МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ДОНЕЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ВАСИЛЯ СТУСА

МИХАЙЛЮК АНЖЕЛІКА РОСТИСЛАВІВНА

Допускається до захисту:
в.о. завідувача кафедри
загального та прикладного
мовознавства і слов'янської
філології,
д.філол.н., доцент
_____ Ситар Г.В.
«____» _____ 2020 р.

АЛГОРИТМ УКЛАДАННЯ КОРПУСУ ТЕКСТІВ НА МАТЕРІАЛІ ТВОРІВ ТАРАСА ПРОХАСЬКА

Спеціальність 035 Філологія

Магістерська робота
Освітньо-професійна програма «Прикладна лінгвістика»

Науковий керівник:
І.Г. Данилюк, доцент
кафедри загального та прикладного
мовознавства і слов'янської філології,
к.філол.н., доцент

Оцінка: ____ / ___ / _____
Голова ЕК: _____

Вінниця 2020

АНОТАЦІЯ

Михайлюк А.Р. Алгоритм укладання корпусу текстів на матеріалі творів Тараса Прохаська. Спеціальність 035 «Філологія», Освітньо-професійна програма «Прикладна лінгвістика». Донецький національний університет імені Василя Стуса, Вінниця, 2020. 65 с.

У кваліфікаційній роботі визначено дистинктивні та типологійні ознаки корпусу текстів; здійснено аналітичний огляд лінгвістичних корпусів текстів; окреслено проблеми корпусної лінгвістики; схарактеризовано поняття «розмітки» в корпусній лінгвістиці; проаналізовано особливості стандартизації оформлення корпусів текстів; здійснено метарозмітку та лінгвістичну розмітку корпусу текстів Тараса Прохаська. Корпус текстів Тараса Прохаська позиціоновано як індивідуально-авторський експериментальний науково-дослідний.

Ключові слова: корпус текстів, корпусна лінгвістика, індивідуально-авторський експериментальний корпус текстів, національний корпус текстів, розмітка, стандартизація.

Табл. 0. Діагр. 0. Рис. 0. Бібліограф.: 75.

Mykhailiuk A.R. Algorithm for Creating a Text Corpus Based on the Works of Taras Prokhasko. Specialty 035 »Philology», Programme «Applied Linguistics». Vasyl' Stus Donetsk National University, Vinnytsia, 2020. 65 p.

In the qualification work the distinctive and typological features of the text corpus are determined; an analytical review of the linguistic corpora of texts was carried out; the problems of corpus linguistics are outlined; the concept of "markup" in corpus linguistics is characterized; the features of standardization of text corpus design are analyzed; meta-markup and linguistic markup of the text corpus by Taras Prokhasko were carried out. The body of texts by Taras Prokhasko is positioned as an individual author's experimental research.

Key words: text corpus, corpus linguistics, individual author's experimental text corpus, national text corpus, markup, standardization.

Table 0. Diag. 0 Fig. 0 Bibliography.: 75.

ЗМІСТ

стор.

ВСТУП 5

РОЗДІЛ 1 КОРПУС ТЕКСТІВ ЯК ОБ'ЄКТ КОРПУСНОЇ ЛІНГВІСТИКИ 8

1.1 Дистинктивні та типологійні ознаки корпусу текстів 8

1.2 Огляд лінгвістичних корпусів текстів 13

1.2.1 Британський національний корпус (англ. British National Corpus, BNC) 14

1.2.2 Французька база Франтекст (фр. Frantext) 15

1.2.3 Український національний лінгвістичний корпус (УНЛК) 16

1.2.4 Національний корпус російської мови (рос. Национальный корпус русского языка, НКРЯ) 18

1.2.5 Інші національні корпуси текстів 20

1.3 Проблеми корпусної лінгвістики 20

Висновок до розділу 1 22

РОЗДІЛ 2 ЕТАПИ СТВОРЕННЯ Й ІНСТРУМЕНТИ РОБОТИ З КОРПУСОМ ТЕКСТІВ 23

2.1 Алгоритм створення корпусу текстів 23

2.2 Поняття «розмітки» в корпусній лінгвістиці 26

2.3 Особливості стандартизації оформлення корпусів текстів: стандарт TEI (англ. Text Encoding Initiative) 29

Висновок до розділу 2 35

РОЗДІЛ 3 ІНДИВІДУАЛЬНО-АВТОРСЬКИЙ ЕКСПЕРИМЕНТАЛЬНИЙ НАУКОВО-ДОСЛІДНИЙ КОРПУС ТАРАСА ПРОХАСЬКА 36

3.1 Тарас Прохасько як лінгвоперсона 36

3.2 Наповнення індивідуально-авторського експериментального науково-дослідного корпусу Тараса Прохаська 38
Висновок до розділу 3 42
ВИСНОВКИ 43
СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ 45
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ 50
СПИСОК ДЖЕРЕЛ ІЛЮСТРАТИВНОГО МАТЕРІАЛУ 51
ДОДАТКИ 52
Додаток А Апробація роботи. Копія сертифікату учасника Міжнародної наукової конференції «Tendances scientifiques de la recherché fondamentale et appliquee» (30 octobre 2020, Strasbourg, Republique Francaise) 52
Додаток Б Апробація роботи. Копія наукової статті «Алгоритм укладання корпусу текстів (на матеріалі творів Тараса Прохаська)» 54
Додаток В Електронний носій корпусу текстів Тараса Прохаська 65

ВСТУП

Створення корпусів текстів розглядається вченими як найважливіше гуманітарне завдання лінгвістики XXI ст. Корпус – це нова своєрідна форма життя мови. На відміну від паперових картотек, які після завершення досліджень зберігаються в архіві, електронний корпус продовжує жити, збагачуватися й активно служити подальшим поколінням філологів.

Нині лінгвістичні корпуси та лексичні бази – найбільш популярні ресурси як для лінгвістів-дослідників, так і для фахівців у сфері інформаційних технологій. Базуючись головним чином на емпіричному підході до аналізу мовного матеріалу, корпусні дослідження дозволяють абстрагуватися від суб'єктивності дослідника й наблизитися до об'єктивного вивчення мови. Опрацювання лінгвістичного матеріалу на основі корпусів дає найбільш достовірні результати й неупереджені висновки щодо стану і проблем досліджуваних мов як на окремому етапі їхнього розвитку, так і у зіставленні. *Актуальність* магістерської роботи мотивована необхідністю створення корпусу текстів на матеріалі творів Тараса Прохаська.

Об'єктом магістерської роботи є корпус текстів, *предметом* – інструменти роботи з корпусом текстів.

Мета дослідження: укласти корпус текстів на матеріалі творів Тараса Прохаська.

Реалізація мети передбачає розв'язання таких *завдань*:

1) визначити дистинктивні та типологійні ознаки корпусу текстів;

2) здійснити огляд лінгвістичних корпусів текстів (Британський національний корпус (англ. British National Corpus, BNC); французька база Франтекст (фр. Frantext); Український національний лінгвістичний корпус (УНЛК); Національний корпус російської мови (рос. Национальный корпус русского языка, НКРЯ) тощо);

3) окреслити проблеми корпусної лінгвістики;

4) узагальнити етапи створення корпусу текстів;

5) схарактеризувати поняття «розмітки» в корпусній лінгвістиці;

6) проаналізувати особливості стандартизації оформлення корпусів текстів (стандарт TEI (англ. Text Encoding Initiative));

7) здійснити метарозмітку та лінгвістичну розмітку індивідуально-авторського експериментального науково-дослідного корпусу текстів Тараса Прохаська.

Методи дослідження. Мета й завдання магістерської роботи зумовили необхідність застосування низки загальнонаукових методів: логіко-поняттєвий метод, метод узагальнення та систематизації, описовий метод. Як спеціальну застосовано *корпусну методологію* з опертям на поняття «розмітка».

Матеріал дослідження становлять твори Тараса Прохаська [73-75]. Обсяг корпусу – понад 10 000 слововживань.

Новизна дослідження визначувана тим, що вперше опрацьовано алгоритм укладання корпусу текстів на матеріалі творів Тараса Прохаська. Визначено дистинктивні та типологійні ознаки корпусу текстів; здійснено аналітичний огляд лінгвістичних корпусів текстів; окреслено проблеми корпусної лінгвістики; схарактеризовано поняття «розмітки» в корпусній лінгвістиці; проаналізовано особливості стандартизації оформлення корпусів текстів (стандарт TEI (англ. Text Encoding Initiative)).

Практичне значення роботи. Результати дослідження можуть знайти практичне використання у вишівських курсах із корпусної лінгвістики. Матеріали магістерської роботи можуть бути використані під час а) лінгвістичних досліджень із лексичної семантики, граматичної та лексичної сполучуваності, стилістики, прагматики, діалектології тощо; б) створення конкордансів, словників слів та словосполучень на основі одномовних корпусів.

Апробація. Результати дослідження апробовано на Міжнародній науковій конференції «Tendances scientifiques de la recherché fondamentale et appliquee»

(30 octobre 2020, Strasbourg, Republique Francaise) (див. Додаток А). За результатами роботи опубліковано 1 наукову статтю [32] (див. також Додаток Б).

Структура роботи. Робота складається зі вступу, трьох розділів із висновками до кожного, висновків, списку використаної літератури (61 позиція), списку використаних джерел (11 позицій), списку джерел ілюстративного матеріалу (3 позиції), трьох додатків (Додаток А Апробація роботи. Копія сертифікату учасника Міжнародної наукової конференції «Tendances scientifiques de la recherché fondamentale et appliquee» (30 octobre 2020, Strasbourg, Republique Francaise); Додаток Б Апробація роботи. Копія наукової статті «Алгоритм укладання корпусу текстів (на матеріалі творів Тараса Прохаська)»; Додаток В Електронний носій корпусу текстів Тараса Прохаська). Загальний обсяг роботи – 60 сторінок, із них наукове дослідження – 45 сторінок.

РОЗДІЛ 1
КОРПУС ТЕКСТІВ ЯК ОБ'ЄКТ КОРПУСНОЇ ЛІНГВІСТИКИ

1.1 Дистинктивні та типологійні ознаки корпусу текстів

Створення і розвиток найрізноманітніших корпусів текстів різних мов постає сьогодні одним із пріоритетних завдань корпусної лінгвістики. Корпуси надають майбутнім поколінням дослідників надійне і доступне джерело даних про функціювання мови в різних сферах і про культуру народу, що говорить цією мовою.

Основою корпусної лінгвістики є розроблення теоретичних засад і практичних прийомів побудови, машинного опрацювання, експлуатації та аналізу мовних даних, оформлених як корпус текстів. Численні розвідки в галузі сучасного корпусного мовознавства [2-3; 5; 7-11; 14-17; 20; 23-25; 29-31; 36; 45-47] відбуваються у двох векторах: 1) перший зосереджений на розробці проблем, що стосуються теорії та практики створення корпусів, іншими словами, концептуалізації корпусу (типологія корпусу, його призначення, обсяг, параметризація предметної галузі, репрезентативність, структурування та принципи відбору базових одиниць, зберігання і т. ін.); 2) другий спрямований на дослідження саме лінгвістичних корпусів, тобто вивчення мови за допомогою корпусних методів. В. Жуковська зауважує, що чіткої межі між зазначеними векторами не існує, «адже практично всі укладачі корпусів в той же час здійснюють і лінгвістичні дослідження на їх основі» [13, с. 10]. Така двовекторність корпусної лінгвістики зумовлюється подвійною природою об'єкта її дослідження – текстового корпусу, який, з одного боку, виступає як вихідний мовленнєвий матеріал для корпусної лінгвістики, а з іншого, є результатом діяльності цього мовознавчого напряму. Предметом

корпусної лінгвістики виступають теоретичні основи і практичні механізми створення та експлуатації мовних корпусів.

За Т. Бобковою, у широкому значенні під корпусом розуміють будь-яке зібрання письмових або усних текстів, використовуване з метою дослідження мови; у вузькому значенні під корпусом розуміють зібрання текстів в електронній формі, що презентує певну мову [2, с. 11].

В англомовній літературі є термін англ. «reference corpus» (‘зразковий корпус’). Англійський учений Дж. Сінклер, автор програмної статті про типологію корпусів, уважає, що зразковий корпус створюється для того, щоб надати вичерпну інформацію про мову [57]. Він повинен бути досить великим, для того щоб репрезентувати всі істотні різновиди цієї мови і її шари лексики, а також служити базою для граматик, словників та іншої надійної довідкової літератури.

О. Демська-Кульчицька наголошує, що корпус формується з реальних уривків писемного або усного мовлення, не передбачаючи модифікації мовленнєвої дійсності, що перетворює його на категорію емпіричну і дозволяє розглядати фактичний корпусний матеріал як емпіричну базу лінгвістичного дослідження [10, с. 41].

У контексті власне корпусної лінгвістики корпус розуміється як велика колекція зразків письмових та усних текстів, доступних у машиночитаній формі, зібраних науково обґрунтованим способом для презентації певного різноманіття або вживання мови [48, с. 3].

Г. Лук’янець акцентує увагу на тому, що лінгвістичні корпуси характеризуються синхронністю та паралельною багатоплановістю, а інколи і мультимодальністю [28, с. 130]. Синхронність найкраще описується я к особливий аспект репрезентативності одиниць всіх рівнів мови, який є ключовим аспектом корпусної лінгвістики. У такий спосіб, лінгвістичний корпус набуває ще ознаки збалансованості, або точніше, є збалансованим у відношенні до оцінок властивостей мовних одиниць.

Н.П. Дарчук описує корпус як «зібрання текстів певною мовою, яке представлене в електронній формі і супроводжується науковим апаратом. Апарат, вбудований у корпус, називається розміткою, або анотацією. Корпус тим кращий, чим повніша і досконаліша його анотація. Наука про корпуси – це, перш за все, наука про те, як зробити хорошу розмітку корпусу» [9, с. 46].

В. Жуковська пропонує таке визначення корпусу текстів – «це машиночитане, збалансоване, репрезентативне зібрання особливо розмічених (анотованих) текстів, відібраних згідно фіксованих параметрів для досягнення визначеної лінгвістичної мети та досліджуваних нелінійно за принципом гіпертексту» [13, с. 58] та називає 5 його дистинктивних ознак: репрезентативність, автентичність, відібраність, збалансованість, машиночитаність [13, с. 55].

І. Мейзерська виділяє такі ознаки лінгвістичного корпусу: 1) фрагментованість; 2) нелінійність, вертикальна організація; 3) наведення конкретних контекстів та мовних структур; 4) зразки загальної мовної практики; 5) відсутність зв’язності; 6) комунікативна мета простежується лише на рівні окремого речення [31, с. 57].

У загальному розумінні, за А. Лучик та І. Остаповою, лінгвістичний корпус – «це представлений у цифровому форматі, великий за обсягом, уніфікований, структурований, розмічений і філологічно компетентний масив текстів природною мовою, доповнений системою керування — універсальними програмними засобами для пошуку та опрацювання різноманітної лінгвістичної інформації» [29, с. 34].

Існує два підходи до організації функціювання корпусів текстів. Перший дозволяє досліднику «скачувати» самі тексти (з розміткою або без неї) і потім працювати з ними, використовуючи як програми експлуатації, рекомендовані укладачами колекції, так і свої власні. Другий підхід полягає в тому, що досліднику надається доступ до програми експлуатації корпусу, але не до самих текстів. Останній підхід, як правило, пояснюваний міркуваннями захисту авторських прав на тексти і їх розмітку.

О. Демська-Кульчицька пропонує таку класифікацію корпусів текстів [11, с. 156-157]:

1) повнотекстові (тексти в корпусі подані повністю) та фрагментарні (подані фрагменти текстів);

2) дослідницькі (застосовують у лінгвістичних дослідженнях із метою формулювання нових теорій, концепцій тощо) й ілюстративні (застосовують для підтвердження уже висловлених теоретичних положень чи гіпотез про мову);

3) дослідницькі (подають тексти як цілісні об’єкти, як факт реалізації мовної системи) й інтерпретаційні (становлять інформаційно-довідкові та дослідницькі системи);

4) діахронні (репрезентують мову в понад часовому зрізі) та синхронні (репрезентують мову або тип тексту певного визначеного часового проміжку);

5) моніторингові (динамічні) (забезпечують можливість відстежувати зміни у мові, враховуючи аспект діахронії) та статичні (засвідчують стан мови на певному синхронному зрізі);

6) загальномовні (репрезентують загальнонародну, національну мову) та спеціалізовані (скеровані на розв’язання часткових, особливих, специфічних науководослідних завдань).

Усі корпуси, за А. Гладковою, поділяються на дві великі групи: «першу формують такі, які містять “чисті” тексти; другу – такі, які містять тексти анотовані, тобто спеціально структуровані та підготовлені таким чином, аби забезпечити досліднику можливість опрацьовувати потрібний саме йому матеріал» [5, с. 329]. Іншими словами, анотовані корпуси містять позатекстову лінгвістичну інформацію, представлену спеціальними тегами (маркерами на позначення релевантної інформації про частиномовні, граматичні, стилістичні, прагматичні та інші особливості тієї чи іншої лексичної одиниці). Для тегування, або ж маркування тексту, прийнято використовувати єдину Стандартну узагальнену маркувальну мову (СУММ) (англ. Standardized Generalized Markup Language, SGML).

І. Мейзерська наголошує, що за типом та призначенням виділяють: *статичні* та *динамічні корпуси* (залежно від можливості їх поповнення новими даними); *одно-, дво- та багатомовні*; *навчальні корпуси* (створені спеціально для потреб іноземців, які вивчають мову); *паралельні* (на матеріалі різних мов); *порівняльні* (тексти певної тематики та проблематики різними мовами); *узгоджені (aligned)* – корпуси з підрядковим перекладом, що укладаються на основі порівняльних. Із дослідницькою метою найзручнішим є використання анотованих корпусів, оскільки вони є потужним інструментом для аналізу лексичних одиниць та різноманітних синтаксичних конструкцій у їх реальному функціонуванні в мовній практиці. Зокрема, така організація корпусу, на думку дослідниці, дозволяє відстежувати варіативні граматичні моделі та одержувати швидкий доступ до ілюстративних прикладів, одержаних із реальних текстів [31, с. 54-55].

Корпуси текстів застосовуються в лінгвістичних дослідженнях різної спрямованості:

1) на корпусах текстів проводяться лексичні дослідження; аналізуються частоти вживання і різні форми слів, виявляються статистичні закономірності і нові слова;

2) вивчення історичного розвитку, діалектів мови, і різних форм слововживань.

Крім вирішення власне-наукових завдань корпус текстів може використовуватися в дидактичних і, навіть, суто практичних цілях, наприклад, у практиці перекладу для встановлення, наскільки «природно» звучить та чи та конструкція і наскільки точно вона відображає вкладений в неї сенс.

Практика викладання іноземної мови базується на правилах і зразках реального вживання тих чи інших граматичних конструкцій. На відміну від традиційних граматик, корпусні враховують стильову специфіку вживання лексичних одиниць та граматичних конструкцій, що є надзвичайно важливим для людей, що вивчають мову.

Перевагами застосування корпусу в мовознавчих студіях, на думку дослідника Я. Свартвіка, є те, що він:

– об’єктивний, оскільки мовці часто не можуть дати точний звіт про те, що вони говорять;

– верифіковний;

– корисний у вивченні мовних варіантів, діалектів, стилів, а також в історичних порівняннях;

– встановлює частотність слововжитку;

– є теоретичним ресурсом;

– корисний для машинного перекладу, розпізнавання та синтезу мовлення, а також програм, пов’язаних зі вживанням мови;

– дає більш репрезентативну картину мови, ніж добірки цитат;

– є єдиним способом вивчати слововживання носіїв інших мов, оскільки жодна інша методика не працює;

– той самий корпус можна використовувати для різних потреб [59, с. 7].

1.2 Огляд лінгвістичних корпусів текстів

Упродовж останніх десятиліть у багатьох країнах здійснюються роботи з формування лінгвістичних корпусів із метою вивчення національних мов. Національний корпус – це великий за обсягом корпус, що прагне до об’єднання у своїй структурі текстів найрізноманітніших жанрів і типів (сучасні технології дозволяють вбудовувати в національні корпуси аудіо- і відоматеріали). Чим більший розмір такого корпусу, чим різноманітніша та точніша розмітка, чим ефективніше програмне забезпечення такого корпусу, тим вища його цінність як лінгвістичного ресурсу [13, с. 63]. Саме тому лінгвісти і програмісти, що зайняті створенням національних корпусів, головним чином вирішують питання матеріального та технічного оснащення, для того щоб матеріали корпусу були доступні для роботи і задоволення запитів користувачів. Важливість та необхідність створення таких корпусів

важко переоцінити, адже, перш за все, вони збирають і зберігають мову для сучасників та майбутніх поколінь, дані цих корпусів дозволяють аналізувати стан багатьох мов світу у синхронічному та діахронічному аспектах (якщо корпус включає історичні підкорпуси, а сама можливість поповнення корпусу вже передбачає діахронію).

Національним корпусам протиставляються спеціальні, які створюються для вирішення конкретних лінгвістичних задач. Спеціалізований корпус – це жанрово чи галузево специфічний корпус, що має на меті відобразити певну підмову.

1.2.1 Британський національний корпус (англ. British National Corpus, BNC). Перші корпуси з'явились у Великобританії в 60-ті рр. XX ст.: Корпус університету Брауна (англ. Brown University Corpus) та Ланкастер / Корпус Осло-Берген (англ. Lancaster / Oslo-Bergen Corpus (LOB)). Цей корпус містить морфологічну розмітку і має приблизно один мільйон слововживань. Зодом у цей корпус було додано також і синтаксичну розмірку.

Одним з найбільш відомих і популярних корпусів англійської мови (однак не єдиним) є Британський національний корпус (англ. British National Corpus, BNC). Цей корпус був створений спільними зусиллями кількох британських університетів і видавництв, а також Британською бібліотекою за період 1991-1994 рр. Корпус включає письмові й устінні тексти британською англійською к. XX ст., що належать до різних жанрів і функційних стилів. Корпус є фрагментарним: тексти обсягом більше 45000 слів, представлені уривками (що дозволяє уникнути впливу індивідуального стилю того чи того автора на результати). Загальний обсяг корпусу становить більше 100 000 000 слововживань. Тексти BNC розміщені в стандартах SGML відповідно до рекомендацій TEI.

Корпус BNC [68] оснащений морфологічною розміткою: кожна словоформа схарактеризована за приналежністю до частини мови, розряду в рамах частини мови та форми словозміни. Ця розмітка робилась автоматично, що призвело до помилок у 1,7% випадків, а 4,7% словоформ не змогли бути

однозначно зінтерпретованими й отримали «подвійний морфологічний код». Фрагмент корпусу, що складає 2% від загального обсягу, був відібраний для більш детальної («ручної») морфолого-синтаксичної розмітки.

Експлуатація корпусу здійснюється за допомогою ряду спеціально створених програм обробки SGML. Обмежений доступ до ресурсу корпусу безкоштовно надається через Інтернет, але для того, щоб користуватися всіма його можливостями, необхідно придбати CD-ROM або придбати платну реєстрацію для доступу в режим «on-line».

Дані BNC широко використовуються під час складання словників, граматик і підручників англійської мови, у лінгвістичних дослідженнях, у роботі штучного інтелекту, а також у практиці викладанні англійської мови.

1.2.2. Французька база Франтекст (фр. Frantext). Однією з перших історично і найбільших на сьогодні електронних колекцій текстів є французька база Франтекст (фр. Frantext). Зауважимо, що це не корпус, проте система його експлуатації дозволяє досліднику формувати свій «робочий корпус» з урахуванням цілого ряду параметрів (автор, дата, жанр, розмір і ін.). Робота зі створення бази почалася в 1957 р. у рамах підготовки 16-томного «Тезаурусу французької мови» (фр. TLFI), проте з часом поповнення і розробка засобів експлуатації корпусу виділилися в самостійну задачу. В створення Франтексту були вкладені великі фінансові кошти: лабораторія Національного центру наукових досліджень Франції (фр. CNRS) у складі 30-50 осіб працювала над ним упродовж півстоліття. Нині Франтекст налічує 3737 текстів XVI-XX ст. (близько 210 000 000 слововживань) і продовжує поповнюватися. Основну масу (близько 80%) становлять літературні тексти, проте в ній також представлені наукові та технічні твори. Більше половини текстів бази (1940 текстів, 127 000 000 слововживань) забезпечені морфо-синтаксичною розміткою.

Зовнішній доступ до Франтексту відкритий з 1992 р. для корпоративних користувачів (бібліотек, університетів і т.п.) і є платним. Вільний доступ

надається до бібліографічної бази та до електронної версії «Тезаурусу французької мови».

В останні роки ведеться робота з поглиблення «історичної перспективи» Франтексту: до нього додані бази текстів старофранцузького (IX-XIII ст.) і середньофранцузького (XIV-XV ст.) періодів, причому цими базами будь-який бажаючий може користуватися безкоштовно.

До переваг Франтексту [64] зараховують його колосальні розміри і тривала історія формування. Разом із тим ці переваги є джерелом його проблем. Розроблені в 60-70-ті рр. формати і системи експлуатації сьогодні сильно застаріли і не відповідають можливостям сучасної техніки і запитам дослідників. Модернізація Франтексту, зокрема, його переведення у стандарт XML і розмітка відповідно до рекомендацій TEI є складним завданням. Проте чинна система експлуатації Франтексту уможливлює вирішення багатьох лінгвістичних і літературознавчих завдань і широко використовувана дослідниками французької мови у всьому світі.

1.2.3 Український національний лінгвістичний корпус (УНЛК). В Українському мовно-інформаційному фонді НАН України з метою дослідження української мови та укладання сучасних словників створено Український національний лінгвістичний корпус (УНЛК). УНЛК розробляється під керівництвом академіка НАН України В.А. Широкова [20, с. 103]. Розпорядженням Кабінету Міністрів України від 11.02.2004 р. №73-р Національну словникову базу Українського мовно-інформаційного фонду НАН України внесено до державного реєстру наукових об’єктів, що становлять національне надбання.

Дослідники називають три основні напрямки УНЛК: 1) надання текстової інформації за певними критеріями; 2) створення вхідних потоків лінгвістичної інформації для різноманітних дослідницьких систем; 3) інтеграція різнопланових лінгвістичнопрограмних засобів обробки текстів у єдиному середовищі.

У системі УНЛК [67] виділяються дві основні функційні підсистеми: бібліографічна та повнотекстова. Бібліографічна частина являє собою електронну бібліотеку — колекцію цифрових ресурсів, яка є основою для розробки будь-якого корпусу. Тому бібліографічна підсистема служить інструментом для збирання, збереження, моделювання й використання природномовної інформації в цифровому вигляді. Обсяг ресурсів не обмежений, тобто корпус постійно поповнюється новими джерелами. Друга можливість — це повнотекстовий пошук. Користувач уводить пошукову фразу, задає максимально бажану кількість слів між пошуковими та обирає додаткові параметри повнотекстового пошуку, а саме:

- урахування порядку слів;
- пошук у певній підмножині об’єктів;
- використання процедури лематизації;
- використання синонімічної лексикографічної бази даних;
- використання певних синонімічних рядів;
- вибір граматичних параметрів для кожного слова, що входить у пошуковий фрагмент.

Результатом повнотекстового пошуку є список бібліографічних описів. Але на відміну від пошуку за бібліографією, користувач отримує прямий доступ до кожної локалізації пошукового фрагмента в тексті, тобто до всіх контекстів, які містять пошуковий фрагмент. Обравши джерело, користувач може переглядати контексти (для зручності пошуковий фрагмент виділено червоним кольором). Розмір (довжину) контексту можна змінювати. Кожне слово з контексту з’єднане зі словниками: граматичним, тлумачним і словником синонімів. Ще одним засобом для створення користувацьких наборів даних у системі є так званий «кошик» (тимчасове сховище). Користувач може відібрати об’єкти збереження, які його цікавлять, з різних пошукових запитів і зберегти образ даних для подальшої роботи в інших сеансах. Користуючись таким інструментом, дослідник може відібрати джерела певного автора, стилю чи жанру і працювати лише з цією частиною

корпусу; така процедура дає можливість виділити із загального корпусу свій власний підкорпус, орієнтований на розв'язання особистих завдань. Такий підхід дозволив відмовитися від занадто суб'єктивною ідеї створення «еталонного корпусу» і надає можливість досліджувати засобами корпусу «нееталонних», аномальних мовних явищ [29, с. 35-36]. Отже, реалізовані в УНЛК функції кошика дозволяють породжувати в режимі реального часу та в рамках єдиного лінгвістичного корпусу віртуальні «підкорпуси», які моделюють ті або інші лінгвістичні ефекти із забезпеченням кожному дослідникові можливостей реалізації його власних уявлень про еталонність, виваженість і збалансованість. Це змінює парадигму корпусних досліджень: ми досліджуємо не наперед заданий набір текстів, а дискурси, незалежно від того, якими текстами вони підтримуються. Система лінгвістичного корпусу є багатоплановою і може мати чимало різноманітних застосувань. В Українському мовно-інформаційному фонді вона, насамперед, використовується як джерельна база лінгвістичної інформації для створення фундаментальної академічної багатотомної лексикографічної системи «Словник української мови». Серед інших застосувань слід відзначити здійснення лінгвістичних досліджень з метою виявлення нових мовних явищ та формалізації наявних. Зрозуміло, що технологія УНЛК надає засоби для граматичного та семантичного маркування текстів. Лінгвістичний корпус може бути і зручним середовищем для статистичного опрацювання текстової інформації. Оскільки основним напрямком УНЛК є надання текстової інформації за певними критеріями, ми звернулись до лінгвістичного корпусу для встановлення синтагматичних властивостей еквівалентів слова, що дало нам можливість вирішувати поставлені задачі на максимально великому за обсягом матеріалі.

1.2.4 Національний корпус російської мови (рос. Национальный корпус русского языка, НКРЯ). Одним із головних корпусів текстів російською мовою є Національний корпус російської мови (рос. Национальный корпус русского языка (НКРЯ)). Корпус містить близько

п’ятисот мільйонів словоформ. У колекції корпусу міститься безліч типів текстів: історичні, літературні, діалектні, письмові, усні, сучасні, перекладні. Корпус оснащений значною кількістю розміток: лексичною, морфологічною, синтаксичною, лексико-семантичною і рядом інших спеціалізованих розміток. Особливістю корпусу є віршована розмітка, яка дозволяє шукати віршовані тексти із завданням різних параметрів.

У корпусі умовно виділяються дві частини – сучасна й діахронічна. Корпус сучасних текстів становлять тексти, що були створені у період 1951-дотепер. Діахронічна частина становить близько 53 млн. слововживань і поєднує тексти XVIII ст. (1,1 млн. слововживань), XIX ст. (23,3 млн. слововживань прозаїчних текстів і 2,5 млн. слововживань у поетичному корпусі) і 1-ї половини XX століття (25,4 млн. слововживань). Основний масив текстів, зібраних у НКРМ, охоплює період в 200 років, тому він найбільш пристосований для вивчення коротких (кілька десятиліть) і середніх мовних змін. Національний корпус російської мови [66] включає такі підкорпуси (субкорпуси) [35]:

1) глибоко анотований корпус, у якому для кожного речення побудована повна морфологічна й синтаксична структура (дерево залежностей);

2) паралельний російсько-англійський корпус текстів, у якому можна знайти всі переклади для певного російського або англійського слова або словосполучення;

3) корпус діалектних текстів, що включає запис діалектного мовлення різних регіонів Росії зі збереженням їх граматичної специфіки; передбачений спеціальний пошук з урахуванням діалектної морфології;

4) корпус поетичних текстів, у якому можливий пошук не тільки за лексичними і граматичними, але й за специфічними для вірша ознаками (пошук певної комбінації в сонетах, в епіграмах, у віршах, написаних амфібрахієм, з певним типом римування й т.п.);

5) навчальний корпус російської мови – корпус зі знятою омонімією, розмітка якого орієнтована на шкільну програму російської мови;

6) корпус усного мовлення включає розшифрування магнітофонних записів публічного й приватного усного мовлення, а також транскрипти кінофільмів 2000-х років.

1.2.5 Інші національні корпуси текстів. У Німеччині найбільшим корпусом є корпус Інституту німецької мови в Маннгеймі. Цей корпус містить близько двох мільйонів слововживань, має морфологічну і синтаксичну розмітки, а також автоматизовану систему пошуку вмісту корпусу за морфологічними ознаками словоформ.

У Чехії є Чеський національний корпус [69], відмінною рисою якого є можливість отримувати всі приклади вживань разом із контекстами, в яких словоформа зустрічається, частотою входження словоформи в корпус. Також є морфологічний аналізатор, який дозволяє проводити морфологічний і контекстний аналізи.

1.3 Проблеми корпусної лінгвістики

На перших етапах розвитку корпусної лінгвістики у корпусах текстів виявилася важлива особливість – їх *вузька спрямованість*. Кожен корпус текстів повинен мати розмітку, а в деяких випадках набір інформації, виходячи з поставлених завдань. Наприклад, для дослідження в області морфології корпус повинен містити дані про характеристики слова (частина мови, рід, число, відмінок – для іменників; вид, час, перехідність – для дієслів тощо), для вивчення виразів потрібне поле для зв’язку слів між собою, для створення класифікаторів необхідне вказати автора та предметної області та т.д.

Корпус текстів, особливо при його використанні для машинного навчання або перевірки якості роботи інших алгоритмів і методів автоматичного аналізу текстів, повинен містити великий обсяг даних. Роботи з реалізації текстового корпусу вимагають багато часу і складаються з декількох підзадач:

- збір інформації з різних джерел і на різні тематики для реалізації принципу репрезентативності;
- обробка зібраної інформації;
- аналіз обробленої та структурованої інформації;
- формування розмітки для корпусу текстів.

У такий спосіб, через специфіку лінгвістичних корпусів і труднощів їх створення існує *проблема невідповідності корпусів після виконання поставлених завдань*, оскільки практично завжди корпус створюється під конкретну задачу. Така проблема властива корпусам меншим за національні, в яких великий арсенал розміток, що робить їх майже універсальними, але в них відсутній програмний інтерфейс або можливість отримання великого обсягу текстів для використання в програмних інструментах. Крім того, нові області застосування інструментів комп'ютерної лінгвістики з'являються швидше за нові тексти або види розмітки у корпусах. При вирішенні складних завдань тексти аналізуються з різних поглядів: чи є тексти в соціальних мережах хибними, яке емоційне забарвлення мають тощо. Рішення таких завдань, особливо засобами машинного навчання, вимагає складання спеціальної розмітки корпусу. Після виконання поставлених завдань корпуси стають не такими затребуваними, як раньше, незважаючи на те, що було створено велику роботу за його реалізацією. Рішенням цієї проблеми може стати розробка комплексу інструментів для створення та розмітки корпусів текстів, орієнтованих на вирішення різних завдань.

Джерелом даних для корпусів можуть слугувати різні електронні бібліотеки, збірки текстів з певної тематики або енциклопедії, новинні ресурси, відкриті дані соціальних мереж тощо. Важливою вимогою для автоматичного створення корпусу і його розмітки є можливість отримання текстів із заздалегідь визначеними дослідниками властивостями, наприклад, тексти за тематикою або за авторами. Таких властивостей може бути багато, тому потрібне створення правил для розширюваної розмітки корпусу з метою забезпечення можливості повторного використання наявних текстів, крім

того, необхідний інструмент, який за відносно невеликий час дозволить автоматично отримувати потрібні тексти.

Висновок до розділу 1

Науковці сприймають корпусну лінгвістику як прикладну мовознавчу дисципліну, зі своєю особливою теоретичною та методологічною базою, яка може використовуватися й у суміжних гуманітарних науках для залучення до досліджень аналізу мовних даних.

Спроба визначення корпусу текстів на базі апріорних детермінативних ознак передбачає широке розуміння терміна з визнанням різних типів корпусів і вузьке, згідно з яким корпусом уважаються лише певні типи корпусів, як-то електронні, анотовані або статичні. Текстовий корпус – це тексти та інформація, зібрані відповідно до визначених принципів, розмічених за певним стандартом і забезпечених пошуковою системою.

Основними сферами застосування корпусів текстів є: 1) наукові лінгвістичні дослідження в галузях лексичної семантики, граматичної та лексичної сполучуваності, стилістики, прагматики, діалектології тощо; 2) лексикографія, зокрема створення конкордансів, словників слів та словосполучень на основі одномовних корпусів, а також багатомовних лексиконів і конкордансів із залученням паралельних текстових масивів із різних мов; 3) лінгводидактика, добір навчального матеріалу та створення ефективних посібників для осіб, що вивчають ту чи іншу мову як іноземну.

РОЗДІЛ 2
ЕТАПИ СТВОРЕННЯ Й ІНСТРУМЕНТИ РОБОТИ З КОРПУСОМ ТЕКСТІВ

2.1 Алгоритм створення корпусу текстів

За В.П. Захаровим [16], формування корпусів відбувається за таким алгоритмом: проєктування; забезпечення надходження текстів відповідно до зазначених джерел; підготовка "технологічного" опису; перетворення в машинозчитувану форму; конвертування й попередня обробка текстів; графематичний аналіз (токенізація); метарозмітка; лінгвістична розмітка (виділення – наше, оскільки саме наявність розмітки різних типів уможливлює оперування корпусу як інформаційно-пошукової системи для вирішення практичних завдань); коригування результатів автоматичної розмітки; завантаження розмічених текстів у структуру корпус-менеджера; забезпечення доступу до корпусу (пошук); створення документального забезпечення.

Технологічний процес створення корпусу текстів передбачає поступову реалізацію восьми основних етапів (за В. Жуковською [13, с. 85-87]): визначення джерел лінгвістичного матеріалу, введення даних, попереднє опрацювання тексту, конвертування й графематичний аналіз, розмітка тексту, коректування результатів автоматичної розмітки, конвертування розмічених текстів у структуру спеціалізованої лінгвістичної інформаційно-пошукової системи, забезпечення доступу до корпусу. Розглянемо ці етапи.

1. Визначення джерел лінгвістичного матеріалу. Основною проблемою, з якою стикнеться дослідник при відборі матеріалу для корпусу – це авторські права. Дотримання авторських прав на матеріали передбачає отримання дозволу на використання текстів для дослідницьких цілей. Закони щодо

авторських прав різняться у різних країнах, тому укладач повинен знати про закони, що охороняють авторські права не лише у своїй країні, але і в інших державах. Найбезпечніший шлях уникнути проблем із законодавством – це використати тексти із відкритих джерел.

2. Введення даних. Існує три способи введення даних у корпус: адаптація даних в електронному форматі, сканування та ручне введення. Готові тексти в електронному форматі є найлегшим способом внесення даних у корпус. Зважаючи на те, що більшість корпусних менеджерів підтримують документи з розширенням .txt, тексти в інших форматах повинні бути переформатовані. Якщо необхідні тексти існують лише у друкованому вигляді, їх конвертують в електронну форму за допомогою сканування. Це можуть бути рідкісні чи старі видання, що не мають цифрових версій. Звичайно, сканування потребує гарного технічного та програмного забезпечення. Скановані версії не позбавлені недоліків, тому після оцифровки тексти необхідно звірити з оригіналом та виправити помилки. Проте найбільш працемістким та часомістким процесом є ручний набір текстів. Такого методу введення даних до корпусу не уникнути, якщо текст знаходиться у такому стані, що сканування є неможливим чи існує лише рукописна версія документа.

3. Попереднє опрацювання тексту. На цьому етапі всі тексти, отримані з різних джерел, проходять етап коректування. Здійснюється також підготовка бібліографічного й екстралінгвістичного опису тексту.

4. Конвертування й графематичний аналіз. Деякі тексти проходять також через один або кілька етапів попередньої машинної обробки, у ході яких здійснюються різного роду перекодування (якщо потрібно), видалення або перетворення нетекстових елементів (малюнки, таблиці, графіки, формули), видалення з тексту переносів, «твердих кінців рядків», забезпечення однакового написання тире та ін. Як правило, ці операції виконуються в автоматичному режимі. Звичайно, на цьому ж етапі здійснюється сегментування тексту на його структурні складові.

5. Розмітка тексту. Розмітка тексту полягає в приписуванні текстам та їх компонентам додаткової інформації (метаданих). Метаопис текстів корпусу включає як змістовні елементи даних (бібліографічні дані, ознаки, що характеризують жанрові й стильові особливості тексту, відомості про автора), так і формальні (ім'я файлу, параметри кодування, версія мови розмітки, виконавці етапів робіт). Ці дані вводяться вручну. Структурна розмітка документа (виділення абзаців, речень, слів) і лінгвістична розмітка здійснюються автоматично.

6. Коректування результатів автоматичної розмітки: виправлення помилок і зняття неоднозначності (вручну або напівавтоматично).

7. Конвертування розмічених текстів у структуру спеціалізованої лінгвістичної інформаційно-пошукової системи (corpus manager), що забезпечує швидкий багатоаспектний пошук і статистичну обробку.

8. Забезпечення доступу до корпусу. Корпус може бути доступний у локальній мережі, тобто лише розробникам та особам, які мають право користування. Наприклад, право на використання корпусу ARCHER має консорціум чотирнадцяти університетів семи країн. Корпус може поширюватися на CD-ROM, як Eearly Modern English Medical Texts та Middle English Medical Texts від видавництва John Benjamins. Також корпус може бути розміщеним у глобальній мережі. Різним категоріям користувачів можуть надаватися різні права й можливості доступу та експлуатації корпусу.

Під час створення корпусу використовується низка процедур і програм, [16, с. 38-41]:

1. Токенізація – це розбиття потоку символів природної мови на окремі значимі одиниці (токени, словоформи).

2. Лематизація – процес утворення початкової форми слова, виходячи з інших його словоформ. У багатьох мовах слово може зустрічатися в декількох формах з різними флексіями. Наприклад, англійське дієслово 'work' має такі форми: 'work', 'worked', 'works', 'working'. Базова форма, 'work', зафіксована в словнику, називається лемою слова. Лематизація – це процес угруповання

різних флективних форм одного слова таким чином, щоб при аналізі вони оброблялись як одне слово.

3. Стеммінг полягає в знаходженні стеми (основи) слова. Стеммер обробляє окреме слово без знання контексту, і, таким чином, не може диференціювати слова, які мають різні значення в силу віднесеності до різних частин мови. Проте стеммери більш прості для реалізації й швидше обробляють дані. Наприклад, токену "better" відповідає лема "good", але це опускається при стеммінзі. Лема "work" єбазовою формою для токена "working", і ця відповідність буде виявлена як при стеммінзі, так і при лематизації.

4. Парсинг – це процес аналізу синтаксичної структури тексту чи частини тексту, що ґрунтується на зіставленні лінійної послідовності лексем (слів, токенів) мови з її формальною граматикою. Результатом є дерево залежностей (синтаксичне дерево). Побудова автоматичних синтаксичних аналізаторів (парсерів) для великих корпусів є однією із найважливіших областей комп'ютерної лінгвістики.

2.2 Поняття «розмітки» в корпусній лінгвістиці

Повноцінний корпус повинен мати, крім власне текстів, косплекс «інструментів» для роботи з ними. Ці інструменти можна поділити на дві категорії:

1) засоби перегляду текстів і запиту даних;

2) засоби збагачення корпусу аналітичною інформацією, яка називається анотацією (англ. annotation), або розміткою (англ. markup, tagging).

Найбільш поширеними способами перегляду тексту є імітація видання (з можливим виділенням об'єктів, що цікавлять дослідника) і конкорданси (список словоформ або словосполучень у контексті). Основна перевага електронного видання перед друкованим полягає в можливості швидкого пошуку необхідних досліднику форм і сполучень. Широта параметрів пошуку

залежить від того, яка аналітична інформація закодована в корпусі. Якщо ми хочемо знайти всі випадки вживання певної словоформи, то це легко зробити в простому текстовому файлі. Якщо ми хочемо знайти всі випадки вживання певної лексеми, представленої рядом словоформ, це складніше, але також можливо. Якщо ж ми хочемо знайти всі випадки вживання певної грамеми (наприклад, орудний відмінок однини іменника), зробити це на нерозмічену корпусі вкрай проблематично.

Розмітка – це збагачення корпусу різного типу аналітичною інформацією. Мінімальна розмітка, як правило, легко проводиться в автоматичному режимі, полягає в оснащенні корпусу покальною інформацією. Іншими словами, коли ми отримуємо відповідь з корпусу на наш запит, ми повинні чітко знати «координати» нашого прикладу («текст / глава / абзац» або «сторінка / рядок»).

Для лінгвістичних досліджень велику цінність має морфологічна розмітка: кожна словоформа співвідноситься з «початковою» («словниковою») формою лексеми, визначається її частиномовна приналежність, грамеми словозмінних категорій.

Для багатьох мов розроблені програми автоматичної морфологічної розмітки, однак всі вони дають певний відсоток похибки (неминучу через мовну омонімію) і вимагають «ручної» перевірки. У ряді випадків можна обмежитися грубими автоматичними даними і враховувати відсоток похибки.

Можливі й інші види лінгвістичної розмітки: синтаксична, семантична, прагматична тощо, проте їх універсальність не настільки очевидна. Якщо в питанні про основні частини мови і складі грамем можна говорити про відносний консенсус серед лінгвістів, то синтаксичні функції і семантичні угруповання різними мовознавчими школами розуміються принципово по-різному.

Крім лінгвістичної, існує і філологічна розмітка. Вона дозволяє включати в корпус варіанти тексту, авторську і редакторську правки, виділяти

іноземні слова, цитати, пряму мову персонажів літературного твору, різного типу стилістичні фігури.

Аналітична розмітка корпусу – досить трудомісткий процес, проте він не позбавлений наукового інтересу. Під час «наклеювання ярликів» на словоформи або синтаксичні конструкції виявляються «слабкі місця» використовуваних класифікацій, звертають на себе увагу цікаві приклади.

Особливий інтерес представляє метарозмітка як необхідний інструмент для роботи з лінгвістичним корпусом. Під метарозміткою Н. Леміш розуміє зовнішні, екстралінгвістичні відомості про автора й відомості про текст: автор, назва, рік і місце видання, жанр, тематика; відомості про автора можуть включати не тільки його ім'я, але й вік, стать, роки життя тощо. За В.Г. Шкляревським, така розмітка, описуючи позамовні властивості тексту, «дає змогу не лише виявити фактори впливу на словесну тканину текстів, виявити залежність мовних (лінгвальних) характеристик текстів від середовища їх побутування, а й окреслити тенденції розбудови тієї чи іншої дисципліни за допомогою аналізу відповідних фахових текстів» [41, с. 300]. Такий підхід також свідчить про розмежування власне розмітки корпусу (надання додаткової екстралінгвістичної інформації про тексти корпусів – позамовної / екстралінгвальної) й анотування текстових елементів (усе, що стосується мовних / лінгвальних параметрів – "лінгвістична розмітка") [41, с. 300]. Значущими є також спостереження В.Г. Шкляревського щодо терміна описова розмітка (descriptive markup) на позначення метарозмітки Браунівського корпусу: "метарозмітка розумілась як загальний опис формальної композиційної структури письмових текстів: "виділення частин тексту, назви, рядка тощо" [Там саме]. Це був проміжний етап між лінгвістичною розміткою та власне метарозміткою. За міжнародний варіант метарозмітки визнають класифікацію Дж. Сінклера-Шарова, що передбачає виділення зовнішніх (позамовних) чинників, серед яких розрізняють три групи – походження тексту, зовнішні ознаки й мету написання тексту, та внутрішніх (змістовних і мовних) чинників – предметної області й стилістичних

особливостей тексту. В результаті укладається паспорт з чотирма групами елементів метаопису (бібліографічні дані, відомості про жанр, автора, інформація про структуру, особливості розмітки й модифікації тексту. Це дійсно так, але для наукових текстів (наприклад, спеціальних з лінгвоантропогенезу) зазначений вище набір елементів не є достатнім для визначення їхньої поняттєвої структури. Слушним у цьому зв'язку є зауваження В. Шкляревського щодо використання "індивідуалізованого стандарту розмітки", або "тематичної метарозмітки".

2.3 Особливості стандартизації оформлення корпусів текстів: стандарт TEI (англ. Text Encoding Initiative)

До сьогодні дослідники говорили про властивості корпусу відсторонено від конкретних технологічних рішень, що дозволяють утілити їх на практиці. Такі рішення можуть бути різними, причому чим більше розмітки містить корпус, чим більше прикладних програм розробляється для його експлуатації, тим різноманітнішими і трудомісткими можуть ставати технічні рішення.

Несумісність стандартів, які використовуються авторами корпусів у різних країнах і дослідницьких центрах, ставить під загрозу таку важливу для корпусної лінгвістики можливість широкого обміну даними, об'єднання і взаємного збагачення корпусів. З сер. 80-х рр. XX ст. починається робота з прийняття певних міжнародних норм оформлення електронних видань текстів. У 1987 р. ця робота організаційно оформилась і отримала назву «Ініціатива з кодування текстів» (англ. Text Encoding Initiative (TEI) [71]). Основним продуктом діяльності TEI є «Рекомендації з кодування й обміну електронними текстами» з використанням стандартів SGML або XML. Підкреслимо, що повна розмітка корпусу текстів із використанням усіх пропонованих TEI елементів не тільки не є обов'язковою, але і навряд чи практично здійснювана загалом. Перевага відкритого формату S/XML полягає якраз в тому, що він дозволяє здійснювати розмітку частково і поетапно: ніякі

«теги» за винятком мінімальної розмітки заголовка і «тіла» тексту не є обов'язковими.

Ідея змістової розмітки електронних текстів (на відміну від кодування формальних друкарських засобів) бере початок із к. 1960-х рр. Уперше висловлена У. Танніклiффом і С. Райсом, вона отримує розвиток у рамах дослідницького проєкту компанії IBM. Розробка власне формату SGML почалася в 1978 р. під егідою Американського національного інституту стандартизації (англ. ANSI), пізніше вона була підтримана міжнародним Інститутом стандартизації (англ. ISO). Назву стандарту англ. Standard General Markup Language перекладаємо як «стандартна узагальнена мова розмітки». Його перший чорновий варіант був запропонований в 1980 рю, а в 1986 р. була офіційно опублікована остаточна версія (стандарт ISO 8879: 1986).

Стандарт SGML був використаний для оформлення документів ряду великих видавництв і промислових корпорацій, однак практика показала, що реалізований у ньому принцип економії розмітки суттєво ускладнює подальшу автоматизовану обробку документів. У сер. 90-х рр. XX ст. провідні компанії-виробники програмного забезпечення делегували своїх представників до робочої групи зі створення більш зручного в експлуатації «підвиду» SGML, який отримав назву англ. Extensible Markup Language (XML, 'розширена мова розмітки'). Одним із завдань, які ставилися на початку розробки XML, було зручність використання цього формату для створення веб-сторінок.

Зберігаючи всі основні переваги SGML (за винятком деяких можливостей скорочення «надмірної» розмітки), XML значно легше піддається обробці з метою візуалізації й аналізу даних. У числі найбільш популярних засобів обробки документів XML слід назвати «стильові листки» (англ. stylesheets) в стандарті XSL. Сьогодні формат XML набуває все більшого поширення в сфері зберігання даних, призначених для передачі через Інтернет.

В основі S/XML лежить уявлення про документ (будь то літературний твір, випуск періодичного видання або збірник законів) як про ієрархічну

структуру вкладених один в одного елементів (наприклад, текст роману складається з трьох томів, кожен з яких включає декілька частин, кожна частина – глави, а кожна глава – абзаци). У тексті твору можуть зустрічатися окремі слова або фрази іноземною мовою, що іноді виділяються шрифтом (курсивом). Ці фрагменти можуть в електронній версії позначатися як особливий елемент, хоча це не обов'язково.

Кожен елемент крім змісту («вкладеного» елемента або тексту) може мати один або кілька атрибутів, що містять додаткову інформацію (наприклад, номер глави або іноземну мову, слово з якої використано в тексті).

Правила того, які елементи може містити це елемент і які атрибути він може мати, записуються в спеціальний файл, який називається 'декларація типу документа' (англ. Document Type Declaration (DTD)). В принципі для кожного документа можна створити власну DTD, однак для зручності обміну текстовими архівами доцільно дотримуватися по можливості єдиних правил. Саме таку універсальну DTD прагне розробити TEI, а її рекомендації можна розглядати як коментар до цієї DTD.

Теги S/XML є певним кодом у кутових дужках. Код складається з назви елемента (одна або кілька латинських букв) і атрибутів, які відділяються пропуском і складаються з назви атрибута, знака рівності і значення в лапках. Зміст елемента міститься між початковим і кінцевими тегами (кінцевий тег містить тільки назву елемента, перед яким ставиться коса риска, або «слеш»).

Отже, в основі XML лежить ідея змістовної розмітки: теги використовуються для кодування не форми подання тих чи тих елементів тексту, а їх функції в структурі тексту. У цьому відмінність XML від HTML або кодів форматування, використовуваних в текстових процесорах (таких, як Microsoft Word). Найважливішими особливостями формату S/XML є його відкритість (можливість поетапної розмітки, включення в неї додаткових елементів) і відсутність «командних кодів», прив'язаних до якогось певного програмного забезпечення: в принципі будь-який документ S/XML може читатися в будь-якому простому текстовому редакторі. Це дає підставу

сподіватися, що використання названого формату забезпечить збереження і доступність даних незалежно від майбутніх змін у програмному забезпеченні і наукових теоріях, на підставі яких проводиться розмітка.

Поряд із перевагами формат XML має певні незручності:

1. Проблеми виникають у зв'язку з жорсткою ієрархічною структурою складових документа. Насправді, будь-який друкований або рукописний документ має як мінімум дві структури – формальну (книга / сторінка / рядок) і змістовну (твір / частина / абзац / речення / слово), які не завжди збігаються. Наприклад, речення або окреме слово можуть починатися на одному рядку або сторінці і закінчуватися на інших. Однак елементи S/XML можуть тільки «вкладатися» один в одного, але не «перетинатися». Іншими словами, ми не можемо почати елемент «речення» всередині одного елемента «сторінка», а закінчити – всередині іншого. Згідно з рекомендаціями TEI, ця проблема може вирішуватися двома шляхами:

1) використання «порожніх елементів» або «міток» (англ. milestone). Порожній елемент не має вмісту, його початковий тег одночасно є кінцевим. Так, наприклад, замість елемента «сторінка», що містить в собі текст, можна скористатися «міткою» «розриву сторінки», який можна помістити всередину абзацу та навіть слова (якщо слово містить перенесення зі сторінки на сторінку). У XML такі елементи позначаються тегами зі скісною рискою перед кінцевою кутовою дужкою (наприклад, на початку третьої сторінки може бути поставлена «мітка» <pb n = "3" />);

2) зв'язок частин «розірваного» елемента за допомогою спеціальних атрибутів. Подібне рішення може, наприклад, використовуватися при розмітці прямої мови, що розповсюджується на кілька абзаців. Практична реалізація обох запропонованих рішень пов'язана з низкою складнощів і обмежень, тому при розробці концепції корпусу слід уважно зважити всі «за» і «проти» прийняття тієї чи тієї системи розмітки пересічних структур тексту.

2. Ще одна проблема, пов'язана з використанням стандарту XML, – це його громіздкість. Чим більш детальною стає розмітка тексту, тим складніше

виявляється його читання за допомогою простого текстового редактора. Якщо кінцевий користувач буде працювати з текстом, поданим в оптимальній для нього формі за допомогою стильових листків, то укладачі корпусу повинні бачити безпосередньо вихідний код з тегами. Для полегшення і прискорення їх роботи можна скористатися двома методами:

1) використання спрощеного коду на стадії введення даних. Наприклад, можна використовувати спеціальні символи або скорочені позначення замість елементів XML на стадії введення даних. Первинну розмітку документа можна проводити в стандарті SGML, а потім скористатися спеціальними програмами для автоматичного перетворення SGML на XML. Цей метод слід використовувати обережно, оскільки він підвищує ризик помилок при подальшій обробці;

2) використання спеціального програмного забезпечення («редакторів XML»). Різні фірми-розробники пропонують цілий ряд таких редакторів, кожен із яких має свої сильні і слабкі сторони. Деякі з них (особливо безкоштовні) виявляються складними у використанні для людей, які не є професійними програмістами, інші коштують досить дорого і при цьому пропонують безліч функцій, яких укладачі корпусів не потребують. Вибір оптимального програмного забезпечення для роботи з XML є окремою важливою проблемою, яку слід вирішити на початку роботи над проєктом корпусу.

Рекомендації TEI не містять повного списку атрибутів і їх значень, які слід використовувати в лінгвістичній розмітці. Це цілком зрозуміло, оскільки багато питань класифікації слів і граматичних категорій, не кажучи вже про синтаксичні конструкції, залишаються до нині дискусійними.

Як уже зазначалося, найбільш затребуваною в різних дослідженнях і найменш залежною від теоретичних розбіжностей між різними лінгвістичними школами може бути елементарна морфологічна розмітка. При складанні протоколу розмітки слід чітко експлікувати зроблений у кожному спірному випадку вибір. Бажано використовувати якомога більш детальну

класифікацію, тому що при подальшій експлуатації корпусу значно легше «нейтралізувати» надмірно детальні опозиції, ніж згодом розмежовувати те, на що не зверталась увага раніше.

Найбільш адекватним методом морфологічної розмітки словоформ у форматі XML є уявлення кожної словоформи як окремого елемента (<w>, згідно з рекомендаціями TEI) і використання атрибутів для запису аналітичної інформації.

Автори монографії «Корпусна лінгвістика» [20, с. 51-53] наголошують на важливості таких критеріїв застосування стандарту:

1) достатність: набір структурних елементів повинен бути достатньо широким, щоб забезпечити хоча б більшість вимог. Водночас бажано, щоб схема розмітки не містила надлишкову інформацію;

2) несуперечливість: схема розмітки має бути сформована на базі несуперечливих правил, які б дозволяли однозначно визначити, які об'єкти належать до тегів, які – до атрибутів, що є вмістом тега тощо;

3) відтворюваність: схема кодування повинна ґрунтуватися на чітко визначених правилах, що дає можливість відтворити вихідний текст за допомогою простих алгоритмів;

4) коректність: за допомогою спеціального програмного забезпечення відбувається перевірка відповідності міток у документах їх структурним специфікаціям;

5) можливість збору даних: збір даних включає безпосереднє накопичення даних (за допомогою ручного вводу або з використанням автоматичного розпізнання тексту) та проведенням кодування даних;

6) технологічність: урахування потреб, пов'язаних з автоматичною обробкою текстів (вибір тексту згідно зі встановленими критеріями, використання спеціальних механізмів, типу міжтекстових покажчиків, поєднання текстів або інших елементів корпусу) тощо;

7) можливість масштабування: важливо, щоб будь-яка створена схема мала можливість поповнюватися.

Висновок до розділу 2

Корпусний аналіз вирізняється низкою характерних ознак: 1) емпіричний підхід до аналізу мовних даних (досліджуються реальні моделі мовної реалізації у природних текстах); 2) використання великих за обсягом, структурованих колекцій природних текстів (корпусів) як основи для аналізу; 3) широке залучення комп'ютерних технологій для дослідження лінгвального матеріалу; 4) застосування кваліативних і квантитативних аналітичних методик, з суттєвою перевагою останніх (вивчення частоти вживання лінгвістичних одиниць, статистичні дослідження сполучуваності і т.ін.). Отримані в результаті корпусного аналізу дані не тільки сприяють формулюванню якісно нових висновків про мову, але й окреслюють такі напрями досліджень, які до появи корпусів не привертали уваги дослідників.

Під час створення корпусів текстів слід орієнтуватися на міжнародні стандарти та рекомендації, покликані забезпечити збереження і доступність даних незалежно від зміни технологій і програмного забезпечення.

РОЗДІЛ 3
ІНДИВІДУАЛЬНО-АВТОРСЬКИЙ ЕКСПЕРИМЕНТАЛЬНИЙ НАУКОВО-ДОСЛІДНИЙ КОРПУС ТАРАСА ПРОХАСЬКА

3.1 Тарас Прохасько як лінгвоперсона

Тарас Прохасько – сучасний український письменник, журналіст, один із представників станіславського феномену. Член Асоціації українських письменників. Автор окремих прозових творів та збірок: «Інші дні Анни» (1998), «FM «Галичина» (2001), «НепрОсті» (2002), «Лексикон таємних знань» (2004), «З цього можна зробити кілька оповідань» (2005), «Порт Франківськ» (2006), «Нічний аеропорт» (2010), «БотакЄ» (2010), «Ознаки зрілості» (2014), «2015 рік» (2016), «Івано-Франківськ. Місто-прогулянка» (2017), «Так, але…» (2019) тощо. Його твори перекладаються іноземними мовами, зокрема польською, англійською та російською.

Тарас Прохасько має відзнаки та нагородження: лауреат видавництва «Смолоскип» (1997); перше місце у номінації «Белетристика» за книгу «З цього можна було б зробити кілька оповідань» за версією журналу «Кореспондент» (2006); третє місце у номінації «Документалістика» за книгу «Порт Франківськ» за версією журналу «Кореспондент» (2007); лауреат літературної премії імені Джозефа Конрада (2007); звання «Книга року» отримав роман «БотакЄ» (2011); лауреат премії імені Юрія Шевельова за книгу «Одної і тої самої» (2013); премією «Книга року ВВС» було відзначено дитячу книжку письменника «Хто зробить сніг», створену разом з Мар'яною Прохасько (2013), а у 2019 році у номінації «Есеїстика» – його книгу есеїв «Так, але...». У 2020 році став лауреатом Шевченківської премії.

Поетика творів письменника є яскраво постмодерністською. Абсолютно чужі прозаїкові епатажність, зверхньо-безапеляційний стиль висловлювання.

Його проза – риторичний апокаліпсис, в якому міститься дискурс про спасіння. У світі Тараса Прохаська просто немає місця для боротьби за першість під сонцем, бо сонця вистачить для всіх: рослини, тварини і люди у його прозі рівноправні й рівновеликі. Адже письменник, за власним зізнанням, безкінечно вірить у Творця: все – Його творіння, все знаходиться під Його опікою, на все є воля Божа, тому щасливими є лише вдячні Богу люди. З огляду на такий ідеалізм критика й читачі сприймають автора як «мандрівного філософа», якому завжди добре, який вміє жити несуєтно, не відчуває до людей агресії, навпаки – завжди чекає від них ліпшого.

Усі дослідники творчості Тараса Прохаська сходяться на тому, що ідіостиль письменника вирізняється ускладненістю організації художніх текстів. У колі філологів стало звичкою називати Т. Прохаська «наскрізь рослинним чоловіком», і це пов'язано не стільки з тим, що письменник за фахом є ботаніком, а передусім з тим, що уся його творчість просякнена рослинними соками. На лексичному рівні організації його творів це виявляється в постійній присутності назв рослин та їх частин. Це додає текстам своєрідного шарму. Герої письменника, ким би вони не працювали, чим би не займалися, обов'язково розуміються на рослинах, збирають гербарії чи доглядають за вазонками [38, с. 127].

Зосередженість на деталях закономірно призводить до домінування в текстах описів. Описів завнішності персонажів у Тараса Прохаська майже немає, зазвичай автор обмежується описом їхнього одягу, однак описи інтер'єрів та пейзажів виписані до подробиць. Каталогізація – не просто стилістичний авторський прийом, це – своєрідна філософія. Ще одна особливість описів письменника – їхня кінематографічність. Незважаючи на те, що персонажі уважно сприймають дійсність, їхнє мислення та сприйняття оточуючого є фрагментарним. Звідси – своєрідна монтажність в описах: чергуються плани, виокремлюються деталі, просторові ракурси тощо. Авторська зосередженість на найдрібніших нюансах нерідко призводить до семантичної надлишковості. Різноманітні повтори, хіазми, ампліфікації та

інші ускладнення структури речення покликані донести до читача естетичні засади феноменологічної творчості письменника. Така синтаксична організація на рівні цілого тексту ускладнює сприйняття, змушує читача читати повільно, щоразу повертатися до початку речення чи абзацу, примушує відчути твір, вхопити «відчуття присутності», за допомогою «чуття» наблизитися до «сутності».

Тарас Прохасько – виразно регіональний письменник, творець малого наративу, приватного епосу, прозаїк, що фіксує емоційні й психічні стани, в поетиці якого домінує деталь. Належачи до письменників-естетів, він не відсторонюється від читача, навпаки – прагне зворушити його, подарувати насолоду від читання, вберегти від страхів та розпачу, внести в його буденне життя елемент цікавості й несподіванки. Без дидактизму, моралізаторства, критичного пафосу й патетики автор нагадує людині про те, що світ задуманий не нею, що людина, як і світ, – це задум Божий.

3.2 Наповнення індивідуально-авторського експериментального науково-дослідного корпусу Тараса Прохаська

Дослідники наголошують на залежності побудови корпусу від зовнішніх чинників, тобто зміст корпусу необхідно відбирати незалежно від його мови. Метатекстова розмітка повинна включати кілька груп за п’ятьма параметрами, а саме:

1) текстові метадані: назва / заголовок тексту (за наявності); точна або приблизна дата створення тексту (не видання); розмір тексту (кількість слів); вказівка на автентичність тексту або перекладний варіант (у цьому випадку зазначення мови тексту оригіналу); канал (письмовий або усний); бібліографічні дані для виданих текстів (місце і рік опублікування, назва видавництва, кількість сторінок; для періодичних видань – том, номер, серія); місце і дата створення для рукописів (якщо відомо), а також форма –

написаний від руки, машинописний, електронний; місце і дата запису, якщо це усне мовлення;

2) дані про автора: повне ім’я автора / авторів (якщо відомо), псевдонім (за наявності); вік автора на момент створення тексту; стать автора (чол., жін., невідома); авторське володіння мовою (носій рідної мови, іноземної);

3) лінгвістичні метадані: зазначення мовного варіанту / діалекту / ідіолекту у випадку мультилінгвального корпусу (з мовним різноманіттям); вказівка на систему письма (різні графічні системи за умови їхніх відмінностей у текстах корпусу);

4) тематичні питання (можуть бути репрезентовані разом з функційними сферами);

5) технічні метадані.

Основними частинами загальної схеми роботи комплексу інструментів управління корпусами текстів є:

- загальний корпус текстів;
- набір інструментів пошуку і вилучення текстів (краулерів) для поповнення корпусу;
- інструменти аналізу текстів для автоматичної розмітки;
- інструмент вибору і розмітки субкорпусов. Інструмент вибору і розмітки субкорпусов дозволяють динамічно додавати різні види розмітки шляхом зміни конфігураційних файлів і додавання інструментів аналізу текстів – розмічати утиліт, на вхід яких у заданому форматі передаються дані про субкорпус для розмітки і необхідних характеристиках.

Як спосіб зберігання даних вибрано використання хмарного рішення Яндекс-диск, що дозволяє, з одного боку мати організований спільний доступ до повного наявного корпусу, з іншого – для користувача – отримувати необхідні субкорпуси без додаткового навантаження на комплекс інструментів загалом.

Інструменти пошуку і вилучення текстів (краулери) служать для збору інформації з веб-ресурсу, на який вони налаштовані, за рядом умов, заданих

користувачем в файлі конфігурації. Важливим критерієм вибору ресурсів – джерел текстів, є їх достовірність. Для наповнення і подальшого розширення корпусу підготовлені інструменти, що дозволяють в автоматичному режимі завантажувати тексти творів Тараса Прохаська. Після отримання інформації тексти конвертуються в XML-документ із розширенням розміткою, що містить інформацію про завантажений матеріал і його вихідний текст. Для збору даних і конфігурування краулера необхідно провести аналіз ресурсу і визначити, чи є фільтри за жанрами чи іншими критеріями, чи потрібні різного типу перевірки, наприклад, чи можливо завантажити текст, чи повна версія роботи представлена, визначення мови тексту тощо. Всі ці можливості включаються в конфігураційний файл інструменту наповнення корпусу. Часто тексти представлені в форматі PDF (Portable Document Format), тому для збереження тексту і реалізації розмітки корпусу в комплексі інструментів реалізовано можливість отримання з них текстів для подальшої обробки та перетворення на XML-файл.

Корпус текстів Тараса Прохаська створений за допомогою вільного корпусного менеджера NoSketch Engine (https://www.sketchengine.eu/), розробленого в університеті Масарика (Брно, Чехія). Цей корпус разом з іншими доступний на сервері кафедри загального та прикладного мовознавства і слов'янської філології ДонНУ імені Василя Стуса за адресою corpora.donnu.edu.ua.

Українськомовна частина корпусу на сьогодні охоплює 54 документи українською мовою, містить блько мільйона токенів.

У створеному корпусі використано екстралінгвістичну й лінгвістичну розмітку. Екстралінгвістична розмітка корпусу об'єднує:

1) метатекстові дані:

- поле area (сфера) з можливими значеннями літературознавство, мовознавство, загальні праці;
- author (автор) - у цьому корпусі тільки Юрій Шевельов, створено для можливості об'єднання з іншими корпусами;

- genre (жанр): есе, стаття, монографія, інтерв'ю, передмова, виступ, доповідь, спогади, стаття, виступ, стаття, доповідь, лист, вступне слово;
- пате (назва твору),
- source (джерело),
- style (стиль): публіцистичний, науковий та епістолярний;
- type (тип): мовою оригіналу, переклад з англійської, переклад з німецької, переклад з французької;
- • year (рік);

2) структурну розмітку.

Корпус містить дані: про межі тексту в тегах; про межі абзацу в тегах ; межі речень у тегах; спеціальний тег позначає розділові знаки, які не відокремлюються пробілом від попереднього токена.

Лінгвістична розмітка аналізованого корпусу текстів Тараса Прохаська сьогодні є результатом автоматичного морфологічного аналізу й лематизації, здійснених за допомогою авторських інструментів. Структура стандартного тегу до кожного токена є такою:

1) на першій позиції стоїть позначка граматичного класу слова,

2) далі – позначки підкласів, усі позначки – односимвольні латиницею або цифрами, за кожним підкласом закріплена позиція, яка не змінюється для різних класів.

У корпусному менеджері доступні типові функції, як-от побудова конкордансу на підставі простого пошуку, пошуку в лемах, пошуку фрази, словоформи, символу або певного шаблону, побудованого за допомогою регулярного виразу.

Запит може спиратися на додатковий пошук у контексті з фільтруванням потрібних лем чи словоформ на відстані до 15 токенів управо чи вліво від головного слова.

Нарешті, пошук можна обмежити різними типами текстів, передбаченими екстралінгвістичною розміткою. У побудованому конкордансі можливі різні типи сортування та фільтрування даних, частотний аналіз

морфологічних позначок чи словоформ для леми тощо. Іншою типовою функцією корпусного менеджера є частотний аналіз із можливістю вибору мінімальної чи максимальної частоти, частиномовних фільтрів, ІЧ-грамів.

Інструмент частотного аналізу дає змогу, крім того, вибрати всі леми та всі словоформи.

Висновок до розділу 3

Комплекс інструментів управління корпусами текстів, що включає в себе інструменти автоматичного наповнення, управління, додаткової розмітки і отримання субкорпусов, здатний істотно полегшити рішення задач корпусної лінгвістики і підготовку даних для розробки нових алгоритмів і інструментів комп'ютерної лінгвістики. Основною відмінною рисою комплексу інструментів є можливість формування власних субкорпусів і додавання додаткової розмітки для обраного субкорпусу. Широкі можливості з автоматизованого наповнення корпусу текстами за рахунок конфігурації інструменту пошуку текстів дозволяють застосовувати корпус у задачах машинного навчання, наприклад класифікації текстів за різними критеріями, де для підвищення точності роботи алгоритмів, потрібна підготовка великої навчальної вибірки за певними критеріями.

Корпус текстів Тараса Прохаська створений за допомогою вільного корпусного менеджера NoSketch Engine (https://www.sketchengine.eu/), розробленого в університеті Масарика (Брно, Чехія). Цей корпус разом з іншими доступний на сервері кафедри загального та прикладного мовознавства і слов'янської філології ДонНУ імені Василя Стуса за адресою corpora.donnu.edu.ua.

Метатекстова розмітка допоможе контролювати процес наповнення корпусу новими текстовими даними та оцінювати баланс корпусу. Для цілей нашого дослідження обрані лише готові автентичні письмові тексти творів Тараса Прохаська.

ВИСНОВКИ

Одним із пріоритетних напрямків сучасних прикладних лінгвістичних досліджень є корпусна лінгвістика. Корпусні студії зосереджуються на аналізі природної мови в умовах реального функціювання з використанням комп'ютерних технологій на основі великих за обсягом, ретельно відібраних та впорядкованих текстових корпусів.

Корпусна лінгвістика пропонує нові можливості для численних інтегрованих з лінгвістикою дисциплін, зокрема психолінгвістики, соціолінгвістики та історичної лінгвістики.

Аналіз наведених у першому розділі дефініцій терміна «корпус текстів» дозволяє виділити низку ознак, що відрізняють сучасний корпус текстів від звичайних колекцій текстів в електронній формі (електронних бібліотек, архівів): репрезентативність, автентичність, відібраність, збалансованість, машиночитаність.

Технологічний процес створення корпусу текстів передбачає поступову реалізацію восьми основних етапів: визначення джерел лінгвістичного матеріалу, введення даних, попереднє опрацювання тексту, конвертування й графематичний аналіз, розмітка тексту, коректування результатів автоматичної розмітки, конвертування розмічених текстів у структуру спеціалізованої лінгвістичної інформаційно-пошукової системи, забезпечення доступу до корпусу. Зауважимо, що у кожному конкретному випадку склад і кількість процедур можуть відрізнятися, а реальна технологія може виявитися набагато складнішою. Під час створення корпусу використовується низка процедур і програм, як-от: токенізація, лематизація, стеммінг, парсинг.

Стандарт TEI забезпечує оптимальну збалансованість між загальною моделлю подання природної мови і нескладною реалізацією кодування. Також TEI оперує великим набором засобів для подання як лінгвальної, так і металінгвальної інформації. Передумовою розроблення системи TEI стало існування великої кількості несумісних систем кодування і розширення сфери

застосування електронних текстів. Базовими принципами системи визначено: а) можливість досягати у тексті ефектів, необхідних для наукових досліджень різного типу; б) простота, чіткість і конкретність; в) нескладність для використання без спеціалізованого програмного забезпечення; г) можливість точного визначення та ефективного програмного оброблення текстів; г) можливість розширень, визначених користувачем; д) узгодженість із чинними і новостворюваними стандартами.

Незважаючи на те, що рекомендації TEI складаються з метою врахувати потреби всіх можливих видів корпусів текстів і електронних видань, на практиці нерідко виявляється, що ці рекомендації або не містять потрібних елементів, або властивості пропонованих елементів не відповідають вимогам укладачів корпусу. Слід зазначити, що TEI постійно веде роботу з удосконалення рекомендацій, і можливість відхилення від рекомендованої схеми не виключається самими розробниками.

Створений корпус Тараса Прохаська є дослідницьким, повнотекстовим і динамічним, має екстралінгвістичну та лінгвістичну розмітку. З-поміж важливих функцій варто виокремити: а) побудову конкордансу на підставі простого пошуку, пошуку в лемах, пошуку фрази, словоформи, символу або певного шаблону, створеного за допомогою регулярного виразу; б) частотний аналіз для словоформ, лем та тегів.

Розвиток корпусних досліджень буде вести до формування відкритої лінгвістичної спільноти, що вільно обмінюється даними і активно розвиває спільні і взаємодоповнюючі дослідні проєкти.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Білозуб А.І. Лексико-семантичні прийоми мовної гри в українському постмодерному тексті [на прикладі творів Ю. Андруховича, Ю. Іздрика, Т. Прохаська та В. Єшкілєва]. *Дослідження з лексикології і граматики української мови*. 2012. Вип. 12. С. 124-132.
2. Бобкова Т.В. Корпус текстів: основні аспекти визначення. *Науковий вісник кафедри Юнеско КНЛУ. Серія Філологія. Педагогіка. Психологія*. Вип. 29. 2014. С. 11-20.
3. Богданова С.Ю. Возможности корпусной методологии в решении лингвистических задач. *Вестник Иркутского государственного лингвистического университета*. 2012. №2 (18). С. 47-53.
4. Бук С. Сучасні методи дослідження мови письменника у слов'янознавстві. *Проблеми слов'янознавства*. 2012. Вип. 61. С. 86-95.
5. Гладкова А.П. Модель анотації текстового корпусу як засіб дослідження художньої картини світу. Studia Linguistica. Вип. 4. К. : КНУ, 2010. С. 524-528.
6. Гребенюк Т.В. Перцептивне заглиблення в повсякденність як домінанта ідіостилю Т. Прохаська. *Актуальні проблеми слов'янської філології. Серія: Лінгвістика і літературознавство* : міжвуз. зб. наук. ст. Бердянськ : БДПУ, 2012. Вип. XXVI, ч. 3. С. 28-35.
7. Данилюк І. Корпус текстів для вивчення граматичної службовості. *Лінгвістичні студії. Лінгвістичні студії* : зб. наук. праць. Донецьк : ДонНУ, 2013. Вип. 26. С. 224-229.
8. Данилюк І., Загнітко А., Ситар Г. Корпус текстів Юрія Шевельова: структура, функції, навігація. *Мова: класичне – модерне – постмодерне*. 2019. Вип. 5. С. 158-169. doi: 10.18523/1стр2522-9281.2019.5.158-169.
9. Дарчук Н.П. Дослідницький корпус української мови: основні засади і перспективи. *Вісник Київського національного університету імені Тараса Шевченка. Серія: Літературознавство. Мовознавство.*

Фольклористика. К. : Видавничо-поліграфічний центр «Київський університет», 2010. №21. С. 45-49.

10. Демська-Кульчицька О. Дещо про класифікацію текстових корпусів. *Наукові записки. Серія: Мовознавство*. 2004. 1(11). С. 153-157.
11. Демська-Кульчицька О. Основи національного корпусу української мови : монографія. Київ, 2005. 218 с.
12. Демська-Кульчицька О.М. Базові поняття корпусної лінгвістики. *Українська мова*. 2003. №1. С. 42-47.
13. Жуковська В.В. Вступ до корпусної лінгвістики : навчальний посібник. Житомир: Вид-во ЖДУ ім. І. Франка, 2013. 142 с.
14. Жуковська В.В. Ресурси корпусної лінгвістики у дослідженні історичної динаміки мови. *Матеріали міжнародної наукової конференції «Слово і речення: синтактика, семантика, прагматика»*. К. : Київ. унт-т ім. Б. Грінченка, 2013. С. 151-156.
15. Загнітко А.П., Данилюк І.Г. Корпус текстів граматичної службовості. *Прикладна лінгвістка та лінгвістичні технології: MegaLing-2012*. Київ : УМІФ, 2013. С. 102-112.
16. Захаров В.П. Корпусная лингвистика : учебно-методическое пособие. СПб. : РОПИ СПб. ун-та, 2005. 48 с.
17. Захаров В.П., Богданова С.Ю. Корпусная лингвистика : учебник для студентов гуманитарных вузов. Иркутск : ИГЛУ, 2011. 161 с.
18. Карпіловська Є.А. Вступ до прикладної лінгвістики: комп'ютерна лінгвістика : підручник. Донецьк : ТОВ «Юго-Восток, Лтд», 2006. 188 с.
19. Коломієць В., Орел В. Корпус анотацій наукових статей із комп'ютерної лінгвістики. *Комп'ютерна лінгвістика: сучасне і майбутнє : матеріали Міжнародної науково-практичної конференції*. К. : Вид. центр КНЛУ, 2012. С. 32-34.
20. Корпусна лінгвістика / В.А. Широков, О.В. Бугаков, Т.О. Грязнухіна та ін. К. : Довіра, 2005. 471 с.

21. Кошка Н.І. Огляд розвитку та стану корпусної лінгвістики в історичному аспекті. *Наукові записки Національного університету «Острозька академія». Сер.: Філологічна*. 2009. Вип. 11. С. 247-252.
22. Кривенко Г.Л. Корпусні дослідження дискурсу: становлення, стан і перспективи. *Вісник КНЛУ. Серія Філологія*. Том 20. №1. 2017. С. 51-63.
23. Кульчицький І. Корпуси текстів як лінгвотехнологічне підґрунтя виявлення змін в українській мові. *ХХ–ХХІ століття: жанрово-стильові й лінгвістичні метаморфози в українській мові та літературі* : монографія / А. Архангельська (голов. ред.), О. Левченко, О. Тищенко [та ін.]. Оломоуць: Ун-т ім. Ф. Палацького, 2016. С. 269-298.
24. Кульчицький І.М., Данчевська Ю.О. Деякі аспекти створення корпусу художніх творів В.С. Стефаника. *Прикладна лінгвістика та лінгвістичні технології. Megaling 2012* : зб. наук. праць. Київ, 2013. С. 143-148.
25. Кульчицький І., Ліхнякевич І., Лотоцька Н. Структурна модель корпусу творів Романа Іваничука. *Науковий вісник Східноєвропейського національного університету імені Лесі Українки*. 2017. № 3. С. 421-430.
26. Лебедєв К. Створення Багатомовного корпусу паралельних текстів. *Комп'ютерна лінгвістика: сучасне і майбутнє : матеріали Міжнародної науково-практичної конференції*. К. : Вид. центр КНЛУ, 2012. С. 36–37.
27. Лендау С.І. Словники: мистецтво та ремесло лексикографії. К. : К.І.С., 2012. 480 с.
28. Лук'янець Г.Г. Основні напрямки сучасних корпусних досліджень мови та перспективи їх подальшого розвитку. *Наукові праці НУХТ*. 2012. №44. С. 127-133.
29. Лучик А., Остапова І., Синтагматична параметризація еквівалентів слова у парадигмі корпусної лінгвістики *Human. Computer. Comunication, (20-22 September, Lviv)* 2017. С. 33-37.

30. Максимів О. Корпус перської мови як джерело матеріалу для навчальних словників-мінімумів. *Вісник Львівського університету. Серія філологічна*. №45. С. 164-169.
31. Мейзерська І.В. Корпусний підхід у сучасній лінгвістиці: перспективи і можливості застосування. *Науковий вісник кафедри Юнеско КНЛУ. Серія Філологія. Педагогіка. Психологія*. Вип. 28. 2014. С. 53-58.
32. Михайлюк А.Р. Алгоритм укладання корпусу текстів (на матеріалі творів Тараса Прохаська). *Tendances scientifiques de la recherché fondamentale et appliquee:* collection de papiers scientifiques «ЛОГОZ» avec des materiaux de la conference scientifique et pratique internationale (30 octobre 2020, Strasbourg, Republique Francaise). 2020. Vol. 3. Pp. 67-69.
33. Перебийніс В.І., Сорокін В.М. Традиційна та комп'ютерна лексикографія : навчальний посібник. К. : Видавничий центр КНЛУ, 2009. 218 с.
34. Плахотнікова О.Ю. Сучасний стан корпусних досліджень в Україні. *Наукові записки Національного університету "Острозька академія". Серія : Філологічна*. 2015. Вип. 56. С. 242-244.
35. Плунгян В.А. Корпус как инструмент и как идеология: о некоторых уроках современной корпусной лингвистики. *Русский язык в научном освещении*. №2 (16). 2008. С. 7-20.
36. Саєнко Н.С. Корпусний підхід у навчанні іноземних мов у технічному університеті. *Педагогічні науки: теорія, історія, інноваційні технології*. 2016. №1(55). С. 142-151.
37. Тарас Прохасько. *Сучасна українська проза: нові імена* : бібліогр. покаж. / КЗ «ЗОУНБ» ЗОР; [уклад.: М. Маслова, А. Маційвська, Ю. Щеглова]. Запоріжжя : Кругозір, 2016. С. 105-120.
38. Тепшич А.І. Роль мовної гри в реалізації феноменологічних експериментів Тараса Прохаська. *Лінгвістичні студії Linguistic Studies* . 2018. Вип. 35. С. 126-130.

39. Шведова М., Січінава Д. Корпусна лінгвістика та лексикограматична типологія. *Українське мовознавство.* №43. 2013. С. 95-103.
40. Широков В.А. Комп'ютерна лексикографія. К. : Наукова думка, 2011. 352 с.
41. Шкляревський В.Г. Метарозмітка корпусів текстів: стандарти і реалізація. *Наукові записки Національного університету Острозька академія. Серія: Філологічна*. 2014. Вип. 49. С. 300-303.
42. Aarts J., Meijs W. Corpus Linguistics: Recent developments in the Use of Computer Corpora in English Language Research. Amsterdam : Rodopi, 1984. 425 p.
43. Altenberg B., Granger S. Lexis in contrast: corpus-based approaches. Amsterdam : John Benjamins Publishing Company, 2002. 337 p.
44. Aston G., Burnard L. The BNC Handbook. Exploring the British National Corpus with SARA. Cambridge : Edinburgh University Press, 1998. 250 p.
45. Biber D. Representativeness in corpus design. *Literary and Linguistic Computing*. 1993. №8(4). P. 243-57.
46. Biber D., Conrad S., Reppen R. Corpus linguistics: Investigating language structure and use. Cambridge : Cambridge University Press, 2001. 312 p.
47. Bobkova T. Corpus of computational linguistic texts. *Computer Treatment of Slavic and East European Languages*. Bratislava : Tribun, 2009. P. 35-40.
48. Dash N.S. Corpus linguistics and language technology: with reference to Indian Languages. *Niladri Sekhar Dash*. New Dehli : Mittal Publications, 2005. 445 p.
49. Francis W. N. Language Corpora B. C. *Directions in Corpus Linguistics* / [ed J. Svartvik]. Berlin and New York : Moutin, 1992. P. 17-34.
50. Gries S.Th. Quantitative corpus linguistics with R : a practical introduction. New York; London : Routledge, 2009. 248 p.
51. Johansson S. Times change and so do corpora. *English corpus linguistics : studies in honour of J. Svartvik* / [ed A. Altenburg]. London : Longman, 1991. P. 305-314.

52. Kennedy G. Introduction to corpus linguistics. London : Longman, 1998. 315 p
53. Leech G. Introducing corpus annotation. *Corpus Annotation: Linguistic Information from Computer Text Corpora.* Addison Wesley Longman, 1997. P. 1-19.
54. Meyer Ch. P. English Corpus Linguistics. An introduction. Cambridge University Press, 2004. 168 p.
55. O'Keeffe A., McCarthy M., Carter R. From Corpus to Classroom. Cambridge: Cambridge University Press, 2007. 315 p.
56. Reppen R. Using Corpora in the Language Classroom / Randi Reppen. New York : Cambridge University Press, 2010. – 118 p.
57. Sinclair J. Corpus, Concordance, Collocation. Oxford : Oxford University Press, 1991. 170 p.
58. Stefanowitsch A., Stefan Th., Gries M. de G. Corpora in Cognitive Linguistics. Berlin, 2006. 360 p.
59. Svartvik J. Corpus Linguistics Comes of Age. *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82*. Berlin : Mouton de Gruyter, 1992. P. 5-21.
60. Svartvik J. Corpus linguistics 25 + years on. *Corpus linguistics 25 years on*. Amsterdam – New York : NY, 2007. P. 11-27
61. Zhukovska V.V. Corpus-based approach to teaching vocabulary and grammar. *XVI TESOL-Ukraine International Conference Current Studies in English «Linguistics and methodology perspectives»*. Zhytomyr, Kamianets-Podilsky, 2011. P. 171-173.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

62. Генеральный Интернет-Корпус Русского Языка [Электронный ресурс]. Режим доступу: http://www.webcorpora.ru/, вільний Заголовок з екрану. (15.01.2020)

63. Корпус текстів Івана Франка. Львів. Режим доступу : http://www.ktf.franko.lviv.ua/ ~andrij/ science/Franko/concordance.html.
64. Корпус FRANTEXT: <http://www.atilf.fr> Краткая история развития SGML: <http ://www.sgmlsource. com/history/sgmlhist.htm>
65. Машинный фонд русского языка: http://www.irlras-cfrl.rema.ru/
66. Национальный корпус русского языка [Электронный ресурс]. Режим доступу: http:// www.ruscorpora.ru/, вільний. Заголовок з екрану. (15.01.2020)
67. Український національний лінгвістичний корпус Українського мовно-інформаційного фонду НАН України [Електронний ресурс]. Режим доступу : http://lcorp.ulif.org.ua/virt_unlc/ <http://www.ilc.cnr.it/EAGLES96/pub/eagles/corpora/corpustyp.ps.gz>
68. British National Corpus [Электронный ресурс]. Режим доступа: http://www.natcorp. ox.ac.uk/, вільний – Заголовок з екрану. (20.01.2020)
69. Český národní korpus [Электронный ресурс]. Режим доступу: https://korpus.cz/, вільний – Заголовок з екрану. (15.01.2020)
70. Corpus Encoding Standard (CES): <http://www.cs.vassar.edu/CES>
71. Text Encoding Initiative (TEI): <http://www.tei-c.org>
72. World Wide Web Consortium (спецификация XML): http://www.w3aorg

СПИСОК ДЖЕРЕЛ ІЛЮСТРАТИВНОГО МАТЕРІАЛУ

73. Прохасько Т. Інші дні Анни : [проза]. Івано-Франківськ : Лілея-НВ, 2010. 95 с.
74. Прохасько Т. Лексикон таємних знань : [новели]. Харків : Фоліо, 2011. 192 с.
75. Прохасько Т. Одної і тої самої : [проза]. Чернівці : Книги – ХХІ : Міжнар. літ. корпорація «Meridian Czernowitz», 2013. 239 с.

ДОДАТКИ

Додаток А

Апробація роботи.

Копія сертифікату учасника Міжнародної наукової конференції «Tendances scientifiques de la recherché fondamentale et appliquee»

(30 octobre 2020, Strasbourg, Republique Francaise)

ДОНЕЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВАСИЛЯ СТУСА
ALMA MATER
NOMEN EST OMEN
ІМ'Я ЗОБОВ'ЯЗУЄ

Додаток Б

Апробація роботи.

Копія наукової статті «Алгоритм укладання корпусу текстів (на матеріалі творів Тараса Прохаська)»

ДОНЕЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВАСИЛЯ СТУСА
ALMA MATER
NOMEN EST OMEN
ІМ'Я ЗОБОВ'ЯЗУЄ

ДОНЕЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВАСИЛЯ СТУСА
ALMA MATER
NOMEN EST OMEN
ІМ'Я ЗОБОВ'ЯЗУЄ

ДОНЕЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВАСИЛЯ СТУСА
ALMA MATER
NOMEN EST OMEN
ІМ'Я ЗОБОВ'ЯЗУЄ

ДОНЕЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВАСИЛЯ СТУСА
ALMA MATER
NOMEN EST OMEN
ІМ'Я ЗОБОВ'ЯЗУЄ

ДОНЕЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВАСИЛЯ СТУСА
ALMA MATER
NOMEN EST OMEN
ІМ'Я ЗОБОВ'ЯЗУЄ

ДОНЕЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВАСИЛЯ СТУСА
ALMA MATER
NOMEN EST OMEN
ІМ'Я ЗОБОВ'ЯЗУЄ

ДОНЕЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВАСИЛЯ СТУСА
ALMA MATER
NOMEN EST OMEN
ІМ'Я ЗОБОВ'ЯЗУЄ

ДОНЕЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВАСИЛЯ СТУСА
ALMA MATER
NOMEN EST OMEN
ІМ'Я ЗОБОВ'ЯЗУЄ

ДОНЕЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВАСИЛЯ СТУСА
ALMA MATER
NOMEN EST OMEN
ІМ'Я ЗОБОВ'ЯЗУЄ

Додаток В

Електронний носій корпусу текстів Тараса Прохаська