УДК 821.161.2-1.08

Ірина Олещук,
Наталя Тандрик,
Ірина Фролова

СЛОВНИК МОВИ ПОЕЗІЇ ЛЕСІ УКРАЇНКИ НА КОМП'ЮТЕРНІЙ ОСНОВІ

Висвітлюється проблема створення словника мови поетичних творів Лесі Українки з використанням комп'ютерних програм опрацювання художнього тексту.

Ключові слова: словник мови поезії, художній текст, комп'ютерна програма, лематизація, параметризована база.

На сучасному етапі розвитку української науки все інтенсивнішими й об'ємнішими стають дослідницькі роботи, що здійснюються у галузі комп'ютерної лінгвістики та лексикографії. Сьогодні комп'ютерна лінгвістика дає можливість аналізувати текст на морфологічному та синтаксичному рівнях. Головним прикладним завданням для українського мовно-інформаційного фонду є створення автоматизованої системи, до складу якої входили б: електронна бібліотека; лексична електронна картотека, призначена для укладання академічних словників; здійснення наукових розвідок із різноманітних питань сучасного мовознавства та організації самого процесу автоматичного укладання словників та їх використання для інших автоматичних систем інтелектуального призначення. Над розробленням та впровадженням державної словникової програми працюють Український мовно-інформаційний фонд НАН України, відділ лексикології, лексикографії та національного корпусу української мови Інституту української мови НАН України, лабораторія комп'ютерної лінгвістики Інституту філології Київського національного університету імені Тараса Шевченка, кафедри прикладної лінгвістики у вузах. В Інституті кібернетики НАН України створено адаптивну лінгвістичну систему АЛІСА. За допомогою цього лінгвістичного процесора стало можливим автоматизоване

укладання словників. Він є прототипом, базою українського автокоректора ТВІР, а також найпопулярнішого на сьогодні українського спелчекера РУТА, розробленого у відділі математичної лінгвістики Інституту мовознавства НАН України імені О. Потебні. За допомогою цих програм стає можливою підготовка активних комп'ютеризованих словників нового типу [2, 85].

Очевидною є вага мовних проблем і важливість організації лексикографічної роботи в Україні на науково-технічному рівні, який відповідав би потребам та можливостям інформаційного суспільства. Адже саме потужні комп'ютерні словники становлять основу всіх інформаційних систем, де використовується природна мова. Отже, саме інформатизація лексикографічних робіт мусить бути ядром системи інформатизації лінгвістичних досліджень.

Створення традиційних паперових словників, як відомо,– надзвичайно трудомісткий процес. А застосування комп'ютерних методів дозволяє скоротити його у багато разів.

Роботи у галузі комп'ютерної лінгвістики та лексикографії здійснюються в усіх країнах світу, в тому числі в Україні. Більшість із них пов'язана з науково-технічною програмою ДКНТ України „Інформатизація та комп'ютеризація гуманітарної сфери", а також програмою під патронатом Президента України „Словники України". Основні центри в цій галузі – Київ, Львів, Харків. Зокрема, у лабораторії комп'ютерної лінгвістики Інституту філології КНУ розроблено принципи створення параметризованої бази даних за поетичними текстами українських письменників.

У травні 2004 року ми у складі групи студентів ВДУ, яку очолювали доцент Н. О. Данилюк і старший викладач В. Ф. Старко, брали участь у семінарі з проблем укладання параметризованої бази даних за творами Лесі Українки. Працівники лабораторії комп'ютерної лінгвістики КНУ виконали частину роботи над аналізом поетичних текстів поетеси на основі автоматичного морфемного сегментатора.

Питання створення словника мови поезії Лесі Українки розглядалися у статтях Н. П. Дарчук, Л. А. Алексієнко, Н. О. Данилюк. У цій публікації ми пропонуємо результати нашої курсової роботи над лематизованими текстами поетеси.

Оскільки поки що не укладено нове академічне видання творів Лесі Українки, за джерельну базу дослідження нами було взято три прижиттєві збірки Лариси Петрівни Косач: „На крилах пісень" (1893), „Думи і мрії" (1899), „Відгуки" (1902).

Підкреслимо, що параметризована база даних – це багатоаспектна і багатофункціональна система даних в електронному вигляді. На думку Н. П. Дарчук і Л. А. Алексієнко, вона повинна включати:

а) текстовий масив, до якого мають увійти твори з нового академічного видання Лесі Українки, що відповідають рукописним варіантам, або прижиттєві видання, максимально наближені до авторського оригіналу;

б) алфавітно-частотний словник поетичного мовлення;

в) словник-конкорданс, який подає лексико-семантичну та стилістичну характеристики реєстрового слова;

г) словники синонімів, антонімів, паронімів;

ґ) словник тропів (епітетів, метонімій, метафор, синекдох, порівнянь тощо);

д) інтегрований словник, який містить граматичну, лексико-семантичну, синтаксичну та стилістичну інформацію [1, 345].

У лабораторії народознавчої лексики Волинського державного університету розпочато створення частотного словника на базі названих збірок. Робота проводиться з використанням розробленого працівниками лабораторії комп'ютерної лінгвістики пакету алгоритмів і програм автоматичного та статистичного аналізів у три етапи:

1) опрацювання результатів автоматичного морфологічного аналізу кожного з текстів із метою розпізнавання граматичної та лексико-граматичної омонімії;

2) редагування та виправлення можливих помилок при автоматичній лематизації;

3) власне створення алфавітно-частотного словника.

Студентам філологічного факультету ВДУ було роздано частини лематизованих текстів. Під час дослідження цього матеріалу ми зафіксували проблемні моменти в роботі автоматичного морфемного сегментатора:

а) неправильне визначення початкових форм дієслів з інфінітивами на -ть, а також деяких займенників, прикметників;

б) нерозпізнавання власних назв;

в) неправильне визначення сполучників;

г) позначення усіх розділових знаків англійським словом BAD;

ґ) нерозпізнавання пестливо-зменшених утворень іменників і прикметників;

д) неправильне визначення лексичного значення слова, зумовленого наявністю омонімії, що є однією з найскладніших ситуацій при проведенні комп`ютерного текстового аналізу.

Тільки в ручному режимі розмежовується лексична та граматична омонімія. Наприклад, автоматично не розпізнаються іменник *мати* і дієслово *мати*, прикметник *мала́* і дієслово *ма́ла*; лексичні омоніми типу: *коса* – сільськогосподарський інструмент або довге волосся, або вузький півострів.

Укладання частотного словника здійснюється у три етапи: перший – аналіз результатів автоматичного морфологічного аналізу кожного із текстів з метою розпізнавання граматичної та лексико-граматичної омонімії; другий – редагування та виправлення можливих помилок при автоматичній лематизації; третій – власне створення алфавітно-частотного словника.

Відповідно до отриманих від працівників лабораторії кодів до комп'ютерного шифрування тексту, зліва подається слово із тексту, за ним у квадратних дужках – коди за частинами мови й граматичними показниками. Справа міститься початкова форма.

МЕНЕ [МР] я НЕ [60] НЕ ЖІНКО [КК] ЖІНКА

Кожна частина мови має свою схему аналізу.

Для іменників зазначаються такі граматичні категорії, як рід, число, відмінок. Для іменників, що мають форму множини, реєстрове слово подається у множині. Для незмінюваних імен-

ників вказується лише їх належність до цієї частини мови. Прикметник об'єднує відмінкові форми всіх родів в однині та множині, а також розпізнається окремо за ступенями порівняння. Вважаючи ці ступені порівняння окремими словами, укладачі словника дотримуються поглядів тих мовознавців, які відносять ступенювання до словотвору, а не до словозміни. За схемою іменника лематизуються деякі займенники й кількісні числівники, дієприкметники. Дієслова об'єднує інфінітив (з *-ся*, без *-ся*), форми часу та наказовий спосіб, що розрізняється за числами і способами. Дієприслівник розрізняється за часом та видом. Для форм на *-но, -то* частина мови не вказується. Слова з часткою, написані через дефіс, вважаються окремими словами.

Отже, за допомогою морфемного аналізу поетичних текстів на основі автоматичного морфемного сегментатора можна буде одержати ілюстрацію особливостей використання слова у поданому матеріалі. Ці дані дадуть чітке уявлення про функціонування того чи іншого слова в мовленні поетеси і можуть слугувати базою для подальших досліджень її творчості.

Словник мови поезії Лесі Українки на комп'ютерній основі стане частиною практичної побудови автоматизованих систем в Українському мовно-інформаційному фонді НАН України, що дасть можливість її розробникам виявляти фактичний матеріал для опису таких теоретичних аспектів комп'ютерної лінгвістики, на яких тримаються багато сучасних інформаційних систем інтелектуального призначення.

Вважаємо, що роботу над словником поезії Лесі Українки слід продовжити в Інституті Лесі Українки, створеному на базі Волинського державного університету.

Література

1. Алексієнко Л. А., Дарчук Н. П. Принципи створення параметризованої бази даних за поетичними текстами Лесі Українки // Леся Українка і сучасність (До 130-річчя від дня народження Лесі Українки): Зб. наук. пр.– Луцьк: Волин. обл. друк., 2004.– С. 344–352.

2. Данилюк Н. Словник мови поезії Лесі Українки на комп'ютерній основі // Філологічні студії.– Луцьк, 2004.– № 4.– С. 85–89.

3. Українка Леся. На крилах пісень: Поезії.– Л., 1893; Перевидання – К.: Веселка, 1994.– 128 с.

4. Перебийніс В. С., Муравицька М. П., Дарчук Н. П. Частотні словники та їх використання.– К.: Наук. думка, 1985.– 204 с.

5. Пещак М. М. Нариси з комп'ютерної лінгвістики.– Ужгород: Закарпаття, 1999.– 200 с.

6. Українська мова: Енциклопедія.– К.: Укр. енцикл., 2000.– 752 с.

7. Широков В. А. Інформаційна теорія лексикографічних систем.– К.: Довіра, 1998.– 331 с.

8. Синтаксический анализ научного текста на ЭВМ / Отв. ред. Т. А. Грязнухина.– К.: Наук. думка, 1999.– 272 с.

Oleschuk I., Tandryk N., Frolova I. Poetry Language Dictionary of Lesya Ukrainka on Computer Basis.

The article is devoted to the problem of creation of poetic works language dictionary of Lesya Ukrainka with the use of the computer programs for working at a text of belles-lettres style.

Key words: dictionary of language of poetry, work of art, computer program, division, parameterized base.