УДК 81’4:001.895:37
DOI https://doi.org/10.32782/tps2663-4880/2024.33.2.25

ІННОВАЦІЙНІ ТЕХНОЛОГІЇ ВИВЧЕННЯ ЛІНГВІСТИЧНИХ КАТЕГОРІЙ ТА СТРУКТУР

INNOVATIVE TECHNOLOGIES FOR STUDYING LINGUISTIC CATEGORIES AND STRUCTURES

Сидоренко Л.М.,
orcid.org/0000-0002-6547-4050
старший викладач кафедри української мови, літератури та культури
Національного технічного університету України
«Київський політехнічний університет імені Ігоря Сікорського»

Ігнатенко І.П.,
orcid.org/0000-0003-3534-3575
старший викладач французької мови кафедри іноземних мов
Національної академії образотворчого мистецтва і архітектури

Плеханова Т.М.,
orcid.org/0000-0003-2639-0396
кандидат філологічних наук, доцент,
завідувач кафедри видавничої справи та редагування
Запорізького національного університету

У лінгвістиці вивчають мовні категорії, структури та закономірності мови, які використовуються для аналізу текстів і мовлення. У цій галузі також розвиваються методи аналізу та синтезу мовних структур, такі як синтаксичний аналіз, фонологія, морфологія тощо. Мета репрезентованої роботи полягає у розгляді лінгвістики крізь призму інноваційних технологій, які вивчають методи обробки та аналізу даних, зокрема текстової інформації. Методика роботи включала аналіз літератури та розробок комп’ютерних програм і алгоритмів для автоматичної обробки й аналізу мови, такі як пошукові системи, машинне навчання та обробка природної мови. Для визначення та відстеження функціонування інноваційних технологій вивчення лінгвістичних категорій і структур та для моделювання й обробки лінгвістичної інформації, в роботі розглянуто: лінгвістику, інформатику та формалізм. Результати роботи засвідчили, що лінгвістика, обмежена визначеними корпусами, вивчає дискурси, які виникають у певних умовах. Мова повинна відповідати матеріальності через чіткі та передбачувані правила функціонування. Це досягається через аналіз або визнання, що автоматично переходить від текстів до формального подання, яке можна використовувати для різноманітних застосувань. Висновок після розгляду чотирьох концептуальних рівнів аналізу (морфологічний, синтаксичний, семантичний та прагматика) – ці рівні дозволяють використовувати аналіз лінгвістичного функціонування для досягнення абстрактної структури, що ґрунтується на лінгвістичних теоріях, відповідних до мови, яку досліджують. Поняття теорії повинні бути деталізовані у формальному контексті для автоматизації аналізу. Аналізатори перетворюють кожну пропозицію у формалізовану одиницю – «вислів», взяту на обробку. Синтаксис конституентів виявляє синтагматичні зв’язки між компонентами та їх інтерпретацію згідно з логікою синтагм. Вибір семантики та прагматики повинен визначати зміст, що передається текстом, і базуватися на об’єктивних правилах і чіткому підході. Отже, використання інноваційних технологій в лінгвістиці передбачає, що процес автоматичного аналізу виконується за допомогою надійних і продуктивних алгоритмів. Ця система ґрунтується на аналізі, який розглядається як цілісна структура, що охоплює лінгвістику, формалізм та інформатику.

Ключові слова: штучний інтелект, машинне навчання, обробка природної мови, нейролінгвістичне програмування, корпус, аналізатори.

In linguistics, language categories, structures, and regularities are studied for the analysis of texts and speech. Methods of analysis and synthesis of language structures, such as syntactic analysis, phonology, morphology, etc., are also developed in this field. The aim of the presented work lies in examining linguistics through the lens of innovative technologies that study methods of data processing and analysis, including textual information. The methodology involved analyzing literature and developing computer programs and algorithms for automatic language processing and analysis, such as search engines, machine learning, and natural language processing. To determine and track the functioning of innovative technologies for studying linguistic categories and structures and for modeling and processing linguistic information, the work considered three areas: linguistics, informatics, and formalism. The results of the work showed that linguistics, limited by specified corpora, studies discourses that arise under certain conditions. Language must correspond to materiality through clear and predictable rules of functioning. This is achieved through analysis or recognition that automatically transitions from texts to formal representations that can be used for various applications. The conclusion after considering four conceptual levels of analysis (morphological, syntactic, semantic, and pragmatic) is that these levels allow for using the analysis of linguistic functioning to achieve an abstract structure based on linguistic theories relevant to the language being studied. The concepts of theory must be detailed in a formal context for analysis automation. Analyzers transform each proposition into a formalized unit «utterance» taken for processing. Constituent syntax reveals syntagmatic relationships between components and their interpretation according to the logic of syntagms. The selection of semantics and pragmat-

ics should determine the meaning conveyed by the text and be based on objective rules and a strict approach. Therefore, the use of innovative technologies in linguistics entails that the process of automatic analysis is carried out through reliable and productive algorithms. This system is based on analysis, which is considered as a holistic structure incorporating linguistics, formalism, and informatics.

Key words: artificial intelligence, machine learning, natural language processing, neurolinguistic programming, corpus, analysers.

Постановка проблеми. Використання інноваційних технологій вивчення лінгвістичних категорій і структур мають в собі застосування штучного інтелекту, машинного навчання, обробки природної мови та інших сучасних методів аналізу та розуміння мови. За допомогою комп'ютера для автоматичної обробки письмових текстів і усних виступів природною мовою визначається сфера обробки природної мови, або нейролінгвістичне програмування. Цю галузь, що знаходиться на стику між лінгвістикою та інформатикою, можна схематично описати вздовж великої осі, або комп'ютерної лінгвістики, або лінгвістичної інформатики, що визначає два різні підходи (рис. 1).

Комп'ютерна лінгвістика розпочинається з аналізу мовних явищ і переформульовує їх у вигляді алгоритмів для роботи з формальними моделями та їх оцінки. Ці алгоритми знайомі лінгвістам від часів Хомського, який описував ієрархію математичних моделей, призначених для опису мовних явищ, починаючи від раціональних граматик (найпростіша модель) і закінчуючи необмеженими граматиками (найпотужніша модель) [1]. Цей підхід сприяв створенню різноманітних комп'ютерних інструментів, таких як xfst, tag, lfg і hpsg. Лінгвістичні обчислення не потребують знання мови або лінгвістики. Вони використовують стохастичні моделі (ймовірнісні, статистичні та нейронні) для аналізу великих текстових корпусів і виявлення закономірностей [2, с. 561], які дозволяють робити обчислення та наводити результати.

Реалізація комп'ютерної лінгвістики потребує значних зусиль та витрат через необхідність аналізу й обробки мовних даних, що вимагає великої кількості робочого часу кваліфікованих фахівців. Тому багато дослідників зосереджують свою увагу на нейролінгвістичному програмуванні, яке використовує подібні програми, як наприклад, розпізнавання мови або машинний переклад, з успіхом в комерційному використанні. Внаслідок наукових відкриттів у сфері комп'ютерної лінгвістики виявився важливий інструмент для реалізації таких програм, що стало новою реальністю у цій галузі.

У сучасних дослідженнях лінгвістики спостерігається зміна в термінології, де поняття «лінгвістичні ресурси» вже не належать до словників або граматик, а до великих обсягів текстів, які мають обмежену мовну інформацію та недостатню достовірність. Наприклад, у Відкритому Американському Національному Корпусі можна знайти приклади складних слів, власних іменників, прислівників тощо [3, с. 66]. Використання граматичних правил для побудови таких корпусів може бути неточним. Тому важливо уникати загальних висновків щодо мовних феноменів. Європейські дослідники також активно працюють над розробкою подібних ресурсів, як

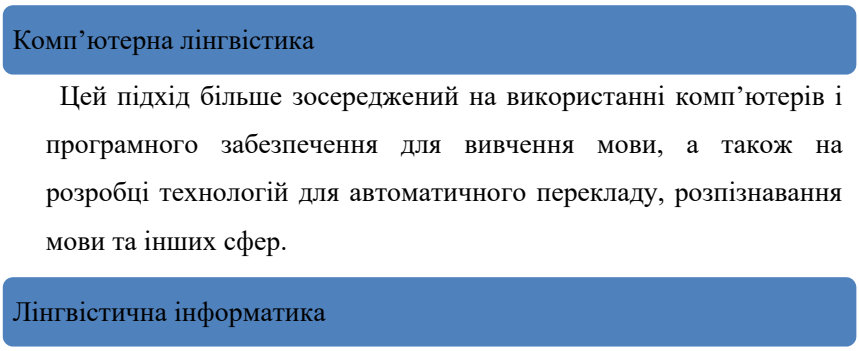

Рис. 1. Інноваційні підходи на перетині лінгвістики та інформатики

Джерело: власна розробка авторів.

SentiWordNet, де значення слів може бути вкрай важливим для визначення їх емоційної забарвленості [4].

Отже, в роботі досліджено стохастичні методи та інноваційні технології, які використовуються для лінгвістичного аналізу. Вони інколи демонструють проблематику, проте викликають велике зацікавлення у дослідників, які прагнуть зрозуміти, як функціонує мова .

Аналіз останніх досліджень та публікацій. Науковий розгляд взаємозв'язків між лінгвістикою та стохастичними методами є новаторським і полемічним. Все частіше автори вказують на важливість формалізації галузі знань для успішного вирішення проблем у науці [5].

У наш час багато досліджень в лінгвістиці використовують стохастичні підходи та аналіз корпусів текстів, які вважаються інноваційними в лінгвістиці, для вивчення різних лінгвістичних явищ [6, с. 168]. Наприклад, дослідники використовують стохастичні методи для аналізу синтаксичних структур речень, вивчення семантичних зв'язків між словами, аналізу вживання слів у різних контекстах та багато іншого [7, с. 1212]. У деяких дослідженнях також використовують стохастичні моделі для розв'язання конкретних лінгвістичних завдань, наприклад, класифікації слів за їхніми граматичними характеристиками, аналізу семантичних асоціацій між словами, виявлення тематичних відтінків у текстах тощо [8, с. 86].

У цьому контексті С. Валліс наполягає на тому, що використання корпусів текстів і стохастичних підходів у лінгвістиці може допомогти відкрити нові знання про мову, її властивості та функціонування. Аналіз корпусів текстів дозволяє отримувати об'єктивні дані про мову, які можуть бути використані для підтвердження або спростування лінгвістичних гіпотез і теорій [9, с. 13]. Водночас цей підхід може забезпечити нові підходи до розв'язання практичних завдань у лінгвістиці, таких як автоматичний аналіз текстів, машинний переклад, створення інтелектуальних асистентів тощо.

Постановка завдання. Проблема використання інноваційних технологій у лінгвістиці полягає в тому, що деякі аспекти мови можуть бути складними для аналізу та вивчення без відповідних програмних інструментів. До цих питань належать аналіз фонетики, фразеології, синтаксису та семантики мови. Мета дослідження полягає в тому, щоби дати оцінку, як використання програмних інструментів може допомогти лінгвістам аналізувати мову на різних рівнях і підвищувати ефективність їхніх досліджень. Для досягнення поставленої мети необхідно проаналізувати використання інноваційних технологій для аналізу лінгвістичних категорій і структур, а також у підкресленні важливості програмних інструментів для лінгвістів, які сприяють вивченню мови на різних рівнях – фонетичному, фразеологічному, синтаксичному та семантичному, шляхом аналізу усних або письмових корпусів.

Виклад основного матеріалу. Мова, характерна для людського виду, є основою пізнання, на яку спираються різні галузі когнітивних наук, такі як лінгвістика, психологія, філософія, нейробіологія та інформатика, кожна з яких зробила свій внесок для розкриття цієї теми.

Співпраця між лінгвістикою та інформатикою виявляється у сферах обробки мови та сигналів, а також в автоматичній обробці мови, яка орієнтована на обробку письмового тексту. Ці зв'язки виникли ще в 1940-х роках, однак, незважаючи на різноманітність завдань, методик та підходів, автоматична обробка мови й досі залишається складним технологічним напрямком [10]. Навіть термінологія цього напрямку неоднозначна. Існують різні підходи до того, що саме ми намагаємося автоматизувати. Говорячи про автоматичну обробку природних мов, ми розуміємо їх відмінність від штучних мов, створених із певною метою.

З появою зв'язку між природними та формальними мовами з середини 1950-х років почав розвиватися напрям лінгвістики, відомий як «комп'ютерна лінгвістика», або «обчислювальна лінгвістика» [11, с. 140]. Головною метою цього напряму є опис функціонування мов у контексті машини та обчислень, зокрема синтаксичних обчислень. Це призвело до пошуку «математичних структур мови», розробки різних типів «формальних граматик», а також спроб розширення цього підходу на семантичному рівні.

Ця галузь лінгвістики, яка варіюється протягом багатьох років, має зв'язки з інформатикою на теоретичному та епістемологічному рівнях. Вона ставить перед собою завдання побудови металінгвістичних концепцій, які відповідають знанням, внутрішньо усвідомленим людьми. Отже, вона досліджує структурну архітектуру мовних знань і входить до складу класичної когнітивістської парадигми, відомої як «комп'ютерно-представницько-символічний підхід», що передбачає обчислення символів для створення концепцій [12]. Однак спроби реалізувати ці обчислення ефективно зазвичай не завжди успішні, що призводить до переважно теоретичного та формаль-

ного напрямку лінгвістики. Використання ІТ може допомогти в ефективній перевірці теоретичних моделей, однак це вимагає більшої співпраці між лінгвістами та спеціалістами з інформатики, щоби вирішувати прикладні проблеми автоматичної обробки мови.

Обробка природної мови та невдачі великих проєктів були спрямовані на автоматичний переклад текстів у політичному контексті холодної війни. Незважаючи на амбіції дослідників, їхні проєкти не дали глобального вирішення всіх проблем, пов'язаних із обробкою текстів природною мовою.

Проте нинішній спад у великих проєктах після розчарувань та невдач перших робіт не може забрати у людства важливих уроків і досягнень «піонерського» періоду [13]. Справді, дослідники швидко зрозуміли проблеми, пов'язані з пошуком глобальних моделей обробки мови, архітектури знань та машинного перекладу. Це створило виклик до співпраці між інформатикою, лінгвістикою та психологією, щоб ефективно вирішувати проблеми автоматичної обробки мови. Співпраця між різними галузями може допомогти досягти нових успіхів у розробці технологій обробки мови та удосконалення машинного перекладу.

Важливі аспекти, які викликають інтерес дослідників і які почали цікавити фахівців у сфері комп'ютерів, – це питання формалізації знань та рівні мови. Межі підходів до мови через призму формальних граматик стали очевидними у цих двох аспектах. Тому автоматична обробка мови поступово відвернулася від певних теоретичних варіантів, які включали різні аспекти когнітивної лінгвістики. Щодо першого пункту, сумніви у відповідності між природною мовою та формальними, логіко-алгебраїчними мовами, призвели до розвитку формалізму, які вважаються більш адекватними для обробки мови. Другий аспект уведення значень лінгвістичних категорій і структур призвів до більшої уваги щодо семантичних і прагматичних знань. Це збільшило інтерес до явищ, таких як неоднозначність, зсув у значенні, референція, неявність, еліпсис, типовість, контекст та ідея множинності рівнів значень.

Великі амбітні проєкти, що розпочалися на початку XXI століття, базувалися на порівнянні розуму з машиною. Багато подальших праць у сфері штучного інтелекту також використовували цю метафору, намагаючись створити програми автоматичної обробки мови, які б не лише виробляли результати, а й мали здатність до відтворення процесів, характерних для людської мовної поведінки. Це вище ніж просто «імітація», це спроба «симулювати» те, що ми знаємо про мовну поведінку. Крім того, ці проєкти ставлять перед собою більш високі завдання, такі як програмування машин для відтворення різноманітності способів розуміння значень, що властиві людському сприйняттю. Та, безсумнівно, такі проєкти мають свої переваги та недоліки (табл. 1).

З розвитком лінгвістичної інженерії та виготовленням операційних інструментів, дослідження лінгвістичних конструкцій за допомогою комп'ютерів вийшло на новий рівень. Зараз акцентується на створенні ефективних інструментів, які можуть допомогти людям у роботі, звільняючи їх від рутинних завдань, що вимагають багато часу та зусиль. Так, учені відмовляються від ідеї «автоматичної» обробки мови, що мала замінити людей на користь «автоматичної» обробки, яка призначена допомагати людям. Цей новий етап характеризується низкою відмінностей у порівнянні з попередніми. Підходи, які раніше вважалися несумісними, – майстрування та теоретична еклектика, – тепер вважаються прийнятними. Замість глибокого вивчення текстів з метою здобуття повного розуміння, нині віддають перевагу «легкому» аналізу, який базується на кількісних і статистичних даних, щоб отримати обмежене розуміння, спрямоване на конкретні цілі. Запізніла потреба у швидкому доступі до інформації з електронних документів обґрунтовує такий підхід.

Таблиця 1

Переваги та недоліки інноваційних технологій у лінгвістиці

Переваги	Недоліки
Створення програм автоматичної обробки мови може значно полегшити роботу людей, які працюють з великими обсягами текстової інформації.	Іноді програми автоматичної обробки мови можуть допускати помилки в інтерпретації або аналізі тексту, що може призвести до неправильних висновків або рішень.
Можливість відтворення процесів, характерних для людської мовної поведінки, може покращити якість комунікації з машинами і збільшити швидкість взаємодії з ними.	Машинне відтворення процесів людської мовної поведінки може призвести до виникнення непередбачених ситуацій або недоліків у спілкуванні з машинами.

Джерело: власна розробка авторів.

Інноваційні технології в лінгвістиці працюють з корпусами – великими колекціями тексту, які містять мільйони слів. Значна увага зосереджена на використання методів машинного навчання для автоматичного виявлення закономірностей у текстових даних, що дає змогу отримати знання з даних.

Лінгвіст розглядає ці нові підходи як можливість мати доступ до оперативних інструментів, які можуть допомогти йому в роботі з великим обсягом даних у природній мові. Комп'ютерні дослідження надають знаряддя, що дозволяють лінгвісту брати участь в емпіричній практиці за допомогою комп'ютера. Цей симбіоз артефакту і людини ґрунтується на їх взаємодії, від якої залежить успішність роботи: швидкість, надійність та гнучкість машини поєднуються з адаптивністю і різноманіттям шляхів людини.

На початковому етапі досліджень вчені, що розробляють нові інструменти для доступу до значень, явно вказують на когнітивну залежність свого підприємства. Вони, подібно до когнітивної лінгвістики, підкреслюють важливість тексту, удосконалення семантики (зазначають про «часткове розуміння» та стверджують неавтономність синтаксису), працюють над контекстом і вказують на існування загальних когнітивних механізмів (категоризація, сприйняття).

Проте в деяких аспектах можна помітити збільшення розриву між комп'ютерними дослідженнями та лінгвістикою. Протягом останніх десяти років інноваційні технології прищеплюють свої підходи, не відкидаючи при цьому традиційні методи передачі досвіду від людини до людини. Хоча для лінгвістів це може бути викликом, оскільки комп'ютерні системи можуть оминати правила, розроблені лінгвістами. Отож, міждисциплінарна співпраця має свою вагомість. Важливо також розуміти, як нові технології впливають на оцінку ефективності та як необхідно оцінювати їхню роботу, особливо коли йдеться про тонку взаємодію між мовою та машинним навчанням.

Висновки. Отже, аналіз використання інноваційних технологій для дослідження лінгвістичних категорій і структур підтверджує їх значущість у вивченні мови на різних рівнях. Програмні інструменти для лінгвістів не лише сприяють аналізу мовних даних, а й полегшують процес досліджень та дозволяють ефективно використовувати великі корпуси текстів для отримання нових знань про природні мови. Розвиток комп'ютерних досліджень в галузі обробки мови дає змогу вирішувати складні проблеми, пов'язані з розумінням природних мов та автоматичною обробкою текстів. Співпраця між різними галузями, такими як інформатика, лінгвістика та психологія, дозволяє досягати нових успіхів у цій сфері. Важливо також зауважити, що сучасні дослідження повинні спрямовуватись на створення інноваційних технологій, які б полегшили та покращили роботу людей, а не заміщували б їх.

Зазначені технології та інструменти відіграють ключову роль у сучасних лінгвістичних дослідженнях, допомагаючи науковцям вирішувати складні проблеми та розкривати нові аспекти мовознавства. Важливо продовжувати розвиток цих технологій та співпрацювати між різними галузями для досягнення нових успіхів у сфері автоматичної обробки мови.

Отже, можна зробити висновок, що співпраця між лінгвістикою та інформатикою є ключовою для досягнення успіхів у галузі автоматичної обробки мови. Розвиток інноваційних технологій у лінгвістиці сприяє створенню ефективних інструментів для роботи з природною мовою, допомагаючи вирішувати складні завдання та збільшуючи швидкість і надійність процесів обробки мови. Важливо враховувати не лише технологічні аспекти, а й когнітивні та психологічні особливості мови, спілкування і розуміння значень. Взаємодія між людиною та машиною відкриває нові можливості для розвитку автоматичної обробки мови й покращення якості комунікації у цифровій епосі.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ:

1. Piantadosi S. Modern language models refute Chomsky's approach to language. *Lingbuzz Preprint.* 2023. URL: https://lingbuzz.net/lingbuzz/007180 (дата звернення: 02.03.2024).

2. Modeling language variation and universals: A survey on typological linguistics for natural language processing / E. M. Ponti et al. *Computational Linguistics*. 2019. Vol. 45. No. 3. P. 559–601. DOI: https://doi.org/10.1162/coli_a_00357 (дата звернення: 02.03.2024).

3. de Andrade G. C., de Paiva Oliveira A., Moreira A. Hybrid semantic annotation: Rule-based and manual annotation of the open American national corpus with top-level ontology. *Abakós*. 2019. Vol. 7. No. 3. P. 64–78. DOI: https://doi.org/10.5752/P.2316-9451.2019v7n3p64-78 (дата звернення: 02.03.2024).

4. A systematic study on the role of SentiWordNet in opinion mining / M. Husnain et al. *Frontiers of Computer Science*. 2021. Vol. 15. No. 4. DOI: https://doi.org/10.1007/s11704-019-9094-0 (дата звернення: 02.03.2024).

5. Breaking the barriers between intelligence, investigation and evaluation: A continuous approach to define the contribution and scope of forensic science / S. Baechler et al. *Forensic science international*. 2020. Vol. 309. DOI: https://doi.org/10.1016/j.forsciint.2020.110213 (дата звернення: 02.03.2024).

6. Analysis of continuous neuronal activity evoked by natural speech with computational corpus linguistics methods / A. Schilling et al. *Language, Cognition and Neuroscience*. 2021. Vol. 36. No. 2. P. 167–186. DOI: https://doi.org/10.1080/23273798.2020.1803375 (дата звернення: 02.03.2024).

7. Kortmann B. Reflecting on the quantitative turn in linguistics. *Linguistics*. 2021. Vol. 59. No. 5. P. 1207–1226. DOI: https://doi.org/10.1515/ling-2019-0046 (дата звернення: 02.03.2024).

8. Kozlova T., Polyezhayev Y. A cognitive-pragmatic study of Australian English phraseology. *AD ALTA: Journal of Interdisciplinary Research*. 2022. Vol. 12. No. 1. P. 85–93. DOI: https://doi.org/10.33543/12018593 (дата звернення: 02.03.2024).

9. Wallis S. *Statistics in corpus linguistics research: A new approach*. New York, NY : Routledge, 2020. 382 p.

10. De Sutter G., Lefer, M. A. On the need for a new research agenda for corpus-based translation studies: A multi-methodological, multifactorial and interdisciplinary approach. *Perspectives*. 2020. Vol. 28. No. 1. P. 1–23. DOI: https://doi.org/10.1080/0907676X.2019.1611891 (дата звернення: 02.03.2024).

11.An energy-based model for word-level autocompletion in computer-aided translation / C. Yang et al. *Transactions of the Association for Computational Linguistics*. 2024. Vol. 12. P. 137–156. DOI: https://doi.org/10.1162/tacl_a_00637 (дата звернення: 02.03.2024).

12. Applying a new framework of connections between mathematical symbols and natural language / U. W. Hultdin et al. *The Journal of Mathematical Behavior*. 2023. Vol. 72. DOI: https://doi.org/10.1016/j.jmathb.2023.101097 (дата звернення: 02.03.2024).

13. Dingli A., Farrugia D. Neuro-symbolic AI: Design transparent and trustworthy systems that understand the world as you do. Birmingham – Mumbai : Packt Publishing, 2023. 196 p.