Київський національний університет імені Тараса Шевченка
Інститут філології

Кафедра сучасної української мови

К.ф.н. доц. Н. П. Дарчук

НАВЧАЛЬНА ПРОГРАМА

з дисципліни
"Корпусна лінгвістика: проблеми, методи, перспективи"

для аспірантів спеціальності 10.02.01 "українська мова"

Затверджено
на засіданні кафедри
сучасної української мови
Протокол № 8
від 14 березня 2013 року
Зав. кафедри проф. А. К. МОЙСІЄНКО

Директор Інституту філології

проф. Г. Ф. СЕМЕНЮК
Затверджено
Вченою Радою Інституту філології
протокол № 8
«26» 03» 2013р.

КИЇВ – 2013

Навчальна програма з дисципліни

"Корпусна лінгвістика: проблеми, методи, перспективи"

Укладач: канд. філол. наук доцент Дарчук Наталія Петрівна

Погоджено з науково-методичною комісією

_____ 2013 року

_____

Невичерпним джерелом для словників та граматик є інтегральний опис мовних фактів. При цьому морфологічні, синтаксичні, семантичні, прагматичні, стилістичні, комунікативні, сполучувальнісні властивості мовних одиниць були і є вивідними безпосередньо з текстів. Фахівці гостро відчувають потребу у якомога більшому числі функціональних характеристик мовних одиниць у різних типах текстів. Необхідну лінгвістичну інформацію для подальшого опрацювання її у філологічних студіях можна отримати з корпусів, розбудова яких є ознакою нашого часу.

Наука про мову поступово наближалася до ідеї текстового корпусу і до самого корпусу в такому вигляді, в якому ми зараз його знаємо: **електронне зібрання текстів природної мови, впорядковане, організоване й оформлене певним чином, призначене для наукового та практичного вивчення мови.**

Що було до корпусу?

До корпусу проблема коректного фактичного матеріалу була чи не найскладнішою. Згодом індивідуальний досвід збирання текстів трансформувався в уніфікований об'єкт – картотеку, а з появою належних передумов, – а саме комп'ютера і комп'ютерних технологій, – відбувся наступний крок до появи нової форми існування текстового дослідницького ресурсу в мовознавстві – корпусу.

З історії корпусів

Поява перших текстових корпусів припадає на 60-і р.р. минулого століття, однак з кінця 80-х – початку 90-х р.р. вони почали активно використовуватися у теоретичній лінгвістиці. В останні десятиліття минулого сторіччя зусилля багатьох країн були спрямовані на створення національних універсальних корпусів текстів з необхідними й достатніми кількісними та якісними параметрами для укладання на їхній основі словників і граматик національних мов. На сьогодні текстові корпусні ресурси вже існують для багатьох не лише

європейських мов, а створення корпусу вважається за обов'язок по відношенню до національної мови (напр., Британський національний корпус (100 млн. слововживань); Великий корпус російської мови (100 млн. слововживань); Корпус австралійської періодики (300 млн. слововживань); Банк англійської мови (320 млн. слововживань); Корпус німецької мови (778 млн. слововживань) тощо), а в маленькій Словаччині існує навіть Інститут корпусного дослідження словацької мови.

В Україні корпусна лінгвістика розпочала своє існування у 2009 р. Виникнення корпусу української мови визначене передумовами двох типів: перші – практичні, другі – теоретичні. До практичних передумов належить наявність емпіричної складової у мовознавстві, картотечної традиції та комп'ютерних текстових ресурсів, а до теоретичних – певний цілісний погляд на мову як явище системне, чи певна наукова парадигма, що може слугувати або ж слугує підгрунтям для корпусної практики і теорії. Від самих початків становлення україністики чимало уваги приділялося спостереженню над мовним матеріалом, що мав форму або фольклорно-етнографічного, або літературного тексту, що було зумовлене істотним впливом порівняльно-історичного мовознавства, у площині якого українське мовознавство значною мірою залишається і сьогодні: «…треба, шукаючи істину, йти не від гіпотез і припущень, а від живої мови, яка є вірогідним матеріалом для умовиводів дорслідника», - писав Г.Удовиченко.

Теоретичні засади створення корпусної лінгвістики

На організацію корпусної мовознавчої емпірики вплинув доробок структуралізму:

1) Празької лінгвістичної школи;
2) глосематичної теорії;
3) американського структуралізму;
4) лондонської лінгвістичної школи.

Усі ці напрями є гілками лінгвістичного структуралізму, а не самостійними одиницями, оскільки усім їм притаманні такі спільні ознаки:

- незгода з младограматизмом;
- визнання структури мови як єдиного об'єкта лінгвістики;
- проголошення синхронного вивчення мови як основного завдання лінгвістики;
- нормалізування лінгвістичного аналізу.

Так чи інакше сьогодні маємо справу з чеською, данською та англосаксонською лінгвістичним традиціями, і саме у цих мовознавчих традиціях (крім данської) найшвидше з'явилися текстові корпуси і почала формуватися корпусна лінгвістика. Приклад – англосаксонська лінгвістична традиція, де маємо низку корпусів англійської мови (Brown Corpus, American National Corpus, British National Corpus, Bank of English, Corpus of Contemporery American English, International Corpus of English), які часто є стандартним зразком для укладання корпусів інших мов, а також розбудовану корпусну теорію та методику дослідження мови. Розбудованість чеської корпусної лінгвістики, а також словацької – наслідок популярності ідей Празької лінгвістичної школи.

Перспективи розвитку корпусу української мови

Робота над українським корпусом розпочалася у листопаді 2009 р. з усвідомлення того, наскільки на сучасному етапі існування нашої науки такий інструмент, як корпус, необхідний лінгвістові у його повсякденній діяльності – і морфологу, і синтаксисту, і лексикографу, і соціологу, і діалектологу.

На жаль, коли це усвідомлення прийшло, стало ясно і те, що готових електронних корпусів української мови значного обсягу немає і найближчим часом не буде. Але був величезний досвід автоматичного опрацювання українського (і російського!) тексту: АМА, АСА, АСемА і був невеличкий корпус поетичних текстів кінця 20 ст.(300 тис. слововживань) і фольклорних текстів (30 тис. слововживань). Вихід був один – зробити власний корпус, придатний для розв'язання таких завдань, які нам здавалися важливими. Ми

керувалися відомим принципом: **«робитимемо для себе – тоді іншим знадобиться».** Тому Корпус робився швидко: вже за один рік він досяг 5 млн слововживань. Ми не могли собі дозволити розкіш обговорювати «теорію», яка базувалася б на безкінечній критиці чужих результатів. Таке обговорення, безумовно, приносить задоволення всім його учасникам, але має істотний недолік: кожний крок нібито до досконалості веде не до створення корпусу, а навпаки – веде до протилежного - корпус чомусь не виникає. Виникають лише численні проекти «найкращого у світі корпусу». На жаль, негативний дослід є. Ми не женемося до найкращого у світі корпусу, нам не потрібно теоретична перевага. Нам треба якомога скоріше одержувати можливість шукати надійні приклади в українських текстах. Таким чином, наш корпус створюється з орієнтацією на виключно прагматичні міркування і в короткий строк. Зрозуміло, він не може бути беззеперечним. Ми стоїмо перед вибором: або він буде мати якісь недоліки, або його не буде ніколи.

Тепер, коли корпус існує і ним можна користуватися для важливих задач – не виходячи з тієї ж прагматичної логіки – намітити певні актуальні напрямки подальшої роботи.

Стратегія корпусу

Основна характеристика будь-якого корпусу, крім, зрозуміло, типів розмітки, стосується кількості і якості представлених у ньому текстів. Зараз присутня відносна незбалансованість, є ряд лакун, які ми намагаємося усунути. До числа них входить:

1) розширення підкорпусів 20 ст;
2) створення підкорпусів 19 ст.;
3) створення кількох нових корпусів, зокрема художньої прози, усних текстів і паралельних текстів.

Для 19 ст. треба комплектувати художніми творами і нехудожніми (публіцистичними, науковими, науково-технічними). Для 20 ст. необхідно хронологічно розширити за періодами. Треба врахувати і те, що в Україні

створилася унікальна ситуація, коли розділилися література на «основний» і «зарубіжний» фонди. Треба включити емігрантську, діаспорну літературу. Плануємо створити корпус діалектних текстів і усних текстів.

Удосконалення інструментів пошуку і програмного забезпечення Корпусу

Одним з основних показників цінності Корпусу є багатство пошукових можливостей, які надаються користувачеві. У цій галузі укладачам корпусу хотілося б удосконалити дуже і дуже багато.

У першу чергу, необхідно удосконалити комплекс засобів, які дозволяють отримувати різного роду статистичні дані про Корпус. Зараз є можливість отримувати дані через побудову частотного слованика слів і словоформ, а також кількість прикладів на явище, яке цікавить користувача. Обидві статистики потребують вдосконалення і постійну підтримку. Але нагальним є завдання удосконалення розмітки Корпусу «навчання» програм автоматично знімати омонімію. Є фрагмент (400 тис.), де граматична і лексико-граматична омонімія знімалася вручну, є фргмент, де вона знімається за допомогою дистрибутивних формул. Стоїть питання і про те, як побудувати зручний для користувача інтерфейс видавання будь-якої лінгвістичної інформації, а також створити програмне забезпечення.

Навіщо потрібний Корпус?

Українська мова, її відмінність від інших мов, її минуле, теперішнє, майбутня доля на межі 20-21 ст. Ці фундаментальні проблеми не можна серйозно ставити без такого інструмента, як Корпус. Але це ефективний і корисний інструмент тільки у тому випадку, коли він великий за обсягом і повний за охопленням матеріалу.

Корпус – це зібрання текстів певною мовою, яке представлене в електронній формі і супроводжується науковим апаратом. Апарат, вбудований у корпус, звичайно називається розміткою, або анотацією.

Корпус тим кращий, чим повніша і досконаліша його анотація. Наука про корпуса – це перш за все наука про те, як зробити хорошу розмітку корпуса.

Хороша розмітка, зокрема, дозволяє швидко і ефективно знайти у корпусі ті слова і конструкції, які потрібні користувачеві. Для цього програма пошуку повинна розуміти як мінімум те, які форми у тексті відносяться до одного й того ж слова (*мати* –ім. жін.р., *мати* – ім. чол. р. мн., *мати* дієсл., інф.), тобто хоча б частково «розуміти» граматичну структуру даної мови. Тим більше це розуміння необхідне, якщо ми хочемо шукати не слова, а форми. Напр., знайти у великому тексті всі форми давального відмінка однини. Ніякий текстовий редактор з таким завданням не впорається. Для того, щоб граматичні форми треба було знайти автоматично у тексті, його треба розмітити, інакше це треба зробити вручну, що є процесом трудомістким.

Зрозуміло, розмітка потрібна і для вирішення багатьох інших завдань. Добре розмічений текст – безцінна знахідка для спеціаліста, адже у своїй дослідницькій роботі лінгвісти залежать перш за все від кількості і якості зібраного матеріалу. Багато лінгвістів пам'ятають ті часи, коли приклади виписувалися на картки. Тепер це вже у минулому, але сам процес вибору прикладів робиться людиною і досить складний для автоматизації. Розмічені корпуси - перший серйозний інструмент, який дозволяє істотно пришвидшити і простити цю процедуру. Іншими словами, те, на що колись у дослідників попередніх поколінь пішли місяці, а іноді роки напруженої праці, за допомогою корпусу можна зробити за хвилини.

Отже, корпус – це електронне зібрання текстів, розмічене так, щоб у ньому можна було швидко знайти слова і конструкції із заданими граматичними та іншими властивостями, цікавими лінгвістові.

Корпус повинен бути представницьким: він повинен містити всі типи текстів, представлені у даній мові в даний історичний період, і при цьому містити їх у правильній пропорції. Він повинен містити твори художньої літератури 19, 20 ст. і не тільки художньої. Він повинен містити і газетні і журнальні статті

різної тематики (від суспільно-політичної до спортивної), і спеціальні тексти (наукові, науково-популярні і учбові з різних галузей), і рекламу, і приватну переписку, і щоденники. Словом, до Корпусу потрапляють взірці практично будь-якого писемного дискурсу – від статті сучасного музичного критика до інструкції з догляду за кактусами, від оповідань віршів Ірванця до довідника з фізики. Укладачі Корпусу також добре розуміють, що для повного і адекватного уявлення про те, що зараз відбувається із сучасною українською мовою, необхідно ще великою мірою розширити рамки Корпусу і включити до нього, наряду з писемними текстами, також записи усного мовлення. Чому треба мати взірці усного мовлення у Корпусі? Люди пишуть не так, як говорять: писемне мовлення завжди більш консервативне. Якщо ми хочемо виявити динамічні структури української мови, якщо хочемо зазирнути у майбутнє української мови, треба звернутися до стихії українського усного мовлення.

Цікавими є тексти «електронної комунікації»: переписка електронною поштою, чати, форуми, електронні щоденники (блоги) – все це своєрідний гібрид усного і писемного мовлення (хоча люди, які пишуть у мережевому середовищі, почувають себе вільними не тільки у відношенні до мови, але й до правил орфографії) (*щас, сьодн*і) . Цікаво, що ті, що пишуть , вдаються до спотворення орфографії свідомо, ніби це художній прийом. Все це як своєрідний сленг, але він цікавий і важливий для лінгвістів, тому що це своєрідна лабораторія мови, де намічаються шляхи майбутнього розвитку мови.

Кому і навіщо потрібний корпус?

Він потрібний професійним лінгвістам, тим, хто так чи інакше має справу з фактами мови, а значить, повинен ці факти збирати і систематизувати. Для лінгвістів корпус – як мінімум неоціненний інструмент, який скорочує час на технічну роботу. Насправді, корпус – це інформаційно-довідкова система із сучасної української. мови, яка дозволяє одержувати відповіді на несподівані

питання, ставити нові проблеми, яких лінгвістика минулого поки що не торкалася. Питання варіативності деяких відмінкових форм, хитання у роді, числі деяких іменників тощо. Простим натисканням кнопки можна одержати дані, на збирання яких в іншій ситуації потрібно було б роки.

Можливим буде і те, що без Корпусу не обійдуться , як зараз не можуть обійтися без словника, а, можливо, словник і граматика поєднаються в один електронний ресурс –БД, на основі яких буде вивчатися мова.

А чи цікавий Корпус не тільки лінгвістам? Він цікавий і програмістам в галузі автоматичної обробки текстів (в тому числі і різного роду пошукових систем). Оскільки програми такого роду мають справу з природною мовою, вони повинні так чи інакше розуміти структуру текстів, написаних тією ж мовою. Програмісти як професіонали зацікавлені у тому, щоб корпуси розвивалися, тому що вони мають природну мовну стихію.Наступна цільова аудиторія – це викладачі, причому не стільки рідної мови, скільки іноземної. Ще одна група, для якої Корпус цікавий – це люди, діяльність яких пов’язана зі словом, напр., редактори газет і журналів – їм доводиться значно частіше вирішувати питання узусу: чи допустиме таке словосполучення чи конструкція? Для яких типів тексту вона характерна? Хто і як першим ужив ту чи іншу конструкцію? Однак варто зупинитися ще на одній проблемі, яку можна вирішувати за допомогою Корпусу – це спостереження над динамікою розвитку мови. Оскільки тексти, які входять до корпусу датовані, то неважко прослідкувати за хронологією поступових мовних змін – за появою або поступовим згасанням певних слів, конструкцій або граматичних форм (типу другого родового відмінка). Це викликає до життя фактично новий напрямок – свого роду «мікроісторичну» лінгвістику, у центрі уваги якої знаходяться не глобальні зміни в історії мови, а зміни менш масштабні, які відбуваються протягом десятиріччя (для історії мови це надзвичайно малий строк). З рочки зору історії української мови сучасний період є надзвичайно важливим і цікавим. Незалежність України і пов’язаний з ним статус української мови як державної, інтеграція України до ЄС – це злам соціальних умов і зміна

самого статусу літературної мови призвели до мовного зламу, тобто українська мова наблизилася до такого стану і ймовірність великих змін в його структурі найближчим часом велика. Стилістичний, жанровий і навіть орфографічний різнобій, потік запозичень, ймовірність змін у її структурі також велика, адже попередній період розвитку української мови був досить стабільним (мова законсервувалася через закритість).