УДК 811.161

Демська-Кильчицька О.М.

POLIGRAP: ЛІНГВІСТИЧНИЙ ПОШУКОВИЙ РЕСУРС

Сучасний культурний простір, крім традиційних сфер існування, активно засвідчує присутність електронно-технічного спектру, що ставить вимогу інтенсифікації досліджень у галузі і'уманістики загалом і лінгвістики зокрема, використовуючи програмні засоби. Одним із напрямків мовознавства, в основі икого лежить ідеологія комп'ютеризації, є корпусна лінгвістика, стрімкий розвиток якої сьогодні засвідчують не лише в англо-с.аксоністиці чи романо-германістиці, а й у славістиці. Це мотивовано одним з найважливіших чинників динаміки розвитку цього лінгвістичного напряму, а саме: "впровадження до лінгвістичного вжитку спеціально приготовленого матеріалу дозволяє не лише оптимізувати і об'єктивізувати лінгвістичні дослідження, але і по-новому окреслити багато традиційних лінгвістичних понять" [1, 185].

Тут на увагу заслуговує польська лінгвістична традиція, яка, хоча й критикує себе за розпорошеність [6, 6], стимулювала створення низки корпусів польської мови, на відміну від більшості слов'янських корпусних досліджень, які обмежилися одним-двома загальномовними текстовими корпусами. Йдеться про такі загальномовні академічні текстові корпуси польської мови: *Polski Korpus Narodowy* (http: / /www1.uni.lodz.pl/pelcra). *Korpus Instytutu Języka Polskiego* PAN (http://www.ijp-pan.krakow.pll. *Korpus Państwowego Wydawnictwa Naukowego* (http://korpus.pwn.pl). *Korpus Instytutu Podstaw Informatyki* PAN (http: //www.ipipan.waw.pl/ -corpus).

Про популярність корпусних досліджень у полоністиці свідчить також і наявність теоретичних праць з проблем створення та експлуатації текстових корпусів та корпусних досліджень польської мови. Це, зокрема, праці М. Баньки [2], П. Баньского [3], Й. Беня [4], Л. Дембовского [5], А. Пшепюрковского [6] та інших.

Практична побудова загальномовних корпусів і перехід до здійснення лінгвістичних досліджень на їхній базі визначають актуальність корпусної проблематики не лише у полоністиці, а й у лінгвоукраїністиці.

Перехід від створення корпусних побудов до їх експлуатації у мовознавстві передовсім передбачає коректні підходи до екстрагування інформації з корпусного текстового ресурсу, що неможливо без створення комп'ютерних пошукових засобів, які власне становлять предмет нашого дослідження. А об'єктом дослідження є **Poligrap** *(POLy interpretation Indexing Query and Retrieval Processor*) - пошукова корпусна програма універсального типу, створена у межах проекту *Корпусу IPI PAN* 3. Криніцкім та Д. Янусою під керівництвом А. Пшепюрковского.

Проект *Корпус IPI PAN* реалізовано в Інституті основ інформатики Національної академії наук Польщі впродовж 2001-2004 років. Цей корпус польської мови є першим корпусом з вільним доступом у мережі Інтернет (http: //www.ipipan.waw.pl/-corpus), обсягом на 100 млн слововживань, частково репрезентативний, морфосинтаксично[1] анотований, збудований згідно із сучасними стандартами і практиками побудови великих корпусів текстів.

Автори розглядають **Poligrap** як універсальну програму, тобто таку, яку можна застосовувати до корпусів довільних мов, у тім числі й українських. Універсальність цієї пошукової програми досягнута низкою технологічних рішень, а саме: "використовуваний у корпусі тегсет не вбудовано в програму, його лише задано через зовнішній конфігураційний файл, натомість внутрішній, використовуваний формат кодування символів - це універсальний формат UTF-8. Тому ніщо не перешкоджає використання **Poligrapu** до роботи з іншими корпусами, в тому й інших мов" [6:41].

Poligrap існує у трьох варіантах:

1) інтернетний (дозволяє працювати з *Корпусом IPI PAN* в режимі on-line);
2) графічний, призначений для операційних систем Windows 2000, Windows XP та GNU/Linux;
3) текстовий, для системи GNU/Linux.

Усі три варіанти характеризуються доволі багатим синтаксисом запиту, який дозволяє сформулювати запит про текстові уривки, парадигматичні форми і морфосинтаксичні параметри слів.

У запиті про текстові сегменти можливі стандартні регулярні формулювання з символами ?, *, +, ., ,, |, {, }, [,], (,), а також числа, записані цифрами, наприклад, шукаємо окремі висловлювання, або варіанти імені:

przyszedł czas
przyszedł em rano

długo szedł

"A1a|E1a" (тобто, шуканим будуть імена A1a або E1a);

Запит щодо форм слів передбачає пошук і вихідної, і парадигматичної форм, наприклад, шукаємо вихідну форму лексеми *korpus,* синтаксично запитання слід оформити так:

[base=korpus].

Найважливішим пошуком є пошук морфосинтаксичних параметрів слів, де можливо добути інформацію не лише про значення парадигматичної форми (через orth), семи (через base) і граматичного розряду (через pos), а й про значення окремих граматичних категорій, наприклад, формулюємо запит на іменники жіночого роду однини:

[pos=noun & gender=f & number=sg].

Важливим позитивом пошукової програми Poligrap є можливість подання морфосинтаксичної багатозначності. Так, "існують конструкції, у яких неможливо однозначно інтерпретувати різні форми одного й того ж слова, або одну форму з різними значеннями. Це ілюструє приклад:

(2.5) *Pamiętam ją pijaną.*

(2.6) *a. Pamiętam go pijanego.*

b. Pamiętam go pijanym." [6:52].

Дійсно, у наведеному прикладі важко однозначно визначити, чи лексема *pijaną* вжита у формі знахідного відмінка, як *pijanego,* чи - орудного, як *pijanym.* Тому в корпусі форма *pijaną* повинна бути анотована двояко, як знахідного і як орудного відмінків. Такий підхід дозволить тримати інформацію про двозначну парадигматичну форму залежно від запиту про відмінкові характеристики, зокрема:

[case=acc]

[case=gen]

[case=acc & case=gen].

Доволі часто працюючи з великими корпусами виникає необхідність накладати певні обмеження на обсяг аналізованого фактичного матеріалу. Наприклад, коли необхідно вивчити реалізаційну специфіку якогось дієслова або препозиційну дистрибуцію якісних прислівників тощо. Формулювання обмежень передбачені синтаксично. Щоби обмежити обсяг шпиту до *Корпусу IPI PAN* необхідно приписати до сформульованого запиту ключове слово within, а після нього s або p, залежно під того чи йдеться про речення *{sentence),* чи абзац *[paragraph).* Так, формулювання запиту про речення, в яких форма *się* виступає після *BYĆ,* але не безпосередньо після:

[base=być] [orth!= się]+ [orth=się] within s.

заголовок і дату написання або публікації твору-джерела контексту. Для переходу між полями розширеного контексту і метаданих служать кнопки *Metadane* і *Kontekst* у відповідних полях.

У графічному варіанті інтерфейсу існує опція сортування результатів запиту. Так, найпростішим типом сортувань є сортування результатів за алфавітом у порядку зростання або навпаки (див. рис. 2).

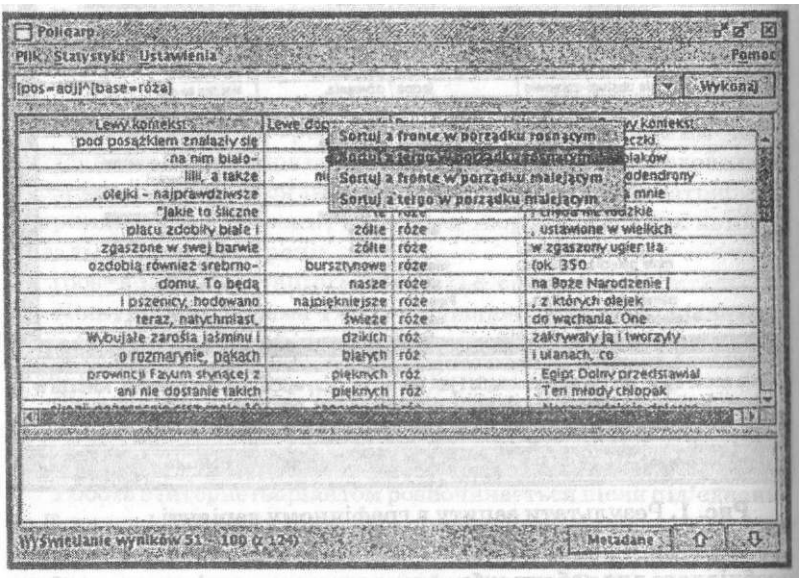

Рис. 2. Сортування *a tergo* за порядком зростання.

Спосіб сортування результатів можна виставити також і в меню: *Ustawiena - Opcje -> Sortowanie.* Крім того, в меню можна виставити запит на статистику контекстного сегмента, яка подаватиметься для кожного результату пошуку.

Важливим атрибутом графічного варіанту інтерфейсу Poligrapu є можливість візуалізації парадигматичних форм слів, їхніх лем і морфосинтаксичних тегів, які можна розділити залежно від стовпчика: лівого, правого контекстів або запитуваного слова. Така інформація вкрай важлива для граматичних, лексичних і особливо лексикографічних досліджень мови.

Крім описаних запитів і результатів, графічний варіант аналізованої програми також дозволяє формулювати запити про так звані аліаси чи скорочення для альтернативних значень певного атрибуту, про метадані або про статистику. Усі ці запити можливо задати через відповідні опції меню.

Текстовий варіант інтерфейсу пошукової програми Poligrap призначений для операційної системи GNU/Linux. Програму запускає команда poligrap korpus, де korpus - це назва субкорпусу (першачастина назв файлівwstępny.cfgiwstepny.poligrap.coipus.image) включно з шляхом до вибраного субкорпусу. Наприклад, якщо субкорпус знаходиться в каталозі ./korpus/ і складається з файлів на взірець wstepny.cfg, wstepny.poligrap.chunk.image тощо, то його можна активізувати командою, сформульованою таким чином: $ poligrap korpus/wstepny і матимемо відповідь (див. рис. 3).

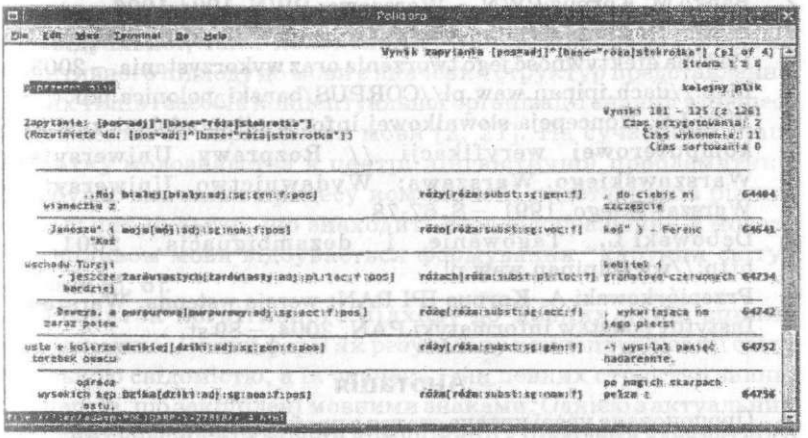

Рис. 3. Результати запиту в текстовому варіанті.

Висновки. Працюючи з *Корпусом IPI PAN* і, відповідно, послуговуючись пошуковою програмою Poligrap, створеною для цього корпусу, маємо підстави стверджувати, що синтаксис запитів, доповнений доволі зручними варіантами інтерфейсів, відповідає поставленому завданню: досягнути універсальності аналізованої програми. Проте такий висновок щодо синтаксису іапитів справедливий лише для користувача-фахівця. Натомість для користувача-початківця він складає певні труднощі, що повинно би стимулювати наступні програмні рішення спрощення роботи з аналізованою пошуковою програмою.

Важливим аспектом розвитку Poligrapu є також її застосування до роботи з українським морфологічно анотованим корпусом текстів, що власне й складає перспективу корпусних програмно-пошукових досліджень.

Примітки

[1] Термін запропонований авторами проекту і йдеться про варіант морфолого-синтаксичної анотації.

Список використаних джерел

1. Рычкова Л.В. Праблема састауных аб'ектау у корпусах славянскі моу і лінгвістичних базах дадзеных // Мовознауства. Література. Культуралогія. Фальклярыстыка. XIII Міжнародны з'езд славыстау. Даклады беларускай дзлегадьіі. - Мінськ. 2003.-С. 184-195.
2. Bańko M. Korpus PWN. - Warszawa: PWN, 1994-1996.
3. Bański P. Anotacjazewnentrzna: wpiyw architektury korpusu IP1 PAN na efektywność jego tworzenia oraz wykorzystania. - 2003. http://dach.ipipan.waw.p1//CORPUS/banski_polonica.pdf.
4. Bień J.S. Koncepcja słownikowej informacji morfologicznej i jej komputerowej weryfikacji // Rozprawy Uniwersytetu Warszawskiego. Warszawa: Wydawnictwo Uniwersytetu Warszawskiego, 1991. -S. 67-78.
5. Dębowski L. Tagowanie i dezambiguacja. 2001. -«і http://www.ipipan.waw.pl
6. Przepiórkowski A. Korpus IPI PAN: wersja wstępna. Warszawa: Instytut podstaw informatyki PAN, 2004. - 89 st.

Анотація

Пропонована увазі читача стаття з проблем корпусної лінгвістики є однією з циклу статей, завдання яких - впровадження корпусного методу в лінгвоукраїністику. У статті розглянуто один із аспектів роботи з текстовими корпусними ресурсами, а саме програмний підхід до екстрагування лінгвальної інформації корпусу. Описано пошукову програму Poligrap, специфіку її синтаксису, а також варіанти інтерфейсів.

Summary

The article continues a cycle of author's publications on corpus linguistics. The main goal of this article is the introduction of corpus method into the Ukrainian linguistics. The article deal with the problem of the computer organization of the empirical material for linguistica analyses and language description. Is described the universal search system Poligrap, specificity of its syntax, and existing variants of interfaces.