РОЗДІЛ ІІ. ІНФОРМАЦІЙНО-ОСВІТНІЙ ПРОСТІР: МЕТОДОЛОГІЧНІ ОСНОВИ ПРОЕКТУВАННЯ, СТВОРЕННЯ, ФОРМУВАННЯ

УДК 81’33

Барвіцька Г. К., Храпач Г. С.

ПРО ЗАСАДИ ЛІНГВІСТИЧНОГО КОРПУСУ ТЕКСТІВ І МОЖЛИВОСТІ ЙОГО ВИКОРИСТАННЯ У НАВЧАЛЬНОМУ ПРОЦЕСІ

У статті розглянуто дидактичні можливості використання лінгвістичного корпусу української мови. Обґрунтовано можливе використання ресурсів лінгвістичного корпусу в процесі підготовки молодого фахівця з галузевої термінології. Продемонстровано можливості застосування лінгвістичного корпусу в навчальному процесі та в позашкільній організації учнівської наукової діяльності.

Ключові слова: *корпусна лінгвістика, лінгвістичний корпус, українська галузева термінологія.*

Постановка проблеми. Останнім часом відзначається бурхливий розвиток прикладної лінгвістики, джерелом якого є постійно зростаюча потреба у застосуванні механізмів природної мови в інформаційно-комп’ютерних та людино-машинних системах. Наслідком цього вже фактично постала така галузь як лінгвістична технологія [15, с. 122]. Перспективи розвитку цієї галузі, так само як і прикладної лінгвістики взагалі, очевидно, не можна не пов’язувати із загальним прогресом комп’ютеризації різних сфер суспільного життя. Лінгвістичний аналіз вимагає опрацювання значної кількості мовних та мовленнєвих явищ, що ілюструють та підтверджують основні положення дослідження української галузевої термінології.

Процедура вибору ілюстрацій із тексту є досить кропіткою працею, і тому актуальним є вивчення можливостей та обґрунтування специфіки використання у дослідженнях великих текстових масивів з лінгвістичних корпусів.

Аналіз останніх досліджень і публікацій. Вагомий внесок у розвиток теорії та практики використання лінгвістичного корпусу внесли І. Г. Данилюк, Н. П. Дарчук, О. М. Костишин, О. Г. Рабулець, Т. І. Резнікова, Н. М. Сидорчук, В. А. Широков та інші. Корпуси текстів, їх переваги та можливості застосування у різних сферах науки і техніки досліджували О. М. Демська-Кульчицька, К. П. Сосніна, С. Н. Бук та інші.

Недостатній рівень дослідження електронних корпусів текстів, упереджене ставлення частини лінгвістів до можливості їх практичного використання у різноманітних галузях та науках, штучне обмеження сфери застосування корпусів текстів до розв'язання окремих прикладних завдань та відсутність базових навичок користування корпусами у переважної частини лінгвістів зумовлюють актуальність статті.

Метою статті є аналіз поняття «лінгвістичний корпус» та визначення його переваг і можливостей застосування у навчальному процесі, зокрема при вивченні української галузевої термінології.

Виклад основного матеріалу. Головним поняттям корпусної лінгвістики є корпус мовленнєвої реалізації мови, що кваліфікується як сформована за певними вимогами вибірка мовленнєвого матеріалу, який може використовуватися для опису й дослідження мови як системи [12, с. 668]. Серед переваг корпусної лінгвістики дослідники називають його здатність наводити мовний матеріал у його реальному оточенні, що, у свою чергу, дозволяє досліджувати лексичну і граматичну структуру мови, а також неперервні процеси мовних змін, що відбуваються в мові впродовж того чи іншого періоду [18, с. 36]. Свого бурхливого розвитку корпусна лінгвістика зазнала в останні десятиліття минулого сторіччя завдяки появі потужних комп'ютерних технологій.

Вільям Френсіс, один із піонерів сучасної корпусної лінгвістики визначає корпус текстів як зібрання текстів, що вважаються репрезентованими по відношенню даної мови, діалекту, чи іншої частини мови та призначене для використання в лінгвістичних дослідженнях [19, с. 22].

Російський дослідник С. О. Шаров обґрунтовує власне розуміння поняття «лінгвістичний корпус». Корпусом може бути будь-яка колекція текстів з певної тематики, які є доступними в електронній формі (*Корпус 1*). Колекція текстів, зібрана у відповідності до явно сформульованих правил і, можливо, розмічена на певному рівні лінгвістичного аналізу (*Корпус 2*). На думку вченого, важливою умовою функціонування лінгвістичного корпусу є колекція текстів, що має бути збалансованою за жанрами та функціональними стилями і мати достатній обсяг вибірки за числом текстів та авторів, для того, щоб слугувати основою статистично достовірних досліджень лінгвістичних феноменів у текстах відповідної тематики [14, с. 10].

О. М. Демська-Кульчицька вважає корпусом електронне зібрання текстів природної мови впорядковане, організоване й оформлене певним чином, призначене для наукового та практичного вивчення мови [7, с. 6].

За визначенням В. А. Широкова, корпус текстів – це вид корпусу даних, одиницями якого є тексти або їх достатньо значні фрагменти, що включають, наприклад, якісь повні фрагменти макроструктури текстів даної проблемної області [17, с. 25].

Для В. П. Захарова «корпус текстів» є великим, уніфікованим, структурованим, розміченим масивом мовних (мовленнєвих) даних в електронному вигляді, призначеним для певних цілей [9, с. 73].

Школа О. С. Герда визначає лінгвістичний корпус текстів як великий за обсягом, представлений в електронному вигляді, уніфікований, структурований, розмічений і філологічно компетентний масив мовних даних, доповнений системою керування даними – універсальними програмними засобами для пошуку різноманітної лінгвістичної інформації та зручного

представлення її широкому колу користувачів [3].

Основою лінгвістичного корпусу за теорією, обґрунтованою в Українському мовно-інформаційному фонді, є електронна бібліотека, збудована за стандартними правилами конструювання електронних бібліотек [16].

Корпусна лінгвістика – це розділ мовознавства, що займається дослідженням, розробкою, створенням та використанням текстових (лінгвістичних) корпусів, а також лінгвістичним аналізом на їх основі [17, с. 10-13].

З появою корпусів значно зросла ефективність лінгвістичної обробки великих масивів текстів. Тепер по суті немає обмежень щодо обсягу матеріалу і швидкості пошуку в ньому інформації. Крім того, у своєму розпорядженні дослідник має численні тексти різних типів. Це суттєво впливає на розвиток його знань про мову: можливість масової – зокрема статистичної – обробки текстів дозволяє виявити в структурі мови нові закономірності, вести спостереження за динамікою розвитку мови.

В Україні найбільший Національний лінгвістичний корпус української мови створено групою вчених Українського мовно-інформаційного фонду (УМІФ) в 2003 році. У корпусі зібрані тексти всіх жанрів, зокрема, публіцистика, наукова література всіх напрямів, офіційно-ділові документи і, звичайно, художні тексти тощо. Корпус забезпечений спеціальним пошуковим апаратом, необхідним для роботи з текстами. За рахунок цього апарату з’являється можливість зручного і швидкого пошуку будь-яких слів і словосполучень з урахуванням параметрів, які цікавлять користувача (починаючи з граматичних і семантичних характеристик і закінчуючи сортуванням текстів за жанрами, стилями, авторам тощо). Така збалансованість безлічі текстів щодо жанрів і стилів, а також наявність достатнього обсягу та вибірки за кількістю текстів та авторів, що дає можливість здійснювати статистично достовірні дослідження в текстах відповідної тематики, називається *репрезентативністю* [14]. Вимога

репрезентативності, на думку О. Демської-Кульчицької, полягає у здатності корпусу відображати всі властивості предметної галузі [6, с. 102]. Але корпус – не просто інструмент для лінгвістичних досліджень – це, фактично, довідково-інформаційна система, яка дозволяє отримувати відповіді на найнесподіваніші запитання і окреслювати нові проблеми [17].

У загальному розумінні *лінгвістичний корпус* – це комплекс універсальних програмних засобів для пошуку різноманітної лінгвістичної інформації. При укладанні словників галузевих термінологій лінгвістичний корпус може стати важливим допоміжним інструментом, адже він забезпечує можливість пошуку нового ілюстративного матеріалу, розширення реєстрового масиву, проведення верифікації словникових статей. Обсяги мовного матеріалу, який залучається до лінгвістичного дослідження, комплексність, оперативність опрацювання зазначеного матеріалу та можливість прямого доступу до великого числа лінгвістичних фактів – це ті переваги, які надає лінгвістичний корпус молодому досліднику.

На нашу думку, цінність корпусу можна вбачати в наступному:

- створений одного разу корпус може багато разів використовуватися у майбутньому;
- корпус показує мовні дані в їх реальному оточенні;
- корпус характеризується збалансованим складом текстів, що дозволяє використовувати його для тестування пошукових машин, машинних морфологій, систем перекладу, а також використовувати його в різних лінгвістичних дослідженнях;
- корпус має важливе значення для викладання мови.

Г. І. Дідук-Ступ'як зазначає, що корпуси текстів слугують інформаційною базою для проведення трьох основних напрямів дослідження: 1) власне лінгвістичних синхронних та діахронних досліджень; 2) статистичних досліджень та 3) досліджень у сфері методики викладання мов [8, с. 106]. Варто наголосити на зміцненні зв'язку між корпусною лінгвістикою та методичними напрацюваннями із проблем підготовки

фахівців певної наукової, фахової галузі.

Створення спеціального корпусу галузевої терміносистеми, на нашу думку, буде важливим поступом у розробці корпусу української мови, а також цікавим для лінгвістів та дослідників галузевої термінології, що стає все більш перспективним для дослідження лінгвостилістики із розвитком корпусної лінгвістики.

Подібний корпус може застосовуватися для лінгвістичного аналізу з метою виявлення основних лексичних та синтаксичних помилок мовців та труднощів, що виникають при вивченні термінології. Це допомагає встановити частотність різних типів мовних помилок, види контекстів в яких вони найчастіше зустрічаються та розробити ефективні план та метод покращення навчання.

З огляду на це, все більше вчених звертають увагу на те, що наявність необхідної лінгвістичної інформації для подальших філологічних досліджень є лише у корпусах текстів, активною розробкою яких займаються практично усі галузеві інституції країн світу [5, с. 45-46].

Корпус текстів з галузевої термінології належить до статичних типів корпусів і може бути як складником загального корпусу певної мови так і окремим утворенням, слугуючи виключно для дослідження певної галузевої терміносистеми. Алгоритм утворення такого корпусу полягає у наступних етапах:

- визначення мети створення корпусу текстів;
- відбір джерел текстів відповідно до типу створюваного корпусу;
- створення електронної бібліотеки текстів створюваного корпусу;
- попередня обробка текстів, підготовка бібліографічної та екстралінгвістичної інформації про тексти корпусу;
- конвертування та автоматична обробка текстів, яка полягає у поділі текстів на структурні сегменти, видалення / переробка таблиць чи малюнків, видалення переносів, тощо;

- розмітка (анотування) текстів. Метаопис здійснюється переважно вручну, структурна та власне лінгвістична розмітки – автоматично. Структурна розмітка повинна при цьому відповідати міжнародним стандартам (TEI, EAGLES, тощо), а лінгвістична здійснюватись за допомогою спеціальних програм-аналізаторів;
- перевірка отриманих результатів, усунення омонімії. В протилежному випадку не уникнути помилок при наступних етапах розмітки (лексичної, семантичної, синтаксичної, тощо);
- створення синтаксичного аналізатора з метою проведення синтаксичної розмітки;
- проведення семантичної розмітки із застосуванням семантичних та тлумачних словників;
- перетворення корпусу у спеціально розроблену інформаційно-пошукову систему (т. з. корпус-менеджер), яка забезпечуватиме багатоаспектний пошук та статистичну обробку;
- забезпечення доступу до корпусу через мережу Інтернет [9, с. 73; 17, с. 467].

Призначення корпусу – показати функціонування лінгвістичних одиниць в їх природньому контекстному оточенні. На основі корпусу можна одержати дані: про частоту словоформ, лексем, граматичних категорій; про зміну частот; про зміни контекстів у різні періоди часу; про поведінку мовних одиниць різних авторів; про особливості їх сполучуваності, управління тощо.

Безпосереднім наслідком упровадження лінгвістичного корпусу як провідного технологічного інструментарію сучасного мовознавчого дослідження стало розширення меж лексикографічного опису галузевої терміносистеми. Не випадково, що навіть сам лінгвістичний корпус у його найпростішому варіанті та безпосередній формі за своєю структурою нагадує певний специфічний словник, де реєстровою одиницею і водночас «лівою частиною словникової статті» виступає певна лексема, а інтерпретаційною, «правою», частиною слугує сума «мікроконтекстів», до яких входить

реєстрова лексема (у всіх її граматичних значеннях) [17, с. 34]. Наприклад, див. рис. 1.

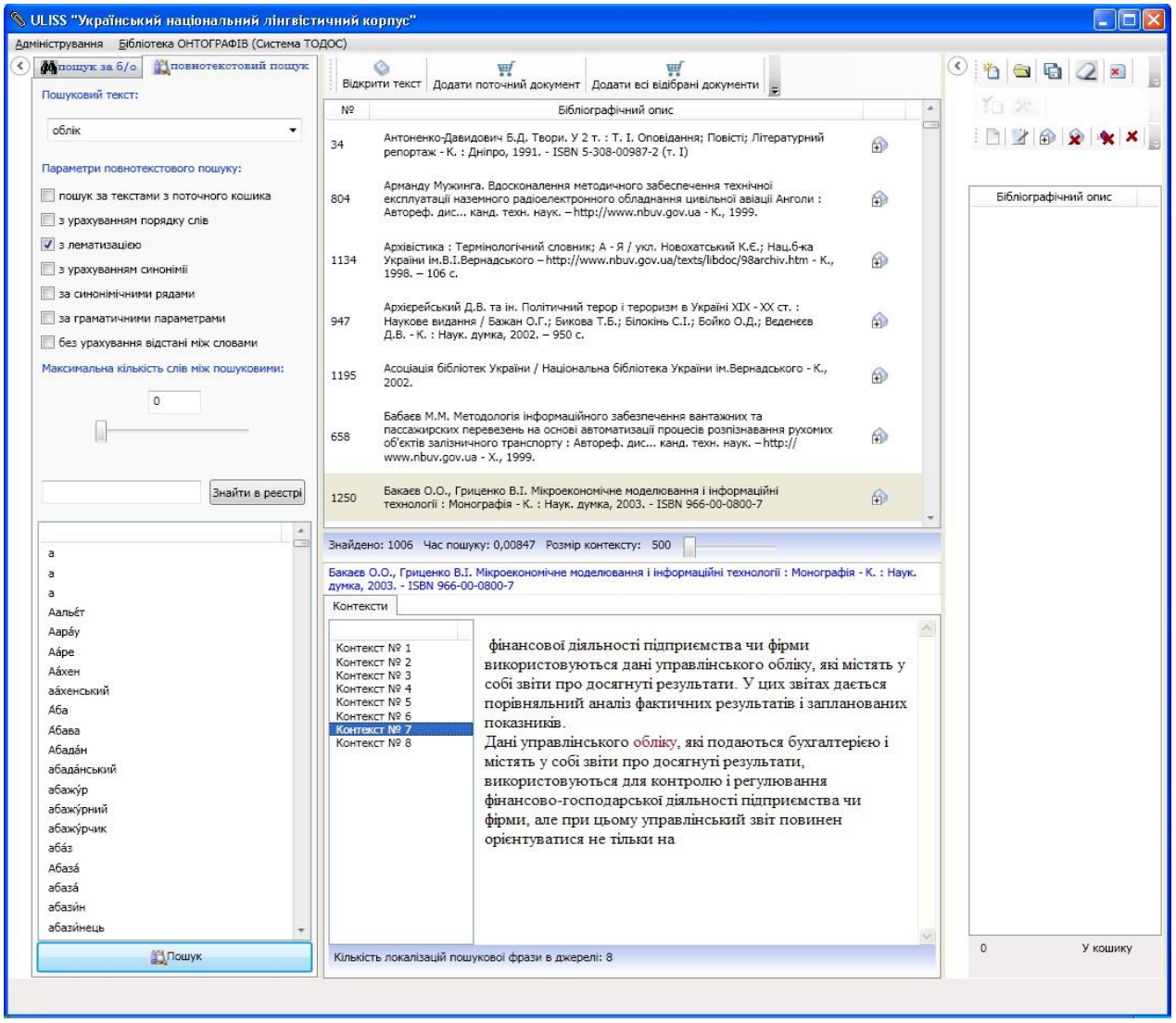

Рис. 1. Інтерфейс результату пошуку

Інтерфейс сайту дозволяє здійснювати пошук того чи іншого слова або фрази, частини мови або ж комбінувати наведені параметри пошуку. Крім того, ресурс дає можливість здійснювати пошук лексичних одиниць разом із оточенням, що допомагає зрозуміти контекстуальне значення тієї чи іншої мовної одиниці та специфіку її функціонування в мовленні.

Висновки. Для подальшого розвитку мовознавчої науки, зважаючи на основні стратегії розвинених лінгвістик світу, вкрай необхідним є створення і використання лінгвістичного корпусу як мовно-інформаційного інструменту, спроможного забезпечити дослідників необхідним набором функцій для

дослідження мовних систем на сучасному науковому рівні.

Запропонований шлях використання лінгвістичного корпусу може застосовуватися для якісного аналізу галузевих понять з метою виявлення суперечностей чи дублювання інформації.

Список використаної літератури

1. Баранов А. Н. «Динамический корпус текстов» как новая технология прикладной лингвистики / А. Н. Баранов, М. Н. Михайлов, Г. О. Сидоров // Труды Международ. Семинара: Диалог-98 по компьютерной лингвистике и ее приложениям. – 1998. Сер. 2. – Т. 2.

2. Бук С. Учнівські корпуси в методиці викладання іноземної мови / С. Бук // Теорія і практика викладання української мови як іноземної, 2007. – № 2. – С. 19-23.

3. Герд А. С. Вступне слово на Міжнародній конференції «Корпусная лингвистика: 2004». [Електронний ресурс]: – Електронні дані – Режим доступу: http://www.phil.pu.ru/news/kllbd/corpling.htm. – Назва з екрану.

4. Данилюк І. Корпус текстів для вивчення граматичної службовості / І. Данилюк // Лінгвістичні студії. – Вип. 26. – Донецьк : ДонНУ, 2013. – С. 224–229.

5. Дарчук Н. Дослідницький корпус української мови: основні засади і перспективи / Н. Дарчук // Вісник Київського Національного університету імені Тараса Шевченка. Серія «Літературознавство. Мовознавство. Фольклористика». – 2010. – № 21. – С. 45 – 49.

6. Демська О. М. Репрезентативність як ознака текстового корпусу // Українська мова. – 2005. – №3. – С. 100–107.

7. Демська О. М. Текстовий корпус : ідея іншої форми / Орися Демська. – К. : ВПЦ НаУКМА, 2011. – 282 с.

8. *Дідук-Ступ'як Г. І.* Лінгводидактичні можливості корпусної лінгвістики / Г. І. Дідук-Ступ'як // Наукові записки ТНПУ ім. В. Гнатюка. сер. Педагогіка. – Тернопіль, 2010. – № 1. – С. 105-109.

9. Захаров В. П. Чешский национальный корпус текстов: организация и способы использования // Доклады научной конференции «Корпусная лингвистика и лингвистические базы данных» 5-7 марта 2002 г. – Санкт-Петербург, 2002. – С. 72-79.

10. Костишин О. М. Системотехнічні та лінгвістичні принципи проектування українського лінгвістичного корпусу / О. М. Костишин, Н. М. Сидорчук // Наукові праці Національної бібліотеки України 47н.. В. І. Вернадського НАН України. Випуск 141. / Національна бібліотека України 47н.. В. І. Вернадського; Редкол. : Онищенко О. С. (гол.) та 47н.. – К., 2005. – С. 190-199.

11. Резникова Т. Славянская корпусная лингвистика: современное состояние ресурсов / Т. Резникова // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб. : Нестор-История, 2009. – С. 402–461.

12. Селіванова О. О. Сучасна лінгвістика: напрями і проблеми / О. О. Селіванова. – Полтава : Довкілля-К, 2008. – 712 с.

13. Соснина Е. П. Корпусная лингвистика и корпусный поход в обучении иностранному языку // Ульян. гос. техн. ун-т. – Ульяновск [Електронний ресурс]: – Електронні дані – Режим доступу:/ http://ling.ulstu.ru/linguistics/resourses/literature/articles/ corpus_ linguistics_language_teaching/ – Назва з екрану.

14. Шаров С. А. Представительный корпус русского языка в контексте мирового опыта // НТИ. – Сер. 2. Информационные системы. – 2003. – №6. – С. 8-18.

15. Широков В. А. Гуманітарна традиція та технологічний статус мови // Мовознавство. – 2001. №3. – С. 120–132.

16. Широков В. А. Інформаційна теорія лексикографічних систем – Київ : Довіра, 1998. – 331 с.

17. Широков В. А. Корпусна лінгвістика / В. А. Широков, О. В. Бугаков, Т. О. Грязнухіна, О. М. Костишин, М. Ю. Кригін; НАН України, Укр. Мов.-інформ. Фонд. – К. : Довіра, 2005. – 472 с.

18. Krajka J. Corpora and language teachers: from ready-made to teacher-made collections / J. Krajka // Computer resources for language learning. – 2007. – №1. – P. 36–55.

19. Francis W. Directions in Corpus Linguistics : proccedings of Nobel Symposium 82. Stockholm, 4-8 August 1991. – Berlin – New Yourk : Mounton de Gruyter. – P. 17-35.

Галина Барвицкая, Анна Храпач. Основы лингвистического корпуса текстов и возможности его использования в учебном процессе.

В статье рассмотрены дидактические возможности использования лингвистического корпуса украинского языка. Обоснованно возможное использование ресурсов лингвистического корпуса в процессе подготовки молодого специалиста по отраслевой терминологии. Продемонстрированы возможности применения лингвистического корпуса в учебном процессе и во внешкольной организации ученической научной деятельности.

Ключевые слова: *корпусная лингвистика, лингвистический корпус, украинская отраслевая терминология.*

Galina Barvytskaya Anna Hrapach. Fundamentals linguistic corpus and opportunities using texts in school.

The article discusses the possibility of using didactic linguistic corpus of the Ukrainian language. Reasonably possible use of the resources of a linguistic corpus in the preparation of young specialist on industry terminology. The possibilities of the use of linguistic corpus in the learning process and in the organization of student extracurricular academic activities are defined.

Key words: *corpus linguistics, linguistic corpus, Ukrainian industry terminology.*