Демська-Кульчицька О.М.

БРИТАНСЬКИЙ НАЦІОНАЛЬНИЙ КОРПУС І НАЦІОНАЛЬНИЙ КОРПУС УКРАЇНСЬКОЇ МОВИ

У пропонованій статті, яка продовжує цикл публікацій автора з корпусної лінгвістики, обговорюється проблема комп'ютерної організації емпіричного матеріалу для лінгвістичного аналізу й опису. Здійснено порівняльний аналіз еталонного в корпусній традиції Британського національного корпусу (British National Corpus) і Національного корпусу української мови.

In the article, which continues a cycle of author's publications on corpus linguistics, the general theoretical problem of the computer organization of the empirical material for linguistic analyses and language description is discussed. The comparative analysis of British National Corpus - the standard corpus in corpus linguistics, and Ukrainian National Corpus is represented.

Внаслідок застосування комп'ютера до збирання й організації мовного матеріалу з метою його наступних вивчення, систематизації, опису тощо, тобто експлуатації у лінгвістичних дослідженнях, у науці про мову впродовж XX ст. виформовується напрямок корпусної лінгвістики, проблематика якої пов'язана з розробленням теоретичних засад і практичних прийомів побудови, машинного опрацювання та експлуатації лінгвальних даних, оформлених як корпус текстів. Об'єктом корпусної лінгвістики є корпус - машиночитане, стандартно подане зібрання репрезентативних дня певної мови, діалекту або іншої підмножини мов писемних або усних текстів, призначених для лінгвістичного аналізу й опису, відібраних і впорядкованих згідно з екстра- та інтралінгвістичними критеріями, а предметом - текст.

Єдиного погляду на витоки цього напрямку все ще немає. Його початки або відносять до 1960-х років, коли був створений перший власне текстовий корпус в Університеті Брауна (США) і названий Браунівським

корпусом, або ж зараховують до корпусної лінгвістики докорпусні дослідження, пов'язані із збиранням, організацією та описом лінгвістичного матеріалу з або без використання машинних ресурсів.

Датування корпусної лінгвістики початком минулого століття, а не серединою чи початком другої половини XX ст. домінує в корпусних дослідженнях. Так, Т. МакЕнері й А. Вилсон зазначають, що „думка про те, що корпуси текстів з'явилися в 1960-х роках і особливо інтенсивно почали розвиватися в 1980-х є помилковою. До появи генеративної граматики у лінгвістиці як раз домінувало вивчення масивів емпіричних даних, тобто корпусів. Інша справа, шо аналіз виконувано вручну, внаслідок чого обсяги даних були надзвичайно обмеженими" [1, 47]. Розвиваючи свою думку, вчені розширюють хронологічні межі аналізованого напрямку, і слушно виділяють два періоди у розвитку корпусних студій (1) так звану ранню і (2) власне корпусну лінгвістику, де рання або протокорпусна лінгвістика припадає на етап формування теоретичного підґрунтя та прагматичних передумов виникнення напрямку і створення текстових зібрань для лінгвістичного дослідження на переважно паперових носіях (1910-60 рр.). З 1960-х р. починається період корпусної лінгвістики чи електронної корпусної лінгвістики, безпосередньо пов'язаної з машинними носіями, в межах якої до 1990-х років чітко сформувалися три напрямки теорії та практики: (1) побудови електронних текстових корпусів, (2) програмного опрацювання текстових корпусів, (3) екстрагування, аналізу й опису, чи дослідної експлуатації корпусних даних. Але слід зазначити, що використання самого терміна *'корпусна лінгвістика'* поширилося лише в останнє двадцятиріччя, після публікації у 1984 році збірника під назвою *„Corpus Linguistics: Recent Developments in the Use of Computer Corpora in English Language Research".*

Актуальність пропонованого дослідження мови важко переоцінити, оскільки йдеться про те, що впровадження до лінгвістичного вжитку спеціально приготовленого матеріалу [корпусу - ОДК], дозволяє не лише оптимізувати і об'єктивізувати лінгвістичні дослідження, але і по-новому окреслити багато традиційних лінгвістичних понять" [2, 185]. Загалом, на сьогодні проблематика корпусного напрямку доволі розгалужена і передбачає опрацювання загальної теорії корпусної лінгвістики, кореляції корпусної лінгвістики та інших лінгвістичних дисциплін, типології корпусів та методики інтерпретації корпусних даних, засад створення текстових корпусів природних мов. Окремим аспектом є теорія і практика програмного оброблення корпусних ресурсів.

У сучасній лінгвоукраїністиці корпусне мовознавство все ще не оформилося ні як метод, ні як напрямок, ні як дисципліна. Існує кілька публікації з корпусної лінгвістики, розділ „Корпусна лінгвістика: предмет дослідження і завдання" у підручнику Є. Карпіловської „Вступ до комп'ютерної лінгвістики" [3], український сегмент у проекті TRACTOR.

Впровадження корпусних досліджень в лінгвоукраїністику, яке передовсім вимагає розроблення засад побудови та створення Національного

корпусу української мови (НКУМ), відбувається, по-перше, шляхом аналізу і застосування теоретичних положеннях, розроблених для інших національних мов, зокрема для англійської, і йдеться про теоретичне опрацювання та узагальнення практики створення Британського національного корпусу (British National Corpus - BNC). І, по-друге, шляхом використання досягнень національної науки в галузі загального мовознавства, теоретичної граматики та комп'ютерної і математичної лінгвістики.

Аналізуючи теоретичні засади та практику створення національних корпусів, перш за все до уваги береться BNC, який вважають еталонним у системі сучасних текстових корпусів.

Зіставний аналіз BNC та НКУМ (див. Табл. 1), теоретичне обґрунтування якого уже завершено автором статті, передовсім передбачав розгляд типологічних характеристик, сфери використання, загального обсягу, структури корпусу, хронологічних меж, структури і наповнення джерельної бази та технологічних аспектів подання корпусних даних. Фактично аналізовано детермінативні корпусні параметри.

Таблиця 1: Зіставний аналіз BNC і НКУМ

	British National Corpus	*Національний корпус української мови*
Типологічні характеристики	фрагментний, синхронний; загальномовний, мономовний; мішаний	дослідницький, фрагментний, мішаний, синхронно-діахронний, загальнонародної мови, мономовний, морфологічно анотований
Сфера використання	видавнича справа; академічні лінгвістичні студії, вивчення мови; штучний інтелект, оброблення природної мови (NLP); пошук інформації	гуманітарні науки, лінгвістичні студії; математичні науки і програмування; методика мови; сфера державного управління
Обсяг	100 млн. слововживань	2 млн. 450 тис. слововживань
Структура корпусу	монокорпус	полікорпус, до складу якого входять підкорпуси
Хронологічні межі	1960- 1993	XII-XXI ст.
Структура джерельної бази	тексти художньої літератури та інформативна проза	тексти художнього, наукового, офіційно-ділового, публіцистичного, конфесійного та епістолярного стилів
Наповнення джерельної бази	книги; періодичні видання; інші опубліковані джерела; неопубліковані джерела; уривки усного мовлення; некласифіковані джерела	книги; періодичні видання; діалектні та фольклорні записи; історичні рукописи; епістолярій; студентські та учнівські твори; поліграфічна реклама, суспільно-політичні виступи; наукові дискусії; парламентські дебати; аудіореклама; побутові розповіді; діалогічне мовлення
Принципи комп'ютерного видання	SGML-синтаксис; TEI-принципи; третій рівень кодування згідно з CES; морфологічна анотація	SGML-синтаксис; TEI-принципи; третій рівень кодування згідно з CES; морфологічна анотація

Типологічні характеристики аналізованих корпусів виявили незначну відмінність. Так, обидва-корпуси кваліфіковані як:

- *фрагментні:* скомпоновані з текстових уривків, для BNC обсягом не більше, ніж 45 000 слововживань, а для НКУМ - 5 000;

-*мономовні:* корпусний матеріал становлять тексти однієї мови, відповідно в BNC - британського варіанту англійської мови, НКУМ - української;

- *мішані:* передбачено введення текстових фрагментів усного і писемного реалізаційних варіантів мови;

- *загальномовні:* передбачають подання текстів, які покривають усі предметні сфери, стилі та жанри мови, а для НКУМ додатково зазначено необхідність урахуванням територіальної специфіки як у межах України, так і за її межами, що певним чином вплине на відмінність джерельних баз аналізованих корпусних об'єктів.

Відмінність типологічних характеристик порівнюваних корпусів полягає у тому, що *BNC* детермінований як *синхронний* корпус, який включає художні тексти з 1960-х років та інформативну прозу з 1975-х і до сьогодні. А *НКУМ* - як *синхронно-діахронний,* який відповідно охоплює текстовий матеріал найдавнішого періоду існування української мови до сьогодення, що безпосередньо мотивує структуру генерального корпусу *НКУМ,* до складу якого входять хронологічні підкорпуси зіставні з хронологічними періодами розвитку української мови та синхронний корпус сучасної української мови.

Крім того, на відміну від *BNC, НКУМ* додатково детермінований як:

- *дослідницький:* орієнтований на широкий клас лінгвістичних завдань;

- *динамічний:* передбачає константне поповнення множини корпусних текстів;

- *морфологічно анотований:* усі текстові дані розмічені до рівня слова і кожне слово передбачає маркування частиномовної належності та відповідних морфологічних значень.

Остання кваліфікація - анотованість, - лише не формалізована у BNC. Не констатуючи факт морфологічної анотованості при визначенні типологічних характеристик BNC, варто зазначити, що остання наявна у ньому і de facto є взірцем для національних анотаційних схем.

На перший погляд відмінними є сфери використання BNC і НКУМ Але по суті ці відмінності формальні. BNC передбачено використовувати у сфері видавничої справи, що не передбачено для НКУМ, але таке використання і не заборонене. Лінгвістичні студії та вивчення мов (рідної або іноземної") з погляду аплікаційних можливостей однакові для обох корпусів. Єдине, що для НКУМ передбачено також і можливість ширшого використання корпусного ресурсу в гуманітарних науках. Аналогічно штучний інтелект і оброблення природної мови (NLP), що в НКУМ сформульовані як математичні науки і програмування. Загалом, аплікативна специфіка BNC виписана як „reference book publishing, academic linguistic research,

language teaching, artificial intelligence, natural language processing, speech processing, information retrieval" з деталізацією лінгвістичних досліджень: .lexical, semantic/pragmatic, syntactic, morphological, graphological/written form/orthographical" [4, 5]. Одним із варіантів окреслення аплікативних сфер і можливостей НКУМ може бути:

1. Лінгвістичні інститути НАН України та кафедри української мови:

- продовження або створення картотеки української мови;
- створення електронних словників та їх паперових варіантів або навпаки;
- написання підручників, посібників і вправників з української мови для середньої та вищої школи та підручників з української мови як іноземної;
- організації довідково-консультативної правописної служби.

2. Науково-дослідні інститути НА НУ та кафедри гуманітарного напрямку:

- використання текстового ресурсу як дослідно-ілюстративного матеріалу в галузевих дослідженнях;
- створення фахових енциклопедій та термінологічних словників.

3. Інститут кібернетики НАНУ, кафедри прикладної математики:

- українізації комп'ютера і програмних продуктів в Україні;
- побудови машинної мовної моделі як технологічної бази для розробок у галузі інформаційних технологій;
- української локалізації та інтернаціоналізації інформаційних технологій;
- створення програм автоматичного розпізнавання і синтезу мовлення;
- забезпечення автоматичних методів перетворення текстової інформації;
- лінгвістичного забезпечення автоматичних систем управління.

4. Система освіти:

- впровадження найсучасніших методик навчання української мови та літератури;
- організації методичної допомоги вчителям української мови та літератури в режимі CD і on-line;
- забезпечення учнів і вчителів середніх шкіл програмними текстами з літератури в автоматизованому вигляді.

Джерельна база обох корпусів з формально погляду є однаковою. Але її теоретичне обгрунтування базується на відмінних підходах. Так, у BNC реалізовано тематичний підхід до кваліфікації текстового матеріалу, а у НКУМ запропоновано стилістичний з урахуванням жанрових текстових аспектів. Що зумовило для BNC глобальне розрізнення текстів художньої літератури та інформативної прози з тематизацією: „Imaginative, Arts, Belief and thought, Commerce, Leisure, Natural science, Applied science, Social science, World affairs, Unclassified" [4, 11]. А для НКУМ - стилістичну диференціацію текстового матеріалу, яка передбачає виділення художнього,

наукового, офіційно-ділового, публіцистичного, конфесійного та епістелярного стилів.

Відмінність підходів - тематичного та стилістичного, - не дали відмінності результатів, оскільки ми таки приходимо до глобального розрізнення художнього і не художнього тексту в обох корпусах.

Суттєва відмінність BNC і НКУМ у розмірі. Перший охоплює 100 млн. слововживань, а другий - 2 млн. 450 тис. Проте, 2 млн. 450 тис. слововживань НКУМ - це нижня статистична межа. Натомість бажаним чи цільовим є розмір BNC, тобто 100 млн. слововживань, що на сьогодні у корпусній лінгвістиці небагато.

Відмінність аналізованих корпусів виявилася у загальнокорпусній структурі. Так, BNC не передбачає субкорпусного поділу, а для НКУМ субкорпусне структурування мотивовано необхідністю охоплення не лише синхронного, а й історичного текстового матеріалу. Тому, генеральний корпус НКУМ складають хронологічні підкорпуси: давньоукраїнського періоду (від XII до кінця XIV ст.), середньоукраїнських періодів (XV ст. до кінця XVIII ст.), сучасної української мови (від останніх років XVIII ст. і до кінця XX ст.), української мови кінця XX - початку XXI ст. Відповідно відмінними є хронологічні межі корпусів: BNC 1960- 1993 рр. і НКУМ XII - XXI ст.

Однією з ідей корпусної лінгвістики є реалізація мовнонезалежного принципу комп'ютерного подання корпусів. З цією метою до комп'ютерного оформлення корпусів різних мов було використано єдиний базовий стандарт SGML (Standard Generalized Markup Language), а в новіших корпусах наступне покоління цього стандарту XML (Extended Markup Language), Принципи TEI (Text Encoding Initiative) та Стандарт кодування корпусу (Corpus Encoding Standard). Цей підхід реалізований у BNC. І не принципово слід реалізувати у НКУМ.

Отже, зіставний аналіз Британського національного корпусу та Національного корпусу української мови не виявив суттєвих відмінностей ж рівні типологічних характеристик, сфери використання, структури корпусу, хронологічних меж, структури і наповнення джерельної бази й особливо технологічних аспектів подання. А формальні відмінності зумовлені національно-мовною специфікою та лінгвістичною традицією.

Суттєва відмінність не на користь НКУМ - це обсяг, який за у м о м и національного статусу корпусу української мови повинен сягати згідно з корпусними вимогами 100 млн. слововживань. Різняться також і хронологічні параметри. Ця відмінність вмотивована типом - синхронний, синхронно-діахронний, - кожного з корпусів.

Як слушно зазначає А. Пшепюрковський: „у багатьох країнах створення національного корпусу є обов'язком щодо рідної мови" [5, 5], та очевидно, що перспективність Національного корпусу української мови з площини вузько наукової переміщується у площину національної відповідальності.

ЛІТЕРАТУРА

1. McEnery T., Wilson A. Corpus Linguistics. – Edinburgh, 1996.
2. Рычкова Л.В. Праблема састаўных аб'ектаў у корпусах славянскімоў і лінгвістычных базах дадзеных // Мовознаўства. Літаратура. Культуралогія. Фальклярыстыка. XIII Міжнародны з'езд славістаў. Даклады беларускай дэлегфцыі. – Мінськ, 2003. – С. 184-195.
3. Карпіловська Н.Є. (2003). Вступ до комп'ютерної лінгвістики. – Донецьк: Юго-Восток, 2003.
4. Burnard L. (1995) User Reference Guide British National Corpus. – Oxford: Oxford University Computing services, 1995.
5. Przepiyrkowski A. (2004). Korpus IPI PAN. wersa wskrpna. – Warszawa: Instytut Podstaw Informatyki PAN, 2004.