«Золотий корпус» творів Лесі Українки

Н. Дарчук

Інститут філології Київського національного університету імені Тараса Шевченка, Київ, Україна
Corresponding author. E-mail: nataliadarchuk@gmail.com

Paper received 20.08.20; Accepted for publication 10.09.20.

https://doi.org/10.31174/SEND-Ph2020-234VIII69-01

Анотація. У статті розглянуто лінгвістичні засади створення електронного корпусу поетичних творів Лесі Українки як бази для дослідження ідіостилю поетки. Корпус текстів і словники розміщено на мовно-інформаційному порталі www.mova.info. Лінгвістичними засадами є комп'ютерна граматика АГАТ, яка зі створеним програмним забезпеченням дозволить швидко одержувати за запитом словники лексем, словоформ, морфів, словосполучень, лексико-семантичних груп із кількісними показниками, які необхідні для вивчення морфологічної, синтаксичної та семантичної структури авторського мовлення

Ключові слова: *Корпус української мови, морфологічна анотація, синтаксична анотація, морфна анотація, семантична анотація.*

Непослабний інтерес до знакової для української культури постаті Лесі Українки, чиє 150-річчя відзначатиметься у 2021 році, заохочує усе нові й нові дослідження її творчості. Оскільки сучасну наукову добу характеризує активне впровадження новітніх комп'ютерних засобів, закономірним кроком є створення відповідного інструментарію для роботи з текстами поетеси. У цьому полягає водночас і **актуальність**, і **новизна** проекту електронного дослідницького корпусу мови Лесі Українки в межах корпусу української мови (КУМ), створеного в лабораторії комп'ютерної лінгвістики Інституту філології Київського національного університету імені Тараса Шевченка.

Матеріалом слугували прижиттєві видання Лесі Українки: «Відгуки (із циклу «Невольницькі пісні», «Легенди», «Ритми», «Хвилини»); «Думи і мрії» («Відгуки», «Мелодії», «Невільничі пісні», «Поеми»); «На крилах пісень» (I та II розділи). Загальний обсяг матеріалу **55 638** слововживань. Корпус розміщено на мовно-інформаційному порталі mova.info.

Мета Корпусу – створення такого лінгвістичного і програмного продукту, який би, з одного боку, надавав широкі відомості про мову Лесі Українки, з іншого – був зручним для використання у різноманітних дослідженнях. Така мета вимагала виконання низки **завдань**, серед яких: лінгвістичний аналіз текстів Лесі Українки; укладання бази даних зафіксованих у них мовних одиниць із їхніми граматичними та кількісними характеристиками; створення зручного користувацького інтерфейсу, за яким можна було б здійснювати пошук, сортувати та статистично опрацьовувати усю зібрану в базі інформацію відповідно до потреб дослідників.

Лінгвістичне опрацювання здійснювалося двома способами: 1) автоматично за допомогою автоматичного морфологічного (для визначення частин мови та їх граматичних форм) та автоматичного синтаксичного аналізу (виділення словосполучень і приписування їм необхідної інформації, побудова дерев залежностей); 2) автоматизовано в результаті роботи лінгвіста, який контролює правильність результатів аналізу, редагує й усуває можливі помилки. Корпус вичитаний, помилки виправлені, тому вважаємо, що Корпус, з лінгвістичної точки зору, є «золотим» на кшалт Браунського корпусу (англ. Brown Corpus), що був створений в університеті Брауна в 60-ті роки).

Анотування здійнювалося за такими параметрами:

– бібліографічна інформація (рік написання, видання, жанр тощо) виконуються вручну за допомогою спеціалізованої програми-редактора, після чого комп'ютер автоматично опрацьовує текст, доповнюючи його лінгвістичною анотацією.

– структурна інформація (поділ текстів на слова, словосполучення, речення, абзаци);

– лінгвістична інформація; цей тип анотації представлений чотирма рівнями:

- морфологічна (частиномовна належність слів, їхні леми й граматичні категорії);
- морфна будова лексеми і словоформи;
- синтаксична (види словосполучень за частинами мови, ролі слів у словосполученнях (головне чи залежне), типи синтаксичних зв'язків і дерево залежностей цілого речення;
- семантична інформація.

Лінгвістична інформація здійснюється за допомогою комп'ютерної граматики АГАТ [1], яка складається з чотирьох модулів – **морфологічного, морфемного, синтаксичного та семантичного. Перший модуль** здійснює аналіз слів за граматичними характеристиками (частиномовною належністю і граматичними категоріями). **Другий модуль** подає інформацію про морфну будову слова / словоформи, омонімію й аломорфію коренів. **Третій модуль** має два рівні опрацювання – синтаксис словосполучень, організований за принципами граматики безпосередніх складників, та синтаксис речень, реалізований у вигляді дерева залежностей. **Четвертий** класифікує лексеми за семантичним кластерами – доменами.

Результати роботи аналізаторів оформлюються у вигляді спеціальних кодів, приписаних словам та їх сполученням, і заносяться до відповідних баз даних.

Корпус творів Лесі Українки є динамічним, функціональні можливості його інтерфейсу забезпечують гнучкий варіативний пошук, який дозволяє дослідникові: **по-перше**, виокремити для роботи певний жанрово-тематичний підкорпус; **по-друге**, здійснювати пошук як за конкретними словами, так і за певними лінгвістичними характеристиками (частини мови, граматичні форми); **по-третє**, виявити закономірності вживання як окремих слів у специфічних граматичних формах (а отже, і в певних синтаксичних функціях), так і загалом певних частин мови; **по-четверте**, побудувати конкорданс певної лексеми або словосполучення, частотні словники як базу комп'ютерної лексикографії. Конкорданс дозволяє

простежити закономірності контекстної реалізації мовних одиниць, вивчити їхню граматичну й лексичну сполучуваність; частотний словник надає інформацію про ступінь їхньої вживаності, допомагає визначити мовне ядро текстів чи їхніх груп, виявити у текстах ключові слова, простежити й порівняти статистичні параметри функціональних або індивідуально-авторських стилів (табл.1).

Табл. 1. Фрагмент бази даних інтегрованого частотного словника всіх прижиттєвих творів Лесі України (загальний обсяг текстів 55 638 слововживань, обсяг словника 5356 лексем)

freq

Код	wrd	cls	xabs	xsred	sigma	v	R	context
1935	і	С	1020	26,842105263	23,367073406	0,8705380288	129	
2902	не	Ь	875	22,236842105	18,607645199	0,8367935119	124	
5506	я	М	866	22,789473684	20,132043338	0,8833922019	102	
773	в	П	810	21,315789474	19,044342483	0,8934382893	129	
2887	на	П	630	16,578947368	13,84955773	0,8353701488	121	
5017	ти	М	477	12,552631579	11,9446148	0,9515626046	68	
4965	той	О	396	10,052631579	8,2637445316	0,8220478853	99	
707	він	М	389	10,236842105	8,7337105992	0,8531645315	55	
256	бути	Г	352	9,2631578947	7,6793423657	0,8290199145	82	
1889	з	П	345	9,0789473684	7,5003693353	0,8261276369	88	
5142	у	П	333	8,7631578947	7,5784447933	0,8648075139	93	
5456	що	М	308	8,1052631579	6,4308596376	0,7934177475	81	
2505	мій	О	296	7,7631578947	6,5448061495	0,8430597752	80	

Область навигации

Запись: 14 из 5356 | Нет фильтра | Поиск

Синтаксична інформація. Оскільки словосполучення водночас є і **номінативною одиницею,** і елементарною **синтаксичною конструкцією,** у словнику вона представлена: моделями словосполучень і лексично. На основі словосполучень формуються члени речення і синтаксичні категорії та явища, розкривається сутність міжслівних зв'язків. До словосполучень зарахувалися: підрядні, сурядні, предикативні сполуки (рис.1). Всі вони демонструють важливу властивість: кожен з компонентів цих конструкцій – носій морфологічних ознак із релятивною функцією, тобто властивістю приєднувати до себе словоформи і приєднуватися до них, утворюючи певний синтаксичний комплекс.

Структура цілого речення представлена у вигляді дерева залежностей (рис.1), теоретичні засади якого сформульовані Л.Теньєром [4] й описані І.Севбо [3]. Концепція аналізу та його послідовність така: 1) спочатку встановлюються зв'язки між словоформами і виділяються словосполучення в простому реченні / предикативній частині; 2) визначаються (ідентифікуються) складники – предикативні частини (у складному реченні); 3) встановлюються зв'язки між предикативними частинами.

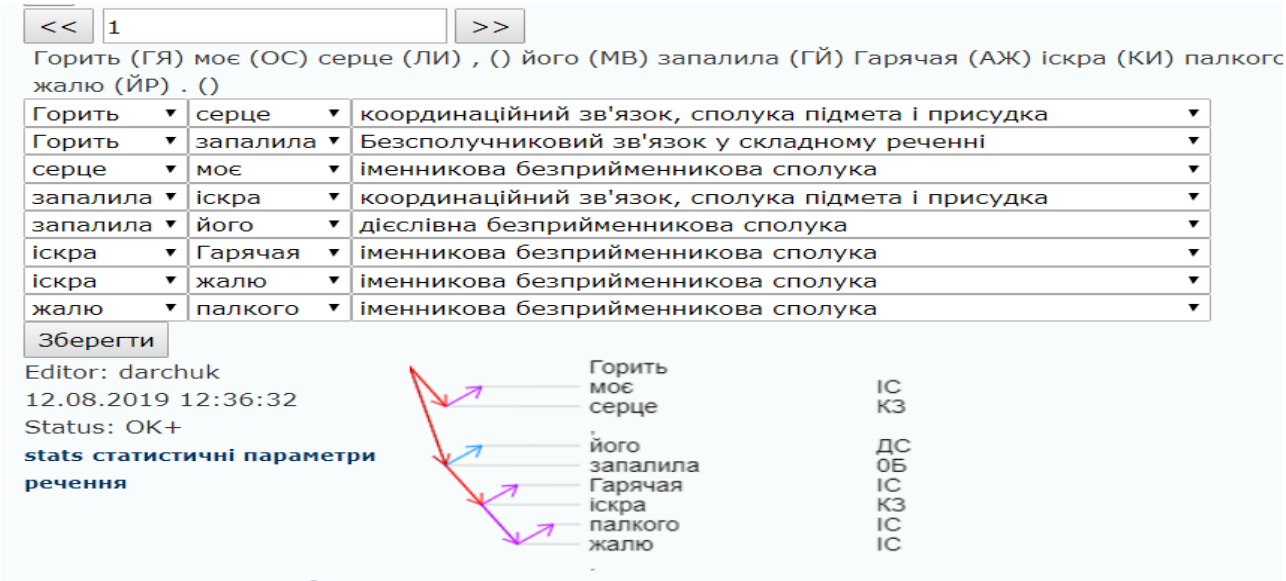

Рис.1 Бінарні сполуки і типи синтаксичних зв'язків з поезії «Горить моє серце» («Мелодії»// «Думи і мрії»)

Автоматично побудовано 3124 дерева залежності, вони автоматизовано відредаговані і складають банк синтаксично розмічених речень. Це уможливлює опис сполучуваності слів за такими граматичними параметрами:

1) **тип сполуки**, до якої входить аналізоване слово (за частиномовною належністю головного слова): іменникова, прикметникова, дієслівна, прислівникова, числівникова, займенникова;

2) **роль слова у сполуці** (тільки для підрядних конструкцій, оскільки в сурядних та у випадку координації обидва члени є рівноправними): ядрова (словосполучення, в яких аналізоване слово є головним); ад'юнктна (словосполучення з аналізованим словом у ролі залежного члена);

3) **тип синтаксичного зв'язку**: підрядний, сурядний, координація.

Дослідження моделей сполучуваності дозволить дати відповіді на такі питання:

- чи можливі випадки, коли певна частина мови у реченні не має залежних від нього елементів, або, навпаки, нічому не підпорядкована;
- якими класами слів може керувати слово (частина мови);
- до яких комплексів і в якій ролі може входити;
- якими комплексами може керувати;
- чи може дане слово встановлювати предикативні зв'язки, і якщо так, то з якими класами слів тощо.

ДЗ дає важливу інформацію для стилістичних висновків (рис.2). Автоматично для кожного ДЗ вираховуються параметри:

- вузлові параметри, або середня кількість вузлів у дереві;
- параметр глибини дерева, або середня кількість рівнів у дереві;
- параметр ширини, або середня кількість вузлів на одному рівні дерева;
- асиметричний параметр, який являє собою співвідношення у кількості вузлів між піддеревами, отриманими поділом навпіл другого рівня;
- гілковий параметр, який вимірюється відношенням числа кінцевих вузлів до числа рівнів;
- параметр кратності – відношення числа кратних вузлів до числа дерев (кратним називаються вузли, у якого кілька підлеглих);
- наскрізний параметр, або середня довжина наскрізного ланцюжка (наскрізний ланцюжок – це шлях у дереві, який веде від кореня до кінцевого вузла).
- Вся перерахована інформація у вигляді параметрів, яка стосується ДЗ, будується для кожного речення конкретного тексту. Такий протокол параметрів одержуємо автоматично шляхом натискання відповідної кнопки «статистичні параметри речення»

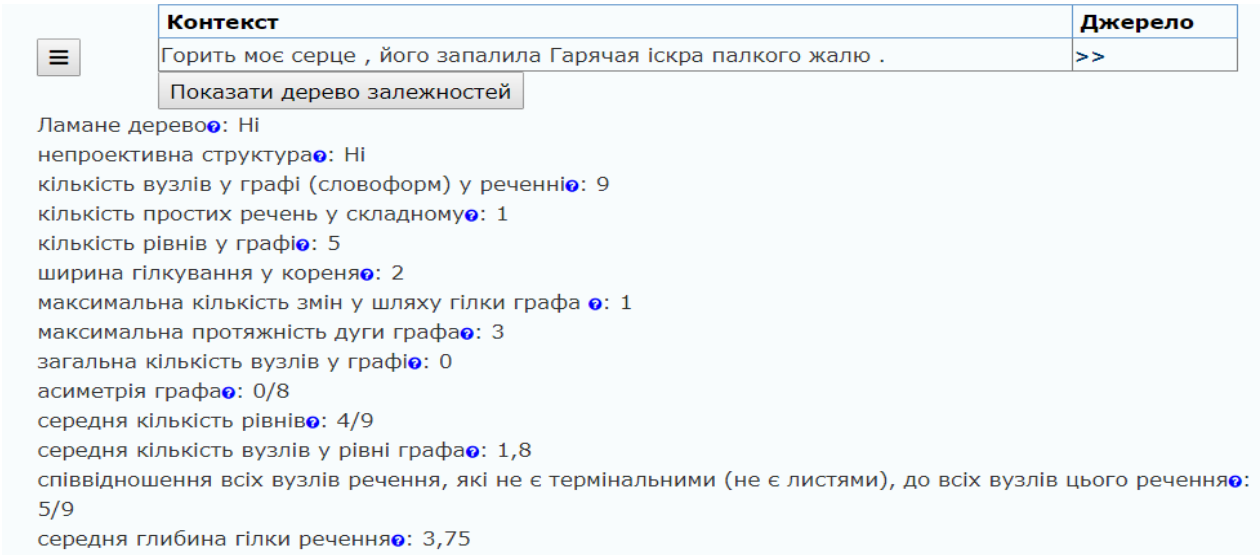

Рис.2. Статистичні параметри дерева залежностей

Така сукупність характеристик дозволила створити параметризовану базу даних, під якою укладачі розуміють багатоаспектну і багатофункціональну систему, яка включає: 1) корпус текстів – джерело різного роду словників; 2) серію алфавітно-частотних словників з усією інформацією про слово: граматичною, лексико-граматичною, стилістичною та статистичною (абсолютна та середня частота, міра коливання середньої частоти, коефіцієнт стабільності); 3) алфавітно-частотні словники слів та слововживань спільної лексики. 4) словники неолексем; 5) словники синтаксичних моделей керування: дієслівних, іменникових, атрибутивних; 6) серію морфемних та словотвірних словників з частотними характеристиками морфа/морфеми, за якими можна вивчати комбінаторно-дистрибутивну будову, словотвірне значення кожної афіксальної морфеми в текстах; 7) моделей багатокомпонентних речень [2]; 8) словники синонімів, антонімів, фразеологізмів, тезауруси; 9) словник-конкорданс як допоміжний інструмент для формування лексико-семантичної, синтаксичної та стилістичної характеристики слова. Ми розглядаємо «Золотий корпус» текстів Лесі Українки як реальну можливість для створення універсального (багатоцільового) електронного словника.

ЛІТЕРАТУРА

1. Дарчук Н.П. Комп'ютерне анотування українського тексту: результати і перспективи / Дарчук Н. П. – К.: Освіта України, 2013. – 543 с.
2. Darchuk N. Compiling of the Electronic Dictionary of Models of the Ukrainian Language Multicomponent Complex Sentences. Ukrainian Linguistics, 2019, 49, 117–129. 10.17721/um/49(2019).117-1293.
3. Севбо И. П. Графическое представление синтаксических структур и стилистическая диагностика / И. П. Севбо. – К.: Наук. думка, 1981. – 192 с.
4. Теньер Л.Основы структурного синтаксиса. М.: Прогресс, 1988. – 656 С.

REFERENCES

1. Darchuk, N. P. (2013). Computer Annotation of Ukrainian Text: Results and Prospects. *Publishing House "Osvita Ukrainy", 543*. ISBN: 978-617-7111-54-1 [in Ukrainian]
3. Sevbo I. Graphic Representation of Syntactical Structures and Stylistical Diagnostic. K.: Nauk.Dumka, 1981. – 192 p.
4. Tesnière Lucien. Elements of Structutal Syntax. M. Progress, 1988. – 656 C.

"The Golden Corpora" of Lesja Ukrayinka's poetry

N. Darchuk

The article reviews linguistic principles of design and development of digital Corpora of Lesja Ukrayinka's poetry as the basis for the study of linguistic peculiarities of the individual writing style. The Corpora with glossaries are located in the World Wide Web resource (www.mova.info). The developed system for the study of morphological, syntactical and semantic structures of the writing style based on computerized grammar checking software will enable to receive immediately under request the lists of lexemes, word forms, morphs, word combinations, lexical-semantic groups with statistical key figures.

Keywords: *The Corpora of Ukrainian texts, morphological annotation, syntactical annotation, morph annotation, semantic annotation.*