Міністерство освіти і науки України
Житомирський державний університет імені Івана Франка

В.В. Жуковська

ВСТУП ДО КОРПУСНОЇ ЛІНГВІСТИКИ

Навчальний посібник

Вид-во ЖДУ ім. І.Франка
Житомир 2013

УДК 81’33 (076)
ББК 81.1 я 7
АЗ Ж 86

Рекомендовано Вченою радою Житомирського державного університету імені Івана Франка
Протокол № 3 від 25 жовтня 2013 р.

Автор:

Жуковська В. В. – кандидат філологічних наук, доцент, завідувач кафедри міжкультурної комунікації

Рецензенти:

Морозова О. І. – доктор філологічних наук, професор, професор кафедри англійської філології Харківського національного університету імені В.Н. Каразіна
Сингаївська А. В. – кандидат філологічних наук, професор, директор Навчально-наукового інституту іноземної філології
Могельницька Л. Ф – кандидат філологічних наук, доцент, завідувач кафедри іноземних мов Житомирського державного технологічного університету

Жуковська В. В.
Вступ до корпусної лінгвістики: навчальний посібник / В. В. Жуковська – Житомир: Вид-во ЖДУ ім. І. Франка, 2013. – 142 с.

Навчальний посібник покликаний ознайомити студентів філологічних спеціальностей з базовими поняттями корпусної лінгвістики, охарактеризувати її мету та завдання, описати методи дослідження в цій галузі, представити основи корпусних технологій та показати теоретичне й практичне значення корпусів для проведення наукових досліджень в галузі лінгвістики та для навчання і вивчення іноземної мови.

Адресований студентам вищих навчальних закладів. Прислужиться всім, хто займається прикладними лінгвістичними дослідженнями та цікавиться питаннями реального функціонування мови.

Зміст

Вступ

У сучасних умовах постійного зростання інформаційних потоків проблема автоматичного опрацювання природномовної інформації набуває всі більшого значення. Комп'ютерні технології, прийшовши в лінгвістику, запропонували свої технічні можливості для опрацювання, збереження і відбору мовного матеріалу. Це дало поштовх до зародження та розвитку нової галуззі мовознавчих досліджень: корпусної лінгвістики. Базуючись на використанні величезних масивів емпіричного матеріалу та залученні комп'ютерних технологій для отримання об'єктивних висновків про функціонування мовних одиниць, лінгвістичні корпуси швидко перетворилися на неоціненний лінгвістичний ресурс. Отримані в результаті корпусного аналізу дані сприяють формулюванню якісно нових висновків про мову та окреслюють такі напрями лінгвістичних досліджень, які до появи корпусів не привертали уваги дослідників. Спираючись головним чином на реальний «живий» мовний матеріал, а не на мовну інтуїцію та інтроспекцію, корпусні дослідження дозволяють абстрагуватися від суб'єктивності дослідника і наблизитися до об'єктивного вивчення мови.

З часу свого становлення до сьогодні корпусна лінгвістика пройшла складний шлях від повного неприйняття у 60-ті роки XX століття до магістрального напряму сучасного мовознавства. Про популярність та інтенсивність розвитку корпусної лінгвістики у світі свідчить як чимале число наукових розвідок у провідних міжнародних наукових виданнях, посібників, підручників, так і самих корпусів різних типів. Вітчизняне корпусне мовознавство ще перебуває на етапі становлення. Відчувається нестача ґрунтовних україномовних навчальних посібників, які б ознайомили майбутніх філологів з базовими

принципами та методами проведення корпусних досліджень. Навчальний посібник «Вступ до корпусної лінгвістики» покликаний забезпечити лекційний курс з навчальної дисципліни «Корпусна лінгвістика», що викладається студентам магістратури спеціальності 8. 02030302 Мова і література (англійська)*.

Метою посібника є ознайомити студентів філологічних спеціальностей з базовими поняттями корпусної лінгвістики, охарактеризувати її мету та завдання, описати методи дослідження в цій галузі, представити основи корпусних технологій та показати теоретичне й практичне значення корпусів для проведення наукових досліджень в галузі лінгвістики та для навчання і вивчення іноземної мови.

Виходячи з поставленої мети, посібник спрямовано на розв'язання таких *завдань*:

1) ознайомити студентів із предметом, завданнями, методами й понятійним (термінологічним) апаратом корпусної лінгвістики;
2) представити історію корпусних досліджень від перших паперових конкордансів до сучасності;
3) надати інформацію про комп'ютерний інструментарій корпусної лінгвістики;
4) проаналізувати основні типи існуючих корпусів;
5) ознайомити студентів із практикою укладання корпусів;
6) сформувати навички роботи із програмними засобами й інформаційними ресурсами корпусної лінгвістики;
7) продемонструвати можливості роботи з корпусними базами даних в лінгвістичних дослідженнях та з метою навчання і вивчення іноземної мови;

Слід зауважити, що запропонований посібник ніяким чином не претендує на повноту та вичерпність опису актуальних

проблем у царині корпусної лінгвістики. Проте автор сподівається, що викладений матеріал сприятиме формуванню у майбутніх філологів більш цілісного та поглибленого свідомлення професійної діяльності, наближенню її до сучасного рівня наукових лінгвістичних знань, розвитку компетенцій самостійного здійснення наукових досліджень та отримання нових знань за допомогою комп'ютерних технологій.

Розділ 1.
Корпусна лінгвістика як галузь прикладного мовознавства

1.1. Сутність, предмет і завдання корпусної лінгвістики

Корпусна лінгвістика є одним із найбільш перспективних напрямків сучасного теоретичного і прикладного мовознавства. Ця відносно нова лінгвістична галузь розпочала своє активне становлення у 60-х роках XX століття у зв'язку із інтенсивним розвитком комп'ютерних технологій. Сам термін «корпусна лінгвістика» міцно ввійшов до наукового вжитку лише в останні десятиліття XX століття з публікацією у 1983 році збірника наукових праць «Corpus Linguisitcs: Recent Developments in the Use of Computer Corpora in English Language Research» за матеріалами конференції ICAME "Conference on the Use of Computer Corpora in English Language Research". Звичайно, застосування комп'ютерів та спеціального програмного забезпечення суттєво змінило спосіб дослідження мови та значно полегшило роботу по збору лінгвістичних даних. Без зусиль, лише за декілька секунд, стало можливим здійснювати пошук у багатомільйонних текстових масивах (лінгвістичних корпусах), будувати конкорданс для будь-якого слова, одержувати дані про частоту словоформ, лексем, граматичних категорій, синтаксичних конструкцій, відстежувати зміни у частоті і контексті мовної одиниці у різні хронологічні періоди, одержувати дані про лексичну (колокацію) та граматичну (колігацію) сполучуваність і т. ін. Створення емпіричної бази дослідження таким способом забезпечує мінімум витрат зусиль і часу на зібрання даних, обсяг яких водночас суттєво збільшується.

Сьогодні дані корпусів масштабно використовуються в лексикографії, стилістиці, судовій лінгвістиці, лінгвістичній варіантології, перекладознавстві, соціолінгвістиці, методиці навчання і вивчення іноземної мови та в багатьох інших лінгвістичних дослідженнях [Baker, 2006: 2-3].

Отож, що ж становить собою корпусна лінгвістика? Передусім, корпусна лінгвістика як галузь прикладного мовознавства займається визначенням загальних принципів побудови, обробки та експлуатації даних лінгвістичних корпусів (корпусів текстів) із використанням сучасних комп'ютерних технологій, розробленням методики збору реальних мовних явищ – писемних та усних текстів, а також способів їх збереження та аналізу. Під ***корпусом текстів*** розуміється значний за обсягом, представлений в електронному вигляді, уніфікований, структурований, розмічений, філологічно компетентний масив мовних даних, створений для вирішення конкретних лінгвістичних завдань [Захаров, 2005: 3].

Корпусний аналіз вирізняється низкою характерних ознак: 1) емпіричний підхід до аналізу мовних даних (досліджуються реальні моделі мовної реалізації у природних текстах); 2) використання великих за обсягом, структурованих колекцій природних текстів (корпусів) як основи для аналізу; 3) широке залучення комп'ютерних технологій для дослідження лінгвального матеріалу; 4) застосування кваліативних і квантитативних аналітичних методик, з суттєвою перевагою останніх (вивчення частоти вживання лінгвістичних одиниць, статистичні дослідження сполучуваності і т.ін.). Отримані в результаті корпусного аналізу дані не тільки сприяють формулюванню якісно нових висновків про мову, але й окреслюють такі напрями досліджень, які до появи корпусів не привертали уваги дослідників. Спираючись головним чином на

реальний «живий» мовний матеріал, а не на мовну інтуїцію та інтроспекцію, корпусні дослідження дозволяють абстрагуватися від суб'єктивності дослідника і наблизитися до об'єктивного вивчення мови. І, як стверджує Л. Ричкова, матеріал корпусу дозволяє не лише оптимізувати і об'єктивізувати лінгвістичні дослідження, але і по-новому окреслити багато традиційних лінгвістичних понять [Ричкова, 2003: 185]. Корпусні розвідки переорієнтовують традиційний підхід до вивчення мови, а результати аналізу даних корпусу сприяють переоцінці низки лінгвістичних теорій [MacEnery, Hardie, 2012: 1].

Власне кажучи, численні розвідки в галузі сучасного корпусного мовознавства відбуваються у двох магістральних напрямках. Перший напрям зосереджений на розробці проблем, що стосуються теорії та практики створення корпусів, іншими словами, концептуалізації корпусу: типологія корпусу, його призначення, обсяг, параметризація предметної галузі, репрезентативність, структурування та принципи відбору базових одиниць, зберігання і т. ін. А другий напрям спрямований на дослідження саме лінгвістичних корпусів, тобто вивчення мови за допомогою корпусних методів [Копотев, Мустайоки, 2008: 12]. Проте слід зауважити, що чіткої межі між зазначеними напрямами не існує, адже практично всі укладачі корпусів в той же час здійснюють і лінгвістичні дослідження на їх основі. Така двовекторність корпусної лінгвістики зумовлюється подвійною природою ***об'єкта*** її дослідження – текстового корпусу, який, з одного боку, виступає в якості вихідного мовленнєвого матеріалу для корпусної лінгвістики, а з іншого, є результатом діяльності цього мовознавчого напряму. ***Предметом*** корпусної лінгвістики виступають теоретичні основи і практичні механізми створення та експлуатації мовних корпусів [Захаров, Богданова , 2011: 10].

Головне ***завдання*** корпусної лінгвістики вбачається у повному й системному відображенні змістовного спілкування мовою. Важливою особливістю цього інформаційно-семіотичного напряму лінгвістичних досліджень є підхід до розгляду прикладних проблем лінгвістики конкретно в комунікативних процесах. При цьому в центрі уваги виявляється не мова як система, і не проблема її формалізації, а процес змістовного спілкування мовою, і по можливості точний її опис, який може бути використаний для розв'язання науково-технічних завдань інформатики [В.В. Риков[1]].

Першочерговою ***метою*** цієї мовознавчої галузі є об'єктивний лінгвістичний опис мовної системи, причому до цього опису корпусна лінгвістика підходить від вивчення конкретної людської комунікації. У якості другорядної цілі розглядається вироблення особливого способу відображення мовного матеріалу в корпусі текстів. Цей спосіб, у свою чергу, може використовуватися й іншими лінгвістичними дисциплінами.

Теоретичним підґрунтям корпусної лінгвістики є структуралізм, який декларує примат реального тексту в лінгвістичному дослідженні. Для корпусних розвідок головним є постулат, що мова як об'єкт дослідження може бути вивчена лише у формі писемних та усних текстів [Демьска 2010: 6]. У тексті, як формі існування мовлення, існує лише те, що передбачено системою мови, тому корпусна лінгвістика розглядає текст як засіб експлікації специфіки міжелементних мовних зв'язків. Дослідження тексту поглиблює розуміння суті природної мови. Для корпусного мовознавства притаманним є дещо відмінна від традиційного трактування рецепція тексту, згідно якої текст є тим, що дано лінгвістові для дослідження

[1] http://corpora.iling.spb.ru/theory.html

мови, мовною дійсністю, реальністю, яка піддається науковому спостереженню, а не особливою мовною одиницею. Корпусний підхід до тексту перетворює процедуру інвентаризації тексту на науковий метод, оскільки базується на принципі, згідно з яким із тексту можна виділити ті риси (ознаки, параметри), на основі котрих реально збудувати певну модель об'єкта дослідження. Корпус трактує текст під такими кутами зору: а) єдине, що дано дослідникові мови як вихідний пункт у своїй нерозчленованій і абсолютній цілісності – це текст*;* б) лінгвістична теорія починається з тексту як єдиного даного й намагається досягти несуперечливого й вичерпного опису цього тексту шляхом аналізу чи послідовного розподілу; в) через те, що тексти природної мови надзвичайно великі числом і тривалістю, доцільно задовольнятися певною вибіркою з них або членуванням текстового континууму на твори окремих авторів, окремі праці, розділи, параграфи тощо, а відтак на складні й прості речення та слова. Тобто корпус, про який ми говоримо, тільки текстовий, текст у ньому може творити окремий корпус (наприклад, корпус Біблії) або, частіше, окремий текст, або його фрагмент є структурним елементом зібрання текстів чи корпусу текстів. Оскільки до корпусу не можуть увійти всі тексти, створені у різні історичні періоди, то застосовують прийом цільового відбору текстів або фрагментів текстів. Корпусний текст відрізняється від так званого класичного тексту, крім іншого, наявністю анотації: спеціальної розмітки структури тексту або текстів, композиційних та мовних одиниць тексту або текстів, і така розмітка неможлива без глибинного лінгвістичного аналізу текстових одиниць, що передбачає ієрархічний поділ тексту або текстів аж до виділення навіть літерного символу [Демьска 2010: 6].

В основі корпусної лінгвістики лежить розуміння мови як повністю соціального явища, яке можна описати заснованими на досвіді даними, тобто в мовленнєвому акті. Як соціальний феномен мова проявляє себе саме в текстах, які можна записати, описати й проаналізувати. Внутрішні, «німі» тексти також є текстами, але їх не можна спостерігати і, отже, вони не є соціальним явищем. Більшість текстів зустрічаються у вигляді мовленнєвих актів, тобто як взаємодія між членами мовного суспільства [Teubert, 2007: 112].

Все вище сказане чітко окреслює дослідницьку програму корпусної лінгвістики, яка, будучи суто емпіричною дисципліною, при аналізі лінгвального матеріалу покладається на реальне функціонування мови з метою встановлення правил та вивчення особливостей продукування мови людиною, на відміну від тих досліджень, які опираються на вигадані приклади чи інтроспекцію. Застосування комп'ютерів дозволяє миттєво обробити величезний обсяг мовного матеріалу і відібрати всі можливі у конкретному корпусі приклади вживання необхідних для аналізу одиниць. У розпорядження лінгвіста надаються об'єктивні кількісні дані, забезпечуючи досягнення більш ґрунтовних та переконливих висновків. Корпусна лінгвістика дозволяє вченим підтвердити або спростувати гіпотези про функціонування мови, а також окреслити нові напрями дослідження, які до застосування корпусних методів не попадали до фокусу уваги дослідників.

1.2. Корпусна лінгвістика в системі мовознавчих наук

Незважаючи на значні досягнення та популярність, статус корпусної лінгвістики у сучасній мовознавчій парадигмі ще не є однозначно визначеним (див. наприклад: [Biber, Conrad, Reppen,

1998: 3; Kennedy, 1998: 2; McEnery, Wilson 2001: 2; Meyer 2002: xi]. Так, Т. МакЕненрі, Е. Вільсон, Р. Ксіао та інші лінгвісти характеризують її як особливу «методологію» (*methodology*) дослідження мовного матеріалу, на противагу таким традиційним галузям лінгвістики, як фонетика, синтаксис, семантика чи граматика. Зазначені мовознавчі галузі досліджують окремий аспект мовної системи, натомість корпусна лінгвістика не обмежується вивченням лише одного аспекту мовної реалізації. Більш того, корпусні методи можуть використовуватися для вивчення мовних одиниць будь-якого мовного рівня. Наприклад, синтаксичні явища можливо дослідити як із застосуванням корпусних методик, так і без них, так само існують корпусні і некорпусні семантичні дослідження [McEnery, Wilson, 2001; McEnery, Xiao, Tono, 2006: 6]. Відзначається, що корпусна лінгвістика як система методів і принципів використання корпусів для вивчення мови та для вивчення/навчання мові має теоретичне підґрунтя, але саме по собі теоретичне підґрунтя ще не є науковою теорією. Звідси, корпусна лінгвістика визнається методологією вивчення мови з широкими можливостями застосування в багатьох лінгвістичних напрямах та теоріях [McEnery, Xiao, Tono, 2006: 7].

Проте більшість вітчизняних та зарубіжних лінгвістів не погоджується з цією точкою зору і розглядає корпусну лінгвістику як самостійну дисципліну прикладного мовознавства з власним предметом, об’єктом і метою дослідження, термінoапаратом, теоретичним і практичним підґрунтям, власними дослідними прийомами [Демська, 2005: 8, 12; Баранов, 2001: 112–117]. На переконання російських лінгвістів В.П. Захарова та С.Ю. Богданової, корпусна лінгвістика має принаймні дві ознаки, що дають їй підставу претендувати на статус самостійної дисципліни: 1) характер

аналізованого словесного матеріалу; 2) специфіка інструментарію [Захаров, Богданова, 2011: 9]. Сучасна корпусна лінгвістика вже «вийшла зі статусу методології» та стала окремою «дисципліною» [Tognini-Bonelli 2001: 1], сформувалася в самостійний науковий напрямок, досягнення якого знаменують новий етап розвитку наукової думки [Колпакова, 2011: 1]. Російський вчений, член-кореспондент РАН, завідувач відділу корпусної лінгвістики Інституту російської мови ім. В.В. Виноградова РАН, професор МДУ ім. М.В. Ломоносова Володимир Олександрович Плунгян переконаний, що корпус – це не просто новий і потужний інструмент: за використанням корпусу стоїть певна ідеологія, основні тенденції якої зародилися ще в класичній філології XIX століття, але значно інтенсифікувалися в останні десятиліття. Головними пріоритетами цієї ідеології є:

1) увага не до слова чи речення, а до *тексту* (дискурсу), тобто до реального інструменту комунікації в цілому, а не до його окремих фрагментів;
2) увага до *квантитативного компонента мови*, тобто врахування в першу чергу більш частотних елементів порівняно з менш частотними, визнання квантитативних відношень суттєвим фактором у мовній еволюції і структурі мовних правил;
3) увага до *синхронічної варіативності мови*, тобто визнання того факту, що не існує єдиної жорсткої системи засобів вираження змісту, а існують її різні реалізації, в тому числі залежні від психологічних, біологічних і соціальних факторів;
4) увага до *діахронічної варіативності мови*, тобто визнання того факту, що мова постійно змінюється у часі і повністю відволіктися від цієї нестабільності не можливо, в кожен

момент часу в мові співіснують «прогресивні» і «консервативні» ділянки;

5) зміна відношення до поняття мовної норми і мовної правильності, тобто межа між «помилкою» та «маргінальним варіантом» визнається більш рухомою та хиткою [Плунгян, 2008: 7–20.].

Таким чином, лінгвіст визнає, що корпусна лінгвістика – це не просто наука про те, як створювати корпуси і як ними користуватися, а певна ідеологія. Корпусне мовознавство пропонує новий погляд на мову, яка, по суті, і сама є корпусом [Плунгян, 2008: 7–20] [див. його онлайн лекцію **«**Почему современная лингвистика должна быть лингвистикой корпусов?"[2]].

Корпусна лінгвістика як емпіричний мовознавчий напрям суттєво відрізняється від традиційної лінгвістики підходами та методами вивчення мовного матеріалу. Ці відмінності систематизовано російським дослідником В.В. Риковим[3] (див. табл.1):

Таблиця 1

Відмінності між корпусною та традиційною лінгвістикою

№	Корпусна лінгвістика	Традиційна лінгвістика
1.	Основна увага – вивчення мовлення	Основна увага – вивчення мови
2.	Мета – опис мови у тому вигляді, як вона проявила себе в мовленні, представленого у вигляді спеціально відібраного корпусу текстів	Мета – опис та пояснення мови

[2] http://www.polit.ru/lectures/2009/10/23/corpus.html

[3] http://www.rykov-cl.narod.ru/

№	Корпусна лінгвістика	Традиційна лінгвістика
3.	У своїх дослідженнях спирається на дані корпусу тексту	У своїх дослідженнях йде від теорії до її пояснення і підтвердження у фактах мовлення
4.	Надає перевагу квантитативним методам	Надає перевагу кваліативним методам
5.	Вбачає себе частиною традицій, що базується на емпіричних методах	Вбачає себе частиною традицій, що базується на раціоналістичних методах
6.	Текст розглядається як певна фізична сутність	Текст розглядається як певна абстракція
7.	Укладення граматики конкретних мов	Вивчає мовні універсалії
8.	Основна увага приділяється формі	Головна увага – не лише формі, але і змісту
9.	Розглядає тексти у глобальній перспективі	Розглядає тексти у локальній перспективі
10.	Фокусує свою увагу на якомога більшому погляді на текст, не обмеженому ніякими догмами	Аналізує певну конкретну, штучно обмежену, проблемну область
11.	У своїх висновках спирається на спостереження мовленнєвої діяльності, що проявляється у вигляді текстів	Спирається на інтуїцію у відборі мовленнєвого матеріалу, у відборі емпіричних матеріалів своїх досліджень
12.	Часто користується ймовірнісними методами і статистикою для первинної обробки мовленнєвого матеріалу	Надає перевагу логічним роздумам

№	Корпусна лінгвістика	Традиційна лінгвістика
13.	Проводиться робота з лінгвістичними даними (слововживаннями) у тому вигляді, в якому вони зустрілися в контексті	Надає перевагу штучним прикладам з ізольованих від тексту слововживань
14.	Надає перевагу індуктивним методам обробки емпіричного словесного матеріалу, вважає їх суттю наукового методу	Надає перевагу дедуктивним методам обробки емпіричного словесного матеріалу
15.	Вірить у наукові відкриття, базовані на обробці емпіричних даних	Вірить у відкриття, базовані на процедурах, оцінках, порівняннях і т.ін., як результат багатовікових досліджень

Як видно з вищенаведених відмінностей, корпусні студії змінюють пріоритети сучасних лінгвістичних досліджень і демонструють виразну переорієнтацію об'єкта дослідження з «системи» на «узус», з «мови» на «мовлення».

Ще одна відмінність у підходах між традиційною лінгвістикою й корпусною полягає в тому, що традиційне мовознавство вивчало *можливість (possibility)* або неможливість якого-небудь лінгвістичного явища. Наприклад, у традиційному підручнику англійської мови буде сказано, що конструкція *I'm not* можлива в літературному англійському, а от конструкція *I ain't* не існує. Корпусна лінгвістика додатково вивчає й *імовірність (probability)* лінгвістичних явищ. Тобто, з погляду корпусної лінгвістики, ми не можемо сказати, що вживання *I ain't* у літературній мові зовсім неможливо. Воно всього лише малоймовірне.

Оскільки корпусні дослідження проводяться із широким залученням комп'ютерів та спеціально розроблених

комп'ютерних програм, корпусна лінгвістика часто ототожнюється з іншою галуззю прикладної лінгвістики – комп'ютерною лінгвістикою. Хоча ці два напрями мають багато спільного, проте вони не ідентичні. А.М. Баранов розмежовує зазначені дисципліни на основі функції мови (комунікативної, епістемічної (гносеологічної), когнітивної, інформаційної чи соціальної), яку вони оптимізують. Так, комп'ютерна лінгвістика, поряд із такими напрямами прикладної лінгвістики як квантитативна лінгвістика, психолінгвістика й афазіологія, лінгвістична кримінологія, покликана оптимізувати когнітивну функцію мови, що пов'язана із мисленням людини та з пізнанням дійсності. Корпусна лінгвістика оптимізує епістемічну функцію, що пов'язана із збереженням і передачею знань, а також із відображенням специфічної національної самосвідомості [Баранов, 2001: 7–8].

Комп'ютерна лінгвістика, як окрема прикладна лінгвістична дисципліна, характеризується обов'язковістю використання комп'ютерних засобів до оброблення лінгвальних даних, це широка сфера використання комп'ютерних інструментів – програм, комп'ютерних технологій організації та обробки даних – для моделювання функціонування мови в різних умовах, ситуаціях, проблемних галузях, а також сфера застосування комп'ютерних моделей не лише в лінгвістиці, а й у суміжних із нею дисциплінах [Баранов, 2001: 13]. Комп'ютерна лінгвістика займається вирішенням таких проблем, як автоматичний переклад, автоматизоване добування інформації з природних текстів, конструювання зручних інтерфейсів між людиною та машиною, кількісний опис спілкування на природних мовах.

Для корпусної лінгвістики застосування комп'ютерних інструментів не є визначальним критерієм, але, на відміну від

комп'ютерної лінгвістики, корпусна лінгвістика покликана не моделювати функціонування мови в різних умовах, ситуаціях, проблемних галузях та послуговуватися цими моделями, а лише фіксувати всі аспекти функціонування мови, зберігаючи як інтра-, так і екстралінгвістичну специфіку, забезпечуючи оптимальну адекватність лінгвальних даних [Демська, 2005: 14].

Процедура корпусного аналізу включає три кроки: 1) ідентифікація мовних даних за допомогою категоріального аналізу; 2) співвідношення мовних даних за допомогою статистичних методів; 3) інтелектуальна інтерпретація результатів. Якщо перші два кроки повинні бути найбільшою мірою автоматизованими, то останній вимагає людської розумової сутності, адже будь-яка інтерпретація є актом залучення розумових здібностей, а тому не може бути переведена в алгоритмічну процедуру. Саме у цьому проявляється головна відмінність між корпусною і комп'ютерною лінгвістикою, що зводить мову до набору процедур [Teubert, 2007: 113].

Однак в певних аспектах комп'ютерна і корпусна лінгвістики тісно взаємодіють і доповнюють одна одну. Так, комп'ютерна лінгвістика створює необхідні інструменти для корпусної лінгвістики, тобто спеціалізовані комп'ютерні програми для аналізу та обробки великих масивів мовних даних. Це так звані парсери, тегери, конкордансери та інш.

1.3. Типологія досліджень у царині корпусного мовознавства

Сьогоднішня корпусна лінгвістика – це гетерогенна область дослідження мови, всередині якої виокремлюються окремі піднапрями, що різняться підходами до конструкції, експлуатації корпусів та аналізу корпусних даних. В основі виділення цих під напрямів знаходяться такі параметри [McEnery, Hardie 2012: 3-21]:

- ***формат представлення текстів у корпусі (mode of communication)***. Корпуси можуть містити тексти, представлені в усній чи писемній формі. В залежності від форми представлення текстів виділяють корпуси усного мовлення, корпуси писемного мовлення та корпуси змішаного типу. Нові типи корпусів, такі як мультимедійні корпуси та корпуси кінетичного мовлення, реєструють також і паралінгвістичні засоби, які супроводжують спілкування або є безпосереднім способом спілкування. Відмінності у формі представлення текстів у корпусі обумовлюють особливі підходи до відбору та обробки лінгвального матеріалу, використання спеціального програмного забезпечення для аналізу даних та виявляють суттєві лінгвістичні відмінності між отриманими даними в результаті аналізу певного типу корпусу. Наприклад, масштабні контрастивні дослідження усного і писемного мовлення показали, що вони суттєво різняться на граматичному рівні (див. дослідження Д.Байбера та інш. (1999) [Biber, Johansson, Leech, Conrad, Finegan 1999], Р. Картера і М. Маккарті (1995) [Carter, McCarthy 1995], Д. Бразіл (1995) [Brazil 1995];
- ***корпуснобазовані (corpus-based)*** vs. ***корпуснокеровані (corpus-driven) дослідження.*** У корпуснобазованих дослідженнях дані корпусу використовуються для доведення, спростування чи уточнення визнаної на певному етапі розвитку

наукової думки теорії чи гіпотези. Цей тип корпусних розвідок трактує корпусну лінгвістику як метод дослідження мови. Натомість корпуснокерована лінгвістика відмовляється від визнання корпусної лінгвістики як методу і стверджує, що корпус сам по собі є єдиним джерелом гіпотез про мову та втілює свою теорію мови;

- ***режим накопичення даних у корпусі (data collection regimes).*** Відбір даних до корпусу та їх кількість безпосередньо залежать від тих лінгвістичних завдань, заради вирішення яких і створюється корпус. Звідси питання обмеженості / необмеженості обсягу корпусу є одним із важливих питань корпусної лінгвістики. Виділяють два широкі підходи до режиму накопичення даних у корпусі: 1) підхід моніторингового корпусу (Дж. Синклер [Sinclair, 1991]), згідно якого розмір корпусу пропорційно зростає з плином часу; 2) підхід статичного корпусу (Д. Байбер [Biber,1993] та Дж. Ліч [Leech, 2007]), згідно якого ретельно організований корпус має кінцевий розмір і засвідчує стан мови на певному синхронному зрізі;

- ***використання анотованих (annotated) / неанотованих (unannotated) корпусів***. Головною відмінністю сучасного корпусу є наявність анотації, тобто спеціальних міток, що приписуються словам у текстах корпусу та позначають різноманітні лінгвістичні категорії, наприклад, граматичні, синтаксичні і т.інш. Анотація може бути внесена безпосередньо до корпусу, а може супроводжувати корпус окремим документом. Слід зауважити, що саме по собі використання анотованих чи неанотованих («сирих» корпусів) не є настільки значущим, щоб розмежовувати окремі напрями корпусних досліджень. Проте той факт, що деякі лінгвісти взагалі несхвально відносяться до процесу анотування корпусу,

особливо до внесення анотації в корпус вручну, надає вказаному критерію особливої ваги. Аргументи проти корпусної анотації зводяться до двох тверджень. По-перше, здійснена анотація певним чином нав'язує користувачеві корпусу готовий лінгвістичний аналіз даних, здійснений на основі певних наукових позицій укладачів. По-друге, будь-яка анотація не позбавлена об'єктивних неточностей та протиріч. Саме через останнє зауваження багато лінгвістів надають перевагу використанню неанотованих корпусів. Хоча це не означає, що ці дослідники не аналізують лінгвістично дані, якими вони користуються, просто вони не вносять цей аналіз до даних корпусу;

- ***повне врахування (total accountability)*** vs ***відбір даних (data selection)***. Згідно цього протиставлення корпусні дослідження варіюються залежно від способу експлуатації корпусу. Принцип повного врахування даних корпусу полягає в тому, що задля об'єктивності результатів перевірки певної лінгвістичної гіпотези/теорії необхідно використовувати весь корпус та аналізувати всі дані, отримані в результаті аналізу корпусу, а не надавати перевагу чи свідомо відбирати лише ті приклади, що підтверджують власну гіпотезу дослідника, а приклади, що суперечать висунутій гіпотезі, не брати до уваги. Іншими словами, принцип повного врахування в корпусному дослідженні полягає у неприпустимості вмотивованого відбору даних із корпусу з метою уникнення фальсифікації відомостей для підтвердження досліджуваної гіпотези/теорії.

Згідно принципу відбору пошук у корпусі здійснюється з метою підбору специфічного прикладу чи низки ретельно відібраних прикладів для спростування висунутої гіпотези. Такий спосіб експлуатації корпусу є сумісним із принципом повного врахування. Тож використання лише певних частин

корпусу в якості банку прикладів для ілюстрації теорії, що розвивається дослідником (так звані корпусноінформовані дослідження (corpus-informed)), без врахування решти даних корпусу може спотворити реальний стан речей і дозволить досліднику маніпулювати даними корпусу заради підтвердження власної (можливо і хибної) теорії;

- ***багатомовні (multilingual)*** vs ***одномовні (monolingual) корпуси.*** Іншим критерієм, що розрізняє типи корпусів є кількість мов, представлених у корпусі. Більшість корпусів є одномовними, в тому сенсі, що вони репрезентують лінгвістичну варіативність певної однієї мови. Для прикладу, the International Corpus of English – це великий одномовний корпус, що надає можливість здійснювати контрастивні дослідження національних варіантів англійської мови. Натомість багатомовні корпуси – це корпуси, що побудовані на матеріалі двох або більше мов. У вузькому розумінні багатомовним корпусом взагалі вважається корпус, укладений принаймні на трьох мовах, а корпуси, що містять дані двох мов, називаються двомовними (*bilingual*) корпусами. Багатомовні корпуси у свою чергу підрозділяються на порівняльні корпуси (*comparable*) та паралельні (*parallel*) корпуси.

Під порівняльним корпусом розуміється такий корпус, що містить компоненти, які відбираються за таким критерієм: однакові вибірки текстів з однакових жанрів з однакових комунікативних сфер з досліджуваних мов за однаковий часовий період. Паралельний корпус визначається як корпус, що містить тексти з мови-джерела (L1) та їх переклади (L2) (the CRATER, EMILLE, MULTEXT, P-ACTRES). Порівняльні та паралельні корпуси використовуються для проведення перекладацьких та контрастивних досліджень. Для порівняльних корпусів вагомим є суворе дотримання

методології відбору текстів з досліджуваних мов: відібрані тексти повинні повністю співпадати за обсягом, жанром і часом створення. Для текстів-перекладів у паралельному корпусі ці параметри не є важливими, оскільки вони є точними перекладами текстів з мови-джерела. Цінною характеристикою паралельного корпусу є так зване вирівнювання (*alignment*) тексту-джерела та його перекладів, що передбачає анотування відповідностей між ними на рівні речення чи слова.

Таким чином, проаналізувавши типологію корпусних досліджень, ми встановили існуючі відмінності у підходах до конструювання та експлуатації корпусів різними корпусними напрямами. Окресленні критерії будуть детальніше висвітлені у наступних розділах посібника.

1.4. Критика корпусних розвідок

Одним із найпоширеніших критичних зауважень, що звучить на адресу корпусних студій, це обсяг корпусу. Зауважується, що сучасні корпуси є занадто масштабними, що не дозволяє досліднику працювати з повним текстом. Проте, як влучно зауважує А. Партінгтон, нарікати на великий обсяг корпусу – це те ж саме, що і скаржитися на телескоп, у який можна побачити лише розташовані на дуже великій відстані об’єкти, а не ті, що знаходяться прямо перед нами, як через мікроскоп [Partington 1998: 144].

А.Д. Шмельов у статті «Языковые факты и корпусные данные», розміщеній на офіційному сайті Російського національного корпусу[4], виступає з доволі гострою критикою публікації російського лінгвіста Н.В. Перцова «К суждениям о фактах русского языка в свете корпусных данных» [Перцов

[4] http://ruscorpora.ru

2006]. У цій публікації обговорюються загальні тенденції розвитку сучасної лінгвістики та важливу роль корпусних даних у цьому розвитку. У доволі різких судженнях А.Д. Шмельов стверджує, що «пафос Н.В. Перцова по відношенню до корпусів текстів здається дещо перебільшеним» (переклад наш – Ж.В.). На підтвердження цьому автор наводить низку аргументів, які по суті зводяться до наступних. Лінгвіст не заперечує доцільності і вагомості емпірично отриманих на основі великих корпусів текстів даних, але застерігає від їх ідеалізації та беззастережної екстраполяції на всю мову. Науковець зазначає, що для деяких лінгвістичних досліджень використання корпусів текстів є необхідним і доцільним, наприклад, для дослідження давніх мов, індивідуального стилю письменника, розмовного мовлення і. т.ін., при яких мовна компетенція (інтуїція) не є надійним критерієм для об'єктивних висновків. Але при дослідженні живих мов немає необхідності обмежуватися лише аналізом прикладів із корпусу. Конструювання висловлювань та оцінка їх правильності на основі мовної компетенції носіїв мови дозволяє безпосередньо отримати мовний матеріал, який може бути відібраний із корпусу лише опосередковано і з суттєво меншою мірою достовірності. Дослідник також стверджує, що у будь-яких реальних текстах є різного роду огріхи (помилки, описки, відхилення від норми, неправильності і тому подібне), які присутні і в корпусах текстів. Звідси, простий факт наявності таких лінгвістичних відхилень у корпусі ще не є гарантією їх правильності та підставою для ствердження, що ці одиниці чи конструкції існують в мові. Для підтвердження правильності чи доцільності використання таких одиниць мови лінгвіст повинен звернутися до мовної компетенції носіїв мови. Крім того, без звернення до мовної компетенції носіїв мови не можливо отримати так званий «негативний мовний матеріал», тобто дані

про те, які мовні одиниці і конструкції є неприйнятними для даної мови. Вирішальне слово про визнання прийнятності чи неприйнятності певної мовної одиниці залишається за носієм (носіями) мови. Наявність чи відсутність того чи іншого явища у корпусі ще нічого не означає, якщо носії мови відмовляються визнавати це явище правильним. Вагомим аргументом щодо потенційних недоліків корпусних досліджень А.Д. Шмельов вважає також і той факт, що використання електронного корпусу сприяє швидкому отриманню даних, але не допомагає у встановленні того, які саме дані для дослідження необхідні, та не надає можливостей аналізу отриманих результатів. Також відсутність ретельної роботи зі збору мовного матеріалу може привести до поспішних висновків.

Таким чином, погоджуючись із думкою А.Д. Шмельова, зазначимо, що використання корпусу може виявитися значно продуктивнішим, якщо лінгвіст критично оцінює дані, отримані з корпусу. Звернення до власної мовної компетенції допоможе досліднику виявити лакуни та неточності в існуючих описах мови. Звідси, орієнтація на дані корпусу та власну мовну інтуїцію не суперечать один одному, а швидше доповнюють один одного.

Визнання того факту, що корпусна методика дослідження лінгвальних даних має як позитивні, так і негативні сторони є важливим, адже лінгвіст повинен чітко усвідомлювати, де її доцільно використовувати, а де застосовувати інші методи.

Запитання і завдання для самоперевірки

1. Поясніть сутність корпусної лінгвістики.
2. Яким чином застосування лінгвістичних корпусів змінило спосіб дослідження мови?

3. Сформулюйте предмет і завдання корпусної лінгвістики як лінгвістичної науки.
4. Охарактеризуйте зв'язки корпусної лінгвістики з іншими мовознавчими науками.
5. Що є теоретичним підґрунтям корпусної лінгвістики?
6. Наведіть аргументи на користь визнання корпусної лінгвістики особливою методологією дослідження лінгвального матеріалу/ самостійною мовознавчою наукою.
7. У чому полягають основні відмінності між корпусною лінгвістикою та традиційним мовознавством; корпусною лінгвістикою і комп'ютерною лінгвістикою?
8. Перелічіть параметри, що лежать в основі виділення піднапрямів у корпусних дослідженнях.
9. Чи погоджуєтеся Ви з наведеними критичними зауваженнями щодо переваг корпусних досліджень? Поясніть вашу думку.

Розділ 2.
Корпусні студії: історична перспектива та сучасний стан

2.1. Історія становлення корпусної лінгвістики: від паперових конкордансів і картотек до перших електронних корпусів

Традиційно вважається, що корпусна лінгвістика почала активно формуватися як самостійна галузь у 60-х роках XX століття, коли з'явилися перші корпуси у комп'ютерному форматі. Наприклад, дослідниця Е. Тогніні-Бонеллі історію корпусних студій пов'язує з безпосереднім розвитком комп'ютерних технологій і виділяє такі етапи:

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

Етап 2 (1980–2000 рр.) поділяється на два періоди по десять років:

а) 1980-ті роки відзначилися появою сканерів, коли навіть із примітивним сканером укладалися корпуси у 20 млн. слововживань;

б) 1990-ті роки ознаменовані розширенням можливостей комп'ютерного набору, що полегшило доступ до великих за обсягом текстових матеріалів в електронному форматі і сприяло значному збільшенню розмірів корпусів.

Етап 3 (з початку 2000-го року і по сьогоднішній день) – це період електронних (віртуальних) текстів, які ніколи не мали матеріальної форми, що надає величезні можливості для створення корпусів будь-якого необмеженого розміру [Tognini-Bonelli, 2010: 16-17].

У. МакЕнері та А. Вільсон значно віддаляють у часі дату зародження корпусної лінгвістики. Вчені виокремлюють два періоди в корпусних розвідках. Перший період – це стадія ранньої корпусної лінгвістики (1910–1960-ті рр.), коли відбувається формування теоретичного підґрунтя та прагматичних передумов виникнення напряму й створення текстових зібрань для лінгвістичного дослідження переважно на паперових носіях. Другий період (починається з 1960 рр.) характеризується інтенсивним піднесенням корпусних студій і безпосередньо пов'язаний із значним розвитком комп'ютерних технологій.

Масово комп'ютеризовані корпуси стали доступними лише з середини 90-х років XX століття, коли лінгвістика раптом перейшла від стану «голоду адекватних даних» [Sinclair, 1991: l] до стану, заповненого даними. Саме у цей період корпусна лінгвістика остаточно сформувалась як окремий розділ науки про мову.

> *While it is natural today to take 'corpus linguistics' to mean 'electronic corpus linguistics', we must not forget that there were language corpora BC, i.e.'before computers'. [Sinclair 2007: 12].*

Проте, якщо корпусна лінгвістика – це вивчення мови, що базується на прикладах реального вживання мови [McEnery, Wilson 2001 :1], то необхідно констатувати, що масштабні емпіричні дослідження мови з використанням великого обсягу реальних мовних даних здійснювалися і в більш ранні періоди розвитку мовознавчої науки. Наука про мову поступово наближалася до ідеї текстового корпусу і до самого корпусу в такому вигляді, в якому ми зараз його знаємо. Як дотепно

висловився Я. Свартвік [Svartvik, 2007: 12], в історії корпусів спочатку був так званий «кам'яний вік», або докомп'ютерний період, коли корпуси укладалися вручну на папері. Ці перші паперові доелектронні корпуси по суті були конкордансами, тобто алфавітними списками всіх ужитих у певному тексті/текстах слів у їх контекстному оточенні. Укладання таких паперових корпусів-конкордансів займало багато часу та вимагало напруженого аналізу, який здійснювався вручну. Найвагомішу роль доелектронні корпуси відіграли в таких лінгвістичних проектах, як укладання конкордансів Біблії й літературних творів, а також написання граматик і словників [Meyer, 2008: 1].

Конкорданси Біблії були першими серйозними корпусними дослідженнями лінгвістичного матеріалу. Дослідники Біблії та їхні помічники уважно проглядали сторінку за сторінкою, рядок за рядком і вручну індексували всі слова. Етимологія латинського слова *concordantia – згода, відповідність* (*con* – від прийменника *cum* означає „з"; *cor* – „сердце") пов'язана з початковим ідеологічним підтекстом цієї трудомісткої діяльності, а саме: підтвердити, що Біблія – це гармонійне божественне послання, а не збірка текстів із розрізнених джерел [McCarthy, O'Keeffe, 2010: 3].

Ім'я францисканця Антонія Падуанського (1195–1231 рр.) пов'язується з першим відомим (анонімним) конкордансом Біблії (*the Concordantiae Morales*), укладеного на основі Вульгати (латинського перекладу Біблії 5 ст.). Відомою працею того ж періоду є конкорданс кардинала Хьюго де С. Каро, який у 1230 році за допомогою 500 монахів-домініканців у монастирі Сент-Джеймс у Парижі уклав покажчик слів Вульгати. З того часу було укладено ще низку конкордансів Біблії: *a Hebrew Concordance*, укладений Ісаком Натаном бен-Калонімусом у XV

столітті, а також конкорданс Александра Крудена (A *Complete Concordance to the Holy Scriptures*) (XVIII століття) та конкорданс Іакова Стронга (*Exhaustive Concordance of the Bible*), опублікований у 1890 році.

З-поміж цих конкордансів за обсягом та доступністю вирізняється конкорданс А. Крудена (1737 р.), що містив 2 370 000 слів, що за кількістю слів більше, ніж сама Біблія. Як не дивно, його укладання зайняло порівняно невеликий проміжок часу. Якщо для укладання конкордансу кардинала Хьюго необхідна була допомога 500 монахів, то А. Круден витратив на укладання конкордансу лише два роки, працюючи по 18 годин на день [Fraser 1996].

Дещо згодом услід за конкордансами Біблії почали з'являтися конкорданси літературних творів. Наприклад, конкорданс праць У. Шекспіра Ендрю Бекета (*A Concordance of Shakespeare*) (1787 р.), конкорданс праць Дж. Чосера, що був укладений у 1871 році, проте, зважаючи на трудомісткість, опублікований лише у 1927 році.

Крім конкордансів значні за обсягом вибірки текстів використовувалися ще для укладання ранніх граматик. Як зазначає Ч. Мейер [Meyer, 2008: 3], навіть перші ранні відомі граматики класичних мов також можна вважати написаними на основі корпусів тестів, наприклад, у IV столітті до н.е. граматика Паніні описувала мову Ведів (санскрит), яка збереглася лише в корпусі ведичних текстів. Аристон Александрійський, грецький учений I століття, у своїй праці «*Неграматичні слова*» описав нерегулярні граматичні конструкції в корпусі текстів Гомера.

Ранні граматики англійської мови також базувалися на класичній традиції використання в якості прикладів цитати з реальних текстів, особливо з художніх творів, як, наприклад,

граматика Роберт Лоуса «*A Short Introduction to English Grammar*» (1762 р.). Одна з найвідоміших граматик цього періоду – семитомна праця Отто Єсперсена (1909–1949 рр.) «*A Modern English Grammar on Historical Principles*» також побудована винятково на прикладах, відібраних з величезної кількості текстів писемного варіанта англійської мови. Отто Єсперсен належав до тієї плеяди лінгвістів, які були переконані, що лінгвістичний опис повинен базуватися не на вигаданих, а на реальних прикладах із текстів реальної мови. Тенденція до цитування літературних творів як прикладів до граматичних правил продовжилася в граматиках кінця XIX – середини XX століття такими авторами, як Дж. Керм, Н. Поутсма, Ч. Фріз.

> *It is impossible for me to put even a remotely accurate number on the quantity of slips I have had or still have: a lot of them have been printed in my books, particularly the four volumes of Modern EnglishGrammar, but at least just as many were scrapped when the books were being drafted, and I still have a considerable number of drawers filled with unused material. I think a total of 3-400,000 will hardly be an exaggeration [Jespersen 1938: 213-215; translation by D. Stoner].*

Проте не всі граматисти притримувалися цієї традиції. Наприклад, у відомій граматиці Генрі Світа «A New English Grammar» (1892 р.) для ілюстрації аналізованих граматичних категорій використовуються вигадані дослідником приклади.

Дослідження великих масивів текстів також здійснювалися з метою укладання доступних словників. Починаючи зі словника Самуеля Джонсона (1755 р.), лексикографи використовували цитати з текстів відомих письменників, щоб проілюструвати значення слів та особливості їх уживання.

Лексикограф зібрав 150 000 ілюстративних цитат для 40 000 заголовних слів словника. Основою Oxford English Dictionary (OED), що створювався під керівництвом Джеймса Муррея (1880 р.), став 5 млн. корпус цитат на картках [McCarthy, O'Keeffe, 2010: 4; Meyer, 2008: 8].

Багато словників мертвих мов використовували цитати із текстів, що представляли слово в контексті, в якості ілюстративних прикладів. У сучасній корпусній лінгвістиці цей метод подібний за формою до конкордансу KWIC (Key Word In Context). Незважаючи на те, що сучасні комп'ютери суттєво спростили пошук та класифікацію прикладів вживання конкретного слова в текстах різних стилів і жанрів, ідеї використання текстів із корпусу все ще дуже схожі з тими, що використовувалися ранніми лексикографами і філологами, що не мали доступу до комп'ютерних технологій.

Інші масштабні лінгвістичні дослідження, що відбувалися на значних за обсягом масивах текстів у період з 1890-х до 1950-х, здійснювалися лише за фінансової підтримки та залучення великої кількості кваліфікованих помічників. Наприклад, з метою вдосконалення методів скоропису в суді Й. Кадінг (1898 р.) залучив значну кількість помічників Прусської державної служби, щоб проаналізувати частоту вживання слів в 11 млн. корпусі німецької мови. З навчальною метою у період 1920–1940 рр. Торндайк та Лодж підрахували частоту слів у корпусі англійської мови обсягом до 18 млн. слів для укладення частотного словника слів. Цей словник використовувався для навчальних цілей. Впливовий (важливий) General Service List, укладений Вестом (1953), також побудований за частотним принципом.

Like most facts of interest and importance … information about the speaker-hearer's competence … is neither presented for direct observation nor extractable from data by inductive procedures of any known sort

Corpus Linguistics does not exist [N. Chomsky цит. за Tognini-Bonelli, 2010: 14]

Найважливішим та найвпливовішим доелектронним корпусом вважається The Survey of English Usage, укладений Рендольфом Квірком у 1959 р. в University College London. Корпус представляв собою велику базу даних на картонних картках, які містили зразки щоденного спілкування (як писемного, так і усного) звичайних громадян. Цей проект став перехідним етапом у становленні корпусної лінгвістики. Спочатку збереження даних в електронному форматі не передбачалося і лише через багато років по тому відбулося перенесення даних корпусу в електронну форму. Сьогодні базу даних корпусу можна переглянути на офіційному сайті проекту[5].

2.2. Корпусна лінгвістика з 60-х років XX ст. до пост 2000-х

Переважна кількість досліджень у царині сучасної корпусної лінгвістики розпочиналася на матеріалі англійської мови. Як зазначає С.А. Шаров [Шаров, 2003: 14], причиною цього був не стільки активний розвиток комп'ютерних технологій у США, скільки інтелектуальний клімат у британській лінгвістиці в 60–80-х роках XX століття. У США

[5] http://www.ucl.ac.uk/english–usage

період формування засад корпусної лінгвістики (60-ті роки XX ст.) був ознаменований розквітом трансформаційної генеративної граматики (ТГГ). Корпусні студії були несхвально сприйняті в тогочасній науковій спільноті та зазнали суттєвої критики від засновника генеративізму Н. Хомського. Дослідник назвав корпусний спосіб накопичення мовних даних неадекватним і хибним для опису породжувальної здатності природної мови, оскільки інтуїція мовця є достатнім джерелом мовного матеріалу [McEnery, Wilson, 2001: 6; Селіванова, 2008: 668; McCarthy, O'Keeffe, 2010: 14].

Інші представники ТГГ також без зацікавлення поставилися до можливості вивчати дані про реальне функціонування мови, і, як наслідок, провідні лінгвісти-теоретики того часу практично не підтримували й не проявляли зацікавленості до нових корпусів. У 1960-80-х роках корпусну лінгвістику взагалі було витіснено на периферію лінгвістичної дослідної парадигми США.

На відміну від США, у британській лінгвістиці були сильними емпіричні тенденції, які передбачали використання реальних прикладів для перевірки лінгвістичних гіпотез (у першу чергу дослідження Джона Фьорса та його учнів – Грегорі, Синклера, Хелідея та інш.) [Шаров, 2003: 15]. Це й привело до створення багатьох корпусів та розробки корпусних досліджень на матеріалі англійської мови.

Проте, незважаючи на непопулярність корпусних досліджень у США, все-таки перший комп'ютеризований корпус, що став стандартом для багатьох інших корпусів, був укладений саме у цій країні.

Корпуси першого покоління. Ідея створення корпусу (уже в сучасному його розумінні) зародилася в 60-х роках XX століття під значним впливом здійснених масштабних емпіричних

досліджень, про які ми вже зазначали. До кінця 1960-х існувало декілька невеликих корпусів, укладених на різних принципах. Саме значні досягнення в сфері комп'ютерних технологій, а не в лінгвістиці дали поштовх до створення перших електронних корпусів. У середині XX століття комп'ютери лише входили у загальний вжиток. Це були перші примітивні машини, з якими було важко працювати, але їхній величезний потенціал було відразу розпізнано і залучено до лінгвістичних досліджень. Комп'ютеризація текстів розпочалася з Father Busa's Index Thomisticus ще до 1950 року (завершено у 1978 р.), продовжуючи наукову традицію укладання конкордансів до праць значимих у культурному аспекті. Перші лінгвістичні корпуси текстів на машинних носіях з'явилися в 60-х роках XX сторіччя. За сучасними стандартами вони були дуже малими, але вирізнялися ретельно продуманою організацією.

На початку 60-х років XX ст. розпочалися два проекти по обидва боки Атлантики по створенню корпусів в електронному форматі. Ці проекти здійснювались одночасно і незалежно один від одного В університеті Единбурга в Шотландії формувався корпус усного мовлення, що включав транскрибовані версії повсякденних розмов носіїв британського варіанту англійської мови. Цей корпус досяг розміру лише у 300 000 слів. Причиною цьому стали трудомісткий процес збору та транскрибування усного мовлення та відсутність комп'ютера в університеті.

У цей же час у Браунівському університеті (США) Нельсон Френсіс та Генрі Кучера розпочали укладання одномільйонного корпусу, який було названо за місцем його створення Браунівським корпусом (the Brown Corpus)[6]. Створення корпусу мало на меті дослідження лінгвістичних особливостей американського варіанту англійської мови. Він містив 500

[6] http://www.helsinki.fi/varieng/CoRD/corpora/BROWN/index.html

текстових уривків обсягом по 2 000 слововживань загальним обсягом біля 1 млн. слів. Тексти були відібрані із п'ятнадцяти найбільш масових жанрів англомовної друкованої прози США, що були опубліковані в 1961 році: газетні статті (репортажі, передові статті, огляди), релігійна література, професійна література, науково-популярна література, белетристика, зразки ділової прози (у тому числі урядові документи), наукова література, художня проза, детективи та наукова фантастика, пригодницька література і вестерни, романтична література, гумористичні розповіді та оповідання (див. табл. 2).

Таблиця 2

Категоризація текстів у Браунівському корпусі

Code	Text category	No. of samples	Proportion (%)
A	Press reportage	44	8.8
B	Press editorials	27	5.4
C	Press reviews	17	3.4
D	Religion	17	3.4
E	Skills, trades and hobbies	38	7.6
F	Popular lore	44	8.8
G	Biographies and essays	77	15.4
H	Miscellaneous (reports, official documents)	30	6.0
J	Science (academic prose)	80	16.0
K	General fiction	29	5.8
L	Mystery and detective fiction	24	4.8
M	Science fiction	6	1.2
N	Western and adventure fiction	29	5.8
P	Romantic fiction	29	5.8
R	Humour	9	1.8
Total		**500**	**100**

Корпус супроводжувався значною кількістю матеріалів його первинної статистичної обробки — частотний і алфавітно-частотний словник, різноманітні статистичні розподіли. У. Френсіс та Г. Кучера ставили собі мету представити корпус текстів, що відповідав ясним і чітким критеріям відбору. Зокрема, укладачами враховувалися такі характеристики, як:

> *In 1962, when I (*W. Nelson Francis*) was in the early stages of collecting the Brown Standard Corpus of American English, I met Professor Robert Lees at a linguistic conference. In response to his query about my current interests, I said that I had a grant from the U.S. Office of Education to compile a million-word corpus of present-day American English for computer use. He looked at me in amazement and asked, 'Why in the world are you doing that?' I said something about finding out the true facts about English grammar. I have never forgotten his reply: 'That is a complete waste of your time and the government's money. You are a native speaker of English; in ten minutes you can produce more illustrations of any point in English grammar than you will find in many millions of words of random text.'[Francis 1982: 7-8].*

1. *походження і склад тексту* (автор повинен був бути уродженим носієм американського варіанту англійської мови, діалогічне мовлення повинно було займати менше половини всього обсягу тексту);

2. *часова віднесеність* (всі відібрані до корпусу тексти були вперше опубліковані у 1961 році);

3. *збалансоване представлення різних жанрів*;

4. *доступність для комп'ютерної обробки* (спеціальні помітки для передачі графічних особливостей тексту і т. п.).

Перша версія корпусу була представлена простим текстовим форматом (з невеликою кількістю структурної розмітки для виділення абзаців, заголовків, цитованих фрагментів і т. ін.). Пізніше у 1980 році корпус був доповнений розміткою частин мови та морфологічних ознак слів. Поява Браунівського корпуса викликала загальний інтерес у науковій спільноті та стимулювала наукову дискусію, в якій обговорювалися критерії відбору текстів та потенційні можливості корпусу для наукового аналізу. Корпус швидко перетворився у популярний об'єкт лінгвістичних досліджень. Поступово в процесі його використання вчені дійшли до розуміння того, що провести певні порівняння і виявити конкретні закономірності можливо лише шляхом аналізу значних за розміром масивів текстів, які організовані за визначеними правилами. Так почали проводитися нові дослідження мови вже на більш високому і надійному рівні в межах нового напряму в лінгвістиці, яким стала корпусна лінгвістика.

Браунівський корпус перетворився у певний стандарт для створення інших аналогічних корпусів. Услід за ним з'явився його британський аналог – Ланкастерсько-Осло-Бергенський корпус (the Lancaster-Oslo-Bergen Corpus[7]), що був названий за місцем його створення. Його укладачі притримувались основних принципів укладання Браунівського корпусу – така ж сама середня довжина текстів, такі ж самі способи його представлення на машинному носії (15 жанрів вже британської масової друкованої продукції – 1961 року видання). Анотована версія корпусу з'явилася у 1985 році.

[7] http://www.helsinki.fi/varieng/CoRD/corpora/LOB/index.html

Комп'ютерна обробка 60 000 слів у 1950 роках займала більше 24 год. Сьогодні – мікросекунди.

Створення Браунівського та Ланкастерського корпусів дало можливість проводити різноаспектні філологічні порівняння двох варіантів англійської мови (американського і британського) на текстах різних жанрів, доступних комп'ютерній обробці.

У 1970-і за моделлю близькою до Браунівського корпусу був створений частотний словник російської мови, укладений на основі корпусу текстів обсягом також в 1 млн. слів, що включав приблизно в однаковому співвідношенні суспільно-політичні, художні, наукові та науково-популярні тексти і драматургію. Корпус містив внесену вручну лематизацію та частиномовну розмітку.

За форматом Браунівського та Ланкастерсько-Осло-Бергенського корпусів з деякими модифікаціями було укладено низку інших корпусів, серед яких the Kolhapur Corpus of Indian English, the Wellington Corpus of Written New Zealand English, the Australian Corpus of English, the Corpus of English-Canadian Writing, the Standard Corpus of Present-day English Language Usage, the London-Lund Corpus (LLC) (детальніше див.: http://www.uow.edu.au/~dlee/CBLLinks.htm).

70-ті роки XX століття були періодом уповільнення темпів корпусних досліджень. Крім несприйняття ідей корпусної лінгвістики представниками панівного на той час лінгвістичного напряму: генеративної граматики, і як наслідок непопулярність корпусних студій серед більшості лінгвістів, не менш вагомою причиною такого стану були обмежені можливості комп'ютерної техніки. Комп'ютери все ще були величезними

калькуляторами з малим об'ємом пам'яті, а програмне забезпечення не відповідало нагальним потребам дослідників. Незважаючи на ці об'єктивні причини, у корпусному мовознавстві все-таки були досягнуті певні результати. Насамперед, це створення корпусів, що перевищили поріг в 1 млн. слововживань. Корпусним дослідникам стало очевидним, що корпус в один мільйон слів є недостатнім для адекватного відображення лексичних і граматичних властивостей мови. Адже, як відомо, зі статистичної точки зору мова є великим набором нечастих подій, тобто слова з низькою частотою вживання є більш чисельними, ніж слова з високою частотою. Для отримання надійних даних про особливості вживання мовної одиниці необхідно проаналізувати принаймні сотню прикладів, звідси дані про функціонування низькочастотних одиниць на невеликих за розміром корпусах можуть бути не достовірними. З цієї причини, а також у зв'язку із ростом технічних потужностей комп'ютерів, які стали здатними обробляти великі обсяги текстів, у 80-ті роки XX століття у світі було здійснено декілька спроб створити корпуси значно більших розмірів.

Корпуси другого покоління – це продукти Інтернету і характеризуються значним обсягом. Так, у кінці 80-х років у Великобританії був створений перший мега-корпус, що задав новий стандарт для представницьких корпусів – Британський національний корпус (British National Corpus[8]). Цей корпус характеризується обсягом 100 млн. слів, використанням повних текстів, а не вибірок з текстів, підкорпусом усного мовлення (10 млн. слів), наявністю частиномовної розмітки та доступом через Інтернет. Для корпусу використовувалася детальна класифікація документів за декількома параметрами: вид мовлення (писемне,

[8] http://www.natcorp.ox.ac.uk/

усне приватне і усне публічне), для писемного за тематикою, типом видання (книги, періодика, машинописні тексти і т.п.), параметром утворення очікуваної аудиторії (високий, середній чи довільний) та складністю мови (складний, середній, простий).

Укладачі BNC для порівняння спробували представити корпус у вигляді звичайної книжкової продукції і одержали вражаючі показники. Якщо видруковувати корпус на тонкому папері з розрахунку 400 слів на сторінку, то весь його обсяг у друкованому вигляді займатиме простір близько 10 м^2. Для того, щоб прочитати цю продукцію зі швидкістю 150 слів на хвилину, витрачаючи на це 8 годин щодня, знадобилося б 4 роки [Карпіловська 2006: 76]

За заданим Британським національним корпусом стандартом були укладені представницькі корпуси багатьох європейських мов. За цією моделлю були створені національні корпуси іспанської, італійської, хорватської, чеської мов.

Подібний проект Банк англійської мови (the Bank of English) розпочався у 1980-і рр. У 1989 році його обсяг був 20 млн. слів, а у 2012 – 650 млн. слів. Банк англійської мови – це так званий моніторинговий корпус, що покликаний відслідковувати мовні зміни шляхом регулярного поповнення новими текстами та порівняння частотних параметрів, наприклад, таких, як зміна частоти слів та граматичних конструкцій, поява нових слів і т.ін. Він охоплює англійське писемне та усне мовлення, а також різні територіальні варіанти англійської мови. Цей корпус є

продовженням COBUILD (Collins Birmingham University International Language Database) корпусу Бермінгемського університету, що був розпочатий раніше – у 1980 р. Цей корпус став основою для словника Collins COBUILD English Dictionary та низки англійських граматик.

Банк англійської мови та Британський національний корпус мали потенційну підтримку від видавців, що використовували корпуси для укладання словників і граматик. Такими ж корпусами є Кембриджський[9] та Лонгманівський[10] корпуси, що є закритими для вільного доступу і використовуються лише авторами та укладачами навчальних матеріалів видавництв.

Ще один вид корпусів представлений Інтернаціональним корпусом англійської мови (the International Corpus of English[11]). Цей проект розпочався у 1990 році з метою здійснення порівняльного дослідження існуючих варіантів англійської мови. Двадцять центрів у всіх куточках світу працюють над створенням підкорпусів власного національного чи територіального варіанту англійської мови, дотримуючись єдиного принципу побудови та граматичної анотації.

Успіх Британського національного корпусу у дослідженні британського варіанту англійської мови дав поштовх для створення американського відповідника. У 1999 році було засновано консорціум Американського національного корпусу (the American National Corpus)[12]. У 2003 році була випущена перша частина (близько 10 млн. слів) корпусу, а за декілька

[9]http://www.cambridge.org/gb/elt/catalogue/subject/item2701617/Cambridge-International-Corpus/?site_locale=en_GB

[10]http://www.cambridge.org/gb/elt/catalogue/subject/item2701617/Cambridge-International-Corpus/?site_locale=en_GB

[11] http://ice-corpora.net/ice/

[12] http://americannationalcorpus.org/

років по тому була випущена і друга частина (близько 20 млн. слів). Особливою прикметою Американського національного корпусу є те, що він містить жанри, які ще не були поширеними за часу створення Британського національного корпусу. Це особисті електронні тести, Інтернет чати та веб сторінки.

У колишньому Радянському Союзі таким мега-проектом став Машинний Фонд російської мови, створений під керівництвом А.П. Єршова та В.М. Андрющенко. Головними цілями цього фонду були укладання представницького корпусу і підкорпусів різних жанрів та розробка комп'ютерних програм опрацювання природної мови, з метою сприяння комплексній інформатизації лінгвістичних досліджень, включаючи створення граматик та словників. На жаль, цей проект не завершився створенням власне представницького корпусу, хоча були зібрані колекції текстів різних типів. Сьогодні з Машинного Фонду доступні деякі тексти XIX століття та газетний корпус за 1997 рік.

У 1992 році була створена організація Європейська корпусна ініціатива (ECI), метою якої були об'єднання і координація зусиль лінгвістів різних країн, що працюють над створенням корпусів текстів на інших, крім англійської, мовах. Під її егідою було створено біля 50 корпусів текстів (кожен обсягом від 12 тисяч до 5 млн. слів) на європейських мовах. Це об'єднання поставило собі за мету не лише укладання представлених корпусів текстів на європейських мовах, але і створення так званих «паралельних» корпусів.

Сучасний розвиток корпусної лінгвістики (пост 2000-і роки) дуже бурхливий, що підтверджується величезною кількістю нових досліджень у галузі. Про популярність та інтенсивність розвитку корпусної лінгвістики свідчить як чимале число теоретичних і технологічних праць, підручників, так і самих

корпусів. Йдеться про дослідження у галузі лексичної граматики [Stubbs 1996; Hunston, Francis 2000; Renouf 2001; Nesselhauf 2005; Exploring the Lexis-Grammar Interface 2009], лексикографії та навчання мові [McEnery, Kifle 2002, Altenberg, Granger 2002; McEnery, Xiao 2004, Максимів 2008], когнітивної лінгвістики [Corpora in Cognitive Linguistics 2006; Gilquin 2003; Gries 2003; Gries, Stefanowitch 2004; Schmidt 2000; Schonefeld 1999], прагматики та дискурс-аналізу [Aijmer and Stentström 2004; Archer 2005; Baker 2005; Baker, McEnery 2005; Hardt-Mautner 1995; Koller, Mautner 2004; McEnery 2005; Orpin 2005; Partington et al. 2004; Vivanco 2005; Wang 2005], стилістики [Burrows 2002; Charteris-Black 2004; Corpus-Based Approaches to Metaphor and Metonymy 2006; Deignan 2005; Semino and Short 2004; Stubbs 2005], перекладознавства [Malmkjær 1998; Zanettin 1998; Incorporating Corpora. The Linguist and the Translator 2008]. Корпусно-базовані дослідження відбуваються для вивчення значення слова [Partington 2004], фразеології [Hunston 2001, Лозинська 2009], синтаксичних властивостей граматичних структур [Duffley 2003], дистрибуції граматичних категорій [Biber 2001] (детальніше див.: Corpus Linguistics and Written Language Resources Bibliography[13]).

У цей час корпуси створені для багатьох мов світу (див. веб сайт Дейвіда Лі[14], на якому розміщено дані про існуючі корпуси та ресурси корпусної лінгвістики).

Найновіші досягнення в царині корпусного мовознавства друкуються у визнаних міжнародних наукових журналах: *Corpus* (2001–) (Nice: Laboratoire "Bases, Corpus, Langage", Université de Nice – Sophia Antipolis)[15]; *Corpus Linguistics and*

[13] liceu.uab.es/~joaquim/language_resources/lang_res/biblio_corpus.html

[14] http://www.uow.edu.au/~dlee/CBLLinks.htm

[15] http://corpus.revues.org/

Linguistic Theory (2005–) (Berlin – New York: Mouton De Gruyter)[16]; *ICAME Journal*, Journal of the International Computer Archive of Modern English (1987–) (Bergen: Norwegian Computer Centre for the Humanities)[17]; *International Journal of Corpus Linguistics* (1996–) (Amsterdam: John Benjamins)[18]; *Language Resources and Evaluation* (2005–) (Dordrecht: Springer)[19]; *Literary and Linguistic Computing* (1986–) (Oxford: Oxford University Press)[20].

Таким чином, підсумовуючи огляд історії становлення корпусних студій, зазначимо, що з часу свого становлення до сьогодні корпусна лінгвістика пройшла складний шлях від повного неприйняття у 60-ті роки XX століття до магістрального напряму сучасного мовознавства. На сьогодні в науковій спільноті також немає однозначного ставлення до таких досліджень: у них є як віддані прибічники, так і суворі опоненти. Проте необхідно пам'ятати, що незважаючи на революційні можливості корпусної лінгвістики, вона є лише частиною численного методологічного інструментарію сучасної лінгвістичної науки. І як зазначив Ч. Філмор [Fillmore 1992: 35], навіть значні за обсягом корпуси не в змозі відобразити все можливе у мові, натомість і невеликі за обсягом корпуси можуть надати інформацію, яку б нереально було отримати, не звертаючись до корпусних даних.

[16] http://www.degruyter.com/view/j/cllt

[17] http://nora.hd.uib.no/journal.html

[18] http://www.benjamins.com/#catalog/journals/ijcl

[19] http://www.springerlink.com/content/1574-020x/

[20] http://llc.oxfordjournals.org

2.3. Корпусні дослідження в Україні

На сьогоднішній день практично для усіх європейських мов існує по кілька корпусів текстів різного типу, обсягу та структури. Створення національного корпусу вважається виявом поваги до рідної мови. Хоча для українського мовознавства укладання національного корпусу і впровадження корпусного мовознавства ще перебуває на етапі становлення, у царині вітчизняної корпусної лінгвістики вже здійснено вагомі доробки. На разі у складі Національної словникової бази Українського мовно-інформаційного фонду НАН України функціонує і постійно розвивається Український національний лінгвістичний корпус (УНЛК), що розробляється під керівництвом академіка НАН України В.А. Широкова [Корпусна лінгвістика 2005: 103]. Розпорядженням Кабінету Міністрів України від 11.02.2004 р. № 73-р Національну словникову базу Українського мовно-інформаційного фонду НАН України внесено до державного реєстру наукових об'єктів, що становлять національне надбання. Створення УНЛК передбачає розв'язання низки вагомих завдань для вітчизняного мовознавства:

1) створення гами українських словників нового покоління, виконаних як у паперовому, так і в електронному варіанті, включаючи й онлайновий (це орфографічні, орфоепічні, тлумачні, етимологічні, фразеологічні, граматичні, синонімів, антонімів, омонімів та багато інших). До речі, вже зараз у серії «Словники України», яка становить основу Національної словникової бази України, налічується близько 50 видань, серед яких слід відзначити двотомні Орфоепічний словник української мови та Словник синонімів української мови, п'ять видань Орфографічного словника української мови,

а також декілька версій першого повномасштабного українського електронного словника на диску – Інтегрованої лексикографічної системи «Словники України», онлайновий варіант якого розміщено на сайті Українського лінгвістичного порталу[21];

2) укладання низки перекладних двомовних словникових систем, орієнтованих переважно на інтелектуальне комп'ютерне опрацювання мови, насамперед, на машинний переклад **з** та **на** мови: англійську, німецьку, французьку, іспанську, турецьку, всі слов'янські мови та ряд інших;

3) розробка та створення комп'ютерних, морфологічних, синтаксичних і семантичних аналізаторів, систем автоматизованого редагування та інтелектуального пошуку в масивах природномовної інформації, а також ряд інших [Корпусна лінгвістика 2005: 121].

Обсяг корпусу понад 100 млн. слововживань. Оскільки об'єктом дослідження УНЛК є сучасна українська літературна мова, то стратегія формування корпусу полягає у максимально повному охопленні літературних українських джерел із початку XIX до початку XXI століття, тобто за двісті років існування та розвитку сучасної української літературної мови. Корпус представлений текстами різних стилів і жанрів без збереження пропорцій. За необхідності дослідник може самостійно генерувати підкорпуси окремих стилів із врахуванням статистичних параметрів.

Наразі до УНЛК ще немає вільного доступу у мережі Інтернет, але детальніше про корпус можна дізнатися на сайті Українського мовно-інформаційного фонду НАН України[22]. Крім того, доступ до корпусу надається в межах наукової

[21] www.ulif.org.ua

[22] http://lcorp.ulif.org.ua/LSlist/

співпраці Фонду з освітніми установами. Національний центр «Мала академія наук» працює над пошуком нових форм та засобів оптимізації навчально-дослідницької та науково-дослідницької діяльності дітей і молоді, з-поміж яких апаратно-програмні засоби та мережеві інструменти відіграють важливу роль. Наприклад, продуктивна співпраця національного центру «Мала академія наук» з колективом Українського мовно-інформаційного фонду (НАН України) надає школі унікальну можливість використовувати ресурси УНЛК в організації навчально-дослідницької діяльності учнів. У 2012 році з ініціативи дирекції національного центру «Мала академія наук України», Українського мовно-інформаційного фонду (НАНУ) та ректорату Житомирського державного університету імені Івана Франка було створено Експериментально-тематичний центр навчальної та науково-дослідницької діяльності. Директором Центру (ЕТЦ) стала доктор педагогічних наук, професор кафедри філології і лінгводидактики ЖДУ імені Івана Франка К. Я. Климова. З самого початку свого заснування Центр став інтелектуальним осередком, об'єднанням викладачів, аспірантів, магістрантів, студентів університету та учителів та учнів членів МАН у м. Житомирі та області. Одним із головних завдань Центру є популяризація та навчання використанню ресурсів Українського національного лінгвістичного корпусу для написання студентських та учнівських наукових робіт та використанню даних корпусу в практиці викладання навчальних дисциплін [див.: Організація навчально-дослідницької роботи школярів із використанням українського національного лінгвістичного корпусу 2012]. У Центрі також ведеться активна діяльність зі створення Національного корпусу навчально-методичної літератури МАН України.

На лінгвістичному порталі Mova.info[23] (Інституту філології Київського університету імені Тараса Шевченка) представлено Дослідницький корпус сучасної української мови обсягом у 3 млн. словоформ, який побудований як інформаційно-довідкова система. Це тексти в електронній формі, призначені для з'ясування різних питань, пов'язаних з українською мовою.

Розробленням методики та процедури організації лексикографічного матеріалу на засадах корпусної лінгвістики займаються науковці відділу лексикології, лексикографії та української мови Інституту української мови НАН України, зокрема, це створення електронного варіанта «Словаря української мови» за ред. Б. Грінченка [Балог, Балог 2007]. Над створенням корпусів працюють також лабораторії комп'ютерної лінгвістики в Національному університеті «Острозька академія»[24] та Київському національному лінгвістичному університеті[25].

В Україні вже декілька років поспіль відбувається щорічна конференція MegaLing[26], присвячена різноаспектним корпусним дослідженням. У центрі уваги конференції знаходяться питання, пов'язані з теоретичними засадами корпусної лінгвістики, лексикографії, термінографії, нотографії, концептографії і т.ін.

З'являються оглядово-ознайомлюючі публікації з питань корпусного мовознавства, а також здійснюються теоретичні та прикладні дослідження у цьому річищі (див. Список використаних джерел).

[23] http://www.mova.info/

[24]http://www.oa.edu.ua/ua/departments/filologist/filol_literature/lexilab/project3

[25] http://www.complinguide.com.ua/Corpora.aspx

[26] http://megaling.ulif.org.ua/

Запитання і завдання для самоперевірки

1. Охарактеризуйте історичний розвиток корпусних досліджень.
2. У чому виявляється специфіка доелектронних корпусів? Назвіть основні лінгвістичні проекти, в яких доелектронні корпуси відіграли значну роль.
3. Поясніть, чому корпусна лінгвістика почала активно розвиватися на матеріалі англійської мови?
4. Як були сприйняті корпусні дослідження в лінгвістичній науковій спільноті 60-80-х років XX ст.?
5. Окресліть головні характеристики корпусів першого покоління, назвіть їх.
6. Чому Браунівський корпус став стандартом створення корпусів на для декількох поколінь?
7. Визначте основні характеристики корпусів другого покоління.
8. Назвіть найвідоміші мегакорпуси сучасності.
9. Охарактеризуйте сучасний етап розвитку корпусних студій у світі та в Україні.
10. Які, на Вашу думку, перспективи розбудови корпусної лінгвістики в українському мовознавстві?

Розділ 3.
Корпус текстів природної мови

3.1. Конститутивні параметри корпусу текстів

Центральним для корпусної лінгвістики є поняття «корпус текстів». Адже, одного боку, він виступає головною ціллю, яку ставить перед собою корпусна лінгвістика, а з іншого, постає об'єктом дослідження цієї мовознавчої галузі.

Взагалі лінгвістичний корпус текстів – це надзвичайна річ, унікальність якої полягає не лише в тому, що це велика за обсягом збірка електронних текстів, але і в тих можливостях, яких набуває корпус, за умови, він добре організований і правильно укладений. Доцільність створення й використання корпусів визначається такими передумовами:

1) досить великий (репрезентативний) обсяг корпусу гарантує типовість даних і забезпечує повноту представлення всього спектру мовних явищ;

2) дані різного типу перебувають у корпусі у своїй природній контекстній формі, що створює можливість їх всебічного й об'єктивного вивчення;

3) одного разу створений і підготовлений масив даних може використовуватися багаторазово, багатьма дослідниками й у різних цілях [Захаров, Богданова 2011: 8].

Перед тим, як увійти до лінгвістичної термінології, слово «корпус» широко використовувалося на позначення збірки або колекції певних текстів. Лише у 1984 році Дж. Аартс та В. Мейс у статті «Corpus Linguistics: Recent Developments in the Use of Computer Corpora in English Language Research» [Aarts, Meijs 1984]. вперше використали його як термін корпусної лінгвістики. До кінця 90-х років термін «корпус» міцно увійшов

до лінгвістичного вжитку вже у його традиційному значенні як сукупність стандартно організованих електронних текстів, призначених для проведення конкретних лінгвістичних досліджень.

На сьогодні в обігу корпусного мовознавства наявні декілька трактувань ***лінгвістичного корпусу текстів*** *(англ.* ***linguistic corpus*** або ***text corpus****)*. З метою виокремлення конститутивних параметрів корпусу текстів наведемо декілька прикладів визначення терміну «корпус» вітчизняними та зарубіжними корпусними лінгвістами:

- корпус – це організована певним чином словесна єдність, елементами якої є цілі тексти чи спеціальним чином відібрані уривки з текстів, що доступні для лінгвістичного аналізу [Meyer 2004: xi];
- корпус – це зібрання текстів, яке вважається репрезентативним стосовно даної мови, діалекту або іншої ділянки мови й призначене для використання в лінгвістичних дослідженнях [Francis 1991];
- корпус – це певне зібрання текстів, в основі яких лежить логічний задум, логічна ідея, що об'єднує ці тексти. Логічна ідея втілюється в правилах організації текстів в корпус, алгоритмі і програмі аналізу корпусу текстів та в пов'язаних з цим ідеологією та методологією. Корпус є четвертою фактурою мовлення (тексти на машинному носії) [Рыков[27]];
- корпус – це значне за обсягом цифрове зібрання текстів і текстових уривків, що слугує репрезентативною вибіркою для певного, обмеженого на основі різних параметрів використання мови, а в загальномовному

[27] http://corpora.iling.spb.ru/theory.htm

лексикографічному контексті виступає цільовою вибіркою з мови в цілому [Asmussen 2007: 123];

- корпус – це машиночитане, стандартно організоване зібрання репрезентативних для певної мови, діалекту або іншої підмножин(и) мов(и) писемних або усних текстів, призначених для лінгвістичного аналізу й опису, відібраних і впорядкованих згідно з експліцитними екстра- та інтралінгвальними критеріями [Демська-Кульчицька 2005].

Отже, аналіз наведених вище дефініцій дозволяє виділити низку дистинктивних ознак, що відрізняють сучасний корпус текстів від звичайних колекцій текстів в електронній формі (електронних бібліотек, архівів): репрезентативність, автентичність, відібраність, збалансованість, машиночитаність. Зупинимося детальніше на цих ознаках.

1. ***Репрезентативність*** полягає в здатності корпусу відображати всі властивості предметної галузі. Під предметною галуззю розуміється рівень реалізації мовної системи, яка містить феномени, що підлягають лінгвістичному описові. Предметна галузь для корпусу може бути як завгодно великою або як завгодно малою. Так, якщо йдеться про авторський корпус і лінгвістичний опис стосується авторської мови, то предметна галузь – авторська мова – не буде надто великою, наприклад, у порівнянні з предметною галуззю загальномовного корпусу. Але якщо завданням є побудова національного корпусу, то, відповідно, предметна галузь буде значно більшою [Демська-Кульчицька 2005: 101-102]. Як зауважують Т. МакЕнері та інш., репрезентативність більшості корпусів великою мірою обумовлюється двома факторами: набір жанрів, включених до корпусу (збалансованість), та критерій відбору текстів кожного жанру (відібраність) [McEnery, Xiao, Tono 2006: 11].

2. Автентичність передбачає відбір реально створеного носієм(ями) мови писемного або усного тексту(ів), уривка(ів) тексту(ів) у процесі реальної комунікації. Дотримання вимоги автентичності є однією зі складових емпіризації фактичного корпусного матеріалу.

3. Відібраність ставить вимогу обмеження фактичного матеріалу шляхом відбору певних фрагментів мови з усього мовного континууму. Навіть найбільший за обсягом корпус природної мови завжди залишається лише крихітним взірцем усіх усних і писемних текстів, створених усіма носіями мови навіть упродовж одного дня, і навіть сучасні інформаційні технології не дають змоги подати весь цей мовний матеріал, тому необхідна певна вибірка, яка передбачає застосування чітких правил екстрагування даних, що відповідають обраній стратегії побудови корпусу, мотивовані типом корпусу і метою його створення.

4. Збалансованість полягає у введенні до корпусу пропорційної кількості текстових ресурсів. На практиці, де традиційно використовують різні методики відбору текстового матеріалу до корпусу, одним із доволі складних завдань є досягнення збалансованості. Для досягнення збалансованості корпусу необхідні мінімальні критерії відбору текстів, які мають включати розрізнення між художньою літературою і нехудожньою літературою; книжкою, журналом або газетою; нормативним і ненормативним варіантом мови; з контролем віку, статі та походження авторів.

5. Машиночитаність є визначальною ознакою до сучасного електронного текстового корпусу природної мови. Крім електронної форми подання, ця вимога передбачає наявність кодування первинних корпусних даних та лінгвістичну анотацію, хоча на сьогодні це вже параметр «за

промовчанням», тобто іншим сучасний корпус не повинен бути. До найбільших переваг комп'ютеризації корпусу відноситься швидкість обробки та легкість оперування великими масивами мовних даних (пошук, відбір, сортування, форматування і т. ін.). По-друге, комп'ютери надають точні та об'єктивні результати обробки машиночитаних даних. По-третє, комп'ютери виключають можливість впливу людського фактору на аналіз даних, що сприяє отриманню об'єктивних результатів. Нарешті, машиночитаність уможливлює подальшу автоматичну обробку даних корпусу, що в свою чергу збагачує корпус різноманітними метаданими (різними типами розміток). Звичайно, без комп'ютерної підтримки багато корпусно-базованих досліджень, що були здійснені за останні два десятиліття, були б просто неможливими Е. Тогніні-Бонеллі [Tognini-Bonelli 2000: 210]переконана, що саме комп'ютер визначив методологічну модель лінгвістичного пошуку.

Корпус формується з реальних уривків писемного або усного мовлення, не передбачаючи модифікації мовленнєвої дійсності, що перетворює його на категорію емпіричну і дозволяє розглядати фактичний корпусний матеріал як емпіричну базу лінгвістичного дослідження [Демська-Кульчицька 2003: 41]. Завданням авторів корпусу є зібрати якомога більшу кількість текстів, що відносяться до тієї мови чи підмножини мови, для вивчення якої корпус створюється. Таким чином, можливо стверджувати, що корпус – це зменшена модель мови чи підмови.

Однією із вагомих проблем сучасної корпусної лінгвістики є визначення обсягу корпусу, достатнього для отримання надійних висновків. Так, найбільш загальноприйнятий підхід стверджує, що мінімальний обсяг корпусу повинен становити 1 мільйон слововживань. Проте деякі корпусні дослідники

зауважують, що будь-який текст чи зібрання текстів теоретично може бути визнаним корпусом (від лат. *corpus* – тіло), а тому корпусний аналіз можливо здійснювати навіть на дуже малих за обсягом текстах (наприклад, дослідження М. Стаббса [Stubbs 1996: 81−100] зосереджено на вивченні лінгвістичних особливостей лише двох листів обсягом у декілька сотень слів) або колективну монографію «Small Corpus Studies and ELT» [Small Corpus Studies and ELT 2001].

Т. МакЕнері та Е. Вілсон констатують, що корпус, як правило, складається з вибірок, що «максимально репрезентують досліджувану область/сферу) [McEnery, Wilson 2001: 24], а це означає, що він повинен бути достатньо великим, щоб надати дані про частоту вживання певних лінгвістичних явищ, надаючи дослідникам можливість визначити, що є типовим, а що рідкісним явищем. Звідси значна кількість теперішніх корпусних досліджень здійснюється як на основі багатомільйонних корпусів [див.: McEnery 2005; Davies 2009; Millar 2009], так і на матеріалі менших спеціалізованих корпусів [див.: Ghadessy and Gao 2001; McEnery and Kifle 2002].

Таким чином, корпус текстів – це машиночитане, збалансоване, репрезентативне зібрання особливо розмічених (анотованих) текстів, відібраних згідно фіксованих параметрів для досягнення визначеної лінгвістичної мети та досліджуваних нелінійно за принципом гіпертексту.

Кількість корпусних досліджень збільшувалася вдвічі кожні п'ять років протягом 1976-1991 р. [Johansson 1991].

3.2. Типологія та огляд існуючих корпусів

Сучасне корпусне мовознавство має у своєму дослідницькому арсеналі значну кількість корпусів різних типів і розмірів. Таке розмаїття існуючих корпусів визначається широким спектром дослідницьких і прикладних завдань, для розв'язання яких вони створюються, та особливостями мовного матеріалу, покладеного в основу. Однак варто відзначити, що на разі не існує єдиної загальноприйнятої типології корпусів, а наявні корпусні систематизації різняться засадничими класифікаційними критеріями. Так, наприклад, Т. МакЕнері, Р. Ксіао та Й. Тоно в класифікації корпусів залежно від дослідницької мети та завдання виділяються такі типи корпусів, як загальні/спеціалізовані корпуси, корпуси писемного/усного мовлення, синхронічні/діахронічні корпуси, учнівські корпуси та моніторингові корпуси [McEnery, Xiao, Tono 2006: 59-70].

У типології корпусів В.В. Рикова виділяються такі типи[28]:

1. За ступенем організації й структурованості:

- електронний архів – це тексти на електронному носії, але форма їх представлення на машинному носії не стандартизована й не уніфікована;
- електронна бібліотека – тексти тут представлені однорідним і стандартизованим способом;
- корпус текстів – форма стандартизована й уніфікована, тексти призначені для відображення частини лінгвістичної реальності;
- субкорпус – це деяка автономна частина корпуса.

2. За хронологічною ознакою:

- синхронічний;
- моніторинговий (відслідковує поточний стан мови);

[28] http://corpora.iling.spb.ru/theory.htm

- діахронічний.

3. За індексацією:

- простий;
- анотований.

4. За мовою:

- одномовний;
- двомовний;
- багатомовний.

5. За способом застосування й використання корпусу:

- дослідницький;
- ілюстративний;
- паралельний.

6. За способом існування корпусу:

- динамічний;
- статичний.

Автор російського посібника «Корпусная лингвистика» В.П. Захаров пропонує власне бачення класифікації типів корпусів на основі набору класифікаційних ознак [Захаров 2005: 13]:

Таблиця 3

Класифікація корпусів за В.П. Захаровим

Ознака	***Типи корпусів***
Тип даних	писемного мовлення усного мовлення змішаного типу
Мова текстів	російська англійська і т.ін.
«Паралельність»	одномовні двомовні багатомовні

Ознака	*Типи корпусів*
«Літературність»	літературні діалектні розмовні термінологічні змішані
Жанр	літературні фольклорні драматургічні публіцистичні
Доступність	у вільному доступі комерційні закриті
Призначення	дослідницькі ілюстративні
Динамічність	динамічні (моніторингові) статичні
Розмітка	розмічені нерозмічені
Характер розмітки	морфологічні синтаксичні семантичні просодичні і т. ін.
Обсяг текстів	повнотекстові «фрагментно текстові»
Хронологічний аспект	синхронічні діахронічні
«Спільність»	загальні одного письменника
Структура	центральні й архівні ядерні й периферійні

Орися Демська-Кульчицька пропонує інший підхід до класифікації корпусів, що поданий на Рис.1 [Демська-Кульчицька 2004: 156-157]:

Рис.1. Класифікація корпусів за О. Демьскою-Кульчицькою

Незважаючи на різноманітність корпусів, можна виділити два основні способи розподілу корпусів на класи: 1) протиставлення корпусів, що укладаються до всієї мови (так звані національні корпуси), корпусам, що укладаються до певної підмови (жанру, стилю, мови певної вікової або соціальної групи, мови письменника або вченого й т.п.); 2) поділ корпусів за типом лінгвістичної розмітки. Хоча існує безліч типів розмітки, більшість реально існуючих корпусів відноситься до корпусів морфологічного або синтаксичного типу (англійською treebanks, що можна перекласти як «банки синтаксичних структур»). При цьому слід підкреслити, що корпус із синтаксичною розміткою експліцитно або імпліцитно містить у собі й морфологічні характеристики лексичних одиниць [Захаров, Богданова 2011: 20-21].

Розглянемо детальніше зазначені типи корпусів. ***Національний корпус*** – це великий за обсягом корпус, що прагне до об'єднання у своїй структурі текстів найрізноманітніших жанрів і типів (сучасні технології дозволяють вбудовувати в національні корпуси аудіо- і відоматеріали, наприклад, Національний корпус російської мови[29]). Чим більший розмір такого корпусу, чим різноманітніша та точніша розмітка, чим ефективніше програмне забезпечення такого корпусу, тим вища його цінність як лінгвістичного ресурсу. Саме тому лінгвісти і програмісти, що зайняті створенням національних корпусів, головним чином вирішують питання матеріального та технічного оснащення, для того щоб матеріали корпусу були доступні для роботи і задоволення запитів користувачів. Важливість та необхідність створення таких корпусів важко переоцінити, адже, перш за все, вони збирають і зберігають мову для сучасників та майбутніх

[29] http://www.ruscorpora.ru/

поколінь, дані цих корпусів дозволяють аналізувати стан багатьох мов світу у синхронічному та діахронічному аспектах (якщо корпус включає історичні підкорпуси, а сама можливість поповнення корпусу вже передбачає діахронію).

Наприклад, Національний корпус російської мови (НКРМ) містить близько 200 млн. слововживань з текстів, які відносяться до всіх основних сфер використання російської мови (наукової, офіційно-ділової, публіцистичної, церковно-богословської, художньої, розмовно-побутової, включаючи усну й електронну комунікацію). Робота зі створення корпусу розпочалася в 2000 році, а у квітні 2004 року корпус вже було розміщено на офіційному сайті[30]. У корпусі умовно виділяються дві частини – сучасна й діахронічна. Корпус сучасних текстів становлять тексти, що були створені у період 1951-2007 років. Обсяг цієї частини корпуса – 97,5 млн. слововживань, причому половина текстів відноситься до періоду після 2000 року. Діахронічна частина становить близько 53 млн. слововживань і поєднує тексти XVIII століття (1,1 млн. слововживань), XIX століття (23,3 млн. слововживань прозаїчних текстів і 2,5 млн. слововживань у поетичному корпусі) і 1-ї половини XX століття (25,4 млн. слововживань).

Основний масив текстів, зібраних у НКРЯ, охоплює період в 200 років, тому він найбільш пристосований для вивчення коротких (кілька десятиліть) і середніх (1-2 сторіччя) мовних змін.

Національний корпус російської мови в цей час включає наступні підкорпуси (субкорпуси) [Плунгян, Рахилина 2009: 21-35]:

[30] http://ruscorpora.ru/

- глибоко анотований корпус, у якому для кожного речення побудована повна морфологічна й синтаксична структура (дерево залежностей);
- паралельний російсько-англійський корпус текстів, у якому можна знайти всі переклади для певного російського або англійського слова або словосполучення;
- корпус діалектних текстів, що включає запис діалектного мовлення різних регіонів Росії зі збереженням їх граматичної специфіки; передбачений спеціальний пошук з урахуванням діалектної морфології;
- корпус поетичних текстів, у якому можливий пошук не тільки за лексичними і граматичними, але й за специфічними для вірша ознаками (пошук певної комбінації в сонетах, в епіграмах, у віршах, написаних амфібрахієм, з певним типом римування й т.п.);
- навчальний корпус російської мови – корпус зі знятою омонімією, розмітка якого орієнтована на шкільну програму російської мови;
- корпус усного мовлення включає розшифрування магнітофонних записів публічного й приватного усного мовлення, а також транскрипти кінофільмів 2000-х років.

Серед відомих національних корпусів також назвемо the British National Corpus (обсяг 100 млн. слововживань)[31], the American National Corpus (22 млн.)[32], the PELCRA Referenc Corpus of Polish Corpus (100 млн.), the Czech National Corpus (більше 100 млн.)[33], the Hungarian National Corpus (187,6 млн.)[34],

[31] http://www.natcorp.ox.ac.uk/

[32] http://americannationalcorpus.org/

[33] http://korpus.cz/english/co_je_korpus.php

[34] http://corpus.nytud.hu/mnsz/index_eng.html

the Hellenic National Corpus (корпус сучасної грецької мови, загальним обсягом 47 млн. слововживань)[35], the DWDS corpus (обсяг 100 млн. слововживань)[36], the Slovak National Corpus (339 млн.)[37], the Modern Chinese Language Corpus (100 млн. знаків)[38]. Інші національні корпуси the FRANTEXT Database for French[39], the Croatian National Corpus (101,3 млн.)[40], Korpus 2000 for Danish (28 млн.)[41], the National Corpus of Irish (30 млн.)[42] та інші [детальніше про інші корпуси див.: http://www.lancaster.ac.uk/fass/projects/corpus/cbls/corpora.asp]. Як ми вже зазначали у п.2.3, на базі Національної словникової бази Українського мовно-інформаційного фонду НАН України також активно розробляється Український національний лінгвістичний корпус (УНЛК).

Національним корпусам протиставляються спеціальні, які створюються для вирішення конкретних лінгвістичних задач. ***Спеціалізований корпус*** – це жанрово чи галузево специфічний корпус, що має на меті відобразити певну підмову. Наприклад, the Guangzhou Petroleum English Corpus містить 411 612 слововживань писемного варіанту англійської мови, відібраних із текстів нафтохімічної тематики. Сьогодні спостерігається підвищений інтерес до створення та використання спеціалізованих корпусів в освітній та професійних сферах. Наприклад, the Corpus of Professional Spoken American English

[35] http://hnc.ilsp.gr/en/default.asp

[36] http://www.dwds.de/

[37] http://korpus.juls.savba.sk/index_en.html

[38] http://rocling.iis.sinica.edu.tw/CKIP/engversion/20corpus.htm

[39] http://www.lib.uchicago.edu/efts/ARTFL/databases/TLF/

[40] http://www.hnk.ffzg.hr/default_en.htm

[41] http://korpus.dsl.dk/korpus2000/engelsk_summary.php?languk

[42] http://www.focloir.ie/corpus/

(CPSA) складається з транскриптів комунікативних ситуацій з академічної та політичної професійних галузей. The Michigan Corpus of Academic Spoken English (MICASE) містить біля 1,7 млн. слововживань (близько 200 годин записів) сучасного усного університетського мовлення, що було записано в Мічиганському університеті.

Всередині цього типу корпусів можливо виділити корпуси одного автора чи одного жанру (наприклад, the Epistolary Corpus of Victorian Women Writers' Letters, the Shakespeare Corpus, Корпус словаря языка Достоевского).

Надзвичайно ефективними для визначення рівня адекватності перекладу, удосконалення систем машинного перекладу і гармонізації термінології є паралельні корпуси, які складаються з оригінальних та перекладних текстів, що вирівняні по реченнях. Наприклад, до складу Національного корпусу російської мови станом на січень 2013 р. входять 8 двомовних паралельних корпусів: англійсько-російський (та, відповідно, російсько-англійський), німецько-російський, французько-російський, іспансько-російський, італійсько-російський, польсько-російський, українсько-російський та білорусько-російський (для кожної мовної пари подані також тексти, що перекладені у зворотному напрямку) [Шведова, Січінава 2013: 95].

Особливий тип спеціальних корпусів – це так звані корпуси учнівського мовлення або ***учнівські корпуси*** (*learner corpora*), які укладаються з усних і/або писемних текстів, спродукованих особами, що вивчають мову як іноземну. Тут англійський термін *learner*, перекладається лексемою «учнівський», похідною від іменника *учень* зі значенням «той, хто вчиться, вивчає щось», тобто під цим прикметником слід розуміти людину, яка навчається, безвідносно до віку. Такі корпуси почали

створюватися ще в кінці 80-х на початку 90-х років XX століття. Виділяють такі типи корпусів учнівського мовлення:

1) *комерційні*, що ініціюються видавничими компаніями, vs *академічні* корпуси, які укладаються навчальними закладами. Хоча академічних корпусів більше, ніж комерційних, останні вирізняються більшим обсягом та більшою представленістю мов, що є рідними для тих, хто вивчає певну мову. Головними комерційними учнівськими корпусами для англійської мови є the Longman Learners' Corpus та the Cambridge Learner Corpus, кожен обсягом понад 10 мільйонів слововживань. Академічні корпуси вирізняються різноманітністю структури та обсягу. Серед них найбільш відомим та представницьким є the International Corpus of Learner English (ICLE), загальним обсягом 2,5 млн. слововживань, в якому представлено есе носіїв 11 національних мов;

2) *великі* vs *малі* за обсягом корпуси учнівського мовлення;

3) учнівські корпуси *на матеріалі англійської* vs *інших мов*. Англійська мова, зрозуміло, домінує над іншими мовами. Серед найбільших учнівських корпусів слід назвати the International Corpus of Learner English та 25-мільйонний the Hong Kong University of Science and Technology Learner Corpus, що містить тексти англійською мовою, створені винятково носіями китайської мови. Меншими за обсягом є учнівські корпуси носіїв норвезької мови (the EVA Corpus of Speech), іспанської мови (the APU Spanish Learner Corpus). Слід зауважити, що кількість учнівських корпусів для інших мов, окрім англійської, невпинно зростає. Для прикладу, вже укладено корпуси для французької, шведської, норвезької, голландської, іспанської, німецької та інших мов;

4) учнівські корпуси *писемного* vs *усного мовлення*;

5) корпуси для *безпосереднього* vs *віддаленого в часі використання для педагогічних цілей* [детальніше про існуючі учнівські корпуси див.: http://www.lancs.ac.uk/fass/projects/corpus/cbls/corpora.asp].

Корпуси учнівського мовлення є надзвичайно корисними не лише для лінгвістів, що цікавляться особливостями інтермови, але і для викладачів-методистів та спеціалістів, які займаються розробкою навчальних програм з англійської мови, оскільки такі корпуси дозволяють відслідкувати, як реально засвоює учень іноземну мову (граматику, синтаксис, лексику), що засвоюється з труднощами або ж не засвоюється взагалі. Такі дані допоможуть скоректувати змістовну частину курсу англійської мови (чому навчати) в школах і на курсах.

Іншим типом корпусів, що виділяються за хронологічним параметром і зорієнтовані на вивчення та аналіз еволюційних процесів у конкретній мові, є ***історичні корпуси***. Такий тип корпусів переважним чином укладається на матеріалі текстів однієї мови, відібраних у різні часові проміжки з метою вивчення історичної динаміки мовних змін [Xiao 2008: 401].

Історичні корпуси поділяються на ***синхронні*** та ***діахронні***. Синхронні історичні корпуси базуються на текстах конкретних історичних періодів, наприклад, the Century of Prose Corpus обмежується текстами 1680-1780 рр. Діахронні історичні корпуси охоплюють довші часові відрізки, як, наприклад, the Helsinki Corpus of English Texts, що включає тексти за десять століть (730-1710 рр.). Зауважимо, що такий поділ історичних корпусів є досить умовним, оскільки підкорпуси діахронних корпусів можуть розцінюватися як синхронні історичні корпуси [Claridge 2008: 242].

В іншій класифікації історичні корпуси підрозділяються на динамічні та статичні. Статичні історичні корпуси мають

кінцевий розмір, а динамічні, навпаки, відкриті для поповнення іншими текстами. Переважна кількість історичних корпусів включають не повнотекстові матеріали, а текстові вибірки з метою ширшої жанрової презентації (наприклад, наукові та юридичні документи, драматичні та художні твори, щоденники, особисті листи і т.ін.). Історичні корпуси також можуть бути анотованими, включаючи частиномовну й синтаксичну розмітку, та супроводжуватися соціолінгвістичною інформацією [Vazquez, Esteban-Segura, Marques-Aguado 2011: 120].

Отож найбільш відомим історичним корпусом англійської мови вважається the Helsinki Corpus of English Texts[43], який було укладено командою дослідників під керівництвом професора Матті Ріссанен (Хельсінський університет) у період з 1984 по 1991 роки [Vazquez, Esteban-Segura, Marques-Aguado 2011: 121]. Корпус, безперечно, вважається основоположним проектом у галузі укладання історичних корпусів, що активізував подальші дослідження та розробки у цій сфері. Корпус містить біля 1,5 млн. слововживань, представлених у 400 текстах VIII-XVIII століть. Матеріал корпусу поділено на три періоди (Old, Middle and Early Modern English) та одинадцять підперіодів приблизно по 100 років. Кожен текстовий документ супроводжується коротким описом, що включає 25 параметрів (наприклад, соціолінгвістична інформація: стать автора, вік та соціальний статус; тип тексту: науковий текст, судовий документ, проповідь, щоденник, п'єса, особистий, офіційний лист і т.ін.). Така різнопланова розмітка розширює сферу використання корпусу від встановлення та опису мовних змін до широких соціолінгвістичних та жанрових розвідок.

The Helsinki Corpus of English Texts має декілька синтаксично анотованих версій, що дозволяє здійснювати

[43] http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/

пошук не лише за окремими словами (лексемами) і словосполученнями, але і за синтаксичними конструкціями: the Brooklyn-Geneva-Amsterdam-Helsinki Corpus of Old English, the Penn-Helsinki Parsed Corpus of Middle English (1150-1500 pp., 1,2 млн. слововживань), the Penn-Helsinki Parsed Corpus of Early Modern English (1500-1700 pp., 1,7 млн. слововживань), the Penn Parsed Corpus of Modern British English (1700-1914 pp., 1 млн. слововживань).

Ішим потужним проектом є A Representative Corpus of Historical English Registers (ARCHER)[44]. Це багатожанровий корпус британського та американського варіантів англійської мови, що містить 1,8 млн. слововживань, представлених у 1037 текстах ранньоновоанглійського та пізньоновоанглійського періодів (1650-1990 pp.). Корпус створений у 1990 році Д. Байбером та Е. Фінеганом. Сьогодні корпус є закритим проектом, що виконується консорціумом чотирнадцяти університетів із семи країн. Корпус є збалансованою репрезентацією семи писемних (щоденники, листи, художні твори, новини, наукові твори і т.ін.) та трьох усних жанрів (діалоги з драматичних та прозових художніх творів, проповіді) британського (дві третини корпусу) та американського (одна третя корпусу) варіантів англійської мови. ARCHER супроводжується детальною розміткою граматичних і функціональних категорій, що уможливлює здійснення широкого кола досліджень історичної динаміки мови, дискурсу та жанру.

Крім the Helsinki Corpus та ARCHER, що охоплюють декілька століть, існує низка відомих історичних корпусів, що зосереджені на окремому періоді чи певному стилі або жанрі.

[44]http://www.anglistik.unifreiburg.de/seminar/abteilungen/sprachwissenschaft/ls_mair/research/projects/archer

Наприклад, The Lampeter Corpus of Early Modern English Tracts[45] – це збалансований корпус англійських текстів за період 1640-1740 рр. загальним обсягом 1,1 млн. слововживань. До корпусу включено тексти з шести галузей знань (релігія, політика, економіка/торгівля, наука, юриспруденція та некласифіковані тексти). Корпус містить екстралінгвістичну розмітку, що надає інформацію про автора тексту (ім'я, вік, стать, місце проживання, освіта, соціальний стан, політична приналежність), видавця, час та місце видання, текстові характеристики та бібліографічні джерела. Оскільки корпус надає у розпорядження дослідника повні тексти, то він може бути використаним і для вивчення текстової організації в ранньоновоанглійський період.

The Corpus of Early English Correspondence (CEEC)[46] містить 96 збірок 6039 особистих листів, написаних 778 особами (20% з яких жінки) між 1417 та 1681 роками, загальним обсягом 2,7 млн. слововживань. Корпус супроводжується соціолінгвістичною інформацією про відправників. The Corpus of Early English Correspondence Extension (CEECE) є продовженням The Corpus of Early English Correspondence, містить 2,2 слововживань та охоплює період з 1681 по 1800 р.

The Zurich English Newspaper Corpus[47] – це корпус газетних текстів, що репрезентує 120 років історії Британської преси з 1671 по 1791 р. загальним обсягом 1,2 млн. слововживань. Корпус включає повнотекстові газетні матеріали, а спеціальна розмітка дозволяє проводити порівняльні дослідження текстів новин XVII-XVIII ст. із сучасними новинними текстами.

[45] http://khnt.hit.uib.no/icame/manuals/LAMPETER/LAMPHOME.HTM

[46] http://www.helsinki.fi/varieng/domains/CEEC.html

[47] http://es-zen.unizh.ch

The Innsbruck Computer Archive of Machine-Readable Texts (ICAMET)[48] містить 500 текстів середньоанглійського періоду загальним обсягом 5,7 млн. слововживань. Корпус підрозділяється на *The Prose Corpus* (129 текстів, написаних у період з 110-1500), *the Letter Corpus* (254 листи датованих 1386-1688 рр.) та *the Prose Varia Corpus* (переважно переклади та нормалізовані версії середньоанглійських текстів). Головною перевагою даного корпусу є включення до бази повнотекстових матеріалів, а не фрагментів, що дозволяє проводити різноаспектні лінгвістичні розвідки.

The Corpus of English Dialogues (CED)[49] (1560-1760) включає 1,2 млн. слововживань із діалогічних текстів ранньоновоанглійського періоду, датованих 1560-1760 рр. Оскільки з об'єктивних причин безпосередній доступ до усного варіанта реалізації англійської мови зазначеного періоду відсутній, то корпус укладено на матеріалі текстів, що репрезентують усне мовлення (судові розгляди, свідчення свідків, драматичні твори, діалоги з художніх творів, дидактичні твори у формі діалогу та некласифіковані тексти).

The Corpus of Late Modern English Texts (CLMET)[50], загальним обсягом понад 10 млн. слововживань, репрезентує тексти різних жанрів за 1710-1920 роки (від особистих листів до художніх і наукових творів, написаних чоловіками та жінками різних соціальних класів).

Слід також зазначити такі історичні корпуси, як the ProQuest Historical Newspapers, A Linguistic Atlas of Early Middle English (LAEME), the Corpus of Middle English Medical Texts (MEMT),

[48] http://www.uibk.ac.at/anglistik/projects/icamet/

[49] http://www.engelska.uu.se/corpus.html

[50] http://www.helsinki.fi/varieng/CoRD/corpora/CLMETEV/

the Corpus of Early Modern English Medical Texts (EMEMT) та інші.

За типом мовних даних корпуси поділяються на корпуси писемного, усного мовлення та змішаного типу. У ***корпусах писемного мовлення*** усний варіант реалізації мовної системи не представлений. Найпершим сучасним корпусом писемного англійського мовлення був корпус американського варіанту англійської мови the Brown University Standard Corpus of Present-Day American English, услід за яким було укладено низку подібних корпусів. Іншими корпусами такого ж типу є the Australian Corpus of English, що репрезентує австралійський писемний варіант англійської мови (1986-). The Wellington Corpus of Written NZ English (WWC) представляє новозеландський писемний варіант англійської мови (1986-1990 рр.), a the Kolhapur Corpus відображає індійський писемний варіант англійської мови (1978-).

Корпуси усного мовлення включають транскрибовані версії реальних усних комунікативних ситуацій. ***Корпуси усного мовлення*** – це спеціальні колекції ретельно відібраних текстових уривків (слів, фраз, речень), вимовлених численними мовцями за різних акустичних умов [Teubert 2007: 126]. Зважаючи на трудомісткість і напруженість збору усних даних порівняно із писемними, ці корпуси є значно меншими за обсягом. Серед корпусів усного мовлення назвемо the London-Lund Corpus (LLC), the Lancaster/IBM Spoken English Corpus (SEC), the Cambridge and Nottingham Corpus of Discourse in English (CANCO DE), the Santa Barbara Corpus of Spoken American English (SBCSAE) та the Wellington Corpus of Spoken New Zealand English (WSC).

Підводячи підсумок огляду існуючих корпусів, зазначимо, що різноманіття наявних на сьогоднішній день корпусів аж

ніяким чином не обмежується наведеними вище даними. З огляду на розширення спектру лінгвістичних завдань, що ставлять перед собою дослідники мови, арсенал корпусної лінгвістики постійно поповнюється новими, більш досконалими розробками.

Запитання і завдання для самоперевірки

1. Дайте визначення терміну «лінгвістичний корпус текстів»?
2. Перерахуйте дистинктивні ознаки, за якими корпус текстів відрізняється від електронної бібліотеки та архіву. Охарактеризуйте кожну з них.
3. Проаналізуйте наведені класифікації типів корпусів. Яка із класифікацій видається Вам найбільш вдалою. Обґрунтуйте свою точку зору.
4. Назвіть і схарактеризуйте існуючі типи корпусів. Наведіть приклади до кожного з типів та коротко опишіть основні характеристики цих корпусів.
5. Відповідно до сфери Ваших наукових інтересів, відвідайте офіційні сторінки згаданих у розділі корпусів, використовуючи наведені у тексті посилання. Проаналізуйте можливості обраних Вами корпусів, зазначивши їх переваги та недоліки.

Розділ 4.
Поняття корпусної розмітки

4.1. Типи корпусної розмітки

Лінгвістичний корпус за визначенням є такою колекцією природно мовних текстів, де здійснено розмітку (маркування) хоча б за одним лінгвістичним параметром. Ця ознака є такою, що вирізняє лінгвістичний корпус з-поміж великого числа інших лінгвістичних інформаційно-інструментальних систем, баз даних та знань [Корпусна лінгвістика 2005: 33]. Процес розмітки (*tagging, annotation*) полягає в приписуванні текстам і їх компонентам спеціальних міток (*tag, tags*):

- *зовнішніх, екстралінгвістичних* (відомості про автора й відомості про текст: автор, назва, рік і місце видання, жанр, тематика; відомості про автора можуть включати не тільки його ім'я, але також вік, стать, роки життя й багато чого іншого (це кодування інформації має назву *метарозмітка*);
- *структурних* (розділ, абзац, речення, словоформа);
- власне *лінгвістичних*, що описують лексичні, граматичні та інші характеристики елементів тексту.

Набір цих метаданих багато в чому визначає можливості, що надаються корпусами дослідникам. Анотація додає цінності корпусу, адже суттєво розширює діапазон дослідницьких питань, що можуть бути досліджені на матеріалі даного корпусу. При виборі цих даних необхідно керуватися цілями дослідження й потребами лінгвістів, а також можливостями по внесенню в текст тих або інших додаткових ознак.

Зауважимо, що у корпусних студіях розрізняють **анотацію (***annotation***)** та структурне **маркування (***markup***)** корпусу/корпусних даних. Розмежування цих двох понять

виходить з визначення анотації Дж. Лічем, згідно якого «процес анотування корпусних даних – це додавання інтерпретованої, лінгвістичної інформації до електронного корпусу усного чи/або писемного мовлення» [Leech 1997: 2]. Під поняттям «інтерпретована лінгвістична інформація» розуміється така анотація, яка в деякій мірі є продуктом розуміння тексту людським розумом (наприклад, анотація за частинами мови). Маркування надає відносно об'єктивну верифіковану інформацію про частини корпусу та структуру кожного тексту [McEnery, Xiao, Tono 2006: 29].

Зупинимося дещо детальніше на *структурній анотації* (*corpus markup*). Терміносполука "структурна анотація" у спеціальній літературі має неоднакові дефініції. Найширше її розуміння знаходимо у Ч. Меєра, який використовує цю терміносполуку на позначення і структури тексту, і зовнішньої стосовно нього інформації (його бібліографічний опис, дані про мовців тощо) [Meyer 2002: 81] . У вужчому значенні "структурну анотацію" описують Ґ. Астон і Л. Бернард: "...корисно вказувати межі глав, розділів, абзаців, речень, і т. д., а також особливу роль заголовків, переліків, приміток, посилань, супровідних підписів, покликів та ін." [Aston , Burnard 1998: 24]. Міжнародний стандарт кодування текстової інформації TEI (*англ.* Text Encoding Initiative – проект кодування текстів), який "репрезентує ті ознаки тексту, які потребують експліцитної ідентифікації з метою сприяти текстовому опрацюванню за допомогою комп'ютерних програм" [TEI]. Під елементами універсальної структури тексту розуміються <head> (заголовок), (частина, розділ), <p> (абзац), <s> (речення), <epigraph> (епіграф), <dateline> (дата), <note> (примітка), <said> (пряма мова), <dedication> (присвята), <l> (рядок, у вірші), <abbr> (скорочення), <num> (число) та ін.

Отже, структурою тексту вважаємо такі його елементи, як назва, розділ, підрозділ, рубрика, присвята, епіграф, поклик, цитата, вживання алфавітів інших писемних систем, цифр тощо. Структурне анотування – це виділення структурних елементів тексту за допомогою певної мови маркування; сукупність маркерів-вказівок на елементи зовнішньої будови тексту.

Під **лінгвістичною анотацією** у корпусній лінгвістиці традиційно розуміють: а) довільну лінгвістичну інформацію про лінгвально релевантні одиниці текстових даних, поданих через формальний код; б) практику введення формалізованої лінгвістичної інформації в електронний текст; в) наявність такої інформації у тексті [Демська-Кульчицька 2004: 26].

Лінгвістична анотація може здійснюватися на різних рівнях та набувати різних форм. Так, серед лінгвістичних типів розмітки виділяються:

- **Морфологічна розмітка**. В іноземній термінології вживається термін *part-of-speech tagging* (POS-tagging), дослівно – частиномовна розмітка. У дійсності морфологічні мітки включають не тільки ознаки частини мови, але й ознаки граматичних категорій, властивих даній частині мови. Це основний тип розмітки, оскільки, по-перше, більшість великих корпусів є саме морфологічно розміченими корпусами, по-друге, морфологічний аналіз розглядається як основа для подальших форм аналізу – синтаксичного й семантичного, і, по-третє, успіхи в комп'ютерній морфології дозволяють автоматично розмічати корпуси великих розмірів. Наведемо приклад частиномовної розмітки з корпусу LOB:

  hospitality_NN is_BEZ an_AT excellent_JJ virtue_NN ,_, but_CC not_XNOT when_WRB the_ATI guests_NNS have_HV to_TO sleep_VB in_IN rows_NNS in_IN the_ATI

cellar_NN !_! the_ATI lovers_NNS ,_, whose_WP$ chief_JJB scene_NN was_BEDZ cut_VBN at_IN the_ATI last_AP moment_NN ,_, had_HVD comparatively_RB little_AP to_TO sing_VB '_' he_PP3A stole_VBD my_PP$ wallet_NN !_! '_' roared_VBD Rollinson_NP ._.

Інший приклад такого типу розмітки з **THE SPOKEN ENGLISH CORPUS:**

[S[N Nemo_NP1 ,_, [N the_AT killer_NN1 whale_NN1 N] ,_, [Fr[N who_PNQS N][V 'd_VHD grown_VVN [J too_RG big_JJ [P for_IF [N his_APP$ pool_NN1 [P on_II [N Clacton_NP1 Pier_NNL1 N]P]N]P]JJ]V]Fr]N] ,_, [V has_VHZ arrived_VVN safely_RR [P at_II [N his_APP$ new_JJ home_NN1 [P in_II [N Windsor_NP1 [safari_NN1 park_NNL1]N]P]N]P]V] ._. S]

[http://ucrel.lancs.ac.uk/annotation.html]

- **Синтаксична розмітка**, що є результатом синтаксичного аналізу, або *парсинга* (*parsing*), виконуваного на основі даних морфологічного аналізу. Цей вид розмітки описує синтаксичні зв'язки між лексичними одиницями й різні синтаксичні конструкції (наприклад, підрядне речення, дієслівне словосполучення й т.п.). Наприклад, речення *The cat sat on the mat* у корпусі буде представлено таким чином:

[S [NP The cat NP] [VP sat [PP on [NP the mat NP] PP] VP] S]

Where:
[S Sentence S]
[NP Noun Phrase NP]
[VP Verb Phrase VP]
[PP Prepositional Phrase PP]

[Baker, Hardie, McEnery 2006: 130]

- **Семантична розмітка**. Хоча для семантики немає єдиної семантичної теорії, найчастіше семантичні теги позначають семантичні категорії, до яких відноситься дане слово або словосполучення, і більш вузькі підкатегорії, що специфікують його значення, наприклад:

  PPIS1 I Z8
  VV0 like E2+
  AT1 a Z5
  JJ particular A4.2+
  NN1 shade O4.3
  IO of Z5
  NN1 lipstick B4

  [http://ucrel.lancs.ac.uk/annotation.html]

- **Анафорична розмітка**. Фіксує референтні зв'язки, наприклад, займенникові:

ANAPHORIC ANNOTATION OF A NEWSWIRE

S.1 The state Supreme Court has refused to release Rahway State Prison inmate James Scott on bail.

S.2 The fighter is serving 30-40 years for a 1975 armed robbery conviction.

S.3 Scott had asked for freedom while he waits for an appeal decision.

S.4 Meanwhile, his promoter, Murad Muhammed, said Wednesday he netted only $15,250 for Scott's nationally televised light heavyweight fight against ranking contender Yaqui Lopez last Saturday.

S.5 The fight, in which Scott won a unanimous decision over Lopez, grossed $135,000 for Muhammed's firm, Triangle Productions of Newar;

S.1 (0) The state Supreme Court has refused to release

{1 [2 Rahway State Prison 2] inmate 1}} (1 James Scott 1) on bail .

S.2 (1 The fighter 1) is serving 30-40 years for a 1975 armed robbery conviction .

S.3 (1 Scott 1) had asked for freedom while <1 he waits for an appeal decision .

S.4 Meanwhile , [3 <1 his promoter 3] , {{3 Murad Muhammed 3} , said Wednesday <3 he netted only $15,250 for (4 [1 Scott 1] 's nationally televised light heavyweight fight against {5 ranking contender 5}} (5 Yaqui Lopez 5) last Saturday 4) .

S.5 (4 The fight , in which [1 Scott 1] won a unanimous decision over (5 Lopez 5) 4) , grossed $135,000 for [6 [3 Muhammed 3] 's firm 6], {{6 Triangle Productions of Newark 6} , <3 he said .

[http://ucrel.lancs.ac.uk/annotation.html]

- **Просодична розмітка**. У просодичних корпусах застосовуються мітки, що описують наголос та інтонацію. У корпусах усного розмовного мовлення просодична розмітка часто супроводжується так званою *дискурсною* розміткою, яка служить для позначення пауз, повторів, застережень, і т.д. Як приклад просодичної розмітки наведено частину **LONDON-LUND CORPUS [**Leech 1997]**:**

  well ^very nice of you to ((come and)) _spare the !t\/ime and #
  ^come and !t\alk # -
  ^tell me a’bout the - !pr\oblems#
  And ^incidentally# .
  ^I [@:] ^do ^do t\ell me#
  ^anything you ‘want about the :college in ”!g\eneral

За даними дослідження Райсона та інш. (1997) мовці з економічно забезпечених прошарків вживають прислівники actually та really частіше ніж мовці з менш забезпечених прошарків, які натомість більш схильні використовувати слова say, said, saying, числа та непристойні слова.

Інший тип інформації також може бути закодований у корпусі. Наприклад, у корпусах розмовного мовлення може бути представлена соціолінгвістична інформація про такі риси як стать, вік, приналежність до соціального прошарку та місце проживання.

4.2. Вимоги до розмітки

Розмітка повинна відповідати низці вимог, семи максимам Дж. Ліча [Leech 1997: 6-7]. Розмітка мусить бути незалежною від тексту: повинна бути можливість прибрати розмітку і переглянути текст без неї, і, навпаки, вичленувати саму лише розмітку. Принципи розмітки, їх розробники та спосіб внесення розмітки в корпус повинні бути відомими кінцевому користувачу. Користувач повинен бути поставлений до відома про те, що розмітка не є безпомилковою, а являє собою лише потенційно корисний інструмент. В основу розмітки повинні бути покладені загальноприйняті і, по можливості, теоретично нейтральні лінгвістичні принципи. І, насамкінець, жодна розмітка не може апріорно вважатися стандартом. Реалізація будь-якого типу анотування передбачає низку процедур:

1. Сегментизація тексту.
2. Формалізація параметрів анотування.

3. Створення тегсету чи набору формальних кодів з відповідною семантикою.

4. Визначення анотаційної схеми та її принципів.

Автори монографії «Корпусна лінгвістика» [Корпусна лінгвістика , 2005: 51-53] зазначають такі критерії застосування стандарту:

1) *Достатність*: набір структурних елементів повинен бути достатньо широким, щоб забезпечити хоча б більшість вимог. Водночас бажано, щоб схема розмітки не містила надлишкову інформацію.

2) *Несуперечливість*: схема розмітки має бути сформована на базі несуперечливих правил, які б дозволяли однозначно визначити, які об'єкти належать до тегів, які – до атрибутів, що є вмістом тега тощо.

3) *Відтворюваність*: схема кодування повинна ґрунтуватися на чітко визначених правилах, що дає можливість відтворити вихідний текст за допомогою простих алгоритмів.

4) *Коректність*: за допомогою спеціального програмного забезпечення відбувається перевірка відповідності міток у документах їх структурним специфікаціям.

5) *Можливість збору даних*: збір даних включає безпосереднє накопичення даних (за допомогою ручного вводу або з використанням автоматичного розпізнання тексту) та проведенням кодування даних.

6) *Технологічність*: урахування потреб, пов'язаних з автоматичною обробкою текстів (вибір тексту згідно зі встановленими критеріями, використання спеціальних механізмів, типу міжтекстових покажчиків, поєднання текстів або інших елементів корпусу) тощо.

7) *Можливість масштабування*: важливо, щоб будь-яка створена схема мала можливість поповнюватися.

8) *Компактність*: проведення розмітки може істотно вплинути на розмір файлу, від чого залежить швидкість обробки даних текстів. Серед можливих методів досягнення компактності називають мінімізацію тегу, наприклад, пропущення або скорочення кінцевого тегу, застосування специфічних кінцевих тегів елементів або відмова від останніх; використання XML схеми розмітки тощо.

9) *Зрозумілість*: коли виникає потреба у безпосередній роботі користувача з текстом без використання спеціального програмного супроводу, прозорість розмітки є досить важливою.

Важливим у корпусній лінгвістиці залишається питання створення засобів автоматичного (чи, принаймні, автоматизованого) анотування текстів за різними критеріями – морфологічними, орфоепічними, семантичними, синтаксичними тощо.

Запитання і завдання для самоперевірки.

1. Поясніть у чому полягає суть корпусної розмітки.
2. У чому полягає відмінність між анотацією та структурним маркуванням корпусних даних?
3. Що включає в себе структурне маркування тексту?
4. Дайте визначення лінгвістичній анотації. Назвіть типи лінгвістичної анотації.
5. Охарактеризуйте головні вимоги до корпусної розмітки.
6. Які на Вашу думку переваги та недоліки анотованих корпусів?

Розділ 5.
Технологія створення корпусів та комп'ютерні програми для роботи з корпусом

5.1. Етапи створення корпусу

Кожного року кількість корпусів для розв'язання різного роду дослідницьких завдань невпинно зростає. Тому перед тим, як починати створювати власний корпус, дослідник повинен впевнитися, що на даний момент не існує корпусу, який би відповідав його специфічній дослідницькій меті. Якщо ж лінгвіст все-таки вирішив укласти власний корпус, то технологічний процес його створення передбачає поступове виконання таких кроків:

1. Визначення джерел лінгвального матеріалу. Дані для корпусу можна відібрати з двох головних джерел: публічно доступних текстів (джерел) та приватно доступних джерел. Публічно доступні дані можна відібрати з низки джерел. Це передусім газети, журнали, інтернет-сайти. Одна з головних проблем, з якою стикнеться дослідник при відборі матеріалу для корпусу – це авторські права. Дотримання авторських прав на матеріали передбачає отримання дозволу на використання текстів для дослідницьких цілей. Закони щодо авторських прав різняться у різних країнах, тому укладач повинен знати про закони, що охороняють авторські права не лише у своїй країні, але і в інших державах. Найбезпечніший шлях уникнути проблем із законодавством – це використати тексти із відкритих джерел, наприклад, текстових архівів, що є у вільному доступі в Інтернеті, як Project Gutenberg (детальніше про джерела текстів у вільному доступі див.: www.copyright.gov/title17/).

Під приватними джерелами розуміються дані, яких немає у вільному доступі, наприклад, особисті листи, документація компаній, твори і т.інш. Такі тексти є власністю конкретної особи чи осіб, компанії, тому доступ до них є обмеженим, а то і закритим. Звідси дослідникові необхідно отримати дозвіл у власників на право користуватися цими документами та оприлюднення даних дослідження.

2. Введення даних. Існує три способи введення даних у корпус: адаптація даних в електронному форматі, сканування та ручне введення.

Готові тексти в електронному форматі є найлегшим способом внесення даних у корпус. Зважаючи на те, що більшість корпусних менеджерів підтримують документи з розширенням .txt, тексти в інших форматах повинні бути переформатовані.

Якщо необхідні тексти існують лише у друкованому вигляді, їх конвертують в електронну форму за допомогою сканування. Це можуть бути рідкісні чи старі видання, що не мають цифрових версій. Звичайно, сканування потребує гарного технічного та програмного забезпечення. Скановані версії не позбавлені недоліків, тому після оцифровки тексти необхідно звірити з оригіналом та виправити помилки. Проте найбільш працемістким та часомістким процесом є ручний набір текстів. Такого методу введення даних до корпусу не уникнути, якщо текст знаходиться у такому стані, що сканування є неможливим чи існує лише рукописна версія документа.

3. Попереднє опрацювання тексту. На цьому етапі всі тексти, отримані з різних джерел, проходять філологічну перевірку й коректування. Здійснюється також підготовка бібліографічного й екстралінгвістичного опису тексту.

4. Конвертування й графематичний аналіз. Деякі тексти проходять також через один або кілька етапів попередньої машинної обробки, у ході яких здійснюються різного роду перекодування (якщо потрібно), видалення або перетворення нетекстових елементів (малюнки, таблиці, графіки, формули), видалення з тексту переносів, «твердих кінців рядків», забезпечення однакового написання тире та ін. Як правило, ці операції виконуються в автоматичному режимі. Звичайно, на цьому ж етапі здійснюється сегментування тексту на його структурні складові.

5. Розмітка тексту. Розмітка тексту полягає в приписуванні текстам та їх компонентам додаткової інформації (метаданих). Метаопис текстів корпусу включає як змістовні елементи даних (бібліографічні дані, ознаки, що характеризують жанрові й стильові особливості тексту, відомості про автора), так і формальні (ім'я файлу, параметри кодування, версія мови розмітки, виконавці етапів робіт). Ці дані вводяться вручну. Структурна розмітка документа (виділення абзаців, речень, слів) і лінгвістична розмітка здійснюються автоматично.

6. На наступному етапі робиться коректування результатів автоматичної розмітки: виправлення помилок і зняття неоднозначності (вручну або напівавтоматично).

7. Заключний етап – конвертування розмічених текстів у структуру спеціалізованої лінгвістичної інформаційно-пошукової системи (corpus manager), що забезпечує швидкий багатоаспектний пошук і статистичну обробку.

8. І, нарешті, забезпечення доступу до корпусу. Корпус може бути доступний у локальній мережі, тобто лише розробникам та особам, які мають право користування. Для прикладу, право на використання корпусу ARCHER має

консорціум чотирнадцяти університетів семи країн[51]. Корпус може поширюватися на CD-ROM, як Eearly Modern English Medical Texts та Middle English Medical Texts від видавництва John Benjamins. Також корпус може бути розміщеним у глобальній мережі. Різним категоріям користувачів можуть надаватися різні права й можливості доступу та експлуатації корпусу.

Звісно, у кожному конкретному випадку склад і кількість процедур можуть відрізнятися від вищеперерахованих, і реальна технологія може виявитися набагато складнішою. Проте оскільки укладання великих довідникових корпусів є занадто коштовним, то можливо нові значні за розміром корпуси не будуть створюватись у найближчому майбутньому. Укладені до цього корпуси будуть лише розширюватися за рахунок малих спеціалізованих корпусів, розроблених окремими дослідниками.

Під час створення корпусу використовується низка процедур і програм, як-от: токенізація, лематизація, стеммінг, парсинг [Захаров 2011: 38-41].

Токенізація – це розбиття потоку символів природної мови на окремі значимі одиниці (токени, словоформи). **Лематизація** – процес утворення початкової форми слова, виходячи з інших його словоформ. У багатьох мовах слово може зустрічатися в декількох формах з різними флексіями. Наприклад, англійське дієслово *'work'* має такі форми: *'work', 'worked', 'works', 'working'*. Базова форма, *'work'*, зафіксована в словнику, називається ***лемою*** слова. Лематизація – це процес угруповання різних флективних форм одного слова таким чином, щоб при аналізі вони оброблялись як одне слово. Для прикладу, друга колонка у запропонованому

51 http://www.alc.manchester.ac.uk/subjects/lel/research/projects/archer/

списку є лематизованою [Glossary of Corpus Linguistics 2006: 105]:

He	he
studied	study
the	the
problem	Problem
for	for
a	a
few	few
seconds	second
and	and
thought	think
of	of
a	a
means	means
by	by
which	which
it	it
might	may
be	be
solved	solve

Процес, дещо відмінний від лематизації, називається **стеммінгом**. Він полягає в знаходженні стеми (основи) слова. Різниця лише в тому, що стеммер обробляє окреме слово без знання контексту, і, таким чином, не може диференціювати слова, які мають різні значення в силу віднесеності до різних частин мови. Проте стеммери більш прості для реалізації й швидше обробляють дані. Наприклад, токену "*better*" відповідає лема "*good*", але це опускається при стеммінзі. Лема "*work*" є

базовою формою для токена "*working*", і ця відповідність буде виявлена як при стеммінзі, так і при лематизації.

Парсинг – це процес аналізу синтаксичної структури тексту чи частини тексту, що ґрунтується на зіставленні лінійної послідовності лексем (слів, токенів) мови з її формальною граматикою. Результатом є дерево залежностей (синтаксичне дерево), наприклад:

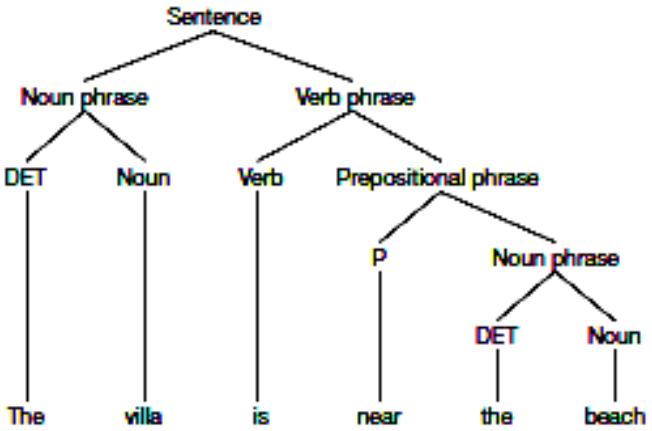


Fig. 3. Visual representation of a treebank:
DET = Determiner, P = Preposition

[Glossary of Corpus Linguistics 2006: 161]

Побудова автоматичних синтаксичних аналізаторів (парсерів) для великих корпусів є однією із найважливіших областей комп'ютерної лінгвістики.

5.2. Формати даних і стандартизація даних корпусу

Зазвичай, корпуси укладаються з метою можливості їх багаторазового використання якісно різними користувачами – лінгвістами vs нелінгвістами, тому, їхня розмітка та програмне забезпечення повинні бути максимально уніфіковані. Загальноприйняті формати представлення даних дозволяють у багатьох випадках використовувати єдине програмне забезпечення й обмінюватися корпусними даними. Стандартизація відносно корпусів, сумісність типів даних важливі також і для можливості порівняння різних корпусів

У цей час на основі міжнародного досвіду виробилися де-факто стандарти представлення метаданих, що базуються на описах текстів у рамках проекту Text Encoding Initiative (TEI) і на рекомендаціях EAGLES (Expert Advisory Group on Language Engineering Standards). Стандарт TEI забезпечує оптимальну збалансованість між загальною моделлю подання природної мови і нескладною реалізацією кодування. Також TEI оперує великим набором засобів для подання як лінгвальної, так і металінгвальної інформації.

Передумовою розроблення системи TEI стало існування великої кількості несумісних систем кодування і розширення сфери застосування електронних текстів. Базовими принципами системи визначено: а) можливість досягати у тексті ефектів, необхідних для наукових досліджень різного типу; б) простота, чіткість і конкретність; в) нескладність для використання без спеціалізованого програмного забезпечення; г) можливість точного визначення та ефективного програмного оброблення текстів; ґ) можливість розширень, визначених користувачем; д) узгодженість із чинними і новостворюваними стандартами[52]

[52] http://www.tei-c.org/index.xml

У якості формальної мови розмітки широко застосовуються мови SGML (Standard Generelised Markup Language) і XML (Extensible Markup Language). У цей час стандарти EAGLES безпосередньо включаються в технологічне середовище мови XML, див., зокрема, розробку стандарту Corpus Encoding Standard for XML (XCES).

У 1996 р. у межах проекту EAGLES було розроблено *Рекомендації для морфолого-синтаксичної анотації корпусу довільної природної мови,* де, власне, запропоновано модель „базові <--> додаткові" граматичні класи і визначено 11 базових лексико-граматичних розрядів слів, плюс пунктуація і спеціальні одиниці, якими можуть бути, наприклад, формули[53].

5.3. Комп'ютерні програми, що використовуються для аналізу даних корпусу

Оскільки корпусна лінгвістика послуговується у своїх дослідженнях великими за обсягом та репрезентативними вибірками текстів природної мови, існує декілька типів програмного забезпечення, що можуть використовуватися під час дослідження: комерційні комп'ютерні програми – конкордансери (н.-д. LEXA, MonoConc, MicroConcord, TACT, WordSmith, WordCruncher, Manatee (Bonito), IMS Corpus Workbench (CQP), XAIRA, LEXA, Virtual Corpus Manager(VMC), EXMARaLDA Corpus-Manager(Co-Ma)), а також комп'ютерні програми, розроблені для специфічних процедур аналізу. Ці типи програм застосовуються для складних граматичних моделей.

Робота з корпусом здійснюється за допомогою спеціальних програмних засобів – конкордансерів (більш простий тип

[53] http://www.ilc.cnr.it/EAGLES/home.html

програм) і корпусних менеджерів, що надають різноманітні можливості по одержанню необхідної інформації з корпуса. **Корпусний менеджер** – це спеціальна пошукова система, що включає програмні засоби для пошуку даних у корпусі, отримання статистичної інформації й надання результатів користувачеві в зручній формі. Результати цієї процедури подаються у вигляді горизонтальних рядків із пошуковим словом посередині. Ця процедура має назву KWIC (Key Word In Context).

Пошукові можливості корпусних менеджерів включають у себе пошук конкретних словоформ, пошук словоформ по лемах, пошук групи словоформ у вигляді розривної або нерозривної синтагми, пошук словоформ по набору морфологічних ознак і т.д., висновок результатів пошуку із вказівкою контексту заданої довжини, збереження відібраних матеріалів в окремому файлі. Використання корпусів дозволяє не тільки вивчати лексичні одиниці в контекстах, але й одержувати дані про частоту словоформ, частоту лексем, граматичних категорій, про спільну зустрічальність лексичних одиниць, особливості їх сполучуваності, управління і т.ін.

Результати пошуку видаються у вигляді конкордансу (тому корпусні менеджери і називають конкордансерами), де аналізована одиниця представлена в її контекстному оточенні та супроводжується статистичною інформацією. Останні можуть фіксувати частотні характеристики окремих мовних одиниць, або грамем, або можуть характеризувати сполучуваність декількох лексичних одиниць. Багато систем дозволяють коректувати формат представлення даних (міняти довжину лівого й правого контексту, задавати обсяг показу й порядок сортування даних, відображати або не відображати лінгвістичні

й екстралінгвістичні характеристики і т.д.). Наведемо приклад конкордансу слова *witnessed* [Baker, Hardie, McEnery 2006: 43]:

1	y told Tom Jones that he had never before	witnessed	a Cabinet scene like it." All who were
2	the early decades of the twentieth century	witnessed	an increase in the power of medical m
3	uld be drawn up carefully and signed and	witnessed	in a particular way. If you write it
4	The first attitude has been	witnessed	in the 1930s and during our more rece
5	nk had recovered from the breakdown we	witnessed	in late 1986 and, despite the months al
6	fought essentially on national issues and it	witnessed	the return not only of a reforming Libe
7	The last year of Ayliffe's Presidency	witnessed	the fulfilment of one of the BDDA's ea
8	eneration after the coming of Cyrus which	witnessed	the most brilliant speculations of the "
9	dirt, gloom and misery as I never before	witnessed	". Queen Victoria had the curtains of h
10	ood that this small Year Niner has been "	witnessed	" to and moves on to his next victim.

Рис. 1. Приклад конкордансу слова witnessed

Розглянемо інший приклад конкордансу слова *invade* з BNC-BYU:

foreign companies are likely to bid for the rare chance to	invade	a new broadcasting market . Until now British TV licences were
the result and one side 's league position . Whereas to	invade	a pitch when a goal is disallowed may be felt magically as
had served under Gordon in Equatoria , had set out to	invade	Abyssinia from Tajura , and been exterminated by the Danakil
rasped , cupping her face , drawing her inexorably back to	invade	again the full softness of her mouth . Her senses felt drugged
instant transformations from one shape to another will start to	invade	all sorts of artwork before very long ! It 's certainly fun
and First Fidelity Into the breach ARE European banks about to	invade	America -- again ? On March 19th Banco de Santander , Spain
need , his mounting passion , his overriding desire to	invade	and conquer her . As his firm , predatory fingers stroked her
by which the malevolent lunacy of powers in the warp could	invade	and ravage worlds ; could corrupt the human race into polluted
has been aptly described as an act of hijack . Viruses	invade	animal , plant and bacterial cells , and commandeer the complex
dirty stories , and what you learn there will mysteriously	invade	any other field in which timing is important , unless some mental
but Winston Churchill did not trust Hitler and defied him to	invade	Britain , Appreciating that Britain had become even more
century the Spanish Armada , before its abortive attempt to	invade	Britain , was ordered to destroy . if nothing else . the

Рис. 2. Приклад конкордансу слова invade

Задля підвищення ефективності досліджень до корпусних менеджерів висувається низка загальних вимог. Корпусний менеджер повинен будувати повні конкордансні списки; шукати не лише окремі слова, але й словосполучення; здійснювати пошук по шаблонах (складні запити); сортувати списки за декількома критеріями, що обираються користувачем; давати можливість відображати знайдені словоформи в розширеному контексті; подавати статистичну інформацію по окремих елементах корпусу; відображати леми, морфологічні характеристики словоформ і метадані (бібліографічні, типологічні), що залежать від ступеня розмітки корпусу; зберігати й роздруковувати результати; працювати як з окремими файлами, так і з корпусами, необмеженими по розміру; швидко обробляти запити й видавати результати; підтримувати різні формати текстових даних (txt, doc, rtf, html, xml ін.); бути легким (інтуїтивно зрозумілим) у використанні як для досвідченого, так і для початківця; і т.ін. [Захаров, Богданова 2011: 55-56].

Як стверджують автори підручника “Corpus Linguistics: Method, Theory and Practice” Т. МакЕнері та Е. Харді [McEnery, Hardie 2012: 37-48], історія розробки конкордансерів (тобто корпусних менеджерів) має тривалу історію, що розпочалася з діяльності Роберто Бузи. У 1951 році він вперше уклав машиночитані корпуси та отримав на їх основі перші автоматичні конкорданси. Хоча Буза і не винайшов конкорданс, оскільки вони укладалися вже протягом довгого часу і до нього (Див розділ), тим не менш дослідник показав, що із незначними зусиллями процедура конкордансування може бути швидко і якісно застосована до електронних текстів. Це стало переломним моментом, коли укладання конкордансу перетворилося з трудомісткого заняття, що застосовувалося для

текстів особливого культурного значення (Біблія, Коран, праці Шекспіра), на технологію, що в принципі може бути застосована до будь-якого тексту. Робота Р. Буза привела до появи конкордансерів першого покоління. Розроблені програми працювали лише на комп'ютері розробників, а тому застосовувати їх на інших комп'ютерах було практично не можливо. Тому кожна дослідницька група розробляла свою власну систему конкордансування і застосовувала її до даних, до яких мала місцевий доступ. Конкордансери першого покоління були функціонально обмеженими. Вони могли укладати лише KWIC конкорданси, а інші маніпуляції з даними, наприклад створення списку усіх вжитих у корпусів слів, виконувалися окремими програмами. Крім того, перші конкордансери не розрізняли деякі символи, зокрема діакритичні знаки. З появою перших конкордансерів як раз і виникла необхідність створення стандартних правил анотування корпусних даних, щоб можна було безпечно передавати корпуси між дослідними центрами та створювати конкордансери, що будуть працювати на різних комп'ютерах та на різних корпусах. Досвід використання перших конкордансерів засвідчив, що замість однофункціональних програм, на зразок генераторів частотних списків, краще розробляти програми-конкондансери, які б поєднували різні операції та дозволяли різноаспектні маніпуляції з корпусними даними.

Конкордансери другого покоління з'явилися з появою машин сумісних з IBM-продукцією. Стало можливим розробляти конкордансери, які могли без проблем працювати на різних комп'ютерах, наприклад, the KAYE (1990), the Longman Mini-Concordance (1989), Micro-OCP (1988). Звичайно, ці програми набагато спростили роботу корпусних лінгвістів. Якщо до цього над корпусом необхідно було працювати у

команді з програмістом, який розробляв необхідне програмне забезпечення, то з ПК-конкордансерами будь-який лінгвіст, який умів включати і працювати з ПК, міг використовувати корпуси. Ефект був миттєвим, і з кінця 1980-х років корпусна лінгвістика почала блискавично розвиватися. Проте слід зауважити, що конкордансери другого покоління теж мали багато недоліків і за деякими функціями були нічим не кращі за конкордансери попереднього покоління. Вони виконували мало функцій за винятком KWIC конкордансування, сортували за алфавітом правий та лівий контекст пошукового слова, продукували списки слів та надавали базові статистичні дані по корпусу. Крім того, були неузгодженості щодо кодування деяких символів, а також вони не могли оперувати великими обсягами даних. Разом з тим зростаючі потужності комп'ютерів сприяли появі конкордансерів третього покоління.

Конкордансери третього покоління включають такі добре відомі системи, як Wordsmith, MonoConc, AntConc, Xiara. Ці конкордансери можуть обробляти величезні масиви даних на ПК (наприклад, 100-мільйонний Британський Національний Корпус оперується конкордансером Xiara). Крім того, ці програми пропонують більше функцій та роблять більш складні статистичні операції. Конкордансери цього покоління можуть обробляти тексти на значно більшій кількості систем письма. Примітно, що конкордансери третього покоління дуже схожі один на одного, особливо в аспекті їх базових функцій: конкорданси, частотні списки, сполучуваність (колокації), аналіз ключових слів.

Що ж стосується конкордансерів четвертого покоління, то вони разюче схожі у функціональному аспекті з конкордансерами третього покоління. Проте конкордансери цього покоління розроблялися не з метою розширення способів

аналізу даних, а для вирішення таких завдань, як обмежена потужність стаціонарних ПК; проблеми, що виникають з несумісними ПК операційними системами; юридичні/правові обмеження на розповсюдження корпусів. З метою дотримання авторських прав на тексти, що включені до корпусу, а також з метою забезпечення швидкого та масштабного доступу сучасні корпуси почали розповсюджувати через спеціальні веб інтерфейси. На свій запит користувачі в межах секунд отримують обмежений обсяг контексту пошукового слова (1-3 речення) (наприклад, BNC, the Polish PELCRA, the Hellenic National Corpus). Саме цей тип передачі даних за моделлю клієнт-сервер через специфічний спосіб – World Wide Web – є визначальною рисою цих корпусів. Як приклад програми аналізу корпусних даних четвертого покоління можна назвати систему, розроблену для Британського Національного корпусу Марком Дейвісом, що була згодом ним розширена для забезпечення доступу до низки значних за обсягом корпусів через сайт corpus.byu.edu [McEnery, Hardie 2012: 37-48].

Для ілюстрації роботи конкондансеру/корпусного менеджеру розглянемо програму AntConc[54] (розробник Laurence Anthony). Ця програма є однією із найзручніших безкоштовних комп'ютерних програм для роботи з корпусом, що може використовуватися у навчальних і в дослідницьких цілях, оскільки конкордансер є ефективним засобом для опанування іноземної мови, сприяючи засвоєнню вокабуляру, сполучуваності, граматики та стилістики. AntConc має зручний і зрозумілий інтерфейс та пропонує опції потужного конкордансера, генератора частотного списку, аналізатора сполучуваності, а також візуалізує входження досліджуваних

[54] http://www.antlab.sci.waseda.ac.jp/

слів у досліджуваному масиві. Ознайомимося із деякими опціями детальніше.

На рис.1 подано скріншот AntConc, на якому представлено результати роботи опції конкордансу. Дана опція дозволяє здійснювати пошук необхідного слова, видаючи кількість знайдених прикладів та контекст, в якому воно вживається у файлах корпусу. Програма надає можливість пошуку за окремими словами, словосполученнями, з урахуванням чи без урахування регістру клавіатури.

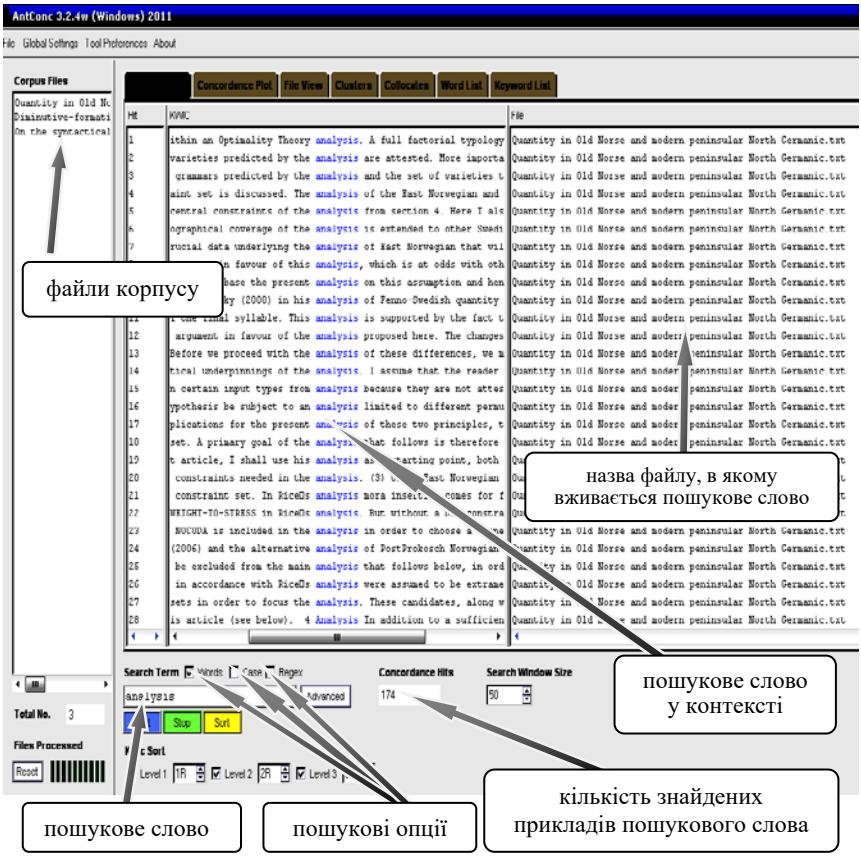


Рис. 1. Опція «concordance» конкондансеру AntConc

Графік розподілу слова у файлах корпусу показує особливості вживання пошукового слова/словосполучення у частинах тексту файлів корпусу. Розглянемо, як розподіляється словосполучення *in this paper* у частинах наукової статті (див. рис.2):

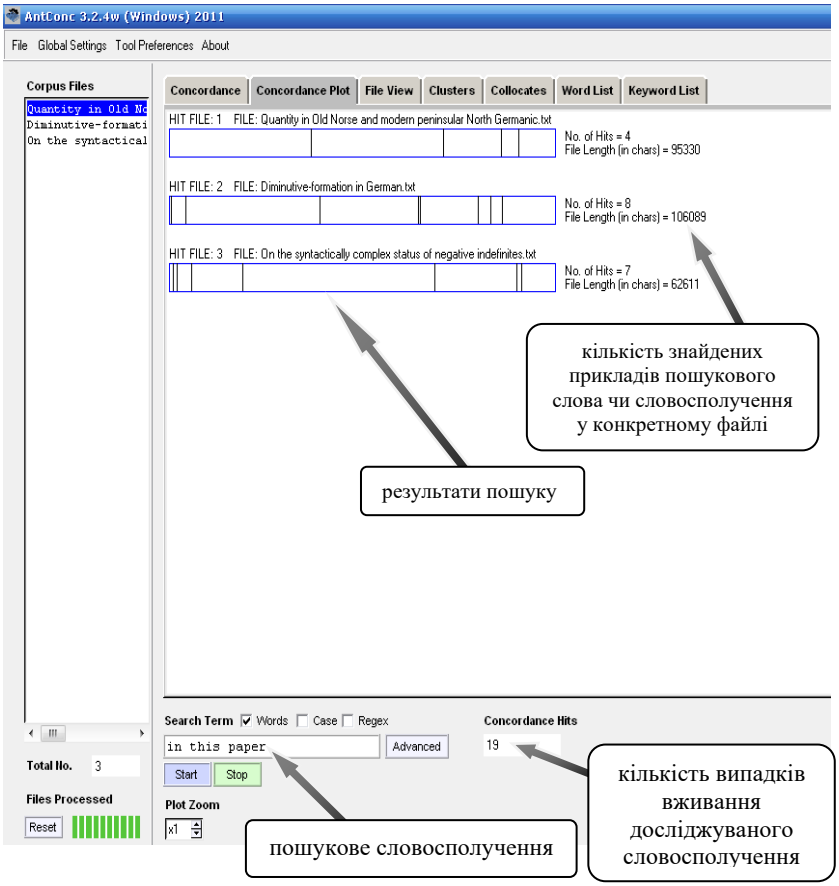


Рис. 2. Результати розподілу словосполучення in this paper у частинах наукової статті

Важливою опцією цієї програми є функція Word Clusters (аналізатор сполучуваності). Ця функція дозволяє прослідкувати особливості оточення заданих слів/словосполучень в корпусі, регулюючи довжину ланцюжка (див. рис.3):

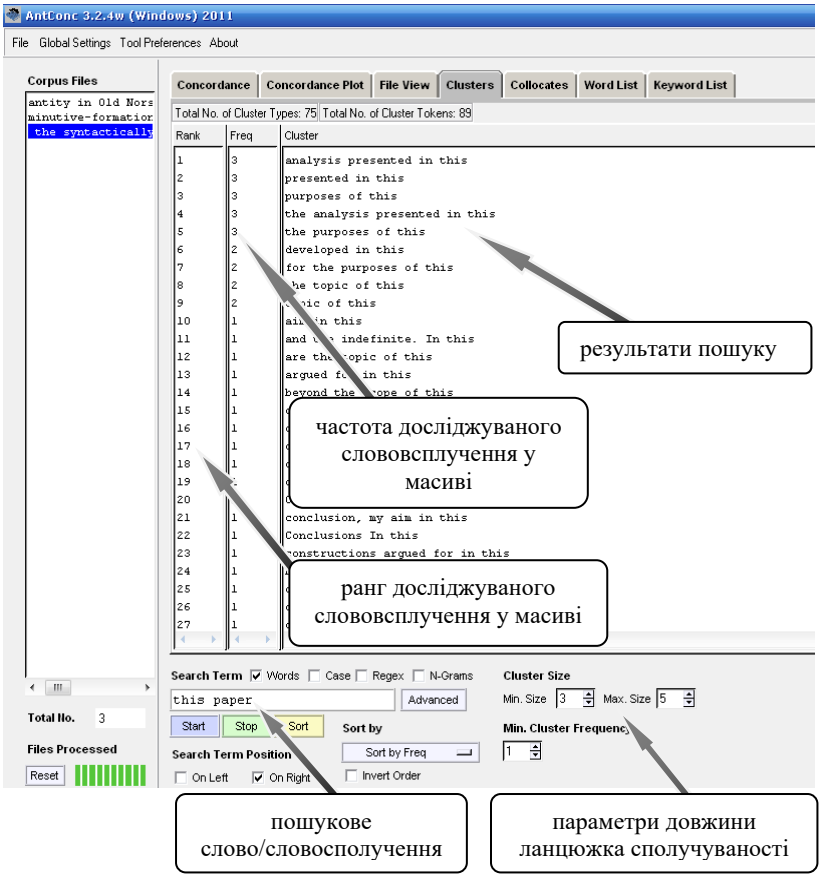


Рис.3. Результати сполучуваності словосполучення this paper у досліджуваному масиві

Отже, конкордансер *AntConc* – це зручна і проста у використанні програма. Хоча вона не пропонує всіх сервісів та функцій, що є у популярних комерційних проектах, ця програма

має низку важливих сервісів, необхідних для аналізу корпусу. Більше про можливості *AntConc* можна дізнатися на сайті розробників[55].

У сучасному розумінні корпус – це завжди комп'ютерна база даних, і в процесі його створення безсумнівним є використання спеціальних програм. Серед цих програм особливе місце займають програми автоматичної розмітки. Розмітка корпусів є трудомісткою операцію, особливо враховуючи розміри сучасних корпусів. Для морфологічного й синтаксичного аналізу існують різні програмні засоби, які прийнято називати відповідно теггери (taggers) і парсери (parsers). У результаті роботи програм автоматичного морфологічного аналізу кожній лексичній одиниці приписуються граматичні характеристики, включаючи частину мови, лему (нормальну форму) і набір грамем (наприклад, рід, число, відмінок, перехідність і т.п.). У результаті роботи програм автоматичного синтаксичного аналізу фіксуються синтаксичні зв'язки між словами й словосполученнями, а синтаксичним одиницям приписуються відповідні характеристики (тип речення, синтаксична функція словосполучення й т.п.).

Однак автоматичний аналіз природньої мови не безпомилковий: він, як правило, дає кілька варіантів аналізу для однієї лексичної одиниці (слова, словосполучення, речення). У цьому випадку говорять про граматичну омонімію. Зняття неоднозначності (морфологічної, синтаксичної) у цілому є однією з найважливіших і найскладніших завдань комп'ютерної лінгвістики. При створенні корпусів для зняття неоднозначності використовуються автоматичні й ручні способи. Корпуси нового покоління включають сотні мільйонів слів, тому висуваються

[55] http://www.antlab.sci.waseda.ac.jp/antconc_index.html

принципи розробки систем, які б мінімізували втручання людини. Автоматичне зняття морфологічної або синтаксичної омонімії, як правило, ґрунтується на використанні інформації більш високого рівня (синтаксичного, семантичного) із застосуванням статистичних методів.

Один із найвідоміших та надійніших тегерів для англійської мови є CLAWS (Constituent-Likelihood Automatic Word Tagging System)[56], розроблений в Ланкастерському університеті. Зазначається, що розмітка тексту вказаним тегером досягає 97% точності. Ця система була використана для анотування Британського Національного Корпусу. Частиномовні тегери розроблені і для інших мов, наприклад: французької, іспанської, німецької, шведської та китайської.

Запитання і завдання для самоперевірки.

1. Охарактеризуйте етапи створення корпусу. Чи можливо опустити певні етапи? За яких умов?
2. Назвіть процедури, що використовуються під час створення корпусу.
3. Наведіть приклади корпусних менеджерів. У чому полягають їх особливості?
4. Поясніть сутність стандартизації даних корпусу.
5. Укладіть міні-корпус англомовних наукових статей з лінгвістики. Для цього з сайту провідного міжнародного Інтернет-видавництва **SAGE** (http://online.sagepub.com/browse/by/discipline) відберіть 5 наукових журналів по лінгвістиці за останні 2-3 роки. З обраних журналів виберіть по 2 статті з будь-яких двох номерів одного журналу. Далі виконайте такі кроки:

[56] http://ucrel.lancs.ac.uk/claws/

Крок 1. Переформатуйте файли PDF в формат txt. У меню "Зберегти як" вибрати TXT файл.

Крок 2. Здійсніть рецензування текстових документів. Видаліть всю інформацію, що не є релевантною (інформацію про інтернет-видавництво SAGE, формули, таблиці, діаграми, малюнки, посилання, список використаних джерел, додатки, якщо є), при цьому залишаючи всі композиційні частини статті: Abstract, Introduction, Methods, Results і т.ін. Текстові файли збережіть у такому вигляді:

1) Автор.
2) Назва статті.
3) Назва журналу, № журналу, № сторінок, на яких представлена стаття.
4) Текст статті.

Крок 3. Збережіть файловий документ, назвавши його по автору і загаловку статті (наприклад, Biber_Variation in Scientific Discourse).

Крок 4. Зберегти файли у папці під назвою журналу.

Розділ 6.
Можливості використання корпусів у лінгвістичних дослідженнях

6.1. Сфери застосування лінгвістичних корпусів

Поява електронних корпусів суттєво змінила спосіб дослідження мовних явищ. Репрезентативні, збалансовані, значні за обсягом корпуси надають інформацію про динаміку процесів зміни лексичного складу мови, функціонування граматичних та синтаксичних одиниць, дозволяють здійснювати аналіз вживання мовних одиниць у різних стилях, жанрах, у різних авторів і т.ін. Корпуси надають недоступні для безпосереднього спостереження дані про реалізацію мовної системи, що в свою чергу відображається у словниках, граматиках та навчальних матеріалах. Якщо ранні корпусні дослідження зосереджувалися головним чином лише на вживаності мовних одиниць (н. д. частота вживання певних іменників, дієслів та прикметників), то з появою потужних комп'ютерних технологій дослідники отримали змогу визначати та аналізувати більш складні лексичні, граматичні та синтаксичні явища.

На сьогоднішньому етапі свого існування корпуси масштабно використовуються як лінгвістичними, так і нелінгвістичними галузями. Так, лінгвісти-теоретики використовують корпуси у якості експериментальної бази для перевірки гіпотез і доведення своїх теорій. Прикладні лінгвісти (викладачі, перекладачі й т.п.) застосовують комп'ютерні корпуси для навчання іноземній мові і розв'язання своїх професійних завдань. Комп'ютерні лінгвісти послуговуються статистичними й лінгвістичними закономірностями, що були

виявлені на матеріалі корпусу, для створення комп'ютерних моделей мови. У соціолінгвістичних дослідженнях корпуси використовуються для вивчення мовного розмаїття (наприклад, соціолектів, регістрів). Інші фахівці з мови (літературознавці, редактори) також у ряді випадків звертаються до даних корпусу. Крім того, корпуси використовуються для аналізу дискурсу, в літературознавстві, перекладознавстві, судовій лінгвістиці. Фахівці із суспільних наук (історики, соціологи) теж можуть вивчати свої об'єкти через мову, використовуючи такі параметри текстів, як період, автор або жанр. Нарешті, корпуси використовуються для розробки й налаштування різних автоматизованих систем (машинний переклад, розпізнавання мовлення, інформаційний пошук).

6.2. Лексикографічні та граматичні дослідження на матеріалі корпусу

У галузі лексикографії корпусна лінгвістика не лише впровадила нові методи аналізу мовного матеріалу, але і значно розширила межі наукового дослідження. Використання корпусу в лексикографічній практиці замістило паперові картки автоматичним пошуком слів у корпусі та суттєво переорієнтувало підхід до відбору слів до сло́вника, замінивши відбір слів, обумовлений особистими перевагами лексикографа, більш об’єктивними критеріями: даними про частоту та дистрибуцію слів у певних текстах чи жанрах. Переважна більшість сучасних словників є корпусоінформованими, а не «укладачеінформованими» [McEnery, Gabrielatos, 2006: 48]. Одним із останніх нововведень у лексикографічній практиці стало розміщення у словниковій статті крім традиційної

інформації про значення та семантичні відношення одиниці також й інформації про її сполучуваність та дистрибуцію.

Ініціатором першого суто корпусобазованого словника англійської мови COBUILD (1987) був Джон Синклер. Сьогодні ж значна кількість словників укладається на основі даних лінгвістичних корпусів. Наприклад, практично всі відомі словники англійської мови (Collins, Webster, Macmillan і т.д.) створюються на матеріалі власних, закритих для загального доступу корпусів, котрі дозволяють зробити словник репрезентативним та відобразити в ньому останні тенденції в мові. Адже вивчення корпусів дозволяє одержувати точні дані про лексичний склад мови, певної підмови, жанру чи індивідуального стилю письменника, виявити останні зміни в лексичному складі мов, різні його варіації (наприклад, поява й зникнення неологізмів) та включити ці дані до словника. Саме сучасність цих словників, побудованих на прикладах із «живих» реальних текстів, сприяє їхньому комерційному успіху.

Вже звичною практикою стало використання корпусів для укладання довідкової літератури – граматик, навчальних словників та довідників як для носіїв мови, так і для тих, хто її вивчає. Граматики все більше звертають увагу на лексику, а словники (особливо навчальні) включають у статті сло́вника й граматичну інформацію. На даний момент вже існує декілька корпусобазованих граматик англійської мови, наприклад, загальних (Collins COBUILD English Grammar, 1990; Longman Grammar of Spoken and Written English, 1999) та таких, що акцентують увагу на певних аспектах (Collins COBUILD Grammar Patterns 1: Verbs, 1996).

Таким чином, хоча це і не швидкий процес, проте мало-помалу все більше навчального матеріалу на сьогоднішній день стає „корпусоінформованим”. Зростає тенденція до вкладання

коштів у розробку значних за обсягом корпусів такими великими видавцями і розробниками навчальних матеріалів, як Cambridge University Press, Oxford University Press, Pearson-Longman, Collins-COBUILD and Macmillan.

6.3. Використання корпусів у навчанні іноземної мови

Використання корпусів стає дедалі популярнішим у навчанні та вивченні мови. Вчителі та учні все частіше звертаються до корпусу для здійснення швидкого різнорівневого пошуку та аналізу найрізноманітнішої лінгвальної інформації, а конкорданс визнається зручним інструментом для ефективного опанування мовою. Корпуси текстів слугують неоціненним ресурсом реальних прикладів вживання тих чи інших лексичних та граматичних одиниць, за умови, що учні і вчителі мають доступ до корпусів та обізнані з технікою роботи з корпусом.

Орієнтоване на дані навчання (data-driven learning) іноземній мові, коли корпус стає головним джерелом та способом отримання лінгвістичних знань й умінь, отримує все більше прихильників серед сучасних студентів та вчителів. При цьому підході базоване на корпусі програмне забезпечення використовується для підтримки інтерактивної навчальної діяльності, а залучення «живого» мовлення з корпусу сприяє пробудженню у студентів інтересу до процесу вивчення мови та заохоченню їх до самостійного вивчення автентичного мовного матеріалу. Студент виконує активну роль дослідника мови, а вчитель лише направляє та сприяє навчанню [Zhukovska 2011, Zhukovska 2012]. Серед найбільш популярних корпусно-базованих програм та освітніх пакетів варто відзначити The

English Collocations CD-ROM компанії Cobuild[57], WordPilot [58], ClozeMaker[59].

Корпуси як джерела емпіричних даних відіграють вагому роль у лінгводидактиці. Визнано, що традиційні шкільні граматики та підручники досить часто проілюстровані штучно створеними чи відредактованими прикладами вживання певної лінгвістичної одиниці чи явища. Такі приклади є малоефективними для успішного навчання учнів використанню мови як засобу комунікації, оскільки вивчене ними може кардинально різнитися від реального функціонування мови. У цьому відношенні дослідження корпусів природньої мови надає більш точний та деталізований опис мови, що у свою чергу відображається у змісті педагогічних матеріалів, граматик та словників, а також дизайні навчальних програм та підручників.

Іншою сферою інновацій, що безпосередньо пов'язана із корпусною лінгвістикою – це укладання та аналіз учнівських корпусів (*learner corpora*), які надають важливу для вчителя інформацію про те, як учні вивчають та засвоюють мову, про типові помилки у вживанні лексичних,граматичних та синтаксичних одиниць. Вагомим аспектом для методики навчання мови є також використання паралельних корпусів, що уможливлюють доступ до безпосередніх перекладних еквівалентів слова, фрази, синтаксичної конструкції. Такий інструмент не тільки полегшує створення перекладних словників, базованих на реальних прикладах перекладацької практики, а і сприяє легшому вивченню та навчанню іноземної мови відповідно до комунікативної стратегії навчання. Застосовуючи двомовні конкорданційні програми, комп'ютерна

[57] http://www.athel.com/cobuild/collscd.html

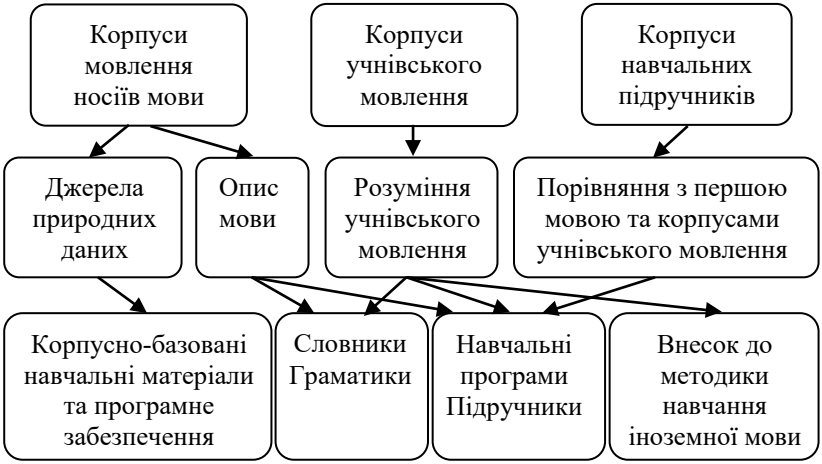
[58] http://www.compulang.com

[59] http://www.edict.com.hk/clozemaker

система висвітлює обидва тексти поряд, і студент легко та швидко перекладає текст, оскільки не змушений шукати кожне невідоме слово у словнику. А от корпуси підручників з іноземної мови надають можливість порівняти використання рідної мови в різних контекстах та мови, з якою мають справу учні у посібниках (див напр.: Harwood 2005; Römer 2004). Аналіз таких невідповідностей допоможе при відборі текстів, які повинні бути включені до навчальних матеріалів. Огляд корпусів начальних текстів також може виявити, якою мірою тексти підручників впливають на усне та писемне мовлення учнів.

Крім того, останнім часом корпуси стали активно використовуватися для створення матеріалів для тестування та навчання вчителів з метою підвищенню рівня їх професіоналізму.

Отож, способи застосування корпусів у методиці навчання іноземної мови можливо графічно зобразити таким чином [McEnery, Gabrielatos, 2006: 51]:



Перелічені вище можливості використання корпусів не вичерпують всі сфери їх експлуатації. Із збільшенням обсягів та

удосконаленням технічного оснащення дослідницький потенціал корпусів невпинно зростає, що робить їх цінним лінгвістичним ресурсом.

Запитання і завдання для самоперевірки.

1. Окресліть основні сфери застосування корпусних даних.
2. Охарактеризуйте можливості експлуатації корпусів у лексикографічних та граматичних дослідженнях. Чи користуєтеся Ви корпусобазованими / корпусоінформованими словниками, граматичними довідниками?
3. Які переваги використання даних корпусів у навчанні та вивченні іноземної мови? Які труднощі/недоліки Ви вбачаєте у впровадженні корпуснобазованого підходу до вивчення іноземної мови?
4. Чи користуєтесь Ви корпусом? Якого типу інформацію Ви шукаєте в корпусі?

Додаток А

Найбільш відомі корпуси та ресурси

1. National corpora

The British National Corpus
The American National Corpus
The Polish National Corpus
The Czech National Corpus
The Hungarian National Corpus
The Russian Reference Corpus
The CORIS corpus
The Hellenic National Corpus
The German National Corpus
The Slovak National Corpus
The Modern Chinese Language Corpus
The Sejong Balanced Corpus
Other National corpora

2. Monitor corpora

The Bank of English
The global English Monitor Corpus

3. Corpora of the Brown family

Brown
Frown
LOB
Pre–LOB
FLOB
Kolhapur
ACE

WWC
LCMC

4. Synchronic corpora

The International Corpus of English
The Longman/Lancaster Corpus
The Longman Written American Corpus
The CREA corpus of Spanish
The LIVAC corpus of Chinese

5. Diachronic corpora

The Helsinki Corpus of English Texts
The ARCHER corpus
The Lampeter Corpus of Early Modern English Tracts
The Dictionary of Old English Corpus in Electronic Form
Early English Books Online
The Corpus of Early English Correspondence
The Zurich English Newspaper Corpus
The Innsbruck Computer Archive of Machine–Readable English Texts
The Corpus of English Dialogues
A Corpus of Late Eighteenth–Century Prose
A Corpus of Late Modern English Prose

6. Spoken corpora

The London–Lund Corpus
SEC, MARSEC and Aix–MARSEC
The Bergen Corpus of London Teenage Language
The Cambridge and Nottingham Corpus of Discourse in English
The Spoken Corpus of the Survey of English Dialects
The Intonational Variation in English Corpus
The Longman British Spoken Corpus

The Longman Spoken American Corpus
The Santa Barbara Corpus of Spoken American English
The Saarbrücken Corpus of Spoken English
The Switchboard Corpus
The Wellington Corpus of Spoken New Zealand English
The Limerick corpus of Irish English
The Hong Kong Corpus of Conversational English

7. Academic and professional English corpora

The Michigan Corpus of Academic Spoken English
The British Academic Spoken English corpus
The Reading Academic Text corpus
The Academic Corpus
The Corpus of Professional Spoken American English
The Corpus of Professional English

8. Parsed corpora

The Lancaster–Leeds Treebank
The Lancaster Parsed Corpus
The SUSANNE corpus
The CHRISTINE corpus
The LUCY corpus
ICE–GB
The Penn Treebank
Parsed historical corpora

9. Developmental and learner corpora

The Child Language Data Exchange System
The Louvain Corpus of Native English Essays
The Polytechnic of Wales corpus
The International Corpus of Learner English
The LINDSEI corpus

The Longman Learners' Corpus
The Cambridge Learner Corpus

10. Multilingual corpora

The Canadian Hansard Corpus
The English–Norwegian Parallel Corpus
The English–Swedish Parallel Corpus
The Oslo Multilingual Corpus
The ET10/63 and ITU/CRATER parallel corpora
The IJS–ELAN Slovene–English Parallel Corpus
The CLUVI parallel corpus
European Corpus Initiative Multilingual Corpus I
The MULTEXT corpora
The PAROLE corpora
Multilingual Corpora for Cooperation
The EMILLE Corpus
The BFSU Chinese–English Parallel Corpus
The Babel Chinese–English Parallel Corpus
Hong Kong Parallel Text

11. Non–English monolingual corpora

The COSMAS corpora
The CETEMPúblico Corpus
The INL corpora
The CEG corpus
The Scottish Corpus of Texts and Speech
The Prague Dependency Treebank
Academia Sinica Balanced Corpus
Sinica Treebank
Penn Chinese Treebank
Spoken Chinese Corpus of Situated Discourse

12. Well–known distributors of corpus resources

CSLU
ELRA
ELSNET
ICAME
OTA
TRACTOR

Список використаних джерел

1. Балог В., Балог О. «Словарь української мови» за ред. Б. Грінченка (1907–1909 рр.): електронна версія / В. Балог, О. Балог // Лексикографічний бюлетень: Зб. наук. пр. – К., 2007. – Вип. 16. – С. 78-82.
2. Баранов А.Н. Введение в прикладную лингвистику / А.Н. Баранов. – Москва, 2001. – 358 с.
3. Демська-Кульчицька О. Дещо про класифікацію текстових корпусів / О. Демська-Кульчицька // Наукові записки. Серія: Мовознавство. – 2004. – 1 (11). – С. 153-157.
4. Демська-Кульчицька О. Корпусна рецепція тексту / О. Демська-Кульчицька // Наукові записки. Т. 111. Сер. Філологічні науки. – 2010. – С.-3-6.
5. Демська-Кульчицька О. Один з аспектів морфологічної анотації (до проблеми побудови тега) / О. Демська-Кульчицька // Українська мова. – 2004. – № 1. – С. 26-38.
6. Демська–Кульчицька О. Основи національного корпусу української мови [монографія] / О. М. Демська-Кульчицька – Київ, 2005. – 218 с.
7. Демська-Кульчицька О.М. Репрезентативність як ознака текстового корпусу / О.М. Демська-Кульчицька. – Українська мова. – №3, 2005. – С. 100-107.
8. Демська–Кульчицька О. М. Базові поняття корпусної лінгвістики / О.М. Демська–Кульчицька // Українська мова. – 2003 – №1. – С. 42-47.
9. Жуковська В.В. Ресурси корпусної лінгвістики у дослідженні історичної динаміки мови / В.В.Жуковська // Матеріали міжнародної наукової конференції «Слово і речення: синтактика, семантика, прагматика» – К.: Київ. ун-т ім. Б.Грінченка, 2013. – С.151-156.

10. Жуковська В.В. Корпусна лінгвістика: історична перспектива та сучасний стан / В.В.Жуковська // Материали за 8-а международна научна практична конференция, «Ключови въпроси в съвременната наука», – 2012. Том 18. Филологични науки. София. «Бял ГРАД-БГ» ООД – 72 с.
11. Захаров В.П. Корпусная лингвистика: Учебно–метод. пособие. / В.П. Захаров – СПб., 2005. – 48 с.
12. Захаров В.П., Богданова С.Ю. Корпусная лингвистика: учебник для студентов гуманитарных вузов / В.П. Захаров, С.Ю. Богданова. – Иркутск: ИГЛУ, 2011. – 161 с.
13. Карпіловська Є.А. Вступ до прикладної лінгвістики: комп'ютерна лінгвістика: Підручник / Є.А.Карпіловська. — Донецьк: ТОВ «Юго-Восток, Лтд», 2006.— 188 с. Колпакова Г.В. Корпусная лингвистика и лексикография / Г.В. Колпакова // Электронный научно-образовательный журнал ВГПУ «Грани познания» – №2 (12). – С. 1-9. [Электронный ресурс]. – Режим доступа: www.grani.vspu.ru
14. Копотев М., Мустайоки А. Современная корпусна русистика / М. Копотев, А. Мустайоки // Инструментарий русистики: корпусные подходы. – Хельсинки, 2008. – С. 7-24.
15. Лозинська О. Фразеологічна одиниця KRĘCIĆ NOSEM у корпусах текстів польської мови / О. Лозинська // Проблеми слов'янознавства. – 2009. – Вип. 58. – С. 221-230.
16. Максимів О. Корпус перської мови як джерело матеріалу для навчальних словників-мінімумів / О. Максимів // Вісник Львівського університету. Серія філологічна. – № 45. – С.164-169.
17. Організація навчально-дослідницької роботи школярів із використанням українського національного лінгвістичного корпусу на матеріалах вивчення функціональних

можливостей сполучника та дослідження концептів «праця», «робота» у текстах творів Тараса Шевченка (методичні рекомендації) / С. О. Довгий, К. Я. Климова, О. Б. Комова, О. В. Лісовий, М. В. Надутенко, Н. М. Сидорчук, О. Є. Стрижак, В. А. Широков, В. В. Шкурко, К. М. Якименко /(Під редакцією доктора технічних наук, кандидата фізико-математичних наук, академіка НАНУ В. А. Широкова і кандидата технічних наук О. Є. Стрижака) – Київ, 2012. – 66 с.

18. Перцов Н.В. К суждениям о фактах русского языка в свете корпусных данных / Н.В. Перцов // Русский язык в научном освещении. – 2006. – № 1 (11). – С. 227-245.
19. Плунгян В.А. Корпус как инструмент и как идеология: о некоторых уроках современной корпусной лингвистики / В.А. Плунгян // Русский язык в научном освещении. – №2 (16), 2008. – С.7-20.
20. Плунгян В.А. Почему современная лингвистика должна быть лингвистикой корпусов?: (публичная лекция, прочитанная 01.10.2009) [Электронный ресурс] / В.А. Плунгян – Режим доступа: http://www.polit.ru/lectures/ 2009/10/23/corpus.html.
21. Плунгян В.А., Рахилина Е.В. Новые возможности национального корпуса русского языка как уникального интернет-ресурса / В.А. Плунгян, Е.В. Рахилина // Русский язык и литература в образовательной системе Армении: проблемы и перспективы. – Ер. «Лингва», 2009. – С. 21-35.
22. Рычкова Л.В. Праблема састауных аб'ектау у корпусах славянскімоу і лінгвістычных базах дадзеных / Л.В. Рычкова // Мовознауства. Літаратура. Культуралогія. Фальклярыстыка. XIII Міжнародньї з'езд славыстау.

Даклады беларускай дзлегацьіі. – Мінськ, 2003. – С. 184-195.

23. Селіванова О.О. Корпусна лінгвістика // О. Селіванова // Сучасна лінгвістика: напрями та проблеми: підручник. – Полтава: Довкілля-К, 2008. – С. 667-669.
24. Шаров С.А. Представительный корпус русского языка в контексте мирового опыта / С.А. Шаров // НТИ. Сер.2. – 2003. – №6. – С. 9-17.
25. Шведова М., Січінава Д. Корпусна лінгвістика та лексико-граматична типологія /М.Шведова, Д.Січінава // Українське мовознавство. Київський національний університет імені Тараса Шевченка – № 43. – 2013. – С. 95-103.
26. Широков В.А., Бугаков О.В., Грязнухіна Т.О. та ін. Корпусна лінгвістика / В.А. Широков, О.В. Бугаков, Т.О. Грязнухіна та ін. – К.: Довіра, 2005. – 471 с.
27. Шмелев А.Д. Языковые факты и корпусные данные / А.Д. Шмелев [Электронный ресурс]. – Режим доступа: http://ruscorpora.ru
28. Aarts J., Meijs. W. Corpus Linguistics: Recent developments in the Use of Computer Corpora in English Language Research / J. Aarts, W. Meijs. – Amsterdam: Rodopi, 1984. – 425 p.
29. Aijmer K., Stenstrom A.-B. Discourse patterns in spoken and written corpora / K. Aijmer, A.-B. Stenstrom. – Amsterdam:John Benjamins Publishing Company, 2004. – 321 p.
30. Altenberg B., Granger S. Lexis in contrast: corpus-based approaches / Altenberg B., Granger S. – Amsterdam: John Benjamins Publishing Company, 2002. – 337 p.
31. Anderman G., Rogers M. Incorporating Corpora. / G. Anderman, M. Rogers // The Linguist and the Translator. –

Multilingual Matters LTD, Clevedon Buffalo Toronto, 2008. – 278 p.

32. Archer D. Questions and answers in the English courtroom (1640–1760: a sociopragmatic analysis) / D. Archer. – Amsterdam: John Benjamins Publishing Company, 2005. – 241 p.
33. Asmussen J. Korpuslinguistische Verfahren zur Optimierung lexikalisch-semantischer Beschreibungen / J. Asmussen // Sprachkorpora – Datenmengen und Erkenntnisfortschritt (Hrsg. von W. Kallmeyer, G. Zifonun). Institut fur Deutscher sprache. Jahrbuch 2006. Berlin – N.J.: Walter de Gruyter, 2007 – S. 123-151.
34. Aston G., Burnard L. The BNC Handbook. Exploring the British National Corpus with SARA / G. Aston, L. Burnard. – Cambridge: Edinburgh University Press, 1998. – 250 p.
35. Baker P. Public discourses of gay men / P. Baker. – London: Routledge, 2005. – 266 p.
36. Baker P., Hardie A., McEnery T. Glossary of Corpus Linguistics / P. Baker, A. Hardie, T. McEnery. – Edinburgh University Press, 2006 – 192 p.
37. Baker P., McEnery T. A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts / P. Baker, T. McEnery // Language and Politics. – 2005. – №4 (2). – P. 197-226.
38. Biber D. Representativeness in corpus design / D. Biber // Literary and Linguistic Computing. – 1993. – № 8 (4). – P. 243-57.
39. Biber D. Using corpus-based methods to investigate grammar and use: some case studies on the use of verbs in English / D. Biber // In Simpson and Swales (eds.). – 2001. – P. 101-115

40. Biber D., Conrad S., Reppen R. Corpus Linguistics: Investigating Language Structure and Use / D. Biber, S. Conrad, R. Reppen. – Cambridge University Press, 1998. [Електронний ресурс]. – Режим доступу: http://books.google.com.ua/books?id=2h5F7TXa6psC&redir_esc=y
41. Biber D., Johansson S., Leech G., Conrad S., Finegan E. Longman Grammar of Spoken and Written English / D. Biber, S. Johansson, G. Leech, S. Conrad, E. Finegan. – London: Longman, 1999. – 613 p.
42. Brazil D. A Grammar of Speech / D. Brazil. – Oxford University Press, 1995. – 264 p.
43. Burrows J. The Englishing of Juvenal: computational stylistics and translated texts / J. Burrows // Style. – 2002. – №36 (4). – P. 677-679.
44. Carter R., McCarthy M. (1999) The English get-passive in spoken discourse: description and implications for an interpersonal grammar / R. Carter, M. McCarthy // English Language and Linguistics. – 1999. – №3 (1). – P. 41-58.
45. Carter R., McCarthy M. Grammar and the spoken language / R. Carter, M. McCarthy // Applied Linguistics – 1995. – №16 (2). – P. 141-158.
46. Charteris-Black J. Corpus approaches to critical metaphor analysis / J. Charteris. –Black Basingstoke: Palgrave-Macmillan, 2004. – 417 p.
47. Davies M. The 385+ million word corpus of contemporary American English (1990– 2008+): design, architecture and linguistic insights / M. Davies // International Journal of Corpus Linguistics. – 2009. – №14 (2). – P. 159-190.
48. Deignan A. Metaphor and corpus linguistics / A. Deignan // Converging Evidence in Language and Communication

Research 6. – Amsterdam: John Benjamins Publishing Company, 2005. – 235 p.

49. Duffley P.J. The gerund and the to-infinitive as subject / P.J. Dufflei // Journal of English Linguistics. – 2003. – №31 (4). – P. 324-352.
50. Fillmore Ch. Corpus linguistics or computer-aided armchair linguistics / Ch. Fillmore // Directions in corpus Linguistics: Proceedings of Nobel Symposium 82. – Berlin, 1992. – P. 35-60.
51. Francis W. Language Corpora B.C. / W. Francis // In: Jan Svartvik (ed.) Directions in Corpus Linguistics. Proceeding of Nobel 82, Stockholm, 4–8. August, 1991, Berlin –New York: Mouton de Gruyter. – P.17-35.
52. Fraser M. (1996), Tools and Techniques for Computer–assisted Biblical Studies / M. Fraser // Paper delivered to the New Testament Research Seminar, Faculty of Theology, University of Oxford, June 1996. [Електронний ресурс]. – Режим доступу: http://users.ox.ac.uk/~mikef/pubs/NT_Seminar_ Oxford_Fraser_1996.html.
53. Ghadessy M., Alex H., Robert L. Small corpus studies and ELT: theory and practice / M. Ghadessy , H. Alex, L. Robert. – John Benjamins Publishing Company, 2001. – 445 p.
54. Ghadessy M., Gao Y. Small corpora and translation: comparing thematic organization in two languages / M. Ghadessy, Y. Gao // In M. Ghadessy, A. Henry and R. L. Roseberry (eds.) Small Corpus Studies and ELT: Theory and Practice. – Amsterdam and Philadelphia: John Benjamins Publishing Company, 2001. – P. 335-359.
55. Gilquin G. (2003) Causative get and have: so close, so different / G. Gilquin // Journal of English Linguistics. – 2003. – №31 (2). – P. 125-148.

56. Gries S. T. (2003) Towards a corpus based identification of prototypical instances of constructions / S.T. Gries // Annual Review of Cognitive Linguistics. – 2003. – №1. – P. 1-27.
57. Gries S., Stefanowitsch, A. Extending collostructional analysis: a corpus-based perspective on 'alternations' / S. Gries, A. Stefanowitsch // International Journal of Corpus Linguistics. – 2004. – №9 (1). – P. 97-129.
58. Hardt-Mautner G. Only connect: critical discourse analysis and corpus linguistics / G. Hardt-Mautner // UCREL Technical Papers 6. – Lancaster University, 1995. – 325 p.
59. Hunston S. Colligation, lexis, pattern and text / S. Hunston // In Scott and Thompson (eds.).– 2001. – P. 13-33.
60. Hunston S., Francis G. Pattern grammar / S. Hunston, G. Francis. – Amsterdam: John Benjamins Publishing Company, 2000. – 288 p.
61. Kennedy G. Introduction to corpus linguistics / G. Kennedy. – Harlow, Essex: Longman, 1998. – 315 p.
62. Koller V., Mautner, G. Computer applications in critical discourse analysis / V. Koller, G. Mautner // In C. Coffin, A. Hewings, and K. O'Halloran (eds.), Applying English grammar: functional and corpus approaches. – London: Hodder and Stoughton, 2004. – P. 216-228.
63. Leech G. Introducing corpus annotation / G. Leech // Corpus Annotation: Linguistic Information from Computer Text Corpora. – Addison Wesley Longman, 1997. – P. 1-19.
64. Leech G. New resources, or just better old ones? / G. Leech // Corpus Linguistics and the Web. – Amsterdam: Rodopi, 2007. – P. 134-149.
65. Malmkjær K. Love thy Neighbour: Will Parallel Corpora Endear Linguists to Translators? / K. Malmkjær // Meta: Translators' Journal. – 1998. – vol. 43. – №4. – P. 534-541.

66. McEnery T., Gabrielatos C. English Corpus Linguistics / T. McEnery, C. Gabrielatos // The Handbook of English Linguistics: McMahon-Blackwell Publishing, 2006 – P. 33-72
67. MacEnery T., Hardie A. Corpus Linguistics: Method, Theory and Practice / T. MacEnery, A. Hardie. – Cambridge University Press, 2012. – 294 p.
68. McCarthy M., O'Keeffe A. What are corpora and how have they evolved? / M. McCarthy, A. O'Keeffe // The Routledge Handbook of Corpus Linguistics. – Routledge, 2010. – P. 3-14.
69. McEnery A., Kifle N. Epistemic modality in argumentative essays of second language writers / A. McEnery, N. Kifle // In Flowerdew (ed.), Academic Discourse. – London: Longman, 2002. – P. 182-195.
70. McEnery A., Xiao Z. Swearing in modern British English: the case of fuck in the BNC / A. McEnery, Z. Xiao // Language and Literature. – 2004. – №13 (3). – P. 237-270.
71. McEnery A.M. Swearing in English: bad language, purity and power from 1586 to the present / A.M. McEnery. – London: Routledge, 2005. – 276 p.
72. McEnery T. Wilson A. Corpus Linguistics An introduction / T. McEnery, A. Wilson. – Edinburgh: Edinburgh University Press, 2001. – 235 p.
73. McEnery T., Xiao R., Tono Y. Corpus-based Language Studies: an Advanced Resource Book / T. McEnery, R. Xiao, Y. Tono. – London: Routledge, 2006. – 386 p.
74. Meyer C.F. English corpus linguistics / C.F. Meyer. – Cambridge: Cambridge University Press, 2002. – 241 p.
75. Meyer Ch. F. Pre-electronic corpora / Ch. F. Meyer // Corpus Linguistics. An International Handbook. Edited by A. Lüdeling, M. Kytö. – 2008. – Volume 1. – P. 1-15.

76. Meyer Ch. P. English Corpus Linguistics. An introduction / Ch.P. Meyer. – Cambridge University Press, 2004. – 168 p.
77. Millar N. Modal verbs in time: frequency changes 1923–2006 / N. Millar // International Journal of Corpus Linguistics. – 2009. – №14 (2). – P. 191-220.
78. Nesselhauf N. Collocations in a Learner Corpus / N. Nesselhauf. – John Benjamins Publishing Company, 2005. – 331 p.
79. Orpin D. Corpus linguistics and critical discourse analysis: examining the ideology of sleaze / D. Orpin // International Journal of Corpus Linguistics. – 2005. – №10 (1). – P. 37-61.
80. Partington A. Patterns and meanings / A. Partington. – Amsterdam:Benjamins, 1998. – 158 p.
81. Partington A. Utterly content in each other's company: semantic prosody and semantic preference / A. Partington // International Journal of Corpus Linguistics. – 2004. – №9 (1). – P. 131-156.
82. Partington A., Morley J., Harman, L. Corpora and discourse / A. Partington, J. Morley, L. Harman // Proceedings of CamConf 2002, Universita degli Studi di Camerino, Centro Linguistico d'Ateneo, September 27-29. – New York: Peter Lang, 2004. – 192 p.
83. Renouf A. Lexical signals of word relations / A. Renouf // In Scott and Thompson (eds.). – Liverpool University, 2001. – 388 p.
84. Romer U., Schulze R. Exploring the Lexis–Grammar Interface / U. Romer, R. Schulze. – John Benjamins Publishing Company, 2009. – 329 p.
85. Schmid H.-J. English abstract nouns as conceptual shells: from corpus to cognition / H.-J. Schmid. – Berlin: Mouton de Gruyter, 2000. – 421 p.

86. Schonefeld D. Corpus linguistics and cognitivism / D. Schonefeld // International Journal of Corpus Linguistics. – 1999. – №4 (1). – P. 71-137.
87. Semino A., Short M. Corpus stylistics / A. Semino, M. Short. – London: Longman, 2004. – 204 p.
88. Sinclair J. Corpus, Concordance, Collocation / J. Sinclair – Oxford: Oxford University Press, 1991. – 170 p.
89. Stefanowitsch A., Stefan Th., Gries M. de G. Corpora in Cognitive Linguistics / A. Stefanowitsch, Th. Stefan, M. de G. Gries. – Berlin, 2006. – 360 p.
90. Stefanowitsch A., Stefan Th., Gries M. de G. Corpus-Based Approaches to Metaphor and Metonymy / A. Stefanowitsch, Th. Stefan, M. de G. Gries. – Berlin, 2006 – 318 p.
91. Stubbs M. Conrad in the computer: examples of quantitative stylistic methods / M. Stubbs // Language and Literature. – 2005. – №14 (1). – P. 5-24.
92. Stubbs M. Text and corpus analysis: computer-assisted studies of language and culture / M. Stubbs. – Oxford: Blackwell, 1996. – 288 p.
93. Svartvik J. Corpus linguistics 25+ years on / J.Svartvik. – Amsterdam, NY 2007. – P. 11-27.
94. TEI: Text Encoding Initiative. P5: Guidelines for Electronic Text Encoding and Interchange. – 2008. – September, 11. [Електроний ресурс] – Режим доступу: http://www.tei-c. org/Guidelines/P5/
95. Teubert W. Corpus linguistics and lexicography / W. Teubert // Text Corpora and Multilingual Lexicography ed. by W. Teubert – John Benjamins Publishing Company – Amsterdam/ Philadelphia, 2007 – P. 109-134.

96. Tognini-Bonelli E. Corpus Linguistics at Work / E. Tognini-Bonelli. – Amsterdam: John Benjamins Publishing Company, 2001. – 224 p.
97. Tognini-Bonelli E. Theoretical overview of the evolution of corpus linguistics / E. Tognini-Bonelli // The Routledge Handbook of Corpus Linguistics / Edited by Anne O'Keeffe and Michael McCarthy. – Routledge, 2010. – 681 p.
98. Vivanco V. The absence of connectives and the maintenance of coherence in publicity texts / V. Vivanco // Journal of Pragmatics. – 2005. – №37 (8). – P. 1233-1249.
99. Wang S. Corpus-based approaches and discourse analysis in relation to reduplication and repetition / S. Wangs // Journal of Pragmatics. – 2005. – №37 (4). – P. 505-540.

Xiao R. Well-known and Influential Corpora / R. Xiao // Corpus Linguistics. An International Handbook. Edited by A. Lüdeling, M. Kytö. – 2008. Volume 1. – P. 383-457.

100. Zanettin F. Bilingual Comparable Corpora and the Training of Translators / F. Zanettin // Meta: Translators' Journal. – 1998. – vol. 43. – №4. – P. 616-630.
101. Zhukovska V.V. Teaching vocabulary through date-driven activities. XVII TESOL-Ukraine International Conference: English in Globalised World. – Kamianets-Podilsky, 2012. – P.201
102. Zhukovska V.V. Corpus-based approach to teaching vocabulary and grammar // XVI TESOL-Ukraine International Conference Current Studies in English «Linguistics and methodology perspectives» – Zhytomyr, Kamianets-Podilsky, 2011. – P.171.

Англо-український глоссарій основних термінів та понять корпусної лінгвістики

alignment	***Alignment*** is the process of adding information about which parts of a text (sentence, word combination, word) in language A in a parallel corpus correspond to the equivalent corresponding text parts in language B.	**автоматичне вирівнювання текстів**	***Автоматичне вирівнювання*** – це процес паралельної сегментації оригінального тексту і його перекладу по реченнях, словосполученням, словам.
annotation	(a) Codes within a corpus that embody one or more linguistic analyses of the language in the corpus. (b) The process of adding such information to a corpus.	**анотація; розмітка**	***Анотація*** – це а) довільна лінгвістична інформація про лінгвально релевантні одиниці текстових даних, подана через формальний код, б) практика введення формалізованої лінгвістичної інформації в електронний текст; в) наявність такої інформації у тексті. ***Процес розмітки*** (tagging, annotation) полягає в приписуванні текстам і їх компонентам спеціальних міток (tag, tags): • зовнішніх, екстралінгвістичних; • структурних; • власне лінгвістичних.

archive	An ***archive*** is a text repository, often huge and opportunistically collected, and normally not structured.	**електронний архів; сховище**	***Електронний архів*** – це тексти на електронному носії, що представлені в не стандартизованій й не уніфікованій формі.
balanced corpus	A ***balanced corpus*** contains texts from a wide range of different language genres and text domains and the relative sizes of each of its subsections have been chosen with the aim of adequately representing the range of language that exists in the population of texts being sampled.	**збалансований корпус**	***Збалансований корпус*** складається з текстів різних стилів і жанрів, що представлені у репрезентативному співвідношенні.
colligation	***Colligation*** is a co-occurrence relationship between a word and a grammatical category or context.	**колігація**	***Колігація*** – це типова граматична сполучуваність слова
collocation	***Collocation*** is a co-occurrence relationship between two words, where one word is more likely to occur in the presence of the other than elsewhere.	**колокація**	***Колокація*** – це типова лексична сполучуваність слова.
concordance	A ***concordance*** is a list of all of the occurrences of a particular search term in a corpus, presented within the context in which they occur - usually a few words to the left and right of the search term.	**конкорданс**	***Конкорданс*** – список усіх уживань даного слова в лівосторонньому та правосторонньому контексті

concordancer	A ***concordancer*** allows to search a corpus and retrieve from it a specific sequence of characters of any length (a word, part of a word, or a phrase).	**конкордансер, корпусний менеджер**	***Конкордансер*** (корпусний менеджер) – спеціальна пошукова система, що включає програмні засоби для пошуку даних у корпусі, отримання статистичної інформації й надання результатів користувачеві в зручній формі. Результати цієї процедури подаються у вигляді горизонтальних рядків із пошуковим словом посередині, що має назву KWIC (Key Word In Context).
corpus	A ***corpus*** is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research[60].	**корпус**	**Корпус текстів** – це машиночитане, збалансоване, репрезентативне зібрання особливо розмічених (анотованих) текстів, відібраних згідно фіксованих параметрів для досягнення визначеної лінгвістичної мети та досліджуваних нелінійно за принципом гіпертексту.

[60] http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter1.htm

corpus linguistics	***Corpus linguistics*** is a branch of applied linguistic concerned with the compilation and analysis of **corpora –** computerized databases created for linguistic research.	**корпусна лінгвістика**	***Корпусна лінгвістика*** – це галузь прикладного мовознавства, що займається визначенням загальних принципів побудови, обробки та експлуатації лінгвістичних корпусів даних (корпусів текстів) із використанням сучасних комп'ютерних технологій, розробленням методики збору реальних мовних явищ – писемних та усних текстів, а також способів їх збереження та аналізу.
corpus-based investigations	***Corpus-based investigations*** use a corpus as a source of examples to check researcher intuition or to examine the frequency and/or plausibility of the language contained within a smaller data set.	**дослідження, базовані на корпусі; корпусобазовані дослідження**	У ***корпусобазованих*** дослідженнях дані корпусу використовуються для доведення, спростування чи уточнення визнаної на певному етапі розвитку наукової думки теорії чи гіпотези.
corpus construction	***Corpus construction*** is the process of designing a corpus, collecting texts, encoding the corpus, assembling and storing the relevant metadata, marking up the texts where necessary and possibly adding linguistic annotation.	**створення корпусу**	Процес ***створення корпусу*** передбачає розробку засад його дизайну, відбору текстів, анотування та розмітки.

corpus driven investigations	***Corpus-driven*** investigations reject the characterisation of corpus linguistics as a method and claims instead that the corpus itself should be the sole source of our hypotheses about language.	**дослідження, що управляються корпусом; корпуснокеровані**	***Корпусокеровані дослідження*** відмовляються від визнання корпусної лінгвістики як методу і стверджують, що корпус сам по собі є єдиним джерелом гіпотез про мову.
data	In corpus linguistics, ***data*** may refer to the texts contained in corpora and to statistics, concordances or collocations extracted from corpora.	**дані корпусу**	***Дані корпусу –*** це тексти, що входять до корпусу, а також різного роду дані, отримані в результаті аналізу корпусу.
data-driven learning	***Data-driven learning*** is a way of using corpora in language teaching that involves the learners being given direct access to the corpus and a tool for searching it, the intention being that their exploration of the corpus helps their learning of the language.	**орієнтоване на дані навчання**	***Орієнтоване на дані навчання*** – це підхід до вивчення мови з використанням даних корпусу. Базоване на корпусі програмне забезпечення використовується для підтримки інтерактивної навчальної діяльності.
diachronic corpus	A ***diachronic corpus*** is a corpus that has been carefully built in order to be representative of a language or language variety over a particular period of time, so that it is possible for researchers to track linguistic changes within it.	**діахронічний корпус**	***Діахронічний корпус –*** тип корпусу, що зорієнтований на вивчення та аналіз еволюційних процесів в конкретній мові; укладається переважним чином на матеріалі текстів однієї мови, відібраних у різні часові проміжки з метою вивчення історичної динаміки мовних змін.

dialect corpus	A ***dialect corpus*** is a specialised spoken corpus, which is compiled in order to carry out studies of regional variation.	**корпус діалектного мовлення**	***Корпус діалектного*** мовлення містить тексти, що репрезентують зразки діалектного мовлення.
dynamic corpus	A ***dynamic corpus*** is continually growing over time, as opposed to a static corpus, which does not change in size once it has been built.	**динамічний корпус**	***Динамічний корпус*** – тип корпусу, що передбачає систематичне збалансоване поповнення новими даними, таким чином забезпечуючи можливість відстежувати зміни у мові із врахуванням аспекту діахронії.
encoding	***Encoding*** is the process of representing a text as a sequence of characters in computer memory.	**кодування**	***Кодування*** – це процес представлення тексту як набору спеціальних символів.
frequency word list	A ***frequency word list*** presents all the words in a given corpus and specifies for each word how many times it occurs in this corpus.	**частотний список слів**	***Частотний список слів*** – це список усіх вжитих у даному корпусі слів, із зазначенням скільки разів конкретне слово вживається у даному корпусі.
key word in context (KWIC)	***Key word in context*** is a format for displaying a concordance where the search result is lined up in a central column, and the columns on either side contain a short chunk of the context preceding and following each result in the corpus.	**ключове слово в контексті**	***Ключове слово в контексті*** становить собою особливий формат представлення даних конкордансу, коли результати запиту подаються у вигляді горизонтальних рядків із пошуковим словом посередині, а по боках представлено ліво- та правосторонній контекст.

learner corpora	A ***learner corpus*** is a special corpus type, consisting of language output produced by learners of a language.	**учнівський корпус; корпус учнівського мовлення**	***Учнівський корпус*** – тип корпусу, що укладається з усних і/або писемних текстів, породжених особами, що вивчають мову як іноземну.
lemmatisation	***Lemmatisation*** is a form of corpus annotation where every token in the corpus is labelled to indicate its lemma.	**лематизація**	***Лематизація*** – процес утворення початкової форми слова, виходячи з інших його словоформ.
representativeness	***Representativeness*** means that a corpus is sampled in such a way that it contains all the types of text, in the correct proportions, that are needed to make the contents of the corpus an accurate reflection of the whole of the language or variety that it samples.	**репрезентативність**	***Репрезентативність*** полягає в здатності корпусу відображати всі властивості предметної галузі.
metadata	***Metadata*** in a corpus mean data about the texts (e.g. the author, date of publication, title and source of a written text, or information about the sex, age and social class of speakers in a spoken text).	**метадані**	***Метадані*** – це дані про дані корпусу (бібліографічні, типологічні).
national corpus	A ***national corpus*** is a large corpus that attempts to represent a range of the language used in a particular national language community.	**національний корпус**	***Національний корпус*** – це великий за обсягом корпус, що прагне до об'єднання у своїй структурі текстів найрізноманітніших жанрів і типів.

parallel corpus	A ***parallel corpus*** consists of two or more corpora that have been sampled in the same way from different languages.	**паралельний корпус**	***Паралельний корпус*** – це корпус, який складається як мінімум з двох підкорпусів, один з яких є вихідним, а інший містить тексти-переклади вихідного корпусу.
parsing	***Parsing*** is the process of analysing the syntactic structure of a text or part of a text (such as a sentence).	**синтаксична розмітка, парсинг**	***Синтаксична розмітка*** описує синтаксичні зв'язки між лексичними одиницями й різні синтаксичні конструкції (наприклад, підрядне речення, дієслівне словосполучення й т.п.).
part-of-speech tagging	**Part-of-speech tagging (POS tagging)** is a form of annotation that lies in adding part-of-speech tags to a text.	**частиномовна розмітка**	***Частиномовна розмітка*** – це тип розмітки, що полягає у внесенні до корпусу морфологічних міток, які включають не тільки ознаки частини мови, але й ознаки граматичних категорій, властивих даній частині мови.
pedagogic corpus	A ***pedagogic corpus*** is used for language teaching and consists of all of the language to which a learner has been exposed in the classroom; for example, the texts and exercises that the teacher has used.	**педагогічний корпус**	***Педагогічний корпус*** використовується для навчання мові та включає навчальні тексти.
raw corpus	A ***raw corpus*** has not been processed in any way, it contains no annotation.	**необроблений (неанотований) корпус**	***Необроблений (неанотований) корпус*** не містить анотації.

sample corpus	***Sample corpus*** is a small section of a corpus that has been constructed or annotated for demonstration purposes, before the full corpus exists or has been annotated or a subset of the texts in the corpus that is released for free or at a very low price is referred to as a sample corpus.	**частина корпусу для демонстрації можливостей всього корпусу**	
semantic tagger	**Semantic tagger** is a piece of software that attaches codes to words based upon their semantic function.	**семантичний тегер**	***Семантичний тегер*** – це комп'ютерна програма, що приписує семантичні теги, які позначають семантичні категорії, до яких відноситься дане слово або словосполучення, а також більш вузькі підкатегорії, що специфікують значення.
specialized corpus	A ***specialized corpus*** has been designed for a particular research project.	**спеціалізований корпус**	***Спеціалізований корпус*** – це жанрово чи галузево специфічний корпус, що має на меті відобразити певну підмову.
spoken corpus	A ***spoken corpus*** consists entirely of transcribed speech (e.g. spontaneous informal conversations, meetings, debates, classroom situations etc.).	**корпус усного мовлення**	***Корпус усного мовлення*** – це спеціальна колекція ретельно відібраних текстових уривків (слів, фраз, речень), вимовлених численними мовцями за різних акустичних умов.

static corpus	A ***static corpus*** is intended to be of a particular size – once that target is reached, no more texts are included in it.	**статичний корпус**	***Статичний корпус*** не передбачає поповнення новими матеріалами.
token	***Token*** is any single, particular instance of an individual word in a text or corpus.	**токен**	***Токен*** – це окрема значима одиниця.
tokenisation	***Tokenisation*** is the automatic process of converting all of a text into separate **token**s.	**токенізація**	***Токенізація*** – це розбиття потоку символів природної мови на окремі значимі одиниці (токени, словоформи).
written corpus	A ***written*** **corpus** contains texts that have been produced or published in written format (e.g. traditional books, novels, textbooks, newspapers, magazines or unpublished letters and diaries etc.).	**корпус писемного мовлення**	***Корпус писемного мовлення*** містить лише тексти писемного варіанту мови.

Предметний покажчик

збалансований к.
діалектного мовлення к.
діахронічний к.
динамічний к. (моніторинговий к.)
дослідницький к.
ілюстративний к.
інтерпретаційний к.
моніторинговий к. (динамічний к.)
необроблений (неанотований) к.
національний к. (загальномовний к.)
повнотекстовий к.
паралельний к.
педагогічний к.
писемного мовлення к.
розмічений к.
синхронічний к.
спеціалізований к.
статичний к.
усного мовленння к.
учнівського мовлення к.
фрагментарний к.

корпусна лінгвістика

лематизація
лема

машиночитаність
метадані

парсинг

репрезентативність
розмітка
анафорична р.
лінгвістична р.
морфологічна р. (частиномовна р.)
семантична р.
синтаксична розмітка, парсинг
структурна р.
просодична р.
частиномовна р. (морфологічна р.)

семантичний тегер
стандартизація
створення корпусу
стеммінг

токен
токенізація

формат даних
частотний список

Навчальне видання

Вікторія Вікторівна Жуковська

ВСТУП ДО КОРПУСНОЇ ЛІНГВІСТИКИ

Навчальний посібник

Комп'ютерна верстка: Гавриловський С. О.

Підписано до друку ___.10.2013 р. Формат 60х90/16. Ум. друк. арк. 8,14.
Обл. вид. арк. 4,81. Друк різографічний.
Гарнітура Times New Roman. Зам. 48. Наклад 100.

Видавництво Житомирського державного університету імені Івана Франка
Свідоцтво про державну реєстрацію:
серія ЖТ №10 від 07.12.04 р.
м. Житомир, вул. Велика Бердичівська, 40
електронна пошта (E-mail): zu@zu.edu.ua