Корпус текстів Тараса Шевченка як джерело для мовознавчих досліджень

Н. Дарчук

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

Paper received 05.03.17; Accepted for publication 15.03.17.

Анотація. У статті розглянуто лінгвістичні засади створення електронного корпусу поетичних творів Тараса Шевченка як бази для дослідження мовних особливостей авторського стилю. Створене програмне забезпечення для вивчення морфологічної, синтаксичної та семантичної структури авторського мовлення дозволить швидко одержувати за запитом словники лексем, словоформ, морфів, словосполучень, лексико-семантичних груп із кількісними показниками. Корпус текстів і словники розміщено на мовно-інформаційному порталі www.mova.info.

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

Стан лінгвістики останніх десятиліть Ю.Апресян [1, с. 13] визначає метафорою «золотий вік лексикографії» у зв'язку з безпрецедентним зростанням кількості й різноманіття словників, удосконаленням методики лексикографування, принципів системного опису лексики й увагу сучасної теоретичної лінгвістики до поглибленого опису окремих лексем. Такий прорив у мікросвіт значення одного слова спонукає філологів до всебічного його лінгвістичного «портретування», а значить – до інтегрального опису мовних фактів. Справді, морфологічні, синтаксичні, семантичні, прагматичні, стилістичні, комунікативні, сполучуванісні властивості лексем є вивідними безпосередньо з текстів, які були і є невичерпним джерелом для словників і граматик. В.Звегінцев зауважив, що «мовленнєва система… не є цілісним утворенням, але являє собою ієрархію мовленнєвих підсистем, які мають границі вживання (соціальні, професійні та інші «мови»). Кожна з цих мовленнєвих систем по-своєму і в своїх потребах перетворює мовленнєву **«значимість»** в мовленнєве **«значення»**, дотримуючись разом з тим смислових (семантичних) меж, які встановлюються даною «значимістю». В результаті й виникає те, що прийнято називати «**багатозначністю**»» [3, с. 280] (виділено нами Н.Д.). Тому вивчення мовленнєвої значимості в різних стилях, дискурсах, підмовах тощо дасть можливість лінгвістам (не тільки лексикографам, а й літературознавцям!) глибоко досліджувати рух і розвиток значення мовних одиниць.

Незаперечним є й той факт, що в останні десятиліття фахівцями галузей, що мають справу з комп'ютерним аналізом текстів, гостро відчувається потреба у якомога більшій кількості функціональних характеристик мовних одиниць у різних типах текстів. Для проведення теоретичних і прикладних, зокрема дидактичних досліджень, спеціалісти в галузях когнітивної лінгвістики, стилістики, семасіології, функціональної граматики, словотвору бракує узагальнених даних про закономірності функціонування мовних одиниць. Необхідну лінгвістичну інформацію для подальшого опрацювання її у філологічних студіях можна одержати з корпусів текстів, розбудова яких є ознакою нашого часу.

В Інституті філології Київського національного університету імені Тараса Шевченка створено Корпус текстів української мови (КУМ) (www.mova.info) з необхідними й достатніми кількісними та якісними параметрами для укладання на їхній основі словників і граматик української мови. Відповідно до класифікації корпусів КУМ є **дослідницьким**, оскільки призначений для вивчення різних аспектів функціонування мовної системи і зорієнтований на широкий клас лінгвістичних завдань, й **статичним**, оскільки відображає певний часовий зріз мовної системи. Водночас корпус дає можливість дослідити мовні явища в динаміці, отже, він є **динамічним**, а також **інтерактивним**, оскільки користувач, здійснюючи дослідження, може виокремити з генерального корпусу робочий корпус, який є частиною генерального.

Головна увага приділяється анотуванню корпусного матеріалу, тобто процесу введення формалізованої лінгвістичної інформації в електронний текст. Аплікативне призначення корпусних даних – морфологічні, синтаксичні, лексикологічні, лексикографічні дослідження – детермінує тип лінгвістичної анотації корпусу. **Морфологічна анотація** передбачає визначення морфологічних параметрів слова: частиномовну приналежність і категорійні ознаки кожної словоформи тексту. **Морфна анотація** полягає у сегментуванні кожної словоформи тексту за типами морфів і подальшому автоматичному укладанні серії морфемних та словотвірних словників із частотними характеристиками морфа/морфеми, за якими можна вивчати комбінаторно-дистрибутивну будову слова. **Синтаксична анотація** пов'язана з автоматичним опрацюванням кожного речення: виокремленням у ньому словосполучень та приписуванням кожному з них інформації про тип (дієслівний, іменниковий тощо), вид синтаксичного зв'язку та семантичного відношення. **Лексико-семантична інформація**, яка приписується кожному слову, відповідає таксономічній класифікації, розробленій відповідно для кожної частини мови.

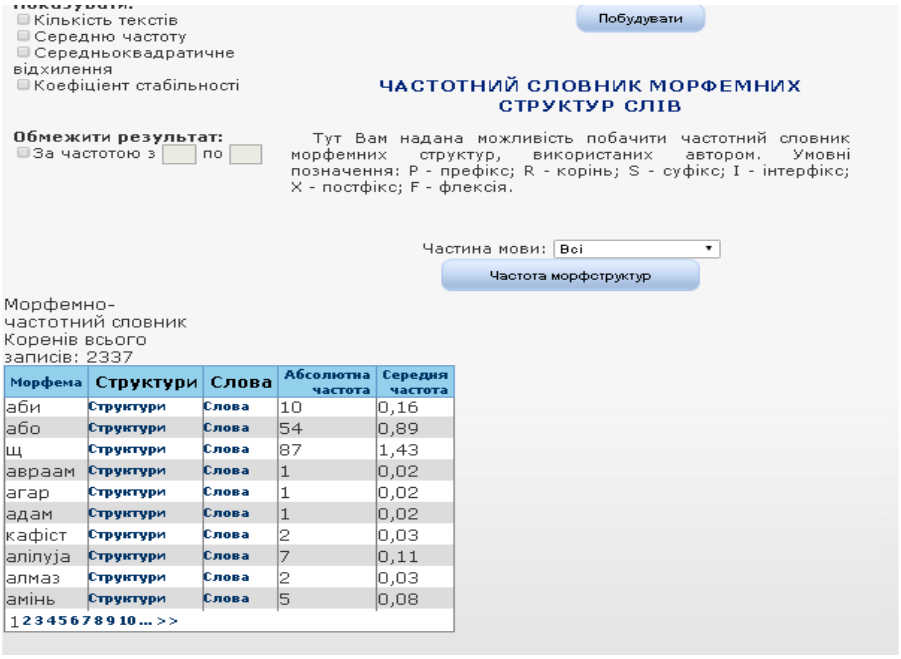
Нескінченний інтерес до постаті Тараса Григоровича Шевченка породжує усе нові й нові дослідження його творчості. Оскільки ж сучасну наукову добу характеризує активне впровадження новітніх комп'ютерних засобів, закономірним кроком є створення відповідного інструментарію для роботи з текстами видатного поета. У цьому полягає водночас і **актуальність**, і **новизна** проекту електронного словника мови Тараса Шевченка.

Метою дослідження було створення такого лінгвістичного і програмного продукту, який би, з одного боку, надавав широкі відомості про мову Тараса Шевченка, з іншого ж, був зручним для використання у

різноманітних дослідженнях. Така мета вимагала виконання низки **завдань**, серед яких: лінгвістичний аналіз текстів Т. Шевченка; 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.

Об'єктом дослідження були оцифровані поетичні тексти Т. Шевченка, **предметом** – їх лексична, морфологічна та синтаксична характеристики.

Укладання словника відбувалося в декілька етапів, за якими опрацьовувався матеріал і лінгвістично, і статистично, в результаті чого укладалася словникова стаття з багатоаспектною інформацією відносно граматичних характеристик з контекстами вживання тієї чи іншої лексичної одиниці, морфеми або словосполучення та абсолютною частотою вживання у досліджуваних текстах. Інтерфейс словника дозволяє робити пошук контекстів не тільки за словоформами, а й за лексемами, а також за такими параметрами, як: морфна модель слова (рис.1), частиномовна належність, категорійні характеристики (рід, число, відмінок, час, особа тощо) (рис. 2), модель словосполучення з кількісними даними з відповідними контекстами вживання (рис. 3).



Морфема	Структури	Слова	Абсолютна частота	Середня частота
аби	Структури	Слова	10	0,16
або	Структури	Слова	54	0,89
щ	Структури	Слова	87	1,43
авраам	Структури	Слова	1	0,02
агар	Структури	Слова	1	0,02
адам	Структури	Слова	1	0,02
кафіст	Структури	Слова	2	0,03
алілуја	Структури	Слова	7	0,11
алмаз	Структури	Слова	2	0,03
амінь	Структури	Слова	5	0,08

1 2 3 4 5 6 7 8 9 10 ... >>

Рис. 1. Інтерфейс програми встановлення морфної структури слів словника Тараса Шевченка

На рис. 1 представлено фрагмент алфавітно-частотного словника коренів, розташованих за алфавітом з частотними характеристиками. Такий словник дає інформацію про всі типи морфів з його оточенням. На рис. 2 представлено інтерфейс програми пошуку лексем/словоформ з кількісними характеристиками і контекстами вживання, а рис. 3 ілюструє словник словосполучень, який автоматично укладається за текстами.

Оскільки синтаксичні і лексичні зв'язки виявляються при взаємодії різної кількості одиниць, але у більшості випадків не виходять за межі речення, то граничним контекстом був вибраний відрізок від крапки до крапки, тобто речення. Мінімальним відрізком був контекст в три позиції 1– X +1 (де X – аналізована словоформа), тобто слово у препозиції та постпозиції до аналізованої словоформи. Надфразові єдності не враховуються. Якщо в окремих випадках зв'язки аналізованого слова виходять за рамки виписаного речення, вони втрачаються. Синтаксична інформація у словнику представлена моделями словосполучень. Опис синтаксису через словосполучення мотивується тим, що це міжрівнева лексико-морфолого-синтаксична одиниця [2]. Словосполучення водночас є і **номінативною одиницею,** і елементарною **синтаксичною конструкцією.** На її основі формуються члени речення і синтаксичні категорії та явища, розкривається сутність міжслівних зв'язків. До словосполучень ми зарахували підрядні, сурядні та предикативні сполуки, оскільки 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.

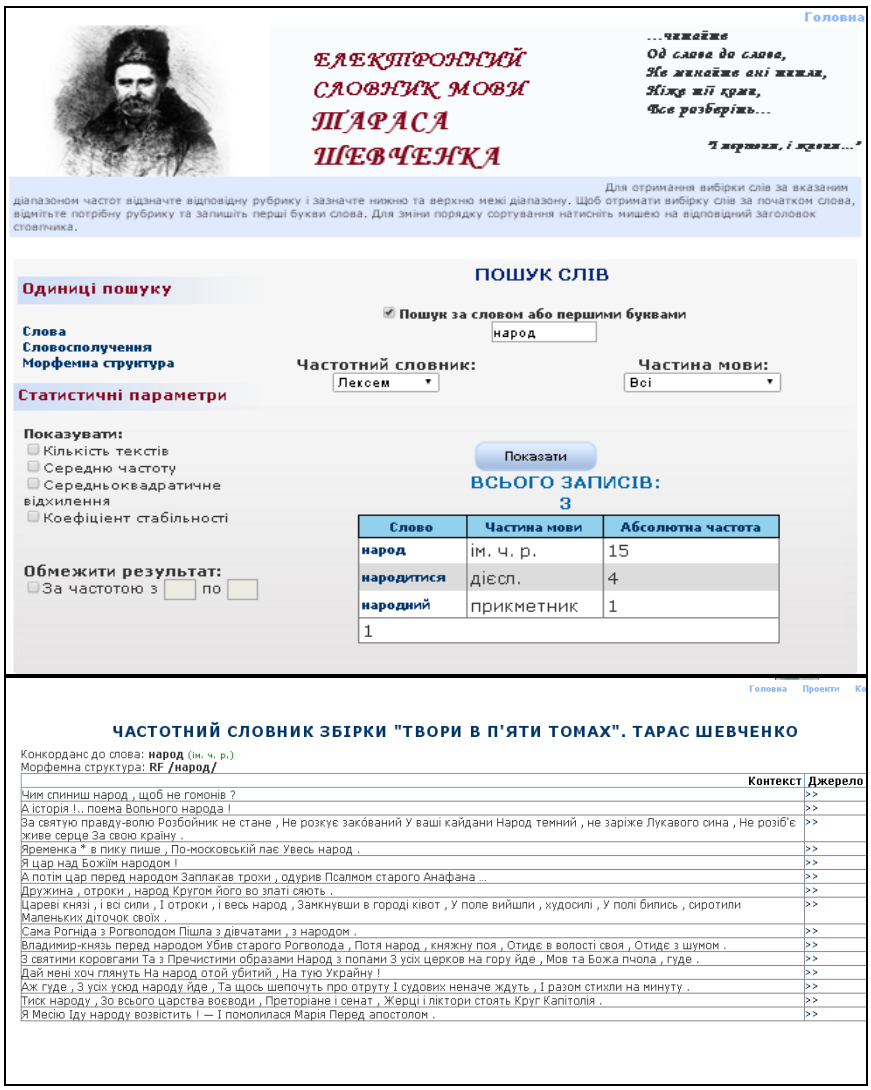


Рис. 2. Інтерфейс програми пошуку слів/словоформ з контекстом вживання на матеріалі словника Тараса Шевченка

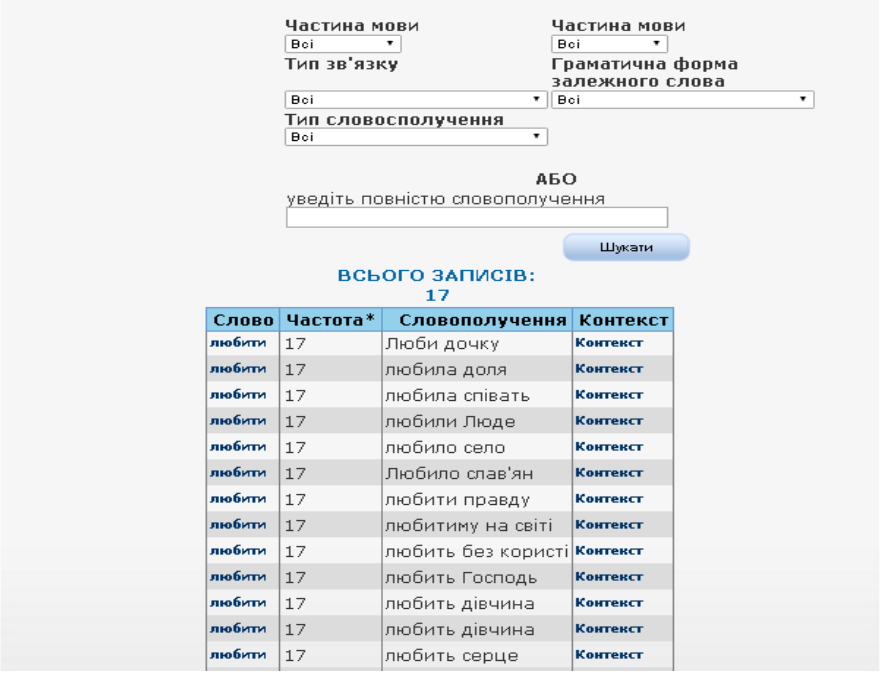


Рис. 3. Інтерфейс програми пошуку словосполучень з контекстом вживання на матеріалі словника Тараса Шевченка

Статистичні характеристики одиниць словника подані у вигляді двох параметрів – абсолютної та середньої частоти. Самої лише абсолютної кількості прикладів може бути недостатньо для відтворення об'єктивної картини, оскільки слова не завжди рівномірно розподілені по текстах. Для врахування такого

явища вводиться поняття середньої частоти, а також середнє квадратичне відхилення абсолютної частоти – параметр, який характеризує рівномірність розподілу входжень слова у текстах. Вся інформація про сполучуваність конкретної словоформи заноситься до бази даних, яка має таку структуру: колонки з номерами слова, речення й тексту, словосполучення, кожне зі слів, що його складають, та їх початкові форми (леми), граматичні класи слів-компонентів словосполучення і тип синтаксичного зв'язку.

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); 2) автоматизовано в результаті роботи лінгвіста, який контролює правильність результатів аналізу, редагує й усуває можливі помилки.

Сполучуваність слів описувалася за такими 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.

Параметри, за якими описані одиниці словника, відкривають широке поле для лінгвістичного аналізу творчості Т. Шевченка. Так, наприклад, 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.

В якості ілюстрації наведемо відомості про отриманий у результаті дослідження синтаксису творів Тараса Шевченка перелік частиномовних моделей сполучуваності:

I. Моделі дієслова

1. Ядрові моделі дієслова

Г И – *благословить дітей;* Г А – *ставати зеленим;* Г П И – *дивлюсь на тебе;* Г Г – *жити хочу.*

2. Ад'юнктні моделі дієслова

Г Г – *ліг одпочить;*

3. Предикативні моделі дієслова

У цьому виді зв'язку дієслово є присудком або підметом:

И Г – *Вітер віє.*

4. Сурядні моделі дієслова

Г Г – *гралися, хвалили.*

II. Моделі іменників

1. Ядрові зв'язки іменників

А И – вольної села; О И - *мій квіте;* И И – день радості;

2. Предикативні зв'язки іменника

И Г – *верба похилилась.*

III. Моделі прикметників

1. Ядрові моделі прикметників

Н А – *дуже цікаве;* А П И – *великая в женах;* А СС И – *червоних як калина*

2. Ад'юнктні моделі прикметників

Г А – *був дужий;* А И – *вольної села.*

3. Предикативні моделі прикметників

И А – *ангелом святим.*

4. Сурядні зв'язки прикметників

А СС А – *зелений і синій кольори*

IV. Моделі прислівників

1. Ядрові зв'язки прислівника. Ядровим вважається прислівник, якщо він не може бути опущений без порушення змісту, тобто є лексично зв'язаним.

И Н – *гріха менше;* Н Н – *досі нудно;* Н П М – *тихо в мене*

2. Ад'юнктні зв'язки прислівників

Н Н – *досі нудно;* Н А – *надто молодую.*

***3.* Предикативні зв'язки прислівників**

И Н – *билини кругом*

4. Сурядні зв'язки прислівників

Н (СС) Н – *любенько та тихо*

Ще одним важливим завданням в галузі лексикології та лексикографії є побудова семантичних авторських словників, зокрема творів Тараса Шевченка. Н. Шведова вказала на надзвичайно важливу роль ідеографічного впорядкування лексики для виявлення особливостей бачення й відображення світу мовцями. Дослідниця слушно зауважила, що підмножини ЛСВ, які наповнюють нижні рівні ідеографічного дерева, не є простими групами одиниць, відібраними на підставі спільної семантики. На думку науковця, вони виконують важливе конструктивне й інформаційне завдання, відкриваючи перед нами певний «шматочок дійсності» [4]. У зв'язку з цим розпочато проект зі створення семантичного словника поетичних творів Тараса Шевченка, в якому лексика групуватиметься в межах певної частини мови за таксономічними класами.

Наведені вище дані є лише прикладом використання електронного словника мови Тараса Шевченка і не вичерпують його можливостей. Коло потенційних досліджень значно ширше й охоплює морфологічні, лексичні, синтаксичні й семантичні й стилістичні розвідки. Цьому сприяє різноманітність лінгвістичної розмітки текстів та гнучкість інтерфейсу користувача, що свідчить про високу ефективність електронних словників і, зокрема, словника мови Тараса Шевченка.

ЛИТЕРАТУРА

1. Апресян Ю. Д. О толковом словаре управления и сочетаемости русского глагола / Ю. Д. Апресян // Словарь. Грамматика. Текст / Рос. Акад. Наук, Отд-ние лит. И яз., Ин-т рус. Яз. Им. В. В. Виноградова. — М., 1998. — С. 13—43.

2. Головин Б. Н. Введение в языкознание / Б. Н. Головин. — М. : Высш. шк., 1966. — 332 с.
3. Звегинцев В. А. Теоретическая и прикладная лингвистика / В. А. Звегинцев. — М. : Просвещение, 1968. — 336 с.
4. Шведова Н.Ю. Русский язык. Избранные работы. – М.: Языки славянской культуры, 2005. – 640с.

REFERENCES

1. Apresyan, Y. O tolkovom slovare upravleniya i sochetaye-mosti russkogo glagola (On Dictionary of government and compatibility of a Russian verb) // Dictionary. Grammar. Text / Rus. Acad. of Sc., Dep. of Lng. and Lit-re., Inst-te of Rus. Lang. of Vinogradov V. – M. 1998. – P.13 – 43
2. Golovin, B. Vvedeniye v Yazykoznanye (Introduction into Linguistics) / Govolin B. – M.: Vysh. Shk. Publ., 1966 – 333 P.
3. Zvegintsev, V. Teoreticheskaya i prikladnaya lingvistika (Theoretical and Applied Linguistics) / Zvegintsev V – M. : Prosveshcheniye Publ., 1968. – 336 P.
4. Shvedova, N. Russkiy Yazyk. Izbrannieye raboty (The Russian Language. Selected Works) – M.: Yazyki slavyanskoy kultury, 2005. – 640 P.

The Corpus of texts of Taras Shevchenko as the source for linguistic studies

N. Darchuk

Abstract. The article reviews linguistic principles of design and development of electronic Corpus of Taras Shevchenko poetry as the basis for study of linguistic pecularities of the writing style. The developed software for study of morphological, syntactical and semantic structures of the writing style will enable to receive immediately under request the lists of lexemes, word forms, morphs, word combinations, lexical-semantic groups with statistical indicators. The text Corpus including dictionaries are located in the World Wide Web resource (www.mova.info)

***Keywords**: The Corpus of Ukrainian texts, morphological annotation, syntactical annotation, morphe annotation, semantic annotation.*

Корпус текстов Тараса Шевченко как источник джерело для языковедческих исследований

Н. Дарчук

Аннотация. В статье рассматриваются лингвистические основы создания электронного корпуса поэтических произведений Тараса Шевченко как базы для исследования языковых особенностей авторского стиля. Созданное программное обеспечение для изучения морфологической, синтаксической и семантической структуры авторской речи позволит быстро получить по запросу словари лексем, словоформ, морфов, словосочетаний, лексико-семантических групп с количественными показателями. Корпус текстов и словари размещены на информационно-языковом портале www.mova.info.

***Ключевые слова**: Корпус текстов украинскоого языка, морфологическая аннотация, синтаксическая аннотация, морфная аннотация, семантическая аннотация.*