30; McEnery, 2006, p. 215; Meyer, 2002, p. 5). Трактування корпусу різниться залежно від:

– предметної галузі лінгвістичного дослідження;

– відповідних складників корпусу;

– галузевої належності відповідних термінів;

– характерних ознак корпусу (Бобкова, 2014, с. 13).

Суттєво, що корпус трактують і як колекцію зразків письмових та усних текстів у машиночитаній формі, відібраних певним науково обґрунтованим методом для встановлення й окреслення особливостей вживання мови (Dash, 2005, p. 3). Відповідно до рекомендацій Консультативної групи експертів із питань мовних технічних стандартів (EAGLES), корпус вважають не лише як прозу чи поезію, але і як реєстри слів, їх упорядкування й лексикографічний опис – словники (Meyer, 2002). Під корпусом розуміють також різновид інформаційно-пошукової системи, певний масив, колекцію машиночитаних текстів, покликаних окреслити мовне розмаїття (Aarts, 2010; Biber, 2010; Kennedy, 1998; Leech, Fligelston, 1992; McEnery, Xiao, 2006). Йому властиві такі ознаки, як машиночитаність, автентичність текстів, добірність і репрезентативність (Демська-Кульчицька, 2005, с. 58).

Важливо пам'ятати, що корпусна організація лінгвальних даних потребує врахування типології текстових корпусів, оскільки це сприяє виробленню стратегії та принципів їх створення, добору фактичного матеріалу, відповідних практичних і теоретичних завдань тощо (Демська-Кульчицька, 2005, с. 58). У ході дослідження нами з'ясовано, що розрізняють такі текстові корпуси:

– повнотекстові, у яких тексти подані повністю, та фрагментні, що вміщують лише уривки текстів;

– дослідницькі, використовувані в лінгвістичних дослідженнях для формулювання гіпотез і теорій, та ілюстративні, що підтверджують або спростовують вже існуючі теорії;

– дослідницькі, що містять тексти як цілісні об'єкти та факти реалізації мовної системи, та інтерпретаційні, що є інформаційно-довідковими й дослідницькими системами;

– моніторингові, які відстежують зміни у мові, та статичні, що засвідчують стан мови на певному часовому зрізі;

– діахронні, що окреслюють мову в понадчасовому зрізі, та синхронні, які описують мову відповідного часового проміжку;

– загальномовні, що описують національну мову, та спеціалізовані, які окреслюють часткові, галузеві та специфічні науково-дослідні завдання.

Висновки. Отже, сучасні корпусні технології допомагають створювати електронні словники, яким властива низка переваг, зокрема охоплення значно ширшого обсягу інформації, тож вони ефективніші за традиційні, тобто за словники на паперових носіях. Корпусні технології доречні в зіставних дослідженнях, оскільки сприяють встановленню спільних та відмінних ознак різних мов і окресленню особливостей їх уживання. Вони допомагають у побудові частотних списків слів, що у подальшому дозволить користувачеві віднайти необхідну лексему, а також сприяють окресленню особливостей вживання певного слова у відповідному контексті. Перспективи подальших досліджень полягають у вивченні та вдосконаленні технологій створення електронного словника.

ЛІТЕРАТУРА

Бобкова Т. В. Корпус текстів : основні аспекти визначення. *Науковий вісник кафедри ЮНЕСКО Київського національного лінгвістичного університету. Філологія, педагогіка, психологія*. 2014. Вип. 29. С. 11–20. URL : http://www.mova. info/ corpus_papers/bobkova-corpus.pdf/.

Демська-Кульчицька О. Основи національного корпусу української мови : [монографія]. Київ : Інститут української мови НАНУ, 2005. 219 с.

Жуковська В. В. Вступ до корпусної лінгвістики : навчальний посібник. Житомир : Вид-во ЖДУ імені Івана Франка, 2013. 142 с.

Лендау С. І. Словники : мистецтво та ремесло лексикографії [пер. з англ.]. Київ : К. І. С., 2012. 480 с.

Aarts J., Meijs. W. Corpus Linguistics : Recent developments in the Use of Computer Corpora in English Language Research. Amsterdam : Rodopi, 1984. 425 p.

Biber D. Corpus-based and corpus-driven analyses of language variation and use. The Oxford Handbook of Linguistic Analysis / eds. B. Heine, H. Narrog. Oxford, 2010. P. 159–191.
Dash N. S. Corpus linguistics and language technology : with reference to Indian Languages. New Dehli : Mittal Publications, 2005. 445 p.
Kennedy G. Introduction to corpus linguistics. London : Longman, 1998. 315 p.
Leech G. Corpora and theories of linguistic performance. *Directions in corpus linguistics* / ed. J. Startvik. Berlin, 1992. P. 105–122.
Leech G., Fligelston S. Computers and corpus analysis. *Computers and written texts* / [ed. C. S. Butler]. Oxford : Blackwell Oxford, 1992. P. 115–140.
McEnery T., Hardie A. Corpus linguistics : method, theory and practice. Cambridge : Cambridge University Press, 2012. 294 p.
McEnery T., Xiao R., Tono Y. Corpus-Based Language Studies : An Advanced Resource Book. London, New York, 2006. 408 p.
Meyer Ch. F. English corpus linguistics. An introduction. Cambridge : Cambridge University Press, 2002. 168 p.
Tognini-Bonelli E. Corpus Linguistics at Work. Amsterdam : John Benjamins, 2001. 219 p.
Francis W. N. Language Corpora B. C. Directions in Corpus Linguistics / [ed J. Svartvik]. Berlin and New York : Moutin, 1992. P. 17–34.

REFERENCES

Aarts, J., Meijs, W. (1984). *Corpus Linguistics: Recent developments in the Use of Computer Corpora in English Language Research*. Amsterdam: Rodopi [in English].
Biber, D. (2010). Corpus-based and corpus-driven analyses of language variation and use. In B. Heine, H. Narrog (Eds.), *The Oxford Handbook of Linguistic Analysis* (pp. 159–191). Oxford [in English].
Bobkova, T.V. (2014). Korpus tekstiv: osnovni aspekty vyznachennia [Corpus of texts: main aspects of designation]. *Naukovyi visnyk kafedry YuNESKO Kyivskoho natsionalnoho linhvistychnoho universytetu. Filolohiia, pedahohika, psykholohiia – Scientific Bulletin of the UNESCO Chair of the Kiev National Linguistic University. Philology, pedagogy, psychology* (Issue 29), (pp. 11–20). Retrieved from http://www.mova. info/ corpus_papers/bobkova-corpus.pdf [in Ukrainian].
Dash, N. S. (2005). *Corpus linguistics and language technology: with reference to Indian Languages*. New Dehli : Mittal Publications [in English].
Demska-Kulchytska, O. (2005). *Osnovy natsionalnoho korpusu ukrainskoi movy [Fundamentals of the National Corpus of Ukrainian Movies]*. Kyiv: Instytut ukrainskoi movy NANU [in Ukrainian].
Francis, W. N. (1992). Language Corpora B. C. In J. Svartvik (Ed.), *Directions in Corpus Linguistics* (pp. 17–34). Berlin and New York: Moutin [in English].
Kennedy, G. (1998). *Introduction to corpus linguistics*. London: Longman [in English].
Leech, G. (1992). Corpora and theories of linguistic performance. In J. Startvik (Ed.), *Directions in Corpus Linguistics* (pp. 105–122). Berlin: Mouton de Gruyter [in English].
Lendau, S. I. (2012). *Slovnyky: mystetstvo ta remeslo leksykohrafii [Dictionaries: art and craft of lexicography]*; [per. z anhl.]. Kyiv: K. I. S. [in Ukrainian].
Leech, G., Fligelston, S. (1992). Computers and corpus analysis. In C. S. Butler, *Computers and written texts* (pp. 115–140). Oxford : Blackwell Oxford [in English].
McEnery, T., Hardie, A. (2012). *Corpus linguistics: method, theory and practice*. Cambridge: Cambridge University Press [in English].
McEnery, T., Xiao, R., Tono, Y. (2006). *Corpus-Based Language Studies: An Advanced Resource Book*. London, New York [in English].
Meyer, Ch. F. (2002*). English corpus linguistics. An introduction* / Charles F. Meyer. Cambridge: Cambridge University Press [in English].
Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam: John Benjamins [in English].
Zhukovska, V. V. (2013). *Vstup do korpusnoi linhvistyky [Entry to Corpus Linguistics]*. Zhytomyr: Vyd-vo ZhDU imeni Ivana Franka [in Ukrainian].