TEKA KOM. POL.-UKR. ZWIĄZ. KULT. – KUL, 2018, 191–206

КОРПУС УКРАЇНСЬКОЇ МОВИ – КОМП'ЮТЕРНА ЕКСПЕРТНА СИСТЕМА ЛІНГВІСТИЧНОГО АНАЛІЗУ УКРАЇНСЬКОМОВНОГО ТЕКСТУ

Оксана Зубань
Київський національний університет імені Тараса Шевченка (Україна), Київ, Україна
ORCID: 0000-0002-2644-3892

Анотація. У статті представлено структуру та засади автоматичного укладання експертної системи лінгвістичного аналізу «Корпус української мови». Методика формалізованого опису мовних одиниць тексту, запропонована у створенні Корпусу, забезпечує проведення автоматичного морфологічного, морфемного, синтаксичного, семантичного аналізів українськомовного тексту, а також автоматичне укладання різноманітних електронних частотних словників за текстовими вибірками.

Ключові слова: Корпус української мови, Електронний частотний словник, база даних, автоматичний лінгвістичний аналіз.

ВСТУП

В останні десятиріччя в центрі наукових досліджень різних сфер гуманітарних знань та інформаційних технологій знаходиться текст як засіб передачі інформації, збереження знань і культури, організації соціальної комунікації, а у філології як об'єкт літературознавчих та лінгвістичних студій. Тому уже на перших етапах вивчення текстів постають завдання: дібрати репрезентативний текстовий матеріал; швидко та ефективно вилучити з текстів необхідну для дослідження інформацію. У сучасній комп'ютерній лінгвістиці ці завдання виконують корпуси текстів.

Метою статті є ознайомлення зі структурою та пошуковими можливостями Корпусу української мови [KUM 2019], який створено колективом лабораторії комп'ютерної лінгвістики Інституту філології Київського національного університету імені Тараса Шевченка під керівництвом доктора філологічних наук, професора кафедри української мови та прикладної лінгвістики Наталії Петрівни Дарчук.

Відома польська дослідниця Наталія Коциба в одній зі своїх статей подає критичний аналіз українських корпусів станом на 2013 р. і зазна-

чає, що в сучасній українській лінгвістиці корпуси мов неефективно використовуються у наукових мовознавчих дослідженнях: "Відсутність лінгвістичних досліджень, проведених на матеріалі корпусів української мови, в загальних рисах свідчить про два важливі моменти: з одного боку, недостатньою є поінформованість потенційних користувачів обох корпусів, що є наслідком їх недостатньої популяризації, а з іншого боку, якість цих корпусів і рівень їх доступності поки що не дозволяє прийняти рішення про проведення серйозних досліджень на матеріалі зазначених корпусів"[1].

На глибоке переконання розробників Корпусу української мови (КУМ), поглиблена лінгвістична параметризація текстового матеріалу на нинішньому етапі створення КУМ відкриває широкі можливості і перспективи для глибоких лінгвістичних розвідок. Тому важливим є популяризація КУМ серед українських філологів та дослідників з інших країн. Пошук у корпусі представлений у вільному доступі в Інтернет-мережі, а крім того, на платформі КУМ розроблені автоматичні системи лінгвістичного аналізу, які працюють у режимі on-line, але мають доступ в Інтернеті за реєстрацією користувача. Розробники КУМ мають можливість у цій статті звернутися до філологічної спільноти із пропозицією про співпрацю: кожен, кого цікавить робота із системами автоматичного лінгвістичного аналізу українськомовного тексту, може отримати доступ через адмін-реєстрацію до аналітичних онлайн-платформ корпусу, звернувшись до автора статті за електронною адресою oxana.mell.zuban@gmail.com.

Українська корпусна лінгвістика має у своєму доробку декілька корпусів. У вільному доступі в мережі Інтернет представлені такі корпуси текстів української мови: Корпус української мови [KUM 2019], Генеральний регіонально анотований корпус української мови [HRAK 2019], Корпуси текстів української мови [KTUM 2019], Браунський корпус української мови [BrUK 2019]. Закритими для доступу широкого користувача є Український національний лінгвістичний корпус[2], Комп'ютерний фонд інновацій (КФІ)[3]. Серед перерахованих корпусів найглибшу лінгвістичну параметризацію має Корпус української мови.

Більшість корпусів мов є ілюстративними, вони ставлять завдання: зібрати тексти, укласти словник-конкорданс за цими текстами і параме-

---

[1] N. Kotsyba, *Praktyczny przewodnik po korpusach języka ukraińskiego*, [в:] *Praktyczny przewodnik po korpusach języków słowiańskich*, red. M. Hebal-Jezierska, Warszawa 2014, с. 182.

[2] В. Широков та ін., *Застосування Українського національного лінгвістичного корпусу в лексикографії та лінгвістичних експертизах*, [в:] *Українська лексикографія в загальнослов'янському контексті: теорія, практика, типологія*, Київ 2011, с. 285-294.

[3] Є. Карпіловська, *Тенденції розвитку сучасного українського лексикону: чинники стабілізації інновацій*, "Українська мова" 2007, 2008, № 4, с. 3-15, № 1, с. 24-35

тризувати, в кращому випадку, морфологічну інформацію (створити лематизатор) і / або лише метатекстову інформацію. У такому розумінні корпус мови виконує функцію фіксації текстів і пошуку текстових прикладів (як правило речень) за словоформою або лемою. Дослідницькі корпуси текстів, до яких належить і Корпус української мови, покликані забезпечити лінгвістичний автоматичний аналіз зібраних текстів на всіх рівнях його організації.

1. КОРПУС УКРАЇНСЬКОЇ МОВИ: СТРУКТУРА, СПОСОБИ АВТОМАТИЧНОЇ СИСТЕМАТИЗАЦІЇ ЛІНГВІСТИЧНИХ ЯВИЩ

У Корпусі української мови можна визначити три взаємопов'язані структурно-функціональні зони: 1) модуль-текст, у якому в електронній формі представлені українські тексти; 2) модуль-аналізатор, який забезпечує автоматичне та автоматизоване оброблення текстової інформації; 3) модуль-словник, в якому результати автоматичного та автоматизованого аналізу тексту систематизуються в електронних словниках, представлених в Інтернеті для користувача. Тобто, тільки 3-ій модуль, як результат роботи всіх систем, бачить користувач.

1.1. Модуль-текст

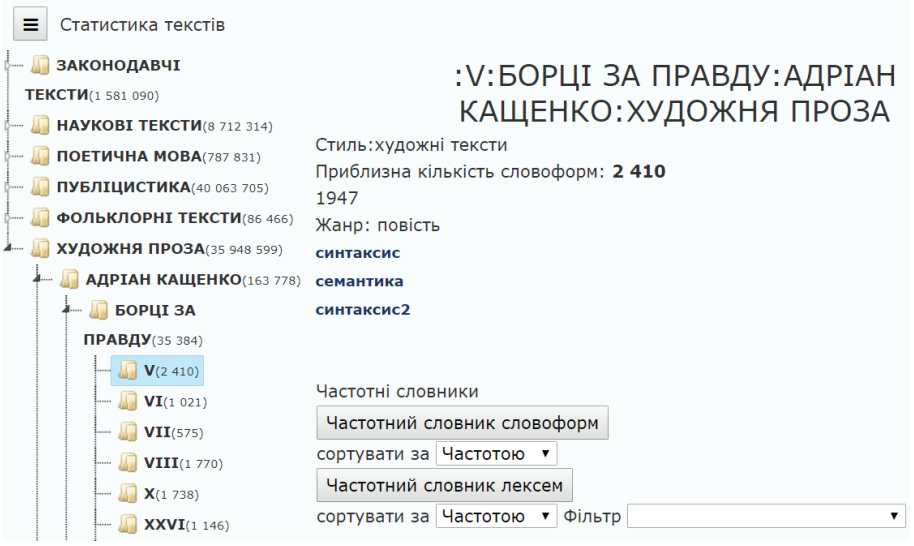

Рис.1. Фрагмент систематизації метатекстової інформації у КУМ

На сьогодні Корпус української мови представляє зібрання текстів обсягом ~ 87 млн. слововживань. Маркування метаінформації текстів корпусу здійснюється, насамперед, за стилем. За стильовими ознаками формується 6 підкорпусів (див. Рис. 1): законодавчі тексти – 1 581 090 слововживань; наукові тексти – 8 712 314 слововживань; поетична мова – 787 831 слововживання; публіцистика – 40 063 705 слововживань; фольклорні тексти – 86 466 слововживань; художня проза – 35 948 599 слововживань.

У межах кожного стильового підкорпусу формуються за ієрархічним принципом підкорпуси за різноманітними ознаками (галузь, тема, періодичне видання, автор, та ін.). Кінцевою ланкою ієрархії є заголовок конкретного тексту (наприклад, художня проза: Андріан Кащенко: Борці за правду: V частина). За умови активації конкретного твору чи частини твору, до нього додається інформація про видавництво, місце видання, жанр тексту, та деяка інша метаінформація. Як показує статистика, корпус текстів вимагає стилістичного збалансування, проте колектив не ставить завдання кількісно вирівняти стильові вибірки, а лише збільшити обсяг текстового модуля до 100 мільйонів слововживань, тому що основна увага на сьогодні зосереджена на поглибленні автоматичного аналізу тексту на базі текстів публіцистичного та художнього стилів.

1.2. Модуль-аналізатор

Модуль-аналізатор – інструмент лінгвістичних досліджень великих текстових масивів, що може виконувати такі функції: 1) забезпечення зв'язку модуля-тексту із лінгвістичними базами даних: морфологічною, морфемною, синтаксичною та семантичною; 2) проведення лематизації текстових слововживань; 3) проведення автоматичного морфологічного, морфемного, синтаксичного, семантичного і статистичного аналізів; 4) забезпечення роботи онлайн-платформ для проведення автоматизованого лінгвістичного аналізу; 4) конструювання словників-конкордансів контекстових слововживань та різних частотних словників.

Тексти Корпусу української мови параметризуються у модулі-аналізаторі за 4-ма рівнями анотації: 1) морфологічна анотація – базовий етап для всіх наступних рівнів: визначення морфологічних характеристик слів (частину мови і граматичні значення кожного слововживання тексту), а також леми слововживань (працює автоматично); 2) морфемна анотація: визначення морфемної будови слововживань тексту та леми лексичного реєстру (працює автоматично); 3) синтаксична анотація: визначення словосполучення, типу і виду синтаксичного зв'язку (працює автоматично); а також дерев структури речень (працює автоматично/автоматизовано); 4) семантична анотація: приписування кожному словов-

живанню/лемі коду семантичного поля таксономічної класифікації (працює автоматично/автоматизовано).

Автоматизація лінгвістичного аналізу на кожному рівні анотації відбувається у два етапи: 1) автоматичне оброблення машиною слововживань/лем; 2) автоматизоване редагування лінгвістом автоматично анотованого тексту. Автоматичне анотування текстів відбувається через зв'язок із великими лінгвістичними базами даних (БД), наприклад, морфологічна БД – 3,5 мл. словоформ, морфемна БД – 200 тис. початкових форм. Бази даних укладалися за розробленою методикою комп'ютерного моделювання одиниць різних мовних рівнів – комп'ютерною граматикою української мови: "Для автоматичного аналізу українського тексту нами створено комп'ютерну граматику, яка є ієрархічним комплексом комп'ютерних моделей: морфемно-словотвірної, морфологічної, синтаксичної моделі, побудованих на основі формальних, точних й однозначних правил. Ці моделі можна вважати дослідницькими, тому що закладені у граматики алгоритмічні правила призводять до виявлення того чи іншого мовного явища (морфів, словоформ з їх частиномовними і категорійними характеристиками, словосполучень, дерев залежностей речень тощо). Алгоритмічно зімітовано діяльність лінгвіста – а саме забезпечено перехід від сукупності текстів до системи, яка лежить в їх основі, встановлено елементарні одиниці і класи елементарних одиниць. Розроблені моделі є моделями аналізу, індуктивними, несемантичними і детерміністськими (структурними)"[4].

Дослідницький Корпус текстів – це лише один спосіб застосування комп'ютерної граматики. Вона може бути використана у різних автоматичних системах оброблення тексту ненаукового спрямування: чатових діалогових системах, системах реферування текстів, системах визначення тематики текстів, пошукових онтологіях, системах перевірки авторства текстів та в інших завданнях, які потребують роботи з текстовими масивами. У такому використанні комп'ютерна граматика є складовою систем штучного інтелекту.

Лінгвістична розмітка у корпусах може проводитися двома способами: 1) суцільна анотація всіх слововживань за введеними текстами на всіх рівнях розмітки і формування великої анотованої бази даних; 2) вибіркова анотація текстових слововживань та словника початкових форм (лем) за обмеженими текстовими вибірками і формування автономних баз даних. У КУМ суцільна анотація використана тільки для морфологічного аналізу[5]. Автоматичне приписування кожному слововживанню тексту грама-

---

[4] Н. Дарчук, *Комп'ютерне анотування тексту: результати і перспективи*, Київ 2013, с. 28.

[5] Докладніше про всі типи анотації можна дізнатися із монографії Н. Дарчук. *Комп'ютерне анотування тексту*….

тичного коду та автоматична лематизація відбувається при введенні тексту у корпус. На всіх інших рівнях анотації параметризація відбувається автоматично/автоматизовано за обмеженими текстовими вибірками. Укладачі корпусу свідомо відмовились від першого способу анотації, тому що розмітка мільйонного масиву тексту на всіх рівнях аналізу вимагає дуже потужного технічного забезпечення, інакше робота з корпусом стає надзвичайно повільною. Ольга Ляшевська, аналізуючи анотацію Національного корпусу російської мови, наводить фрагмент XML-представлення розмітки фрагмента тексту, в якому три слововживання (*Цены в них*) анотовані 79 рядками розмітки: лексико-граматичні теги (lex и gramm) і лексико-семантичні теги (sem), не враховуючи метарозмітки[6]. Цей приклад демонструє, який обсяг інформації систематизує сформована у такий спосіб база даних.

Анотація текстових слововживань у КУМ здійснюється на двох рівнях текстової розмітки: морфологічному та синтаксичному. На морфемному та семантичному рівні параметризується словник початкових форм, який формується як результат лематизації морфологічної анотації.

Визначення морфемної будови слів здійснюється автоматично за допомогою морфемно-словотвірної бази даних, у якій кожному слову приписана програмна процедура сегментації, наприклад, *заледеніти* PCRFSHSIFK, де кожен морф моделюється двома символами PC/RF/SH/SI/FK: перша латинська літера позначає тип морфа P – префікс, R – корінь, S – суфікс, F – флексія, I – інтерфікс, X – постфікс, а друга – межі морфа через порядковий номер (із початку слова) кінцевої графеми кожного морфа. Графемно-цифрові межі морфів подані у БД через латинську літеру за порядковим номером у спрощеній алфавітній системі: P2R5I7S8F10 = RC(2)RF(5)SH(8)FK(10). Зіставлення резидентного словника морфемної БД зі списком лем (початкових форм), укладеним за текстовою вибіркою, за допомогою спеціального програмного забезпечення здійснює автоматичну морфемну сегментацію кожної початкової форми[7].

На семантичному рівні анотація проводиться у два етапи: 1) автоматично за реєстром словника початкових форм, укладеного за обмеженою текстовою вибіркою: кожній лемі тексту приписується код семантичного класу за БД семантичних таксонів, укладеної за лексико-семантичними варіантами (ЛСВ) лексем публіцистичного стилю[8]. 2) автоматизовано за

---

[6] О. Ляшевская, *Корпусные инструменты в грамматических исследованиях русского языка*, Москва 2016, с. 15.

[7] Докладніше про автоматичний аналіз у КУМ див.: O. Zuban, *Automatic Morphemic Analysis in the Corpus of the Ukrainian Language: Results and Prospects*, "Jazykovedný časopis" 2017, Vol 68, № 2, с. 415.

[8] Докладніше про БД семаничних таксонів див.: Н. Дарчук, О. Зубань, М. Лангенбах, Я. Ходаківська, *АГАТ- семантика: семантична розмітка Корпусу української мови*, "Українське мовознавство" 2016, Вип. 1 (46), с. 3.

слововживаннями текстової вибірки: перед лінгвістом стоїть завдання до кожного слововживання, якому на 1-му етапі приписано код семантичних таксонів всіх ЛСВ слова, вибрати той таксон, до якого належить ЛСВ, актуалізований у реченні. Це завдання виконується на базі аналітичної платформи, див. Рис.2.

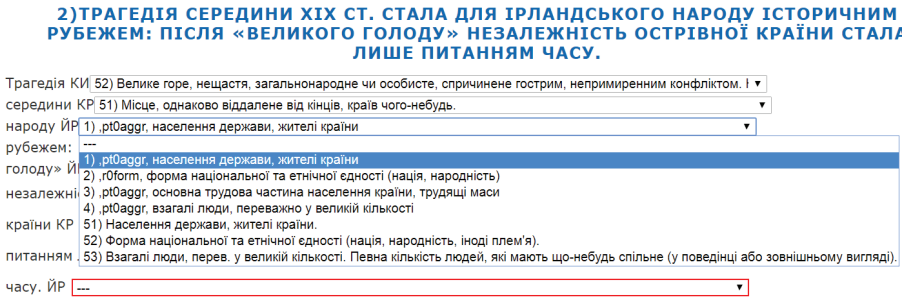

Рис.2. Фрагмент автоматизованого семантичного аналіз у

На синтаксичному рівні проводиться автоматичний синтаксичний аналіз словосполучень та речень. Словосполучення визначаються автоматично за граматикою валентності[9] 3-ох частин мови (іменника, дієслова та прикметника) і фразеологізмів, за правилами якої автоматично/автоматизовано укладені БД словосполучень (див. Табл.1).

Таблиця 1. Фрагмент БД словосполучень: валентність дієслова *щезати*

Код	лема	прийменник	Граматичні (двосимвольні) коди слововживань, якими керує дієслово
29139	щезати	з	ЙРКРЛРЙЕКЕЛЕИЕйРкРлРиЕМРМЗМЧМЕЧРЧЗЧЕ
29140	щезати	зі	ЙРКРЛРЙЕКЕЛЕИЕйРкРлРиЕМРМЗМЧМЕЧРЧЗЧЕ
29141	щезати	із	ЙРКРЛРЙЕКЕЛЕИЕйРкРлРиЕМРМЗМЧМЕЧРЧЗЧЕ
29142	щезати	в	ЙПКПЛПЙЯКЯЛЯИЯйПкПлПМПМНМЯЧПЧНЧЯ
29143	щезати	у	ЙПКПЛПЙЯКЯЛЯИЯйПкПлПМПМНМЯЧПЧНЧЯ

Синтаксична анотація речення також проводиться автоматично/автоматизовано. На вході синтаксичної анотації – речення, в якому кожному слововживанню на етапі морфологічної анотації приписано граматичний код, наприклад: *Трагедія(КИ) середини(КР) XIX(U) ст.(ББ) стала(ГЙ) для(ПР) ірландського(АР) народу(ЙР) історичним(АТ) рубежем:(ЙТ) після(ПР) «великого(АР) голоду»(ЙР) незалежність(КИ) острівної(АЗ) країни(КР)*

[9] Докладніше про принципи виокремлення словосполучень із тексту див.: Н. Дарчук, *Комп'ютерне анотування тексту*…, цит. праця, с. 119.

стала(ГЙ) лише(Ь0) питанням(ЛТ) часу.(ЙР). Синтаксичні зв'язки у реченнях визначаються автоматично за БД словосполучень, а потім в автоматизованому режимі лінгвіст перевіряє синтаксичні зв'язки (див. Рис. 3).

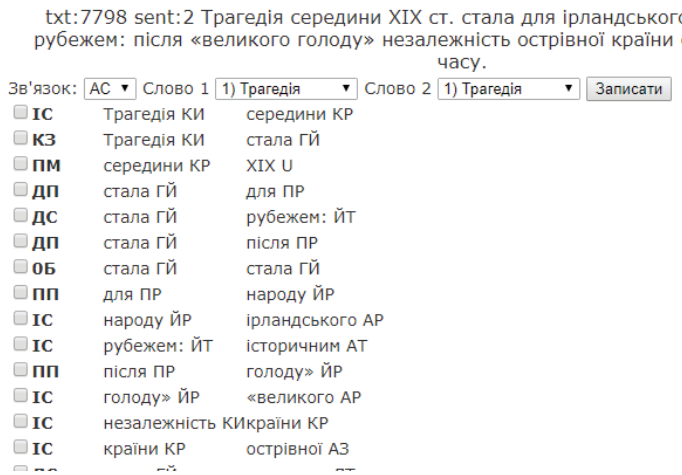

Рис. 3. Фрагмент робочої картки перевірки анотації синтаксичних зв'язків

За редагованою базою даних синтаксичних зв'язків машина будує дерево залежностей, яке також редагується лінгвістом автоматизовано (див. Рис. 4).

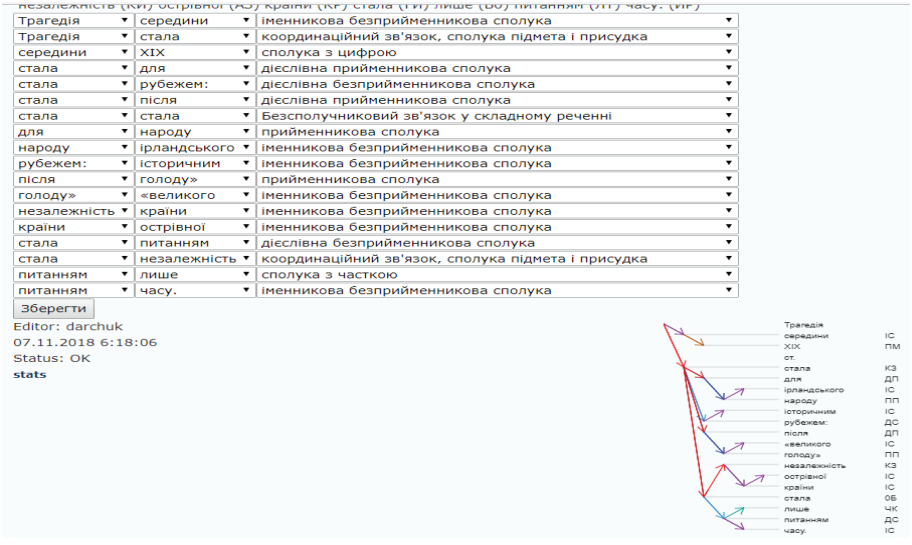

Рис. 4. Робоча картка автоматизованого редагування синтаксичного дерева

1.3. Модуль-словник

У поєднанні роботи двох модулів – модуля-текстів та модуля-аналізатора – за запитом користувача автоматично укладаються різні типи словників: 1) словники-конкорданси; 2) частотні словники (ЧС) слів (лексем), словоформ, морфем, морфемних структур слів, словосполучень, семантичних таксонів, n-грам. Інфологічна модель кожного типу словника та його структура визначалися специфікою електронного характеру та лінгвістичними особливостями представлених одиниць[10].

Словники-конкорданси автоматично укладаються за опцією "Пошук у корпусі" у межах підкорпусу текстів обраного стилю (див. Рис. 5). Укладання може здійснюватися за такими пошуковими параметрами: 1) пошук контекстів до одного слова за заданою конкретною лексемою (всі словоформи лексеми) або словоформою – перше у другому рядку діалогове вікно; 2) пошук контекстів до всіх словоформ стилю за обраною морфологічною характеристикою (друге діалогове вікно у другому рядку, опція "Морфологічні ознаки").

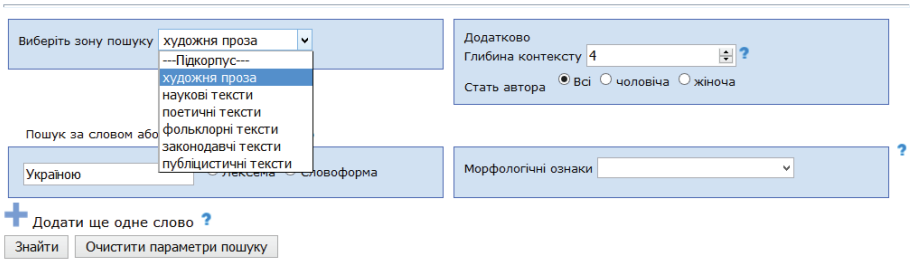

Рис. 5. Параметри автоматичного укладання словника-конкорданса словоформи *Україною*

Словник-конкорданс лексеми/словоформи можна автоматично укласти, записавши слово у перше знизу діалогове вікно: на рис. 5 задано пошук контекстів до словоформи *Україною* (орудний відмінок однини) у межах підкорпусу художньої прози).

При активації кнопки "Знайти" автоматично будується конкорданс до заданої словоформи (див. Рис. 6) з урахуванням двох додаткових параметрів, що вибираються у другому зверху діалоговому вікні: 1) глибини

[10] Докладніше про структуру електронних словників у КУМ див.: О. Зубань, *Електронні частотні морфемні словники в Корпусі української мови*, "Науковий вісник Східноєвропейського національного університету імені Лесі Українки". Серія: Філологічні науки 2015, № 3 (304), с. 315; О. Зубань, *Електронні словники у Корпусі української мови: параметри пошуку та систематизації мовних одиниць*, "Мовні і концептуальні картини світу" 2016, Вип. 54, с. 190.

контексту (кількості слововживань правобічного та лівобічного оточення словоформи у реченні); 2) статі автора. Як показує приклад словника-конкорданса словоформи *Україною*, контекст до аналізованої словоформи може бути розширений у двох напрямках дистрибуції слова до межі речення за допомогою активації позначок – «; – ».

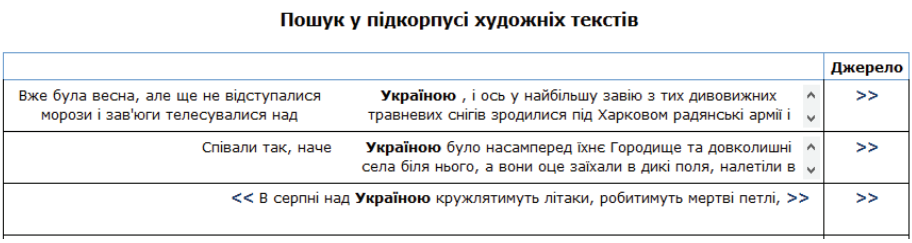

Рис.6. Фрагмент словника-конкорданса словоформи *Україною*

До кожного текстового слововживання подається індекс джерела (див. Рис. 7), з якого взято текстовий фрагмент, наприклад, перше речення – *Вже була весна…* – взято із роману *Диво (глава: «1966 рік Весна. Київ.»)* Павла Загребельного. Навігація до джерела здійснюється автоматично за допомогою позначки – » – у колонці "Джерело". Виведення результатів пошуку в побудові конкордансу можливе і в режимі цитування.

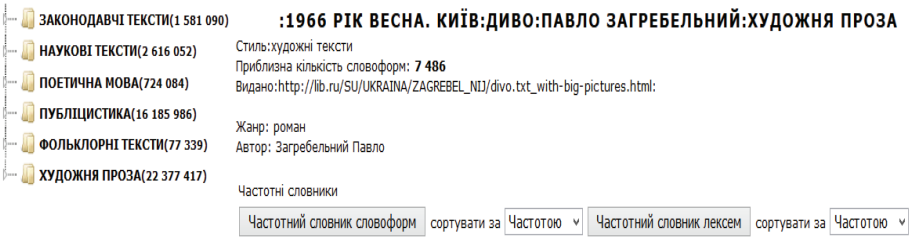

Рис. 7. Визначення джерела текстового фрагмента

Словник-конкорданс за параметром пошуку контекстів до всіх словоформ стилю за обраною морфологічною характеристикою (друге діалогове вікно у другому рядку – див. Рис. 5) будується за вибором морфологічних ознак кожної частини мови у випадному списку опції "Морфологічні ознаки". Наприклад, за параметрами пошуку (частина мови – іменник, рід – жіночий, число – однина, відмінок – орудний) можна автоматично укласти конкорданс до всіх іменників жіночого роду, орудного відмінка однини, які вживаються у текстах художньої прози (див. Рис. 8).

Пошук у підкорпусі художніх текстів

	Джерело
Дівчата хоча притомлені, та водночас і вдоволені щойно пережитою **напругою** , ніби й справді їм вдалося когось порятувати.	>>
<< доводиться в духотняві, яма налита **спекою** . >>	>>
<< З **місією** Червоного Хреста в далекій південній країні була >>	>>
<< , головою поводить, стежить за **танцівницею** , що зовсім близько перед нею теж >>	>>
<< Липка, задушна ніч, повалені **холерою** люди стогнуть за брезентом твого намету, >>	>>
<< Костянтинівна, що, посріблена тепер **сивиною** , з поглядом пригаслим, сидить серед >>	>>

Рис. 8. Фрагмент словника-конкорданса за вибором морфологічних ознак

Рубрика "Статистика текстів" відкриває діалогове вікно зі стилістичною параметризацією корпусу за підкорпусами стилів (див. Рис. 1). Розгортаючи дерево кожного підкорпусу до кінцевої ланки – конкретного тексту, користувач на базі цього тексту може автоматично укласти ЧС лексем та словоформ із визначенням абсолютної частоти вживання. На сьогодні у Корпусі української мови в режимі on-line автоматично укладаються ≈ 40 тис. таких частотних словників.

:1966 РІК ВЕСНА. КИЇВ:ДИВО:ПАВЛО ЗАГРЕБЕЛЬНИЙ:ХУДОЖНЯ ПРОЗА

Стиль:художні тексти
Приблизна кількість словоформ: **7 486**
Видано:http://lib.ru/SU/UKRAINA/ZAGREBEL_NIJ/divo.txt_with-big-pictures.html:

Жанр: роман
Автор: Загребельний Павло

Частотні словники

Частотний словник словоформ сортувати за Алфавітом Всього словоформ:3234

Словоформа	Абс.частота
Але	88
альпіністами	1
алюмінієвий	1
аніж	1
анатомія	1
Андре	1
ансамбль	1

Частотний словник лексем сортувати за Частотою Всього лексем:2410

Лексема	Абс.частота
навіть	23
та	23
б	23
час	23
ми	23
колоти	22
бузина	22

Рис. 9. Фрагмент частотного словника лексем та словоформ глави «1966 рік Весна. Київ» роману «Диво» П.Загребельного

На рис.9. показано фрагмент двох частотних словників: ЧС лексем та ЧС словоформ. Словники укладаються за вибором параметра формування реєстру одиниць: алфавітом або рейтингом абсолютних частот (за спадом або ростом частот при активації опції "Абс. Частота".

За цією текстовою вибіркою також можливе автоматичне укладання семантичного словника, у якому до кожної реєстрової одиниці подано семантичні таксони ЛСВ лексеми (див. Рис. 10).

Лексема	Клас	Абс.частота	Сема
бути	Г	84	Дієслово буттєва сфера існування//Дієслово посесивна сфера//Дієслово буттєва сфера
могти	Г	49	Дієслово буттєва сфера
сказати	Г	29	Дієслово мовлення
колоти	Г	22	Дієслово фізичний вплив//Дієслово мовлення//Дієслово ментальна сфера
піти	Г	20	Дієслово переміщення об'єкта//Дієслово якісний стан
мати	Г	16	Дієслово посесивна сфера//Дієслово якісний стан//Дієслово міжособистісні відношення
знавати	Г	13	Дієслово ментальна сфера
знати	Г	12	Дієслово ментальна сфера//Дієслово буттєва сфера
здавати	Г	12	Дієслово соціальна діяльність//Дієслово посесивна сфера//Дієслово буттєва сфера//Дієслово ментальна сфера//Дієслово якісний стан

Рис. 10. Фрагмент семантичного словника дієслівних лексем

Рубрика “N-грами” відкриває діалогове вікно для укладання частотних словників n-грам (2-грам, 3-грам, 4-грам, 5-грам) за вибором текстової вибірки трьох стилів: наукового, публіцистичного та наукового. На рис. 11 подано фрагмент частотного словника 3-грам, укладеного за вибіркою текстів наукового стилю.

перше слово	друге слово	третє слово	частота
на	відміну	від	1242
з	одного	боку	964
майбутніх	фахівців	з	946
У	зв'язку	з	937

Рис. 11. Фрагмент частотного словника n-грам

Частотні словники за стилями, розділами, авторами, збірками і т. ін. (рубрика “Частотні словники”) із метою оптимізації пошуку на базі великого обсягу текстової інформації представлено у КУМ, як автономні електронні лексикографічні системи. На сьогодні на замовлення користувачів укладено 20 таких словників.

Наприклад, в *Електронному словнику мови Тараса Шевченка* [CZS 2019] користувач може автоматично укласти такі алфавітно-частотні словники: 1) Чс словоформ за заданою буквою або словом; 2) ЧС всіх словоформ всіх частин мови або за вибраною морфологічною характеристикою; 3) ЧС лексем всіх частин мови або за вибраною морфологічною характеристикою; 4) ЧС словосполучень за 9 параметрами; 5) ЧС морфем (префіксів, коренів, суфіксів, інтерфіксів) всіх слів або за вибраною морфологічною характеристикою слів; 6) ЧС морфемних структур слів (початкових форм) усіх частин мови або за вибраними морфологічними ознаками. Інформація в алфавітно-частотних словниках структур-

но розподіляється на 3 зони: 1. Інвентар одиниць за вибраним типом; 2. Реалізація одиниць (морфем) у словах, (слів) у реченнях з інформацією про частотні характеристики; 3. Контексти (речення) вживання аналізованої одиниці. Для прикладу, продемонструємо *Частотний словник морфемних структур слів*, який систематизує статистичні дані про реалізацію моделей морфемної структури слів у текстах Т. Шевченка. Фрагмент словника, поданий на рис. 12, демонструє реалізацію моделі морфемної структури слова PRSF: ця модель вживається у текстовій вибірці 3804 рази (абсолютна частота) і реалізована у 95 лексемах, список яких подається у другій таблиці.

Всього записів: 39

Структура	Абсолютна частота	Середня частота
RF	16323	267,59
RSF	7482	122,66
R	6605	108,28
PRSF	3804	62,36
PRF	1631	26,74
RSSF	898	14,72
RS	604	9,90
PRS	499	8,18
RSS	359	5,89

Частотний словник по морфструктурі:**PRSF**, частина мови:Іменник Всього записів: 95 Всього записів: 95 Всього записів: 95

Слово	Частина мови	Абсолютна частота	Джерело	Середня частота	Середньоквадратичне відхилення	Коефіцієнт стабільності
пожар	ім. ч. р.	20	11	0,33	0,74	2,26
невольник	ім. ч. р.	15	11	0,25	0,64	2,62
пророк	ім. ч. р.	12	7	0,20	0,70	3,54
порада	ім. ж. р.	11	10	0,18	0,42	2,36

Рис. 12. Фрагмент Частотного словника морфемних структур

У третій зоні *Частотного словника морфемних структур* подаються конкорданс до вибраного слова та джерело, з якого взято речення конкорданса (див. Рис. 13).

ЧАСТОТНИЙ СЛОВНИК ЗБІРКИ "ТВОРИ В П'ЯТИ ТОМАХ". ТАРАС ШЕВЧЕНКО

Конкорданс до слова: **пророк** (ім. ч. р.)
Морфемна структура: **PRSF /про/рок/**

Контекст	Джерело
ПРОРОК	>>
Неначе праведних дітей , Господь , любя отих людей , Послав на землю їм пророка — Свою любов благовістить !	>>
Полюбили Того пророка , скрізь ходили За ним і сльози , знай , лили Навчені люди .	>>
кроткого пророка ... Царя вам повелів надать !	>>
Нівроку , До Божого царя-пророка Сама Версавія прийшла , І повечеряла , й сикеру З пророком випила , й пішла Спочити трохи по вечері З своїм царем .	>>
Давид , святий пророк і цар , Не дуже був благочестивий .	>>
Пророка , Свого неситого царя , Кленуть Давида-сподаря .	>>
Пророче Божий ?	>>
А він Нехай лютує на землі , Нехай пророка побиває , Нехай усіх нас розпинає ; Уже внучата зачались , І виростуть вони колись ...	>>
Восплач , Пророче , Сине Божий !	>>
Господь послав Тебе нам , кроткого пророка І обличителя жестоких Людей неситих .	>>
Пророче наш !	>>

Рис. 13. Конкорданс до слова *пророк* за текстами Т. Шевченка

Два типи словників у Корпусі української мови – конкорданси та частотні – поєднані між собою взаємозворотньою й інформаційно доповнювальною навігацією: 1) конкорданс → частотний словник: словники-конкорданси через опцію "Джерело" поєднуються із алфавітно-частотними словниками тексту, з якого взято речення; 2) частотний словник → конкорданс: якщо користувач працює із частотними словни-

ками стилів, авторів, збірок, то через опцію "Контекст" або активацію конкретного слова він може перейти до конкордансу обраного слова або словосполучення.

ВИСНОВКИ

Запропонована методика створення корпусу українських текстів є узагальненням комплексу теоретичних і прикладних ідей сучасного мовознавства. Багатоаспектна систематизація лінгвістичної інформації у Корпусі української мови, встановлення статистичних закономірностей функціонування різнорівневих мовних одиниць у різних типах текстів формують лексикографічну систему нового покоління, яка розглядається як універсальна довідкова система з української мови для учителя, журналіста або пересічного користувача, а для філолога-дослідника, викладача – як лінгвістична база знань.

Технологія конструювання корпусу робить її надзвичайно ефективним та раціональним інструментом для спеціалістів-філологів різного профілю, тому що передбачає роботу в режимі on-line. Статистична та лінгвістична інформація про організацію українських текстів на різних рівнях мовної системи, представлена в електронних словниках КУМ дає можливість вивчати закономірності функціонування мовних одиниць у різних стилях, комплексно досліджувати мовні особливості ідіостилів[11]. Електронні частотні словники такого типу можуть бути укладені для всіх авторів і тестів Корпусу української мови за запитом користувача.

Аналітичні платформи автоматизованого аналізу у КУМ, які працюють в on-line, сьогодні ефективно використовуються викладачами КНУ ім. Т. Шевченка для читання курсів та проведення лабораторних робіт з автоматичного лінгвістичного аналізу тексту для студентів спеціальності 035.10 – філологія: прикладна лінгвістика. Також Корпус української мови – навчальна база для проходження практик студентами: не тільки наші студенти, а й студенти інших закладів вищої освіти України ефективно використовують експертну лінгвістичну систему "Корпус української мови" для здобуття практичних фахових навичок як прикладного лінгвіста, так й інформатика, зокрема у цьому навчальному році на базі КУМ практику проходили студенти Черкаського національ-

[11] О. Зубань, *Стилеметричні ознаки морфемних структур слів у поетичному мовленні Т. Шевченка (на матеріалі Корпусу української мови)*, "Мовні і концептуальні картини світу", Вип. 48, с. 165-179; О. Зубань, *Частотні морфемні словники в Корпусі української мови – джерело стилеметричних досліджень*, "Acta Universitatis Palackianae Olomucensis Philologica" 2016, т. UCRAINICA VII: Současná ukrajinistika Problémy jazyka, literatury a kultury, с. 224-231.

ного університету імені Богдана Хмельницького та факультету інформаційних технологій КНУ ім. Т. Шевченка.

ЛІТЕРАТУРА

BrUK: Brauns′kij korpus ukraïns′koï movi [Браунський корпус української мови. https://r2u.org.ua/corpus. [Доступ 07.03.2019].

CZS: Častotnij slovnik movi T. Ševčenka [ЧС: Частотний словник мови Т. Шевченка. http://www.mova.info/cfqsh_2.aspx. [Доступ 07.03.2019].

GRAK: General′nij regìonal′no anotovanij korpus ukraïns′koï movi. [ГРАК: Генеральний регіонально анотований корпус української мови. http://uacorpus.org/. [Доступ 07.03.2019].

KTUM: Korpusi tekstìv ukraïns′koï movi. [КТУМ: Корпуси текстів української мови. http://corpora.donnu.edu.ua/. [Доступ 07.03.2019].

KUM: Korpus ukraïns′koï movi [КУМ: Корпус української мови. http://www.mova.info/corpus.aspx. [Доступ 07.03.2019].

Darčuk Natalìâ, Zuban′ Oksana ta ìn. 2016. *AGAT-semantika: semantična rozmìtka Korpusu ukraïns'koï movi*. "Ukraïns′ke movoznavstvo" № 1 (46): 3-10 [Дарчук Наталія, Зубань Оксана та ін. 2016. *АГАТ-семантика: семантична розмітка Корпусу української мови.* "Українське мовознавство" № 1 (46): 3-10].

Darčuk Natalìâ. 2013. *Komp`ûterne anotuvannâ tekstu: rezul'tati ì perspektivi*. Kiïv: Osvìta Ukraïni [Дарчук Наталія. 2013. *Комп'ютерне анотування тексту: результати і перспективи*. Київ: Освіта України].

Karpìlovs′ka Êvgenìâ. 2007, 2008. *Tendencìï rozvitku sučasnogo ukraïns'kogo leksikonu: činniki stabìlìzacìï ìnnovacìj*. "Ukraïns′ka mova" № 4: 3-15; № 1: 24-35 [Карпіловська Євгенія. 2007, 2008. *Тенденції розвитку сучасного українського лексикону: чинники стабілізації інновацій*. "Українська мова" № 4: 3-15; № 1: 24-35].

Kotsyba Natalia. 2013. *Praktyczny przewodnik po korpusach języków słowiańskich*. W: http://www.domeczek.pl/~natko/papers/przewodnik-korp-ukr2013.pdf. [Dostęp 07.03.2019].

Lâševskaâ Ol′ga. 2016. *Korpusnye instrumenty v grammatičeskih issledovaniâh russkogo âzyka*. Moskva: Izdatel′skij Dom ÂSK [Ляшевская Ольга. 2016. *Корпусные инструменты в грамматических исследованиях русского языка.* Москва: Издательский Дом ЯСК].

Širokov Volodimir ta ìn. 2011. *Zastosuvannâ Ukraïns'kogo nacìonal'nogo lìngvìstičnogo korpusu v leksikografìï ta lìngvìstičnih ekspertizah*. V: *Ukraïns'ka leksikografìâ v zagal'noslov'ân's'komu kontekstì: teorìâ, praktika, tipologìâ*. Kiïv: Vidavničij dìm dmitra Burago: 285-294 [Широков Володимир та ін. 2011. *Застосування Українського національного лінгвістичного корпусу в лексикографії та лінгвістичних експертизах.* В: *Українська лексикографія в загальнослов'янському контексті: теорія, практика, типологія*. Київ: Видавничий дім дмитра Бураго: 285-294].

Zuban′ Oksana. 2014. *Stilemetričnì oznaki morfemnih struktur slìv u poetičnomu movlennì T. Ševčenka (na materìalì Korpusu ukraïns'koï movi)*. "Movnì ì konceptual′nì kartini svìtu" № 48: 165-179 [Зубань Оксана. 2014. *Стилеметричні ознаки морфемних структур слів у поетичному мовленні Т. Шевченка (на матеріалі Корпусу української мови)*. "Мовні і концептуальні картини світу" № 48: 165-179].

Zuban′ Oksana. 2015. *Elektronnì častotnì morfemnì slovniki v Korpusì ukraïns'koï movi*. "Naukovij vìsnik Shìdnoêvropejs′kogo nacìonal′nogo unìversitetu ìmenì Lesì Ukraïnki". Serìâ: Fìlologìčnì nauki № 3 (304): 315-320 [Зубань Оксана. 2015. *Елек-*

тронні частотні морфемні словники в Корпусі української мови. "Науковий вісник Східноєвропейського національного університету імені Лесі Українки". Серія: Філологічні науки № 3 (304): 315-320].

Zuban′ Oksana. 2016. *Častotnì morfemnì slovniki v Korpusì ukraïns'koï movi – džerelo stilemetričnih doslìdžen'.* "Acta Universitatis Palackianae Olomucensis Philologica" № UCRAINICA VII: Současná ukrajinistika Problémy jazyka, literatury a kultury: 224-231 [Зубань Оксана. 2016. *Частотні морфемні словники в Корпусі української мови – джерело стилеметричних досліджень*. "Acta Universitatis Palackianae Olomucensis Philologica" № UCRAINICA VII: Současná ukrajinistika Problémy jazyka, literatury a kultury: 224-231].

Zuban′ Oksana. 2016. *Elektronnì slovniki u Korpusì ukraïns'koï movi: parametri pošuku ta sistematizacìï movnih odinic'.* "Movnì ìkonceptual′nì kartini svìtu". Vip. 54: 190-201 [Зубань Оксана. 2016. *Електронні словники у Корпусі української мови: параметри пошуку та систематизації мовних одиниць*. "Мовні і концептуальні картини світу". Вип. 54: 190-201].

Zuban Oksana. 2017. *Automatic Morphemic Analysis in the Corpus of the Ukrainian Language: Results and Prospects*. "Jazykovedný časopis" vol. 68, № 2: 415-426.

UKRAINIAN LANGUAGE CORPUS – COMPUTER EXPERT SYSTEM OF LINGUISTIC ANALYSIS OF UKRAINIAN TEXT

Summary. The article deals with the structure and the principles of automatic compiling of expert linguistic analysis system called Ukrainian Language Corpus. The methodology of formalised description of language text units, suggested in creating the Corpus, carries out automatic morphological, morphemic, syntactic, and semantic analyses of Ukrainian texts as well as automatically compiling different Frequency Dictionaries according to text samples.

Key words: Ukrainian Language Corpus, Electronic Frequency Dictionary, Data Base, automatic linguistic analysis

KORPUS JĘZYKA UKRAIŃSKIEGO – KOMPUTEROWY EKSPERCKI SYSTEM ANALIZY JĘZYKOWEJ TEKSTU UKRAIŃSKOJĘZYCZNEGO

Streszczenie. Celem niniejszego artykułu jest przedstawienie struktur i zasad automatycznego tworzenia eksperckiego systemu analizy lingwistycznej „Korpus Języka Ukraińskiego". Zaproponowana podczas tworzenia Korpusu metodologia sformalizowanego opisu językowych jednostek tekstu zapewnia możliwość przeprowadzenia automatycznej morfologicznej, morfemowej, syntaktycznej i semantycznej analizy tekstu ukraińskojęzycznego, jak również automatyczne tworzenie różnorakich elektronicznych słowników frekwencyjnych z wyborem tekstów.

Słowa kluczowe: Korpus języka ukraińskiego, Elektroniczny słownik frekwencyjny, baza danych, automatyczna analiza lingwistyczna.