МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ДОНЕЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ВАСИЛЯ СТУСА

КАРПЕНКО АЛІНА АНДРІЇВНА

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

# КОРПУС ТЕКСТІВ УКРАЇНСЬКОЇ ФАНТАСТИКИ

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,
кандидат філол. наук, доцент
_____
(підпис)

Оцінка: ____ / ___ / _____
Голова ЕК: _____

Вінниця 2020

# АНОТАЦІЯ

**Карпенко А. А.** Корпус текстів української фантастики. Спеціальність 035 «Філологія», спеціалізація 035.10 «Прикладна лінгвістика», освітня програма «Прикладна лінгвістика». Донецький національний університет імені Василя Стуса, Вінниця, 2020.

У кваліфікаційній роботі проаналізовано основні поняття корпусної лінгвістики, створено корпус текстів української фантастики за допомогою використання вільного корпусного менеджера NoSketch Engine. Розглянуто етапи створення корпусу текстів української фантастики. Виявлено сфери та актуальність застосування лінгвістичних корпусів. Встановлено корпус текстів української фантастики з екстралінгвістичною та лінгвістичною розмітками.

Ключові слова: створення корпусу, українська фантастика, корпус текстів, цифрові гуманітарні науки, представлення тексту, вибірка письменників, методологія, лематизація.

60 с., 0 табл., 9 рис., 0 дод., 57 джерел.

Табл. 0. Рис. 9. Бібліограф.: 54 найм.

**Karpenko A.** Text corpus of Ukrainian fiction. Specialty 035.10 «Applied linguistics», Programme «Applied linguistics». Vasyl’ Stus Donetsk National University, Vinnytsia, 2020.

In the qualification work, were analyzed the basic concepts of corpus linguistics, a corpus of Ukrainian fiction texts was created using the free corpus manager NoSketch Engine. The stages of creating a corpus of Ukrainian fiction texts are considered. The scope and relevance of the use of linguistic corpora are revealed. A corpus of Ukrainian science fiction texts with extralinguistic and linguistic markup has been established.

Keywords: corpora creation, Ukrainian fiction, the corpus of texts, digital humanities, text representation, selection of writers, methodology, lemmatization.

Tabl. 0 Fig. 8 Bibliography: 57

# ЗМІСТ

# ВСТУП

У різних галузях знань та науках, щоб якісно дослідити матеріал потрібен певний масив даних, аби надалі можна було його аналізувати, перевіряти теорії та гіпотези, проводити дослідження. Для збору інформації потрібні текстові корпуси – структуровані набори даних та інформації, що використовуються для вимірювання валідності, точності та ефективності інструментів, методів та систем.

Робота з корпусами текстів допомагає науковцям та дослідникам уникнути суб'єктивності в судженнях, а також надає можливість швидкого пошуку та аналізу інформації, матеріалів для досліджень. Також у сфері автоматичної обробки природної мови (АОПМ), зокрема статистичної АОПМ, існує потреба в навчанні моделі або алгоритму з великою кількістю даних. Для цього дослідники збирають безліч корпусів текстів.

Використання цих корпусів формує важливе направлення в лінгвістиці – корпусну лінгвістику. Корпусна лінгвістика як окремий розділ мовознавства остаточно сформувалась у 90-х роках XX ст. На сьогодні вона тільки пришвидшує розвиток інформаційних технологій, також дисципліна збирає навколо себе безліч досліджень, які базуються на її методах. Корпусна лінгвістика не тільки пришвидшує та оптимізує трудомісткий процес збору мовного матеріалу, але й призводить до зміни наукової парадигми в лінгвістиці. Це вже не напрямок певного ярусу мовної системи, не конкретна теорія або аспект аналізу, а вже скоріше ідеологія, згідно якої результати лінгвістичних досліджень повинні спиратись передусім на аналіз текстів (усних чи писемних), а не на інтуїцію дослідника.

Однак довільне зібрання машинних текстів природної мови ще не може називатися корпусом. Для цього тексти мають бути відібрані згідно з визначеними критеріями, відповідати певним вимогам, бути систематизованими, закодованими й організованими відповідно до вимог стандарту кодування корпусу.

В сучасній науці існують різні точки зору щодо оцінки функціональності та значущості корпусного дослідження мови. Одні дослідники визнають за корпусною лінгвістикою статус основної емпіричної лінгвістичної парадигми, інші вважають за краще користуватися корпусом виключно як джерелом прикладів для ілюстрації положень своїх теорій. Проте корпуси текстів є актуальною темою як серед дослідників, так і серед студентів та користувачів мережі Інтернет.

Двоякий характер корпусної лінгвістики (націленість, як на створення, так і на використання корпусів текстів) обумовлюється двояким характером її об'єкта - корпусу текстів, який, з одного боку, являє собою вихідний мовний матеріал для корпусної лінгвістики і для інших лінгвістичних дисциплін; з іншого боку, є результатом діяльності корпусної лінгвістики [1,10].

***Предмет корпусної лінгвістики*** – теоретичні основи і практичні механізми створення і використання представлених масивів мовних даних, які призначені для лінгвістичних досліджень в інтересах широкого кола користувачів.

Необхідно відзначити, що корпусна лінгвістика як система методів і принципів використання корпусів для дослідження мови і для вивчення/навчання мови має теоретичні основи, але теоретичні основи ще не є науковою теорією. Корпусна лінгвістика застосовується в методології вивчення мови з широкими можливостями в багатьох лінгвістичних напрямках і теоріях.

***Актуальність*** – даної випускної кваліфікаційної роботи полягає в тому, що в ній застосовуються методи корпусної лінгвістики як засобу для збору матеріалу з друкованих видань за допомогою використання вільного корпусного менеджера NoSketch Engine (https://www.sketchengine.eu/).

***Об’єкт дослідження*** – тексти творів української фантастики написаних в оригіналі українською мовою в період з 1934-х до 2018-го р.

***Предмет дослідження*** – корпус текстів з розміткою в додатку NoSketch Engine.

***Мета дослідження*** – описати етапи створення корпусу текстів української фантастики, визначити ключові напрями та подальші перспективи лінгвістичних корпусних досліджень, систематизувати знання та досвід дослідників та лінгвістів.

***Завдання*** – розглянути основні поняття корпусної лінгвістики, виявити сфери та актуальність застосування лінгвістичних корпусів, створити корпус текстів української фантастики з екстралінгвістичною та лінгвістичною розмітками тексту.

***Методи дослідження*** – методи корпусного дослідження.

***Наукова новизна дослідження та практичне значення отриманих результатів*** – електронний лінгвістичний корпус конкретного жанру дає змогу використовувати корпус, як для навчання під час розв’язання інших задач аналізу тексту, так і для автоматизації перевірки отриманих результатів при дослідженні різних методів корпусної лінгвістики. Дані в корпусі також знаходяться у своїй природній контекстній формі, то ж надалі є можливість їх всебічного та об’єктивного вивчення. Історично та стилістично репрезентативний об’єм корпусу текстів гарантує типовість представлення мовної інформації.

Для реалізації мети та поставлених задач у процесі роботи були використані загально логічні та гіпотетико-дедуктивні та індуктивні методи. Для унаочнення матеріалу були використані рисунки.

***Апробація результатів дослідження*** – публікація у міжнародній науковій конференції «Науковий простір: актуальні питання, досягнення та інновації», тема роботи: «Загальні вимоги щодо проєктування корпусів текстів», науковий керівник: Данилюк Ілля Григорович, канд. філол. наук, докторант, доцент кафедри загального та прикладного мовознавства і слов'янської філології, Донецький національний університет імені Василя Стуса.

Робота складається зі вступу, трьох розділів, списку використаної літератури ( 00 джерел) та висновків. Обсяг роботи – 60 сторінок

У першому розділі «Аналіз предметної області: корпусна лінгвістика в системі мовознавчих наук» розкрито теоретичне та практичне значення корпусної лінгвістики, схарактеризовано визначення корпусної лінгвістики та корпусу тексту, обґрунтовано власні думки та думки вчених щодо концепції Корпусної лінгвістики та самостійність дисципліни, описано передумови виникнення корпусної лінгвістики, описано технології та основні напрямки даної лінгвістичної дисципліни та поєднання її з комп'ютерною лінгвістикою.

У другому розділі «Міжнародні стандарти та вимоги щодо проєктування корпусів текстів» визначено етапи створення корпусу, засоби лінгвістичної розмітки, розглянуто проблеми повноти та репрезентативності корпусу, типи корпусів та їх застосування у науці, описано міжнародні стандарти для представлення корпусу текстів, описано принцип роботи інструментів корпусів та роботу корпусної пошукової систем Sketch Engine та NoSketch Engine.

У третьому розділі «Створення корпусу текстів української фантастики» акцентовано увагу на методологію та практичну роботу над корпусом текстів української фантастики.

# РОЗДІЛ 1. Аналіз предметної області: корпусна лінгвістика в системі мовознавчих наук

## 1.1. Історія становлення та основи сучасної корпусної лінгвістики

Корпусна лінгвістика - це багатовимірна область з широким спектром, що охоплює всю різноманітність використання мови в усіх сферах мовної взаємодії, спілкування і розуміння. Корпусна лінгвістика - «розділ комп'ютерної лінгвістики, що займається розробкою загальних принципів побудови і використання лінгвістичних корпусів (корпусів текстів) із застосуванням комп'ютерних технологій. Сьогодні корпусна лінгвістика часто розуміється як відносно новий підхід в лінгвістиці, який має справу з вивченням використання мови в «реальному житті» за допомогою комп'ютерів і електронних ресурсів, що є одним із завдань корпусної лінгвістики». [1, 7].

Як зауважують Т. Макінері та А. Вільсон [2, 2-4], корпусні методи виникли задовго до появи електронних корпусів. Введення корпусу в вивчення і застосування мов внесло новий вимір в мовознавство. Сфера корпусної лінгвістики зазвичай розглядається як новий підхід до лінгвістики, який розвинувся і став популярним за останні сорок років - з моменту появи комп'ютерів. Однак, як і щодо всіх нових областей, його коріння лежить в більш ранніх формах дисципліни.

В принципі, корпусна лінгвістика - це підхід, який спрямований на дослідження мови та всіх її властивостей шляхом аналізу великих колекцій зразків текстів. Даний метод роботи століттями використовувався в ряді дослідницьких областей: від описового вивчення мови до мовної освіти, лексикографії тощо. Це в цілому стосується вичерпного аналізу будь-якої значної кількості автентичних, усних та/або написаних зразків тексту. Загалом, він охоплює великий обсяг машинозчитуваних даних про фактичне використання мови, що включає колекції зразків літературних та

нелітературних текстів для відображення як синхронних, так і діахронічних аспектів мови.

Серед теоретиків достатньо відкритих дискусій про те, що таке корпусна лінгвістика або чим вона повинна бути. Однак в той самий час випадковому спостерігачеві або новачкові може здатися що існує вражаюче розмаїття визначень і описів. Аарц, один із батьків-засновників корпусної лінгвістики, схоже, передбачав це.

Сам термін «корпусна лінгвістика» міцно ввійшов до наукового вжитку лише в останні десятиліття XX століття з публікацією у 1984 році збірника наукових праць «Corpus Linguisitcs: Recent Developments in the Use of Computer Corpora in English Language Research» за матеріалами конференції ICAME “Conference on the Use of Computer Corpora in English Language Research”. [3,8] авторства Aarts and Meijs (1984). [4] Хоча цей термін фактично використовувався раніше, наприклад, в Aarts and van den Heuvel (1982). [5]

У «Списку корпусів» Аарц, як повідомляється, коментує, що цей термін був придуманий з деякою невпевненістю, «тому що ми думали (і я все ще думаю), що це не дуже хороша назва: це дивна дисципліна, яку назвали за іменем її основного дослідницького інструменту і джерел даних. Можливо, цей термін до теперішнього часу віджив своє». [6] Це викликає одну з періодичних проблем щодо розмов про корпусну лінгвістику і, можливо, пояснює перевагу альтернатив.

В 1992 році Ліч стверджував, що «комп'ютерна корпусна лінгвістика визначає не просто нову методологію дослідження мови, а насправді новий філософський підхід до предмету», і продовжує описувати характеристики комп'ютерної лінгвістики корпусу як нову парадигму [7, 106] Подібним чином Стаббс 1993 [8] відкидає обмежене визначення корпусної лінгвістики як методології і, коментуючи Сінклера 1991 року, відзначає, що «в цьому баченні предмета корпус - це не просто інструмент лінгвістичного аналізу, а й важливе поняття в лінгвістичній теорії» [8, 23-24]. Teubert (2005) також

підкреслює теоретичну концептуалізацію і описує корпусну лінгвістику як «теоретичний підхід для опанування мови». [9,2]

На даний час сфера корпусної лінгвістики вважається розділом мовознавства, який значною мірою розвинувся з початку 1960-х років, коли у 1964 році вчені Генрі Кучера та В. Нельсон Френсіс з факультету лінгвістики Браунівського університету публікують машиночитаємий загальний корпус, аби допомогти лінгвістичним дослідженням сучасної англійської мови. Корпус нараховував 1 млн слів. В ньому було вміщено 500 текстів 15 найпопулярніших жанрів англомовної прози США по 2 000 слів в кожному. До корпусу додавались вказівник частотності та алфавітно-частотний указчик, а також деякі статистичні розподілення. Оновлені та переглянуті видання з’являються у 1971 та 1979 роках. У 1999-му році корпус з анотаціями був включений у Treebank-3.

Поява Браунівського корпусу викликала загальний інтерес і жваві дискусії. Перш за все вони торкнулися принципів відбору текстів і складу потенційно розв'язуваних на такому корпусі завдань.

Консорціум лінгвістичних даних (LDC) був створений у 1992 році для того, щоб слугувати сховищем для ресурсів автоматичної обробки природної мови. Він розміщений в Пенсільванському університеті.

В період з 1991 по 1994 рік був зібраний 100-мільйонний корпус британської англійської мови під назвою BNC (Британський Національний Корпус). Варто зазначити, що корпус збалансований за жанрами. У 2014-му стартував додатковий проєкт BNC2014, який в подальшому допоможе зрозуміти як розвивається мова. Розмовний BNC2014 був випущений в вересні 2017 року, а письмовий у 2019-му.

1999-го відбувся реліз Penn Treebank-3. Корпус заснований на оригінальному Treebank (1992) та його переробленій версії Treebank II (1995). Робота над ним почалась ще в 1989 році в Пенсільванському університеті. Новий Treebank-3 вміщує Браунівський корпус, який пройшов попередній синтаксичний аналіз та був протегований, а також 1 млн слів з матеріалу WSJ

1989 року, який також був анотований в стилі Treebank II, протегований зразок ATIS-3, також розмічений та синтаксично проаналізований корпус Switchboard. Окрім POS тегів, корпус вміщує теги фрагментів, відношень та прив'язки. BLLIP 1987-89 WSJ Corpus Release 1 вміщує 30 млн слів та доповнює WSJ розділ у Treebank-3.

Корпус сучасної американської мови (COCA), який був в процесі збору матеріалу протягом 1990-2007 років, зібрав в собі 365 млн слів та у 2008-му презентував користувачам свої здобутки. У грудні 2017 року він збільшив колекцію до 560-ти млн слів, і подалі кожного року до збірки додається 20 млн слів. Корпус показує хороший баланс усної, художньої літератури, популярних журналів, газет та академічних текстів. Було відмічено що COCA вміщує багато загальних слів, які відсутні в Американському національному корпусі (ANC), корпусі з 22-х млн слів.

Протягом 1980-90-х років українські дослідники у відділі структурно-математичної лінгвістики Інституту мовознавства НАНУ створювали системи логіко-семантичного, синтаксичного та морфологічного аналізу текстів українською та російською мовами, згодом вони були закладені в основу української версії Windows. З кінця 90-х років проводиться робота над розробкою Українського національного лінгвістичного корпусу, але він не забезпечений вільним доступом. В мережі Інтернет розміщено анотований Корпус української мови обсягом понад 13 млн слововживань, він був укладений дослідниками під керівництвом Н. Дарчук у лабораторії комп'ютерної лінгвістики Київського університету.

Проект з розвитку паралельних українсько-російських та російсько-українських корпусів розробляється з 2010 р. спільним колективом українських і російських учасників, які представляють різні науково-дослідні організації (Інститут української мови та Інститут мовознавства НАНУ, Київський лінгвістичний університет, Інститут російської мови та Інститут мовознавства РАН). [10]

Консорціумом лінгвістичних даних (LDC) було випущено 5-те видання англійського Gigaword в червні 2011-го року. Дані для корпусу поступають з 7-ми англійських стрічок новин. Він вміщує 4 млрд слів та займає 26 ГБ без стиснення. Перше видання з'явилось ще в 2003 році. В листопаді 2012-го року дослідники з університету Джона Хопкінса додали в цей корпус синтаксичні анотації та анотації структури дискурсу після розбору понад 183 млн речень.

В липні 2012-го року на базі оцифрованих книг Google випускає 2-гу версію Google Books Ngrams. Перша версія була випущена ще в липні 2009-го року. Були вміщені тільки ті n-грами, які частотно перевищують 40 вживань в мові. Корпус включає від 1-грама до 5-грам, а також багато неанглійських мов. Для експериментів з невеликими наборами фраз, дослідники можуть спробувати онлайн версію застосунку Google Books Ngram Viewer.

Консорціум лінгвістичних даних (LDC) випустив English Web Treebank (EWT), як корпус для неформальних жанрів, у серпні 2012-го. Він вміщує контент з веб-журналів, оглядів, відгуків, питань-відповідей, груп новин та електронної пошти, разом з електронними листами корпорації Enron за 1999-2002 роки. Ця інформація загалом має 250 тисяч токенів рівня слів та 16 тисяч токенів рівня речень, анотована POS-тегами та синтаксично структурована. В 2014-му році Silveira et al. надасть анотацію синтаксичних залежностей для цього корпусу, яку можна в подальшому використати для навчання аналізатора залежностей.

На разі у складі Національної словникової бази Українського мовно-інформаційного фонду НАН України функціонує і постійно розвивається Український національний лінгвістичний корпус (УНЛК), що розробляється під керівництвом академіка НАН України В.А. Широкова [3, 103]

В вересні 2019-го року Common Crawl публікує 240 ТіБ нестиснутих даних з 2,55 млрд веб-сторінок. З них 1 млрд URL не був присутній на попередніх скануваннях. Common Crawl стартував ще в 2008-му році. В

2013-му вони перейшли з формату ARC на Web ARChive (WARC). Якщо файли WAT вміщують метадані, то файли WET вміщують відкритий текст файлів WARC.

Зараз корпусна лінгвістика регулярно представляється як така, що розвивається на противагу поточній домінуючій формалістичній версії лінгвістики, яка відділяла ідеалізовану "компетентність" від фактичної "продуктивності" і яка постійно підкреслювала дефектний характер виконання, розглядаючи вивчення компетенції як свою мету. Одним із наслідків зосередженості формалістів на компетентності стало прикре (на думку корпусних лінгвістів) нехтування та знецінення фактичного використання мови (та зразків фактичного використання мови), під час створення своїх описів та теоретичних положень.

В. Жуковська, зауважує, що «Важливою особливістю інформаційно-семіотичного напряму лінгвістичних досліджень (корпусної лінгвістики) є підхід до розгляду прикладних проблем лінгвістики в комунікативних процесах». [3]

Унікальність корпусної лінгвістики полягає в способі використання сучасних комп'ютерних технологій, пов'язаних зі збором мовних даних, методи, що використовуються при обробці мовних баз даних, прийоми, що використовуються в мовних даних та пошуку інформації, а також стратегії, що використовуються при їх застосуванні у всіх видах мовної науково-дослідної діяльності.

Електронний (цифровий) мовної корпус - це нове явище. Його історія налічує майже півстоліття, тому ми ще не прийшли до єдиної думки щодо того, що вважається корпусом, і як його слід розробляти, класифікувати, обробляти та використовувати.

Основна філософія корпусної лінгвістики має дві гілки: (а) ми маємо пізнавальний потяг дізнатись як люди використовують мову у своїй щоденній комунікативній діяльності, і (б) чи можливо створити інтелектуальні системи, які можуть ефективно взаємодіяти з людьми. З цією

мотивацією як комп'ютерні вчені, так і лінгвісти об'єднались для розробки мовного корпусу, який можна використовувати для проєктування інтелектуальних систем (наприклад, системи машинного перекладу, системи обробки мови, системи розуміння мови, системи аналізу та розуміння тексту, системи автоматизованого навчання і т.д.). Задля користі мовної спільноти в цілому.

Усі галузі лінгвістики та мовних технологій можуть отримати вигоду з результатів, отриманих при аналізі корпусів. Таким чином, опис та аналіз лінгвістичних властивостей, зібраних з корпусу, набувають першорядного значення в усіх багатьох областях людського знання і застосування.

## 1.2. Особливості терміносистеми та ключові поняття корпусної лінгвістики

Професор лінгвістики Грайс дотримується концепції корпусної лінгвістики як парадигми, але вважає за краще методологічну концептуалізацію, оскільки він заявляє, що «за останні кілька десятиліть корпусна лінгвістика стала основною методологічною парадигмою в прикладній та теоретичній лінгвістиці». [11, 191] У 2001 році Тоннініі-Бонеллі назвав корпусну лінгвістику «методологією підготовки до застосування», яка володіє «теоретичним статусом». [12, 1]

Точно так само Мальберг описує корпусну лінгвістику як «підхід до опису англійської мови з її власними теоретичними рамками» [13, 2] і, щоб підкреслити це, використовує термін «корпусно-теоретичний підхід». [14] При безпосередньому зверненні до цієї проблеми вона бачить відмінність у сприйнятті, як те, що випливає з типу корпусної лінгвістики, який практикують дослідники: «[t] тут все ще існують розбіжності з приводу того, чи є корпусна лінгвістика в основному методологією або потребує власної теоретичної бази. [13]

Томпсон і Ханстон [15] заявляють, що « основна корпусна лінгвістика - це методологія, яка може бути узгоджена з будь-яким теоретичним підходом до мови» [15,8]. Однак вони продовжують описувати дві основні теорії, які

мають витоки з корпусної лінгвістики. Перш за все, це значення знаходиться не в окремих словах, а в «одиницях значення» в термінології Сінклера, і, отже, цей комунікативний дискурс розгортається в основному в вигляді серії напівфіксованих фраз [15, 11-12].

Корпусна лінгвістика, принаймні, має дві риси, що дають їй підставу претендувати на становище самостійної дисципліни – характер використовуваного словесного матеріалу та специфіка інструментарію.

Можна виокремити такі передумови виникнення корпусної лінгвістики:

1) розвиток інформаційно-комунікаційних технологій;

2) звернення до проблем прикладного характеру в лінгвістиці;

3) перенесення акценту з слова і пропозиції на текст, який, на думку М. М. Бахтіна, і «є первинна даність (реальність) будь-якої гуманітарної дисципліни» [16, 289]. Варто також зазначити, що «корпусна лінгвістика і метод корпусного аналізу - не просто данина технічному прогресу і більш зручний інструмент для пошуку ілюстративного матеріалу, але і прикмета нової ідеології вивчення мови, для якої мова, власне кажучи, і є корпус» [16, 289].

Корпусна лінгвістика використовує методи квантитативної лінгвістики та досліджує реальні емпіричні дані в формі корпусів текстів. Варто зазначити, що корпусна лінгвістика часто розглядається як одна з форм практичного застосування квантитативної лінгвістики. [17, 1].

Як вже було згадано, багато технологій, які зараз використовуються при побудові корпусів, були винайдені задовго до появи комп'ютерів і електронних ресурсів. Деякі з них використовувалися ще в XVIII - XIX століттях, коли лінгвістику почали вважати самостійною і незалежною науковою дисципліною. Захаров В.П. в підручнику «Корпусна лінгвістика» називає технології, які вплинули на створення корпусів. Він виділяє три основні області лінгвістичних досліджень, які увійшли в основу корпусної лінгвістики, хоча і зазначає, що їх було набагато більше [18, С. 25].

Першою такою областю він виділяє порівняльно-історичне мовознавство. Вчені, що працюють в цій галузі, завжди зверталися до величезної кількості різних текстів. Застосування технологій з реконструкції прамови можна зустріти і в сучасній лінгвістиці. Друга область, якій Захаров В.П. віддає перевагу, є складання граматик, словників і навчання мови. Дійсно, будь-яке граматичне правило необхідно проілюструвати. І в цьому випадку приклади з тексту відмінно зможуть в цьому допомогти. Корпуси як джерела емпіричних даних відіграють важливу роль під час навчання іноземної мови. Останньою областю, що вплинула на розвиток корпусів, є соціолінгвістика. Ще в XIX столітті вчені починають розробляти діалектні карти і складати збірники діалектних відносин. При цьому необхідно було враховувати різні критерії при складанні посібників з діалектів. Всі ці фактори і послужили початком становлення корпусної лінгвістики.

Якщо такі розділи лінгвістики як синтаксис, семантика і соціолінгвістика мають на меті опис або оцінку мовної структури або мовного використання, то корпусна лінгвістика є більш широким поняттям, методологією, яку можна застосувати до багатьох аспектів мовних досліджень [1, 9].

Наприклад, діахронічні дослідження спирались та спираються на деякий, часто дуже обмежений «корпус текстів», в основі якого робляться висновки про те чи інше мовне явище. Позбавлений можливості продукувати інтроспективні висновки, історик мови використовує задану сукупність текстів. І перехід на електронні форми збереження діахронного матеріалу по суті не змінює методологію дослідження, оскільки опора на інтуїцію у вказаному сенсі при дослідженні історичного матеріалу за даних причин неможлива. Такий самий підхід властивий, як правило, будь-яким польовим дослідникам, наприклад діалектологам, які працюють з природнім, але часто чужим для них мовним матеріалом, тобто з мовою носіїв конкретного діалекту.

В середині XIX ст. в науковому середовищі панував раціональний підхід, заснований «на лінгвістичній інтуїції, яка проводить відмінність між правильними і неправильними конструкціями» [19, С. 14]. На противагу такому підходу з'являється емпіричний підхід, який пропонує розглядати мову як «ресурс, що забезпечує набір можливості для комунікації» [1, С. 14].

Якщо окрім мовного матеріалу в розпорядженні дослідника є і мовна інтуїція, роль першого знижується. Закономірно запитати, чи виправдана опора на мовний матеріал тільки в тому випадку, коли немає можливості спиратись на власне відчуття мови або ж взагалі на думку носіїв мови? Думки вчених при відповіді на це питання варіюються, однак в цілому можна виокремити 4 підходи:

1. Мовний корпус досліджується, вважаючись самостійним та вичерпним об’єктом дослідження. Будь-який корпус можна досліджувати і просто як корпус, який сам себе репрезентує. При такому підході будь-яке зафіксоване явище визнається легітимним, незалежно від степеню його «правильності», нормативності. Такий підхід, зокрема, підходить і для ідіолектів.
2. Мовний матеріал використовується у дослідженні окремих мовних явищ як джерело опису мови в підсвідомості носія. Даний підхід зводиться до того, що корпусні дані використовуються в якості доказу існування в мові того чи іншого явища. Якщо приймається положення про те, що мова в підсвідомості носія «знаходиться» в підсвідомості людини і на цьому рівні може бути регламентована категоріями «можна-неможна», «правильно-неправильно», а не поняттями «може бути використано – не може бути використано», то, корпус не може слугувати джерелом для вивчення мови у підсвідомому. Однак багато вчених не розділяють такого жорсткого відношення до цієї методологічної дилеми. І в захист більш м’якого рішення цього питання наводяться вагомі аргументи.

3. Мовний матеріал (корпус) як предмет інтересу лінгвіста повністю заперечується. Тотальне заперечення значення мовного матеріалу як джерела для висновків та узагальнень відносно мови можна знайти передусім у Н. Хомського. Так, в інтерв'ю І. Андору він стверджує: "Корпусна лінгвістика нічого не значить. Це наче сказати <…> припустімо фізики та хіміки вирішать, що замість того щоб покладатись на експерименти, вони почнуть записувати на камеру усе що трапляється в світі та зберуть величезну колекцію матеріалу всього того, що відбувається, та на основі цього вони дійдуть до певних узагальнень та прозрінь. Ви ж знаєте що в науці так не заведено.". [20] Твердження Ноама Хомського викликало бурхливу дискусію серед спеціалістів з корпусної лінгвістики. Така реакція зрозуміла, оскільки в своїй аргументації Н. Хомський забуває про дві речі. По-перше, коли фізик повторює експеримент, який зробив його колега (відтворює, який вже був проведений), він виходить з того, що результати обов'язково повинні бути однаковими. В лінгвістиці ж думки носіїв мови про правильність якої-небудь мовної одиниці або правила не завжди співпадають. Через це експерименти лінгвістів, що ґрунтуються на своїй мовній інтуїції (або інтуїції інших носіїв мови), можуть призвести до різних, навіть діаметрально протилежних висновків. По-друге, Н. Хомський звужує уявлення про методологію природничих наук. Фізик не тільки робить експерименти, але й спостерігає за тим, що відбувається в світі, збирає інформацію про дійсність. За допомогою телескопів астрофізики спостерігають за тим, як рухаються небесні тіла; складні сучасні апарати дозволяють проводити збір та аналіз, наприклад, найменших аерозольних часточок. Спостереження за навколишнім середовищем не тільки дозволяє підтверджувати достовірність теоретичних роздумів, але й слугують стимулом для створення нових наукових концепцій.

4. Корпус не тільки не приймається до уваги, але й створення лінгвістичної теорії та системи понять засновується саме на корпусному підході. Як було згадано вище, опонентами Н. Хомського виступили лінгвісти, які роблять фундаментальні висновки про мову, спираючись не на інтуїцію, а на корпусні дані. В багатьох концепціях, які засновані на цьому принципі, особливе значення надається одиницям, більшим, ніж слово в традиційному розумінні, але меншим, ніж речення та висловлювання. В своїх найбільш радикальних підходах прихильники цього напрямку заперечують існування мови в підсвідомості носія.

Корпусна лінгвістика по-перше це теорія та методика, а по-друге корпусні дослідження. Чіткої межі між ними не існує, та практично всі засновники-укладачі корпусів в той самий час проводять також лінгвістичні дослідження.

Сучасна корпусна лінгвістика, незважаючи на відносно коротку історію існування, є добре розробленим напрямком мовознавства, тісно пов’язаним з комп’ютерною та когнітивною лінгвістикою. З першою вона пов’язана технологією та інструментами обробки мовного матеріалу, з другою співпадає в базовій передумові: як когнітивна, так і корпусна лінгвістика цікавляться мовною діяльністю, яка представлена в безкінечній кількості текстів [21]. В певному сенсі корпусна лінгвістика змінює пріоритети дослідження: об’єктом вивчення стає мовлення, яке не зводиться до мовної абстракції, нормам літературної мови, судженням про правильність/неправильність в мові, заснованим виключно на інтуїції досвідченого дослідника.

В основі корпусної лінгвістики лежить ідея того, що мова – повністю соціальне явище і її можна описати даними, які засновані на досвіді, тобто під час мовлення. Така емпірична орієнтованість корпусної лінгвістики

змушує вчених досі сперечатись про те, чи є корпусна лінгвістика самостійною наукою або всього-на-всього методом дослідження.

Хоча застосування корпусів дає лінгвістам можливість отримати ґрунтовну емпіричну базу для досліджень, але, власне, корпус не пояснює мовні явища. Зрештою, слід пам'ятати, що корпусна лінгвістика – при всій революційності тих можливостей, які вона відкриває, лише частина з широкого методологічного інструментарію сучасної лінгвістики. Твердження можна проілюструвати наступними словами Ч. Філлмора: «Я не думаю, що можуть існувати будь-які корпуси, навіть великі, які містять інформацію про всі області англійської лексики і граматики, які я хочу дослідити. <...> але кожен корпус, який мені довелося досліджувати, навіть невеликий, відкрив мені факти, про які я не міг дізнатись іншим способом.» [22].

Таким чином, будь-який великий корпус дивує нас неочікуваними відкриттями, важко вловлюваними без звернення до реального мовного матеріалу, з іншого боку, навіть найбільші корпуси не в стані відобразити все можливе в мові.

«Мета корпусної лінгвістики - опис мови в тому вигляді, в якому вона проявила себе в мові, представленої у вигляді спеціально підібраного корпусу текстів» [16, 290]. Ми теж вважаємо, що корпусна лінгвістика вбачає в якості своєї головної цілі об'єктивний лінгвістичний опис мовної системи, до того ж, корпусна лінгвістика до опису підходить з позицій дослідження конкретних людських комунікацій, реальних текстів.

***Основні напрямки наукової діяльності в рамках корпусної лінгвістики:***

1. Корпуси використовуються в граматичних та лексикологічних дослідженнях, частотні списки і списки ключових слів, дослідження колокацій (тобто поєднання лексем) є на сьогодні однією з самих популярних тем корпусних досліджень. Дослідження корпусів дозволяє отримувати точні дані про лексичний склад мов, про відносну частотність вживаних слів. За допомогою корпусної лінгвістики був

доведений закон Ципфа, суть якого полягає в тому, що якщо в будь-якій природній мові всі слова впорядкувати за спаданням частоти їх використання, то частота будь-якого слова в такому списку виявиться приблизно зворотно пропорційною його порядковому номеру (рангу слова). Наприклад, друге за частотністю слово зустрічається приблизно в два рази рідше за перше, третє – в три рази рідше за перше, і т. д. В результаті застосування статистичного аналізу, у поєднанні з методом, що лежить в основі закону Ципфа, до Рукопису Войнича, написаного невідомою мовою, було доведено, що цей рукопис містить осмислену інформацію. [23]

2. Наявність електронних текстів, які належать одному автору, дають можливість розширити коло задач, які традиційно вирішуються стилістикою та авторською стилеметрією.
3. Корпусна лінгвістика займається проблемами машинного перекладу, для чого й створюються та використовуються багатомовні паралельні корпуси, в яких в кожній фразі однією мовою зіставлений її еквівалент іншою мовою. Окрім машинного перекладу такий корпус можна використовувати для досліджень, які пов’язані з порівнянням оригінальних та перекладених текстів. Хоча створення багатомовних паралельних корпусів пов’язане з різними педагогічними задачами , однак вони також мають і власне лінгвістичне значення.
4. Ще одна задача, яка успішно вирішується за допомогою корпусних методів, це встановлення плагіату та прихованого цитування.
5. Дослідження норми та узусу – хоча дослідження норми зазвичай не входить в задачу корпусних лінгвістів, більшість гострих, затребуваних суспільством мовних питань може бути вирішено на основі не суб’єктивних оцінок, а з залученням статистично більш представницького матеріалу.
6. Корпусна лінгвістика досліджує граматики природніх мов, зокрема – сполучуваності тих чи інших граматичних явищ один з одним.

Звичайно, що дані, отримані з живого мовлення набагато актуальніші ніж теоретичні граматики традиційної лінгвістики.

7. Корпусні методи застосовуються для вирішення задач судово-лінгвістичної експертизи.
8. Корпусна лінгвістика з самого початку створення була тісно пов'язана з викладанням мови іноземним студентам. Тож корпуси активно використовуються в лінгводидактиці. Аби знати чому саме потрібно навчити учня необхідні точні кількісні дані про мову, яка викладається, а саме – склад найбільш вживаної лексики, ймовірнісність вживання граматичних конструкцій і т. д.
9. Відносно новою областю є створення корпусів текстів школярів, які дозволяють класифікувати типи помилок та враховувати їх під час навчання.
10. Корпуси текстів дозволяють проводити лексикографічні дослідження, створювати словники. Майже усі сучасні словники англійської мови (MacMillan, Collins, Webster) видаються на основі величезних корпусів, які дозволяють зробити словник репрезентативним.
11. Корпусна лінгвістика займається дослідженням текстів. Наприклад, використовуючи корпуси текстів можливо навчитись визначати функціональний стиль тексту через його статистичні характеристики – середню довжину слова та речення, характерні поєднання слів і т. д. Дані методи використовуються в автоматичному реферуванні та тематичному пошуку.
12. Корпусні методи з самого виникнення активно використовувались в соціолінгвістичних дослідженнях.
13. Хибно вважати, що корпусна лінгвістика працює тільки з писемними текстами. Окремою областю, яка постійно перебуває в процесі розробки, стало створення та вивчення\дослідження корпусів усної мови.

14. Дослідження змін в лексичному складі мов та різноманітних їх варіацій, наприклад, появу та зникнення неологізмів.

## 1.3. Корпус як особливий лінгвістичний та мовний ресурс

Як і випливає з назви, корпусна лінгвістика вимагає використання корпусу. Корпус – це просто тіло тексту; проте в контексті корпусної лінгвістики визначення корпусу набуло більш спеціалізованих значень. Згідно Боукеру і Пірсону (2002) [24], корпус можна описати як велику колекцію автентичних текстів, зібраних в електронній формі згідно з певним набором критеріїв. «Корпус вважається представником того мовного різновиду, який він повинен представляти, якщо результати на базі його змісту можна узагальнити до зазначеного мовного різновиду» [25].

Ключове спостереження, зроблене Кеннеді (1998) [26], полягає в тому, що замість того, щоб ініціювати дослідження корпусу, розвиток інформаційних технологій змінив спосіб роботи з корпусом, так що лінгвістика корпусу тепер нерозривно пов'язана з комп'ютером. За допомогою комп'ютера дослідники можуть зберігати величезні обсяги тексту, швидко і вичерпно витягувати певні екземпляри слів або фраз, а також сортувати і відображати ці текстові дані по-різному, полегшуючи інтерпретацію. Сучасні розмічені електронні корпуси – це дуже потужний і гнучкий інструмент, який дозволяє задавати найрізноманітніші питання про склад мови і миттєво отримувати на них відповіді у вигляді багатьох десятків або навіть сотень речень.

Таким чином, в кожному з представлених визначень поняття «корпус» підкреслюється наступне:

1) Безліч текстів повинно бути представлено в електронному вигляді (в мережі Інтернет або ж на диску);
2) Мовні дані повинні бути розмічені для аналізу в лінгвістичних цілях;

3) В результаті проведеного аналізу повинна існувати можливість різноманітного розподілення отриманого мовного матеріалу (за жанром, роком створення тексту, тематиці і т. д.)

Якщо розглядати перший пункт, то тут істотним критерієм виступає доступність корпусу текстів в електронному вигляді. Всю існуючу множину корпусів текстів можна розділити на три великі категорії: 1) які знаходяться у вільному доступі; 2) які знаходяться в частковому доступі; 3) комерційні. До першої категорії відноситься доволі обмежена кількість з існуючих на даний момент корпусів текстів. Більшість існуючих корпусів відноситься до другої категорії, однак для вирішення конкретних лінгвістичних задач такий частковий доступ є більш ніж достатнім.

У книзі Михайла В. Копотєва "Введення в корпусну лінгвістику" неодноразово повторюється, що з появою корпусів суттєво змінилися цілі і завдання лінгвістики, що текст і тільки текст є об'єктом і предметом дослідження лінгвіста і що можливість звернення до необмеженої кількості фактів тексту надають сьогодні саме корпуси. [27] Після тривалого періоду відходу лінгвістики XX століття від тексту як об'єкта, предмета і мети спостережень, аналізу і висновків відбулося повернення до нього.

Ю. Д. Апресян, один з прихильників зваженого, обережного звернення до корпусних даних, особливо при зверненні до Wikipedia Corpus або Google Books, вважає, що деякі лінгвісти сліпо слідують «моді на корпус» [28], і це повальне захоплення часто призводить до фальсифікації результатів та зловживанню кількісними даними. Сирий частотний підрахунок вживання слів не може виступати критерієм істинності тверджень про функціонування лінгвістичного об'єкта. Дане питання якості отриманих результатів повністю пов'язаний з професійною компетенцією дослідника. Проявам дилетантства і кустарювання в середовищі лінгвістів-корпусників частково «сприяють» і несталий термінологічний апарат і не налагоджена методологія корпусного дослідження, але такі обставини ніяк не применшують значущості корпусних технологій.

Одна з найважливіших ідеологічних та методологічних проблем корпусної лінгвістики стосується сутності корпусу – чи відображає корпус сутність мови взагалі і в якій мірі. Оскільки представники корпусної лінгвістики ототожнюють «існуюче в мові» з «засвідченим в корпусі», необхідно, аби корпус відповідав певним вимогам. Наприклад, американський дослідник С. Гріс позначає таку властивість корпусу як «репрезентативність». Вважаючи «збалансованість» другим найважливішим засобом корпусу, С. Гріс має на увазі, що в співвідношенні текстів повинні зберігатися пропорції. При цьому вчений вказує, що точне процентне співвідношення є скоріше якимсь теоретичним ідеалом і точно виміряти мовленнєві явища за допомогою математичних критеріїв неможливо. Третім важливим критерієм вчений вважає природне походження текстів корпусу, тобто їх виникнення в автентичній комунікативної ситуації [11: 7-8]. В. Плунгян визначає одну з головних вимог до корпусу його обсяг: «Корпус повинен бути великим. Відображати якщо не всі тексти, написані цією мовою, хоча таке завдання теж можна поставити, але найбільш важливі, найбільш представницькі, пропорційно влаштовані. Скажімо, корпуси сучасних мов повинні досліджувати не тільки художню літературу, а й газетні тексти, блоги і т.д.» [29, 11].

В. Захаров вважає, що досить великий (репрезентативний) обсяг корпусу гарантує типовість даних і забезпечує повноту уявлення всього спектру мовних явищ. На його думку корпус повинен містити не менше 100 мільйонів слововживань, розуміючи під репрезентативністю «необхідне, достатнє і пропорційне представлення в корпусі текстів різних періодів, жанрів, стилів, авторів і т. д., тобто здатність відображати всі властивості проблемної області ». Однак, дослідник також відзначає складність введення математичних критеріїв для визначення репрезентативності та збалансованості: «Це поняття неможливо розрахувати і описати суто математично, однак цього можна і потрібно прагнути, як на етапі проектування корпусу, так і на етапі його експлуатації» [18, 18].

Інший вчений, німецький лінгвіст А. Клоза, вказує, що репрезентативність корпусу не можна трактувати в статистичному сенсі: корпус – зібрання текстів, складене спеціально для проведення лінгвістичного аналізу, яке, як передбачається, є репрезентативним для певної мови. «Дієвою є вимога до природного існування висловлювань (частіше в письмовій, ніж в усній формі), сукупність яких була скомпільована на основі встановлених критеріїв відбору для досягнення певної мети. Ця сукупність текстів і повинна репрезентувати природну мову». [30, 106-107].

А. Баранов додає до існуючих критеріїв також критерій економічності. «Корпус текстів повинен економити зусилля дослідника при дослідженні проблемної області. Зокрема, він (корпус) повинен бути не просто серйозною підмножиною текстів проблемної області, але, по можливості, істотно відрізнятися від неї за обсягом» [31, 103]. Окрім всіх перерахованих критеріїв, важливо також щоб одного разу створений і підготовлений масив даних міг би використовуватися багаторазово, різними дослідниками і в різних цілях.

Незважаючи на різноманітність корпусів, можна виділити два основних способи їхнього поділу на класи [31]**:**

1. зіставлення корпусів, що стосуються всієї мови (або мови певного періоду);
2. поділ корпусів, що належать до певної мовознавчої галузі (жанру, стилю, мови певної вікової або соціальної групи);
3. поділ корпусів за типом лінгвістичної розмітки. Більшість з них усе ж таки належить до корпусів морфологічного або синтаксичного типу (так званого treebanks, «банку синтаксичних структур»). Натомість корпус із синтаксичною розміткою вміщує в себе й морфологічні характеристики лексичних одиниць. [32]

Можна погодитись із С. Я. Вадяєвим, що корпус тексту, який становить добірку текстів предметної галузі знань, можна назвати спеціальним, і використовуватись він має з метою, для якої спроектований. [33] Проте під час дослідження спеціального корпусу варто враховувати, що спеціальний корпус не завжди може відображати об'єктивну дійсність; він призначений для використання лише в галузевих структурах. [32]

За степенем організації та структурованості корпуси поділяються на:

Електронний архів – це тексти на електронному носії, але їх форма представлена на машинному носії не стандартизована та не уніфікована.

Електронна бібліотека – тексти представлені у однорідному та стандартизованому вигляді.

Корпус текстів – форма стандартизована та уніфікована, тексти призначені для відображення частини лінгвістичної реальності.

Субкорпус – певна автономна частина корпусу.

Корпуси даних також поділяються на:

Дослідницький корпус – корпус призначений для дослідження різноманітних аспектів функціонування мовної системи. Такі корпуси укладаються перед проведенням певного дослідження.

Ілюстративний корпус – корпус, який створюється після проведення наукового дослідження, ціль не стільки виявити нові факти, як підтвердити та обґрунтувати вже отримані результати. Ілюстративні корпуси не є статистично правильним відображенням проблемної галузі, адже вони включають лише те, що є достатнім для ілюстрування описуваного феномену.

Статичний корпус – відображає конкретний стан мовної системи у певному часовому проміжку. Типовий представник такого виду корпусів – авторські корпуси, тобто колекції текстів письменників.

Динамічний (моніторний) корпус – відрізняється від статичного тим, що не являє собою незмінний заданий набір текстів, а постійно оновлюється та доповнюється з метою моніторингу стану проблемної галузі та динаміки її

змін. За допомогою такого корпусу користувач при проведенні досліджень може виокремити з загального головного корпусу робочий корпус, який буде вміщувати лише частину текстів з основного корпусу.

Тож корпуси поділяються на групи:

1. За хронологічними ознаками: синхронічний, моніторний (відслідковує поточний стан мови), діахронічний;
2. По індексації: простий або анотований;
3. За жанром: літературні, публіцистичні, фантастичні і т. д.;
4. За мовою: одномовний, двомовний, багатомовний;
5. За метою: багатоцільові та спеціалізовані;
6. За способом застосування та використання корпусу: дослідницький, ілюстративний, паралельний;
7. За наявністю розмітки: розмічені та нерозмічені;
8. За способом існування корпусу: динамічний або статичний

Застосунки аналізу корпусу:

Застосунки для укладання конкордансів

Застосунки для індексування та анотування

Конкорданс – список словоформ, які зустрічаються в тексті, розташовані в алфавітному порядку. В противагу словнику – слово подається разом з його мовним оточенням.

Критерії відбору текстів визначають залежно від специфіки створюваного корпусу: усієї мови (загальномовний, національний корпус), корпус періодики (охоплює тільки тексти періодичних видань, написані літературною мовою), діалектний (репрезентує діалектні тексти), історичний (репрезентує тексти певного історичного періоду існування мови) тощо. [34]

Багато лінгвістів використовують корпус як «банк прикладів», тобто намагаються знайти емпіричну підтримку для своїх гіпотез, принципів і правил, над якими вони працюють. Приклади, звичайно, можуть бути придумані або ж знайдені випадково, але підхід корпусної лінгвістики

забезпечує репрезентативність і збалансованість мовного матеріалу, а також пошуковий інструмент, який зазвичай дає можливість хорошої вибірки в певному корпусі [1, 95]. «Найпростіше використання корпусу - підбір контекстів: підібрав під корпус, натиснув «знайти слово» і отримав всі приклади вживання слова X у письменника N. Але лексикологічні дослідження можуть бути і набагато більш складними ». [27,56]

Можна виділити таке застосування корпусів текстів в дослідженні мови:

1. Підбір потрібного корпусу текстів: доступність, достатність словесного матеріалу, чи є корпус досить представницьким для поставленої задачі, яким чином були відібрані тексти, чи є достовірним представлення індексів (якщо він індексований)
2. Наскільки є необхідним дане дослідження (адаптація цілей та задач дослідження під наявний корпус текстів)
3. Практичні рекомендації: аналізувати те, що ясно та явно представлене в машинній формі, шукати те, що легко знайти, підраховувати те, що легко підрахувати.

В цілому, для корпусних методів характерно:

- зміщення дослідницької стратегії з вивчення норми («як правильно») на вивчення узусу («як розмовляють/пишуть»);

- автоматичне вилучення інформації за допомогою пошукових запитів, що може призвести до отримання об’ємного і не завжди релевантного матеріалу;

- розповсюдженість «формально-морфологічного» підходу, за якого пошук прикладів заснований на морфологічній (або ж просто буквенній) формі;

- використання квантитативних методів, які дозволяють враховувати частотні характеристики досліджуваних одиниць, и заміна інтроспективних оцінок матеріалу точними кількісними даними про частотність вживання;

- опора на автоматичне анотування, не позбавлене, з точки зору традиційної лінгвістики, визначених неточностей та спрощень;

- увага до контексту в широкому сенсі (дослідження колокацій, ключових слів, конструкцій передбачає врахування оточення досліджуваної одиниці).

**Висновки до розділу 1**

На сучасному етапі корпусна лінгвістика як наукова методологія і галузь мовознавства розвинена досить нерівномірно. Для деяких мов, таких як англійська, німецька, китайська та ін. створені великі і репрезентативні анотовані корпуси, в той час як для інших мов, включаючи і українську, процес створення повноцінних корпусів, що відповідають основним вимогам, переживає період формування.

Ціль корпусної лінгвістики – опис мови в тому вигляді, в якому вона проявила себе у мовленні, представлена у вигляді спеціально зібраного корпусу текстів.

В своїх дослідженнях спирається на дані корпусу, але віддає перевагу квантитативним методам. Базується на емпіричних методах. Текст розглядається як певна фізична сутність у глобальній перспективі. Фокусує увагу на якомога ширшому погляді на текст, який не обмежений догмами. У своїх висновках спирається на спостереження мовленнєвої діяльності, яка була відображена у вигляді текстів. КЛ часто користується ймовірнісними методами та статистикою для первинної обробки мовленнєвого матеріалу. Проводиться робота з лінгвістичними даними (слововживаннями) в тому вигляді, в якому вони були вжиті в контексті. Надає перевагу індуктивним методам обробки емпіричного словесного матеріалу, вважає його сутністю наукового методу.

Корпусній лінгвістиці властива певна термінологічна нечіткість, адже досі немає узгодженості по відношенню до терміна «корпус». Кожний

дослідник має своє тлумачення терміну та бачення головних властивостей корпусу. Зі зміною технологій, термін постійно обновлювався…..

Також дисципліна корпусної лінгвістики має справу з вже зібраними матеріалами:

1. Необхідно представити структуру мовленнєвої діяльності
2. Виявити, які матеріальні обмеження є на створення корпусу
3. Відбір текстів та укладання корпусу текстів

# РОЗДІЛ 2 Міжнародні стандарти та вимоги щодо проєктування корпусів текстів

## 2.1 Технологія, види та застосунки для мовної розмітки корпусу

Загалом, корпуси можуть бути розмічені та нерозмічені, але розмітка вагомий інструмент, адже для вирішення лінгвістичних задач недостатньо простого масиву текстів.

Н. П. Дарчук, говорячи про корпуси, зазначає: «Наука про корпуси – це, перш за все, наука про те, як зробити хорошу розмітку корпусу» [35], тим самим роблячи акцент на тому, що якість корпусу залежить від розмітки та підкреслює їх неподільність.

Роззначення полягає у приписуванні текстам і їх компонентам спеціальних позначок: зовнішніх, або екстралінгвістичних (відомості про автора й текст); структурних (розділ, абзац, речення, словоформа); власне лінгвістичних, що описують лексичні, граматичні та інші характеристики елементів тексту. [36]

«В розміченому корпусі словам та реченням присвоюються позначки (теги), відповідно до характеру розмітки: морфологічні, синтаксичні, семантичні, просодичні та ін». [1, 25]. «До первинної розмітки текстів відносяться етапи, обов'язкові для кожного корпусу: токенізація, лематизації, морфологічний аналіз. З'явилися застосунки, які вміють витягати з ланцюжка літер морфологічні характеристики тексту форми. Додатки, які автоматично аналізують морфологію слів, назвали лінгвістичними аннотаторами, або тегерами (від англ. Tagger). Отримуючи на вході тексту форму, така програма видає повну морфологічну характеристику у вигляді набору тегів, або тегсету (англ. tagset)» [27, 45].

В. В. Жуковська поділяє такі мітки на три різновиди:

– 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 також вік, стать, роки життя й багато чого іншого (це кодування інформації має назву метарозмітка);

– структурні (розділ, абзац, речення, словоформа);

– власне лінгвістичні, що описують лексичні, граматичні та інші характеристики елементів тексту [36, с. 76].

Не існує загальноприйнятого набору тегів для розмічування текстів, отже автори робіт використовують різноманітні набори тегів в залежності від задач дослідження та ступеню структурованості текстів. Брідж [37] в якості тегів використовував заголовки полів, які заповнюються по мірі опису вразливості в NVD. Набір тегів в роботі Джоши [38] був сформований дослідниками в результаті аналізу текстової колекції. Лім [39] під час роботи брав словник для опису вразливостей MAEC в якості джерела тегів.

Під лінгвістичною анотацією у корпусній лінгвістиці традиційно розуміють: а) довільну лінгвістичну інформацію про лінгвально релевантні одиниці текстових даних, поданих через формальний код; б) практику введення формалізованої лінгвістичної інформації в електронний текст; в) наявність такої інформації у тексті. [34]

Існують такі типи корпусної розмітки: морфологічна, синтаксична, семантична, анафорична, просодична, лематизація, токенізація, стемінг, парсинг. І окрема – метарозмітка, яка презентує відомості про автора та самий текст. В українській корпусній лінгвістиці проблема встановлення вимог до метаданих уперше постала на початку 2000 рр. Вимоги стосовно мовних даних корпусу та інформації про них визначалися потребами користувача – лінгвістів різної спеціалізації [40, 89–90].

Лоу Бернард виділяє 4 типи метаданих: редакторські (editorial metadata), аналітичні (analytic metadata), дескриптивні/описові (descriptive metadata), адміністративні (administrative metadata) [41]. Редакторські метадані – інформація про відношення складових корпусу та їх оригінальних

джерел; аналітичні – шляхи інтерпретації та аналізу складових корпусу; дескриптивні – класифікаційна інформація із внутрішніх або зовнішніх характеристик складових корпусу; адміністративні – документальна інформація про власне корпус (назва/заголовок, наявність, статус перегляду тощо). [42]

Вид морфологічної розмітки найбільш поширений в існуючих корпусах, він визначається як базовий тип розмітки, при цьому враховується не тільки ознака частини мови, але і ознаки граматичних категорій. Для даного типу розмітки важливим є зняття лексико-граматичної та граматичної омонімії (на 94% здійснюється автоматично), це забезпечує достовірність роботи всіх етапів автоматичного опрацювання текстів корпусу.

Мета синтаксичної розмітки – змоделювати синтаксичну структуру вхідного речення на рівні словосполучень, а також приписати інформацію про типи синтаксичних зв’язків і побудувати дерево залежностей речення.

Анафорична розмітка. Фіксує референтні зв'язки, наприклад, займенників. Просодична розмітка: У просодичних корпусах застосовуються мітки, що описують наголос та інтонацію. У корпусах усної розмовної мови просодична розмітка часто супроводжується так званою дискурсною розміткою, яка слугує для позначення пауз, повторів, застережень, і т. д.

На сьогодні на базі міжнародного досвіду виробилися де-факто стандарти представлення метаданих, що базуються на описах текстів в рамках проекту Text Encoding Initiative (TEI) і на рекомендаціях EAGLES (Expert Advisory Group on Language Engineering Standards). В якості формальної мови розмітки широко застосовуються мови SGML і XML. В даний час стандарти EAGLES безпосередньо включаються в технологічне середовище мови XML. Зокрема, розробку стандарту Corpus Encoding Standard for XML (XCES).

Рекомендації TEI передбачають багаторівневий склад метаданих [43]. Зокрема розробник корпусу текстів має обов'язково подати таку інформацію про корпусні дані:

1) назва корпусу, розробник або назва компанії, спосіб доступу до корпусу та особа, відповідальна за забезпечення доступу, контактні дані, наявність корпусу у вільному доступі;
2) вихідні дані корпусу з детальним описом кожного джерела текстів;
3) бібліографічний опис корпусу текстів, виконаний у звичайному форматі (автор, назва, видавництво, дата, ISBN тощо) або за стандартами цитування TEI, BibTeX. [44]

На відміну від TEI стандарт метаданих IMDI [45] пропонує менш деталізовану схему паспортизації корпусу із зазначенням таких параметрів:

1) період роботи над проектом (ім'я виконавця роботи на кожному етапі укладання корпусу, назва проекту, дата);
2) адреса організації (назва, контактні дані виконавців, примітки),
3) укладач корпусу – особа / організація, відповідальна за укладання корпусу (ім'я, контакти);
4) елементи корпусу (контекст, жанр тексту, мета написання, мови, використовувані коди лінгвістичної розмітки);
5) дані учасників проекту (тип виконуваної роботи, ім'я, посада, мова, етнічна група, вік, стать, освіта, анонімність участі);
6) джерела (веб-посилання, розмір джерел, тип, формат, якість, доступ, примітки);
7) анотування (посилання на джерело, дата, тип, формат, мова розмітки, примітки).

Корпусний менеджер – спеціалізований застосунок, призначений для створення, редагування, анотування та пошуку текстів в лінгвістичному корпусі.

Як відзначають В. П. Захаров і С. Ю. Богданова, серед всіх функцій корпусного менеджера, широко поширені наступні функції:

1) пошук не тільки окремих слів, але і словосполучень;
2) пошук по шаблонах (складні запити);
3) сортування списків за кількома критеріями, що обираються користувачем;
4) можливість відображати знайдені словоформи в необмеженому контексті;
5) надання статистичної інформації по окремим елементам корпусу;
6) відображення лем, морфологічних характеристик словоформ і метаданих (бібліографічних, типологічних), що залежить від ступеню розмічування корпусу;
7) збереження і роздрукування результатів;
8) робота як з окремими файлами, так і з корпусами, які не обмежені за розміром;
9) швидка обробка запиту та видача по ньому результату;
10) підтримка різних форматів текстових даних (txt., Doc., rtf., html., xml. та ін.);
11) легкість у використанні (інтуїтивно зрозумілий) як досвідченому користувачу, так і початківцю. [1, 55]

Найбільш відомі такі універсальні корпусні менеджери, як SARA, XAIRA (BNC), Manatee / Bonito, CQP, DDC. Для обробки корпусних даних можуть розроблятися менеджери на основі систем управління базами даних (СКБД) або пошукових систем [1, 56].

Мова запитів URL – штучна мова, використовується для створення запитів до баз даних та інших інформаційних систем.

Мови запитів корпусних менеджерів, представлені в тій чи іншій формі (формалізована мова запитів або віконний інтерфейс), як правило, базуються на формалізмі, який отримав назву «мова регулярних виразів [1, 64].

Регулярні вирази - це строкові записи, що задають правила пошуку на спеціальній мові. Іншими словами регулярні вирази/шаблони (англ. Regular expressions) – зразок для пошуку, що вміщує як звичайні символи (наприклад, літери алфавіту), так і символи підстановки», які заміняють групу символів. Наприклад, вираз «стіл*» дозволить знайти всі слова, що містять літери с-т-і-л і будь-яку кількість символів після них (наприклад, столу, столом, столовий) [27, 116].

## 2.2 Вимоги щодо конструювання та застосування корпусів

За В. П. Захаровим [18], формування корпусів відбувається за таким алгоритмом: проектування; 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.

Перед конструюванням корпусу потрібно подумати над такими проблемними питаннями та вимогами:

1. Хто користувач корпусу: індивід, група, лінгвістичні співтовариства.

2. Яка логічна ідея покладена в основу корпусу?

3. З яким обсягом даних ми будемо працювати при укладанні корпусу? На скільки це необхідно та реалістично обумовлено?

4. Використовуємо уривки текстів, повні тексти, або ж те і інше.

5. Процедура відбору текстів в корпусі. Для різних цілей процес відрізняється: обстеження мовного матеріалу, сканування текстів, остаточне формування, укладання корпусу.

6. Стандартизоване представлення корпусу на рівні стандартів у галузі, тобто представлення корпусу як кінечного продукту: анотація усього тексту в цілому, уніфіковане представлення словесного матеріалу тексту.

7. Анотування, індексація словесного матеріалу текстів.

Також хороший корпус або конкорданс повинен мати такі характеристики:

Інтенсивність даних: наприклад, список слів повинен містити верхні 60 тис. слів, а не лише верхні 3 тис. слів.

Актуальність (сучасний контент): погано коли корпус на основі застарілих текстів не відповідає поточним завданням.

Метадані: в метаданих повинні бути вказані джерела, припущення, обмеження та те, що входить до складу корпусу.

Жанр: якщо корпус не зібраний для конкретних завдань, він повинен включати різні жанри, такі як газети, журнали, блоги, академічні журнали тощо.

Розмір: корпус із півмільйона слів або більше гарантує, що низькочастотні слова також будуть адекватно представлені.

Чистота: список слів, що містить словоформи одного і того ж слова, може бути занадто безладним для обробки. Кращий за чистотою корпус буде включати тільки лему і частину мови.

Доцільність створення, цінність корпусу та сенс використання визначається такими передумовами:

1) досить великий (репрезентативний) обсяг корпусу гарантує типовість даних і забезпечує повноту уявлення всього діапазону мовних явищ;

2) показність та збалансованість складу текстів, що дозволяє використовувати його для тестування пошукових застосунків, автоматичних морфологій, систем перекладу, а також використовувати в лінгвістичних дослідженнях;

3) дані різного типу знаходяться в корпусі у своїй первинній контекстній формі, що уможливлює їх всебічне та об'єктивне вивчення;

4) одного разу створений і підготовлений масив даних може використовуватися багаторазово, багатьма дослідниками й в різних цілях.

У загальному корпусі загальний баланс передбачає діапазон жанрів, що входять до корпусу та їх пропорція, а вибірка – це як текст розбивається на блоки відповідно для кожного жанру. У спеціалізованих корпусах важливими є ступені завершеності та концентрованості стосовно певної лінгвістичної ознаки (наприклад, обсягу лексикону).

Репрезентативність є головною проблемою при розробці корпусу, і вона зумовлена ідентифікацією конкретної сукупності або центральної точки дослідження. Репрезентативність відноситься до «міри, в якій вибірка вміщує повний спектр мінливості в популяції». [46] Отже, репрезентативність достатньо податливе поняття, яке тісно пов’язане з питаннями щодо вашого дослідження.

Репрезентативність загальних корпусів та таких, які залежні від предметної області або жанру, тобто спеціалізованих корпусів досягається та вимірюється різними способами.

Загальні корпуси:

Баланс: діапазон жанрів, що входять до складу корпусу, та їх частка

Вибірка: як відбираються фрагменти тексту для кожного жанру

Спеціалізовані корпуси:

Ступінь закритості / насиченості (наповнення/наповненості): Закритість/ насиченість для певної лінгвістичної ознаки (наприклад, розміру лексикону) різноманітної мови (наприклад, комп’ютерних посібників)

означає, що ця функція, як видається, є кінцевою або може зазнавати дуже обмежених змін за певний момент, тобто крива лексичного зростання вирівнюється

«За критеріями репрезентативності та відбору текстів, розрізняються два основних типи корпусів: корпуси, що стосуються усієї мови і свідомо зміщені корпуси, що відносяться до якої-небудь підмови (жанр, стиль, мова певної соціальної групи і т. д.)» [47, 105]. Корпуси, які підходять під першу класифікацію, будуються на базисі принципу дедукції – рух від загального до відображуваного, це властиво приватному корпусу текстів. Вони універсальні та їх ціль – відображення усієї багатогранності мовленнєвої діяльності, яка існує незалежно від дослідника. Натомість, корпуси другого класу збираються спеціально для відображення деякого лінгвістичного або ж культурного феномену.

Також репрезентативність різнобічна, оскільки деякі тексти, наприклад, відбираються із сукупності всіх наявних текстів (що саме являє собою підмножину всіх можливих текстів), а потім у багатьох випадках з цих текстів отримують менші зразки. Важлива не тільки репрезентативність усієї мови, але й те, наскільки репрезентативними є джерела, з яких взяті зразки.

Головні етапи прийняття рішень, що стосуються репрезентативності: на основі бажаної функції корпусу розробляється стратегія відбору реальних даних; дані відбираються та перевіряються на предмет репрезентативності, після чого може знадобитися додаткова вибірка; застосовується рівень впевненості або помилки; далі дані нормалізуються; і корпус створений. Як і в моделі Тогніні-Бонеллі, відтворюваність та узагальнення залежать від ступеня репрезентативності, а доступність ресурсів є ключовим обмеженням.

Автентичність передбачає відбір реально створеного носієм (ями) мови писемного або усного тексту(ів), уривка(ів) тексту(ів) у процесі реальної

комунікації. Дотримання вимоги автентичності є однією зі складових емпіризації фактичного корпусного матеріалу. [48]

Вибірка та статистика також важливі наукові основи дизайну корпусу. Багато членів лінгвістичної спільноти, включаючи Енгволла, Аткінса, Кліра та Остлера, скептично ставляться до можливості робити достовірні статистичні вибірки лінгвістичного тексту через неточно визначені сукупності певних груп населення та критерії одиниць. Вони критично ставляться до ідеї, що “збалансований” корпус є необхідним перед початком досліджень, аргументуючи це тим, що репрезентативність - це ітераційний процес, який оптимізується на основі зворотного зв’язку у міру використання корпусу.

Деякі дослідники підтримують позицію, що, коли конкретний метод вибірки корпусу невідомий, лінгвісти повинні вважати, що "вибірка була проведена теоретично "правильно". Але це суперечить науковому методу, де популяції суворо визначаються заздалегідь, а методи розкриваються повністю.

Також поширений скептицизм з боку дослідників, що використовують корпуси, будова яких невідома, та заохочення розробників корпусів використовувати ймовірнісну вибірку у своїх проєктах та розробці. «Випадкова» вибірка може бути проблематичною; Тогніні-Бонеллі (2001) [49] зазначає, що "декілька лінгвістичних особливостей тексту розподіляються рівномірно по всьому тексту". Бібер у своїх працях допускає, що загалом стратифікована вибірка дає більш репрезентативний корпус, ніж пропорційна. Він також зазначає, що з точки зору аналізу багато досліджень, заснованих на корпусах, є одноваріантними за своєю суттю, і припускає, що багатовимірні методи, такі як факторний аналіз та кластерний аналіз, корисні для метааналізу репрезентативності корпусу.

Відтворюваність експерименту є ознакою реального наукового дослідження. Здатність відтворювати власні (а також чужі) експерименти важлива не тільки з позиції підтвердження результатів, але і для загального прогресу дисципліни. Однак в інформатиці, як і в інших дисциплінах, «окремі проекти, як правило, працюють із власними даними […]. Це надзвичайно ускладнює порівняння результатів, отриманих різними проектами […]; тому сукупний прогрес у розумінні того, як працюють пошукові системи, за допомогою співвідношення ряду результатів, є низьким ». [50] У випадку конференцій TREC наявність стандартизованого тестового корпусу забезпечує відносний рівень порівнянності результатів, навіть вважаючи, що конкретні підходи різних дослідників відрізняються. Відтворюваність та узагальнення залежать від репрезентативності. І навпаки, виникає питання про пристосованість до змін навколишнього середовища та еволюції знань (тобто необхідність додавання нових спостережень та зміни фізичної структури завдяки вдосконаленим інструментам (реляційний файл у порівнянні з плоским файлом і т. д.)

Існує велика кількість систематичних помилок, які можна ввести в корпусні експерименти: теоретичні, приладові (наприклад, калібрування) та операторські, але враховуючи центральність репрезентативності, найважливішою є, мабуть, помилка вибірки. Помилки вибірки можуть бути спричинені випадковістю чи рідкістю обраного екземпляру, або більш серйозними проблемами, що виникають внаслідок несистематичної побудови корпусу. Оцінювана похибка обернено пропорційна розміру вибірки (тобто чим менша вибірка, тим менш репрезентативна вона для сукупності в цілому, а отже, і збільшується шанс на допущення помилки). [51] Нормалізація даних необхідна на етапі побудови щодо формату вибірки, а також інших атрибутів вибірки. Помилка впливає на всі аспекти експериментального процесу: обґрунтованість, надійність та ефективність.

## 2.3 Корпусна пошукова система Sketch Engine та її відкрита версія NoSketch Engine

Sketch Engine - це корпусна пошукова система (Corpus Query System), що дозволяє користувачеві переглядати сполучуваність слова, отримувати список семантично пов'язаних слів за допомогою функції «Тезаурус», а також порівнювати слова, водночас з іншими функціями, які є типовими для пошукових систем даного виду. Дані про сполучуваність одиниці інтегровані з функцією конкордансу. Проект системи Sketch Engine був запущений в 2004 р компанією Lexical Computing Limited, заснованої британським лексикографом, корпусним лінгвістом А. Кілгаріфом. Основні функції даної системи розроблялися А. Кілгаріфом у співпраці з П. Рихлі, розробником корпус-менеджера Manatee, співробітником Лабораторії обробки природної мови (Natural Language Processing Laboratory) факультету інформатики Університету ім. Масарика в Брно, Чехія. Мережева версія системи Sketch Engine (the.sketchengine.co.uk) включає велику кількість готових до використання корпусів на кількох десятках мов, а також інструменти для створення, установки і управління власним корпусом [52, 17-18].

Ця мережева платформа служить для збору, побудови і дослідження веб-корпусів, при цьому «скетчі» утворюють її центральне ядро. Вона не тільки містить понад 400 корпусів на більш ніж 80 мовах, але також дає користувачеві можливість будувати корпусу самостійно (автоматично, на основі матеріалу з інтернет-сторінок). Система може не тільки просто показувати «скетчі», а й порівнювати їх, показуючи загальні і розбіжні граматичні зв'язки, а також "дистрибутивний тезаурус", дозволяє відображати слова зі схожими граматичними зв'язками у вигляді хмари (word cloud).

Система Sketch Engine розроблена на основі корпусного менеджера Bonito. Корпусний менеджер Bonito – програмне забезпечення для роботи з

корпусами текстів. Система Bonito складається з двох частин: сервера і графічного користувацького інтерфейсу.

Основними особливостями мови запитів даного корпусного менеджера є:

1) пошук окремих атрибутів (словоформа, лема, тег);
2) використання регулярних виразів;
3) логічні оператори;
4) засоби запиту структури;
5) швидка обробка складних запитів;
6) шаблони.

Розглянемо основні інструменти системи Sketch Engine, а саме Word List, Word Sketch, Concordancer, Thesaurus, Sketch Diff.

World list (список слів) дозволяє витягувати з корпусу частотні списки одиниць різних типів (словоформ, лем і ін.).

Пошук в системі Sketch Engine може здійснюватися за словоформою, лемою, поєднанням і морфосинтаксичній мітці в різних комбінаціях на мові CQL (Corpus Query Language). «У корпусній лінгвістиці під лемою зазвичай розуміється графічна одиниця без урахування лексичних відмінностей. Таким чином, лемою стала називатися «початкова форма», а процес автоматичного приписування тексту форми до певної «початкової форми» отримала в корпусній лінгвістиці назву лематизації». [27, 43].

Опція Конкордансер (англ. Concordancer) – застосунок для автоматичного створення конкордансів.

Конкордансі використовуються для вирішення наступних лінгвістичних завдань [53]:

1) порівняння різних використань одного слова;
2) аналіз ключових слів;

3) аналіз частотності слів і словосполучень;
4) пошук і дослідження фраз і ідіом;
5) пошук перекладу, наприклад, термінології;
6) створення списків слів.

У конкордансі мовного корпусу «вказується абсолютна частота лексеми, тобто кількість всіх її вживань у всіх розглянутих текстах [53, 62]. За необхідності дається розбивка за основними типами/жанрами текстів. Багато сучасних корпусів пропонують конкорданс в якості додаткової можливості виведення знайденої інформації на екран. Такий формат виведення називається KWIC (англ. Key word in context). «KWIC - формат виведення конкордансу на екран, таким чином, що шукані слова розташовуються по центру в один стовпчик, що полегшує їх швидкий перегляд і аналіз. Взагалі, конкорданс - не стільки корпус, скільки формат показу результатів». [27, 10]

Інструмент Sketch Engine - Word Sketch надає дані про граматичну і лексичну сполучуваність слова [52, 9].

На основі морфологічно розміченого корпусу, дана система породжує списки слів, в яких міститься інформація про їх «лінгвістичну поведінку», тобто сполучуваність з іншими словами з кількісним зазначенням сили зв'язку, яка розраховується на основі відомих мір асоціації [54, 2].

З Word Sketch відсутня необхідність переглядати один за одним знайдені в корпусі приклади. Це універсальний інструмент, що дозволяє користувачам заощаджувати свій час [Sketch Engine, URL].

Перейдемо до корпуснії пошукової системи Sketch Engine, за допомогою якої виконана практична частина нашого дослідження. Особливість системи Sketch Engine полягає в тому, що в ній є засоби, що реалізують методику дистрибутивно-статистичного аналізу - тезаурус, кластеризація і диференціація. Кластеризацією називають угруповання слів в

опціях дистрибутивного тезауруса або в Word Sketch. Вони виявляють парадигматичні зв'язки між термінами з кількісним зазначенням сили зв'язку. За допомогою дистрибутивного тезауруса є можливість вирішити величезну кількість лінгвістичних завдань. Дистрибутивний тезаурус дозволяє побачити, які слова зустрічаються з тими ж словами, що і обране ключове слово. Таким чином, користувач отримує дані про те, які слова в корпусі мають дистрибуцію, схожу з заданим словом [52, 15]. Для обчислення подібності слів розглядаються набори списків сполучуваності для слів X1 і X2. Схожість дистрибуції слів вираховується на основі статистичної міри logDice. Неінформативні випадки, для яких значення міри є негативним, відкидаються. Розглядаються словосполучення, в яких слова X1 і X2 зустрічаються в однакових граматичних контекстах.

«Міра стійкості (англ. Lexical association measure) - величина, що показує, наскільки випадкові або стійкі поєднання одиниць в складі колокації або коллігації. Для її обчислення використовується безліч різних методів, наприклад: MIscore, t-score та ін. Варто також відзначити, що «в зазначеній системі доступні наступні міри – t-score (t- test), MI (mutual information), MI3, MI -log-prod (MI.log_f), minimum sensitivity, loglikelihood ratio, Dice (logDice), MI.log_f.

«Токенізація (англ. Tokenization) - виділення в текстовому потоці мінімальних фрагментів для подальшого аналізу (в корпусних лінгвістиці їх прийнято називати токени (англ. token) »[27, 40].

**Висновки до розділу 2**

Перед створенням корпусу потрібно вирішити, чи відповідає корпус конкретному дослідницькому питанню. Також потрібно дбати про репрезентативність корпусу. З обережністю інтерпретувати результати

дослідження корпусу, враховуючи, чи ввідні дані корпусу та метод, використаний у дослідженні, відповідають вимогам.

Формат кодування інформації для корпусу текстів має відповідати ряду принципів. А саме, це зрозумілість та прозорість викладу – потрібно формулювати чіткі вимоги до метаданих, які залучені в корпусах, доступно класифікувати параметри розмітки для опрацювання широкого спектру текстової інформації.

Документувати критерії проєктування та зробити документацію доступною для користувачів корпусу. Компактність – відсутність надлишкової розмітки, важливо чітко розуміти які критерії розмітки краще допоможуть розкрити тему дослідження. Також має бути присутня узгодженість з використаним програмним забезпеченням та легкість конвертування в інші формати.

# РОЗДІЛ 3 СТВОРЕННЯ КОРПУСУ ТЕКСТІВ УКРАЇНСЬКОЇ ФАНТАСТИКИ

## 3.1 Характеристика та збір матеріалу

Фантастика – жанр художніх творів, в якому за допомогою вигаданих уявних елементів створюються нові світи, раси, прилади, суспільні правила і порядки, відмінні від реальності.

Як відзначає Т. Чернишова, "фантастика не суперечить жодному літературному методу, вона може “поступити на службу" і до романтизму, і до реалізму, і до модернізму" [55, с. 21].

Найчастіше у фантастичних романах, на думку О. С. Бочкової, зустрічаються терміни, що позначають. [56, 8–10]

- приміщення та їх складові частини, місто та його частини. Висока частотність в текстах фантастики цієї групи термінів природна для наукової фантастики як жанру. Детальний опис приміщень пов'язаний з урбанізованим ландшафтом, оскільки основні події в досліджуваних текстах наукової фантастики розгортаються в «інтер'єрах», а не в «пейзажах». Детальна розробка простору використовується авторами для створення ефекту реалістичності, достовірності описуваного;
- технічні пристрої і транспорт. Характерною стилістичною особливістю текстів наукової фантастики слугують номінації різних технічних пристроїв. Саме вони багато в чому створюють «внутрішню логічність реальності» даних текстів. В текстах наукової фантастики вживаються назви існуючих у нашій дійсності засобів транспорту і технічних пристроїв (космічний корабель, літак, тролейбус, схеми-тригери) і вигаданих (рухомі доріжки, труба швидкісного підйому, комунікаційна трубка, нейтралізатор, машина-матка);
- топоніми. Ще в древніх міфах географічна і етнографічна конкретність була їх невід'ємною частиною. У цю групу входять як реально існуючі топоніми: Нью-Йорк, Владивосток, так і вигадані для створення картини фантастичного світу. Арканар, Гикаючий ліс. Топоніми в

тексті виконують текстотворчу функцію. З одного боку, автор таким чином конкретизує описувані події, наближаючи описуваний світ до дійсності, роблячи його замкнутим простором, а з іншого боку, топонім є згорнутим «текстом у тексті», в даному випадку будучи знаком фантастичного, ще не освоєного людиною і незнайомого йому світу:

- частини людського тіла. Ця лексична група, що належить до основного словарного фонду мови, численна у всіх досліджуваних текстах наукової фантастики: «Він носив окуляри, тому що його очі не переносили звичайних контактних лінз. Тільки звикнувши до окулярів, можна було розгледіти риси його пересічної, невиразної особи» (Азімов, «Сталеві печери»). З точки зору О. С. Бочкової, зовнішні межі тіла й душі людини - центр просторового розташування і ціннісного осмислення зображуваних у творі зовнішніх предметів. [56]

Для початку був проведений збір матеріалу, за основу взяті тексти творів жанру фантастики, які були написані в оригіналі українською. Хронологія творів охоплює 2 періоди: українсько-радянський період (1933-1991) та новітній український (з 1991 року). Найстаріший текст – 1934, найновіший – 2018.

Відібрані для корпусу твори являють собою різноманітні жанри фантастики: жорстка наукова фантастика, антиутопія, кіберпанк, постапокаліптика, містична проза, темне фентезі, фантастика жахів, наукова фантастика, соціальна фантастика, пригодницька фантастика, гумористична фантастика, окрім жанру дитячого фентезі.

Обрані автори:

Авраменко Олег Євгенович і Авраменко Валентин Євгенович;

Андрощук Іван Кузьмович;

Антипович Тарас Георгійович;

Арєнєв Володимир Костянтинович;

Бакалець Ярослава і Яріш Ярослав;

Базь Любов Олександрівна;

Бедзик Дмитро Іванович;

Бедзик Юрій Дмитрович;

Бердник Громовиця;

Бердник Олесь Павлович;

Бережний Василь Павлович;

Бжехва Ян;

Білий Дмитро Дмитрович;

Валентинов Андрій;

Верховський Валерій Дмитрович;

Винничук Юрій Павлович;

Владко Володимир Миколайович;

Волков Віталій;

Волков Олексій та Олександр;

Воронина Леся;

Вухналь Юрій;

Гайдамака Наталія Лук'янівна;

Гранецька Вікторія Леонідівна;

Грибенко Володимир Іванович;

Ганапольський Матвій;

Громов Дмитро і Ладиженський Олег;

Герасименко Юрій Георгійович;

Дев’ятко Наталія Володимирівна;

Декань Олексій Іванович;

Дереш Любко;

Дімаров Анатолій Андрійович;

Дубинянська Яна Юріївна;

Дурєєв Олександр Михайлович;

Дмитрук Андрій;

Дяченки Марина та Сергій;

Єфрємов Іван;

Єшкілєв Володимир Львович;

Завара Олександр Васильович;

Зима Олександр Вікторович;

Золотько Олександр Карлович;

Ільченко Олександр Єлисейович;

Ірванець Олександр Васильович;

Кай Ольга;

Капій Мирослав Дмитрович;

Кацай Олексій Опанасович;

Каторож Ярина Славомирівна;

Копань Лариса Юріївна;

Котляревський Іван Петрович;

Кожеленко Василь Дмитрович;

Корепанов, Дурєєв, Покальчук;

Корній Дара;

Кралюк Петро Михайлович;

Крижевський Андрій;

Кузьменко Володимир Леонідович;

Кацай Олексій;

Ларін Михайло Васильович;

Легеза Сергій Валерійович;

Литовченко Олена та Тимур Литовченко;

Малина Маріанна;

Мельник Ярослав;

Микитчак Тарас і Микитчак Галина;

Михановський Володимир;

Михед Олександр Павлович;

Назаренко Михайло Йосипович;

Околітенко Наталя Іванівна;

Олініченко Олександр Валерійович;

Осійчук Ореста;

Павловський Станіслав Степанович;

Положій Віктор Іванов;

Полонський Радій Федорович;

Потаніна Ірина Сергіївна;

Радутний Радій Володимирович;

Росохватський Ігор Маркович;

Романчук Олег;

Руденко Микола;

Савченко Віктор Васильович;

Соколюк Остап;

Соколян Марина Іванівна;

Смолич Юрій;

Тендюк Леонід Михайлович;

Тисовська Наталя;

Уткін Володимир Сергійович;

Шевчук Валерій;

Чемерис Валентин;

Чигиринська Ольга Олександрівна;

Ющук Іван Пилипович;

Ячейкін Юрій Дмитрович.

### 3.2 Попередня обробка тексту та створення корпусу на сервері NoSketch Engine

Після того, як джерела були визначені, а тексти зібрані у електронних або на паперових носіях, настав час оцифрувати всю літературу.

Паперові джерела було розпізнано за допомогою додатку ABBYY Finereader 12 та збережено у форматі *.docs, а щодо електронних джерел – усі тексти у форматах pdf, djvu, epub, fb2 також було конвертовано у формат *.docs.

Попередня обробка тексту полягала в тому, що матеріал пройшов філологічну перевірку, коректування тексту, видалення нетекстових елементів, а саме зображень та «шуму», видалення з тексту переносів, забезпечення однакового написання тире. Також була здійснена підготовка бібліографічного та екстралінгвістичного опису тексту та сегментування тексту на його структурні складники.


Рис.1

Усі «чисті» тексти було підготовлено у вигляді вертикальних файлів, тобто усі слова та розділові знаки, за допомогою макросів заміни, були розставлені у стовпчики. У додаток EmEditor ми додали збережені тексти у форматі *.txt у кодуванні Юнікод (UTF-8): intermezzo_vert.txt. Було використано набір регулярних виразів для обробки текстів в EmEditor.

Етапи обробки тексту в додатку EmEditor:

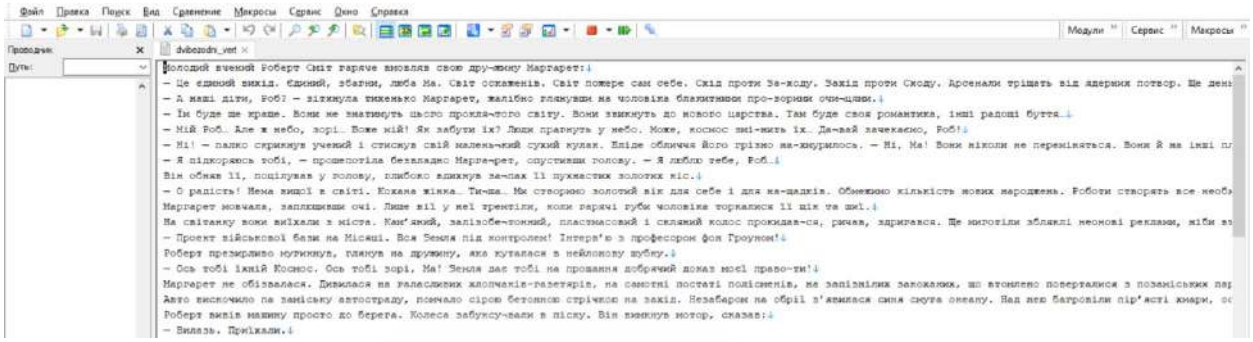

Рис. 2

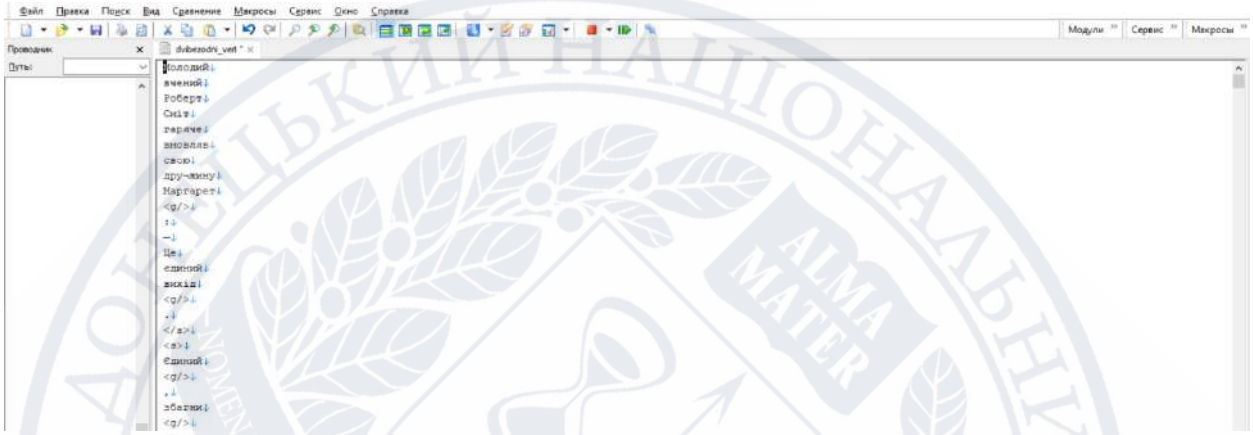

Рис. 3

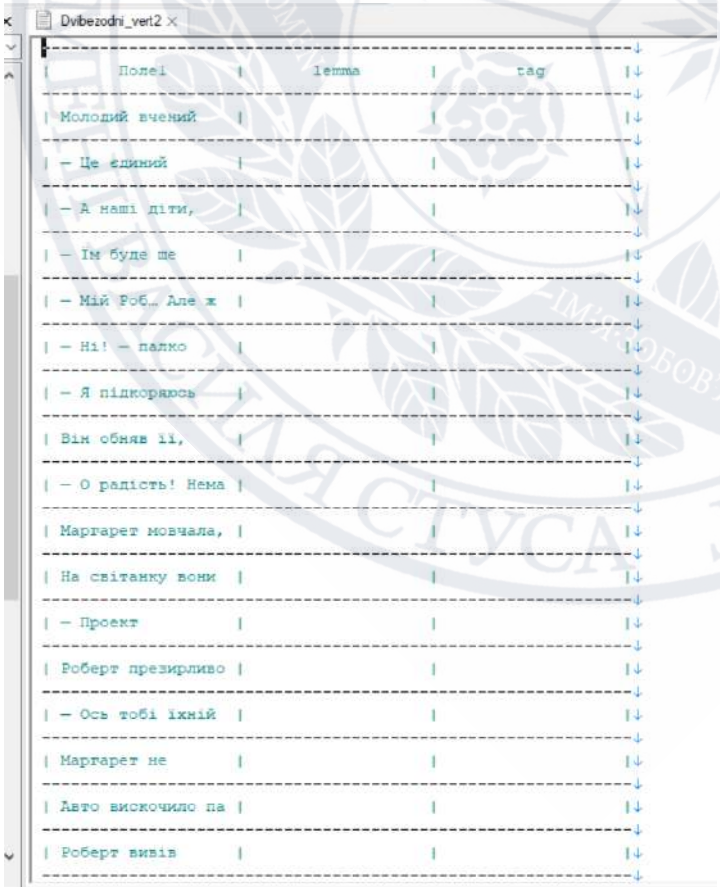


Рис. 4

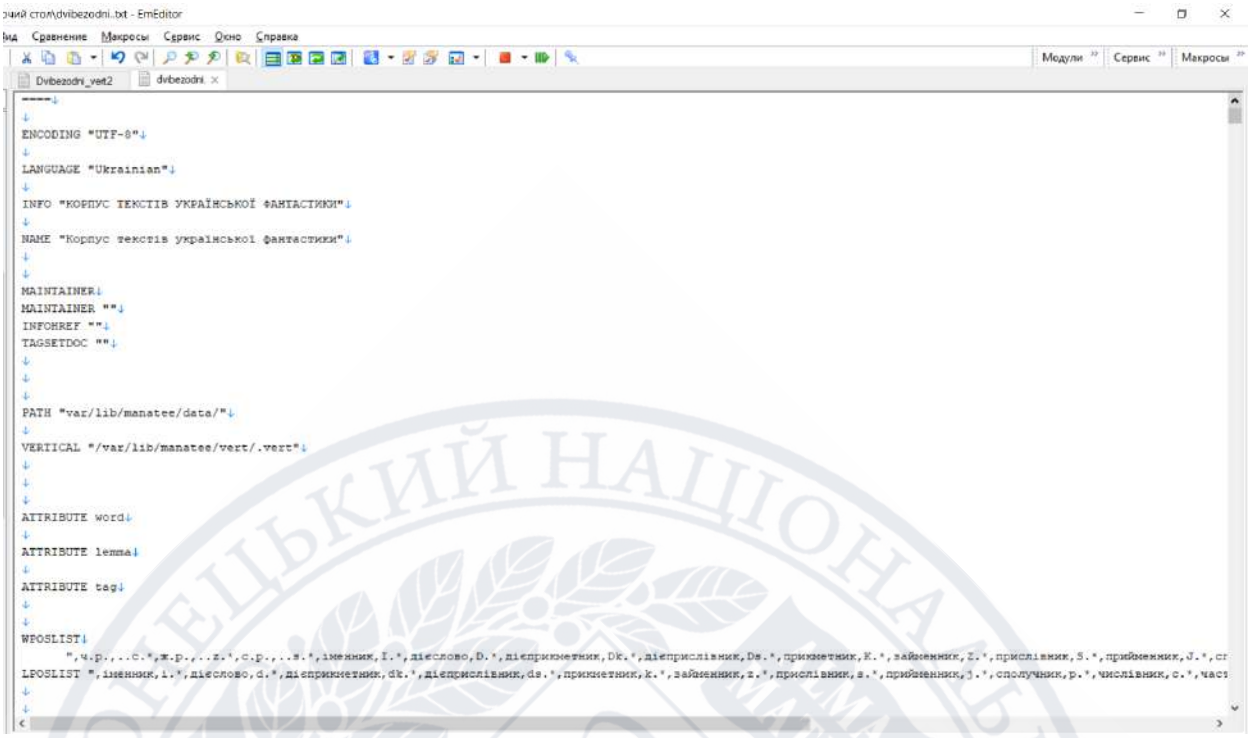


Рис. 5

Згодом було зроблено автоматичний морфологічний аналіз (тегування) у БД ЛематизаторПовний2_3.accdb. Ми імпортували вертикальний файл у базу даних, без обмежувачів, кодування Юнікод (UTF-8), з роздільником – табуляцією. Редагували таблицю додавши поля з іменам lemma і tag, потім оновили відповідними полями з таблиці Слова за допомогою запиту оновлення:

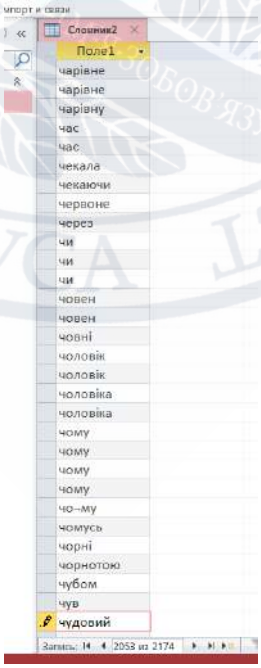


Рис. 6

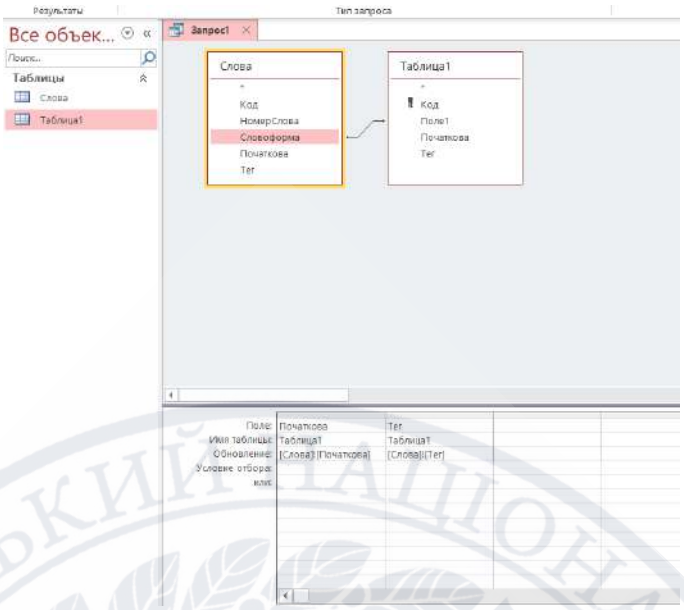


Рис. 7

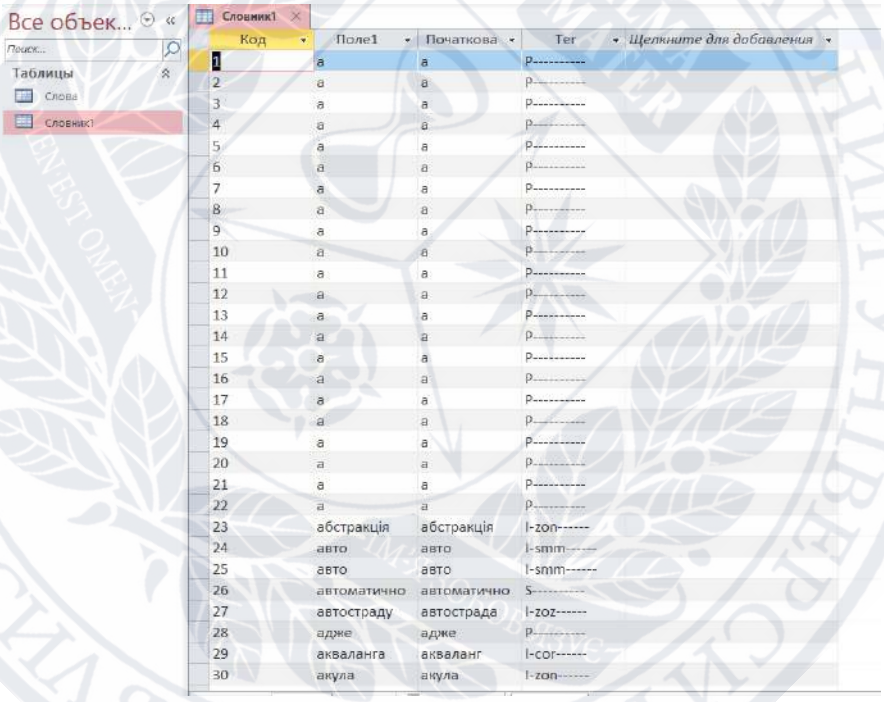


Рис. 8

Експортували таблицю у текстовий файл та додали перед кожним текстом у корпусі рядки з потрібними атрибутами (метарозміткою): автор, назву твору, рік видання, період, жанр, піджанр, стиль.

<doc author="Олесь Бердник" date="1990" genre="сучасне фантастичне оповідання">

<s>

</s>

</doc>

====

Структурна розмітка КТ граматичної службовості наразі включає дані тільки про межі тексту, це тег <doc> – </doc>, і межі речень у ньому тег <s> – </s>. Спеціальний тег <g/> позначає розділові знаки, перед якими не ставиться пробіл – це всі, крім тире. У перспективі маємо намір додати автоматично виставлені позначки абзаців, параграфів, розділів, частин тощо. [57]

====

ENCODING "UTF-8"

LANGUAGE "Ukrainian"

INFO "КОРПУС МІКРОТЕКСТІВ НА ПРИКЛАДІ ЗБІРКИ Подвиг Вайвасвати: Фантастичні повісті та новели. – К.: Молодь, 1967. – 244 с.

Зміст: Подвиг Вайвасати: Роман-легенда. – Вогняний вершник: Космічна повість. – Дві безодні: Новела. – Хор елементів: Новела. / Укладач: Карпенко Аліна"

NAME "Корпус текстів української фантастики"

====

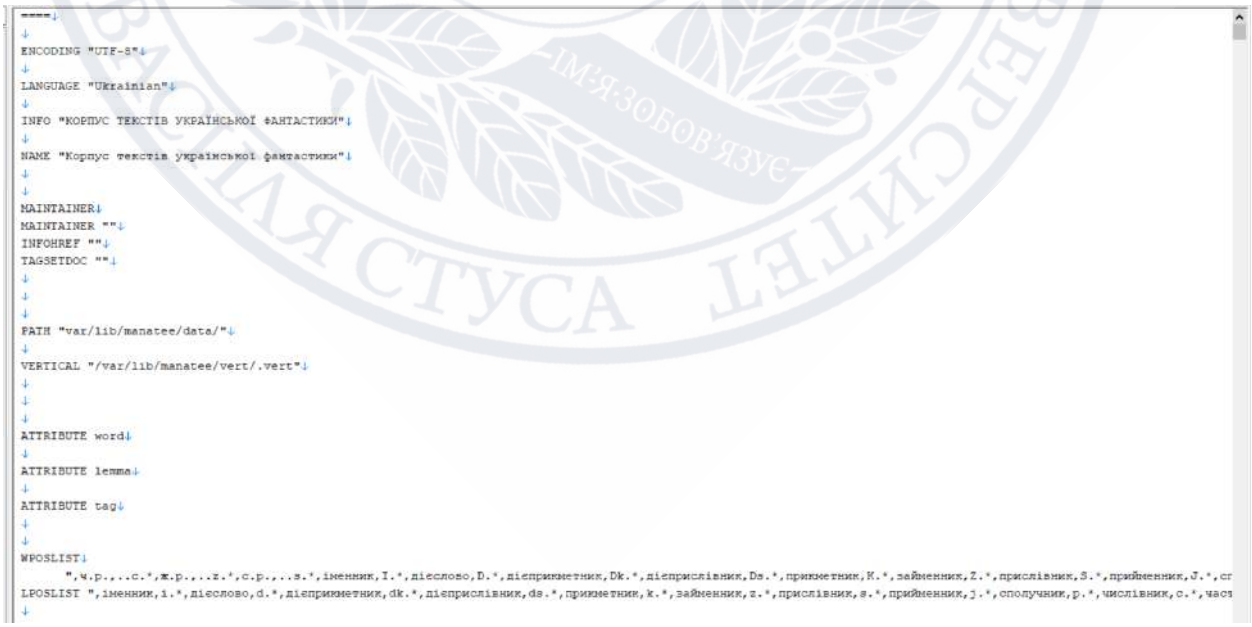


Рис. 9

На сервері NoSketch Engine ми розташували файли налаштувань у /nlp/corpora/manatee/registry, /var/lib/manatee/registry. А вертикальний файл у /var/lib/manatee/vert.

Було відредаговано файл var/www/bonito2/run.cgi

the line corplist = ['susanne', 'proba1'] contains a list of available corpora

the line corpname = 'bnc' sets the default corpus

the line os.environ['MANATEE_REGISTRY'] = '/corpora/registry' is the path to the directory with corpus configuration files

Також було використано спеціальні роззначення у тексті:

<s>...</s> - початок і кінець речення;

<q>...</q>- початок і кінець мовлення персонажів; початок і кінець цитати.

<poetry>...</poetry> — початок і кінець поезії.

При роззначенні треба пам'ятати, що жодні позначки не повинні перехрещуватись, а бути лише вкладені одна в одну - <tag1> <tag2> ... <tag3>...</tag3> ... <tag3> ... </tag3> … </tag2>...</tagl> … </tag2> … </tag3> ... <tag2>... </tag1>...

**Висновки до розділу 3**

Корпусна лінгвістика - одна з методик, яка найбільш стрімко розвивається в сучасному теоретичному і прикладному мовознавстві.

Корпус текстів української фантастики добре «пошарово» презентує уявлення людей про сьогодення та майбутнє, демонструє показність слововживань, популярність слів, вживання неологізмів. Показним є те, що у українсько-радянський період багато вжито слів на тему сім’ї, був досить консервативний погляд на майбутнє, але з проблисками надії, поширеними були наукова фантастика та антиутопія. В період 1980-2000 теж досить часто вживана романтика, часто вживані епітети щодо смутку, а з 2000 по 2018-2019-ті роки збільшилась кількість пригодницького, містичного та комедійного фентезі.

Велика перевага корпусних менеджерів в тому, що у порівнянні з окремими, самостійно зробленими, корпусами текстів, працювати з ними

набагато простіше, адже не потрібно опановувати специфічну символьну мову довільного корпусу. Достатньо засвоїти стандартизовані спрощені команди та помітки. NoSketch Engine – потужний інструмент для створення свого власного корпусу текстів (підкорпусу), або ж для завантаження існуючих масивів даних. Система дає можливість сформувати частотний словник та згрупувати лексичні одиниці в лексико-семантичні поля.

# ВИСНОВКИ

На сучасному етапі корпусна лінгвістика не достатньо розвинена як самостійна наука, але має глобальні перспективи в подальших дослідженнях. Оскільки створення повноцінних корпусів, які б відповідали основним вимогам знаходиться на стадії формування, особливо це стосується корпусів текстів української мови.

В своїх дослідженнях корпусна лінгвістика спирається на дані корпусу та квантитативні методи. Створення корпусу базується й на емпіричних методах. Текст розглядається як певна фізична сутність у глобальній перспективі. Фокусує увагу на якомога ширшому погляді на текст, який не обмежений догмами. Проводиться робота з лінгвістичними даними (слововживаннями) в тому вигляді, в якому вони були вжиті в контексті. Надає перевагу індуктивним методам обробки емпіричного словесного матеріалу, вважає його сутністю наукового методу.

Можна зробити висновки, що корпусна лінгвістика - одна з методик, яка найбільш стрімко розвивається в сучасному теоретичному і прикладному мовознавстві. Це відносно новий лінгвістичний напрямок, який почав свій активний вплив в 60-х роках XX століття в зв'язку з інтенсивним розвитком комп'ютерних технологій.

У корпусній лінгвістиці використовується вже зібраний корпус. Для реалізації корпусів потрібно: представити структуру мовленнєвої діяльності, виявити, які матеріальні обмеження є на стадії створення корпусу, відібрати тексти та укладання корпусу текстів.

Також з’ясовано, що головною ознакою корпусу є його наочність, що належить до різноманітності мови, зображеної в відповідних, природних пропорціях. Колекція текстів повинна бути збалансованою і мати достатню вибірку по числу текстів та авторів, щоб служити основою для статистично перевірених досліджень лінгвістичних феноменів. Хоча мова динамічна та нескінченна, але корпус все ж повинен бути обмежений за розміром, тож ми

беремо вибірку та пропорційно вводимо широкий спектр типів тексту, щоб забезпечити максимальний баланс та репрезентативність.

У дослідженні виявлено, що формат кодування інформації для корпусу текстів має відповідати ряду принципів. А саме, це зрозумілість та прозорість викладу – потрібно формулювати чіткі вимоги до метаданих, які залучені в корпусах, доступно класифікувати параметри розмітки для опрацювання широкого спектру текстової інформації. Важливим принципом є й компактність (відсутність надлишкової розмітки, важливо чітко розуміти які критерії розмітки краще допоможуть розкрити тему дослідження). Також має бути присутня узгодженість з використаним програмним забезпеченням та легкість конвертування в інші формати.

Для роботи з корпусом текстів української фантастики зібрано матеріал на основі творів жанру фантастики, що написані в оригіналі українською мовою та має таку періодику: 1933-2018 років. У корпусі текстів розміщено розподіл різноманітних жанрів фантастики, а саме: антиутопія, наукова фантастика, містична проза, кіберпанк, постапокаліптика, жорстка наукова фантастика, пригодницька фантастика, темне фентезі, фантастика жахів, соціальна та гумористична фантастика.

Для створення корпусу української фантастики було обрано твори таких українських авторів: Авраменко О. Є., Авраменко В. Є., Андрощук І. К., Антипович Т. Г., Арєнєв В. К., Бакалець Я., Яріш Я., Базь Л. О., Бедзик Д. І., Бедзик Ю. Д., Бердник Громовиця, Бердник О. П., Бережний В. П., Бжехва Я., Білий Д. Д., Валентинов А., Верховський В. Д., Винничук Ю. П., Владко В. М., Волков В., Волков Олексій та Олександр, Воронина Л., Вухналь Ю., Гайдамака Н. Л., Гранецька В. Л., Грибенко В. І., Ганапольський М., Громов Д., Ладиженський О., Герасименко Ю. Г., Дев’ятко Н. В., Декань О. І., Дереш Любко, Дімаров А. А., Дубинянська Я. Ю., Дурєєв О. М., Дмитрук А., Дяченко М., Дяченко С., Єфрємов І., Єшкілєв В. Л., Завара О. В., Зима О. В., Золотько О. К., Ільченко О. Є., Ірванець О. В., Кай О., Капій М. Д., Кацай О. О., Каторож Я. С., Копань Л. Ю., Котляревський І. П., Кожеленко В. Д.,

Корепанов, Дурєєв, Покальчук, Корній Дара, Кралюк П. М., Крижевський А., Кузьменко В. Л.,

Кацай О., Ларін М. В., Легеза С. В., Литовченко О., та Литовченко Т., Малина М., Мельник Я., Микитчак Т., Микитчак Г., Михановський В., Михед О. П., Назаренко М. Й., Околітенко Н. І. Олініченко О. В., Осійчук О., Павловський С. С., Положій В., Іванов П., Радій Федорович, Потаніна І. С., Радутний Р. В., Росохватський І. М., Романчук О., Руденко М., Савченко В. В., Соколюк О., Чемерис В., Чигиринська О. О., Ющук І. П., Ячейкін Ю. Д.

Алгоритм створення корпусу текстів української фантастики включає такі етапи: 1) попередня обробка тексту та створення вертикальних файлів в EmEditor; 2) автоматичний морфологічний аналіз (тегування) у БД ЛематизаторПовний2_3.accdb; 3) створення корпусу на сервері NoSketch Engine.

Алгоритм реалізовано за допомогою додатку ABBYY Finereader 12, додатку EmEditor та серверу NoSketch Engine. Також допоміжним у створенні корпусу текстів для проведення автоматичного морфологічного аналізу послугувала БД ЛематизаторПовний2_3.accdb.

## ВИКОРИСТАНІ ДЖЕРЕЛА


1. Захаров В. П., Богданова С. Ю. Корпусная лингвистика. Иркутск: ИГЛУ, 2011. 161 с.
2. McEnery, A. M. and Wilson, A. (2001). Corpus linguistics: an introduction. Edinburgh University Press.
3. Жуковська В. В. Вступ до корпусної лінгвістики: навч. посіб. Житомир: Вид-во ЖДУ імені І. Франка, 2013. 142 с.
4. Aarts, Jan and Willem Meijs (eds.). 1984. Corpus linguistics: Recent develop-ments in the use of computer corpora in English language research.Amsterdam: Rodopi.
5. Aarts, Jan and Theo van den Heuvel. 1982. Grammars and intuitions in corpuslinguistics. In S. Johansson (ed.). Computer corpora in English languageresearch, 66–84. Bergen: Norwegian Computing Centre for Humanities.
6. BELMORE@vax2.concordia.ca: "Corpora: First use of the term 'corpus linguistics'"
7. Leech, Geoffrey. 1992. Corpora and theories of linguistic performance. In J.Svartvik (ed.). 105–122.
8. Stubbs, Michael. 1993. British traditions in text analysis: From Firth to Sinclair.In M. Baker, F. Francis and E. Tognini-Bonelli (eds.). Text and technology:In honour of John Sinclair, 1–36. Amsterdam: John Benjamins.
9. Teubert, Wolfgang. 2005. My version of corpus linguistics. International Jour-nal of Corpus Linguistics 10(1): 1–13.
10. Січінава Д. В. Паралельні українсько-російський та російсько-український корпуси / Д. В. Січінава, О. О. Тищенко-Монастирська, М. О. Шведова // Лексикографічний бюлетень, 20. – К., 2011. – С. 35-38.
11. Gries, Stefan Th. 2006a. Some proposals towards more rigorous corpus linguis-tics. Zeitschrift für Anglistik und Amerikanistik 54(2): 191–202.
12. Tognini-Bonelli, Elena. 2001. Corpus linguistics at work. Amsterdam: JohnBenjamins.)

13. Mahlberg, Michaela. 2005. English general nouns: A corpus theoreticalapproach. Amsterdam: John Benjamins
14. Mahlberg, Michaela. 2006. Lexical cohesion: Corpus linguistic theory and itsapplication in English language teaching. International Journal of CorpusLinguistics 11(3): 363–383.
15. Thompson, Geoffrey and Susan Hunston (eds.). 2006. System and corpus:Exploring connections. London: Equinox.
16. Корпусная лингвистика vs лингвистика стереотипного и творческого СЛ Мишланова, ТМ Пермякова, ЕЮ Аликина - Стереотипность и творчество в тексте, 2011.
17. Малышева Н. В. Квантитативная лингвистика в современной научной парадигме // Научно-методический электронный журнал «Концепт». – 2014. – Т. 20. – С. 546–550. – URL: http://e-koncept.ru/2014/54373.htm.
18. Захаров В.П. Корпусная лингвистика: учеб.-метод. Пособие / В.П. Захаров. – СПб., 2005. – 48 с.)
19. Шаров С.А. Представительный корпус русского языка в контексте мирового опыта / С.А. Шаров // Научно–техническая информация. – Сер. 2. – – № 6. – С. 12–16.)
20. Andor 2004: 97 the master and his performance.
21. Gonzalez-Marquez M, Becker R, Cutting J. An introduction to experimental methods for language researchers. In: Gonzalez-Marquez M, Mittelberg I, Coulson S, Spivey M, eds. Methods in cognitive linguistics. 2007: 53-86.
22. Fillmore 1992: 35 Corpus linguistics or computer-aided armchair
23. Кочерган М. П. Загальне мовознавство: підручник / Михайло Петрович Кочерган. — Київ: Академія, 2003. — С. 398.]
24. Working with specialized language: A practical guide to using corpora* L Bowker, J Pearson – 2002
25. Leech, G. 1991.”The state of the art in corpus linguistics”. En K. Aijmer & B. Altenberg (eds.) 1991. English Corpus Linguistics. London: Longman; 8-29.

26. Kennedy, G. (1998). An Introduction to Corpus Linguistics: Studies in Language and Linguistics. London: Longman.
27. Копотев 2014: Копотев, М. В. Введение в корпусную лингвистику: Учебное пособие для студентов филологических и лингвистических специальностей университетов. Praha, 2014. Accessed 15. 09. 2015. https://play.google.com/books/reader?printsec=frontcover&output=reader&id=9IxUBQAAQBAJ&pg=GBS.PP1.).
28. Апресян, Ю.Д. Избранные труды. [Текст] / Д.Ю. Апресян. – М.: Школа «Языки русской культуры», 1995. – Т. II: Интегральное описание языка и системная лексикография. – 767 с.
29. Плунгян В. А. Корпус как инструмент и как идеология: о некоторых уроках современной корпусной лингвистики // Русский язык в научном освещении. — 2008. — № 16 (2). — С. 7–20.
30. Klosa, A. (2007). Korpusgestützte Lexikographie: besser, schneller, umfangreicher. In W. Kallmeyer & G. Zifonun (Eds.), Sprachkorpora. Datenmengen und Erkenntnisfortschritt (pp. 105-122). Walter de Gruyter.
31. Баранов А.Н. Проблема репрезентативности корпуса данных (на примере политической метафорики) // Труды Международного семинара Диалог ‘2001 по компьютерной лингвистике и ее приложениям. – Аксаково, 2001 г.
32. Різновиди корпусу текстів у процесі перекладу документів офіційно-ділового стилю / Ю. І. Дем’янчук // Науковий вісник Дрогобицького державного педагогічного університету імені Івана Франка. Сер. : Філологічні науки (мовознавство). - 2016. - № 5(1). - С. 104-107. - Режим доступу: http://nbuv.gov.ua/UJRN/nvddpufm_2016_5%281%29__27
33. Вадяев С Е Электронная лексикография и корпусная лингвистика // Аспекты становления и функционирования запад1югерманских языков - Самара Изд-во «Самарский университет», 2003 - С 83-92

34. Демська-Кульчицька О. М. Що таке корпус текстів? / Орися Демська-Кульчицька // Культура слова. - 2004. - № 64. - С. 35-38. http://www.ekmair.ukma.edu.ua/handle/123456789/1690)
35. Дарчук Н. П. Корпусна лінгвістика: проблеми, методи, перспективи (робоча навчальна програма для аспірантів) / Н. П. Дарчук. – К. : КНУ імені Тараса Шевченка, 2013. – 11 с.
36. Zhukovs'ka V. V. Vstup do korpusnoyi linhvistyky: navchal'nyy posibnyk [Introduction to Core Linguistics: Textbook] / V. V. Zhukovs'ka. – Zhytomyr : Vyd-vo ZHDU im. I. Franka, 2013. – 142 s.)
37. Jones C.L., Bridges R.A., Huffer K.M., Goodall J.R. (2015), Towards a relation extraction framework for cyber-security concepts. In Proceedings of the 10th Annual Cyber and Information Security Research Conference, p. 11.
38. Joshi A., Lal R., Finin T., Joshi A. (2013), Extracting cybersecurity related linked data from text. In Semantic Computing (ICSC), 2013 IEEE Seventh International Conference, pp. 252–259.
39. Lim S.K., Muis A.O., Lu W., Ong C.H. (2017), MalwareTextDb: A database for annotated malware articles. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vol. 1, pp. 1557–1567.
40. Демська-Кульчицька О. Основи національного корпусу української мови : [монографія] / Орися Демська-Кульчицька. – К. : Інститут української мови НАНУ, 2005. – 219 с.
41. Burnard Lou. Metadata for corpus work [Electronic Resource]. – 2004. – Mode of Access : http://users.ox.ac.uk/~lou/wip/metadata.html
42. Леміш Н. Є. Корпусна метарозмітка спеціальних текстів з лінгвоантропогенезу / Н. Є. Леміш // Науковий часопис Національного педагогічного університету імені М. П. Драгоманова. Серія 9 : Сучасні тенденції розвитку мов. - 2017. - Вип. 15. - С. 139-152. - Режим доступу: http://nbuv.gov.ua/UJRN/Nchnpu_9_2017_15_18.

43. Burnard L. Metadata for Corpus Work / L. Burnard. – [Access Mode] : http://users.ox.ac.uk/~ lou/wip/ metadata.html
44. Сучасні корпуси текстів: вимоги до метаданих / В. Г. Шкляревський // Науковий вісник кафедри ЮНЕСКО Київського національного лінгвістичного університету. Філологія, педагогіка, психологія. - 2015. - Вип. 30. - С. 241-246. - Режим доступу: http://nbuv.gov.ua/UJRN/Nvkyu_2015_30_37 )
45. Wittenburg P. Metadata Proposals for Corpora and Lexica / P. Wittenburg, W. Peters, B. Broeder // LREC, 2002. – Max-Planck-Institute for Psycholinguistics. – [Access Mode] : http://www.mpi.nl/IMDI/documents/2002%20LREC/Metadata%20Proposals%20for%20Corpora%20and%20Lexica.pdf
46. Biber, Douglas (1993). Representativeness in Corpus Design. Literary & Linguistic Computing 8:4:243-257.
47. Ганиева И.Ф. Об использовании корпусов в лингвистических исследованиях. Башкирский государственный университет. 2006. № 4. С.104–106.
48. Демська-Кульчицька О.М. Репрезентативність як ознака текстового корпусу / О.М. Демська-Кульчицька. – Українська мова. – №3, 2005. – С. 100-107.
49. Tognini-Bonelli, Elena (2001). Corpus Linguistics at Work. Amsterdam: John Benjamins
50. Spärck Jones and Van Rijsbergen, 1976).( Spärck Jones, Karen and Van Rijsbergen, C. J. (1976). Information Retrieval Test Collections. Journal of Documentation 32:1:59-75
51. Friedman, Charles P. and Wyatt, Jeremy C. (1997). Evaluation Methods in Medical Informatics. NY: Springer. (Computers in Medicine series
52. Kilgarriff A. The Sketch Engine: Ten Years On / A. Kilgarriff, V. Baisa, J. Bušta, M. Jakubíček, V. Kovář, J. Michelfeit, P. Rychlý, V. Suchomel // Lexicography ASIALEX. 2014. Vol. 1. Pp. 7-36. URL: http://link.springer.com/article/10.1007/s40607-014-0009-9
53. Добрынина К.С. О методике работы над конкордансами //Филологические науки. Вопросы теории и практики, № 1 (12) 2012.
54. Хохлова М.В. Лексико-синтаксические шаблоны как инструмент выявления специальной лексики предметной области [Электронный

ресурс] // Материалы ежегодной Международной конференции «Диалог» (Бекасово, 30 мая – 3 июня 2012 г.). URL: http://www.dialog21.ru/digests/dialog2012/materials/pdf/Хохлова МВ.pdf

55. Чернишева Т. А. Природа фантастики / Татьяна Чернишева. Иркутск, 1985.
56. Бочкова О. С. Некоторые тексто-лингвистические особенности научной фантастики. Саратов Саратовский гос. социально-экономический ун-т, 2002. С 8-10.)
57. Корпус текстів для вивчення граматичної службовості / І. Данилюк // Лінгвістичні студії. - 2013. - Вип. 26. - С. 224-230. - Режим доступу: http://nbuv.gov.ua/UJRN/lingst_2013_26_46

# ДОДАТКИ

Додаток А

Апробація роботи.

Копія сертифікату учасника Міжнародної наукової конференції «Науковий простір: актуальні питання, досягнення та інновації» (Харків, 2020)

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

Додаток Б

Апробація роботи.

Копія наукової статті

Карпенко, А. Загальні вимоги щодо проєктування корпусу текстів. Матеріали конференцій МЦГД, жовтень 2020, с. 46-48

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 української фантастики