Маргарита Надутенко,
Максим Надутенко
Український мовно-інформаційний фонд НАН України

МЕТОДИ ЦИФРОВИХ ЛІНГВІСТИЧНИХ ДОСЛІДЖЕНЬ: КОРПУСНІ ТЕХНОЛОГІЇ ТА ШТУЧНИЙ ІНТЕЛЕКТ

Обґрунтовано тезу про основні методи цифрових лінгвістичних досліджень, які доцільно використовувати для дослідження мовного матеріалу. У групі презентованих методів виділено основні: статистичний, корпусний, лексикографічний метод дослідження, в основі якого теорія лексикографічних систем, та метод штучного інтелекту: машинне навчання, глибоке навчання, нейромережі.

Ключові слова: *цифрові методи лінгвістичних досліджень, корпусні технології, мультимедійний словник, теорія лексикографічних систем, штучний інтелект, машинне навчання, нейромережі.*

The research substantiates the thesis about the main methods of digital linguistics which are expedient to use for research language material. In the group of presented methods main ones are highlighted: statistical, corpora, lexicographic methods of research (based on the theory of lexicographic systems) and the methods of artificial intelligence: machine learning, deep learning, neural networks.

Key words*: digital methods of linguistic research, corpus technologies, multimedia dictionary, theory of lexicographic systems, artificial intelligence, machine learning, neural networks.*

У сучасному суспільстві знання стають одним із найважливіших чинників культурного, економічного, технологічного та

гуманітарного розвитку будь-якої держави. Основну частину нових знань люди отримують безпосередньо із текстів, які створені природною мовою (мовами). ***Тексти*** є носіями лінгвістичної інформації. Колосальне збільшення обсягів накопиченої інформації та висока швидкість надходження нових даних зумовлюють потребу в високоефективних інформаційних технологіях та інтелектуальних системах пошуку нових способів автоматичної обробки інформації, систематизації, відповідної класифікації та презентації за вимогою користувача. Виникає необхідність пошуку нових методів дослідження лінгвістичної інформації. Однією з ознак виникнення певного напряму у науці є наявність власного методу. Саме **метод** формує підходи до аналізу мовних і мовленнєвих явищ. Не існує єдиної загальновизнаної класифікації лінгвістичних методів, проте існує нагальна потреба до аналізу великих масивів цифрової інформації та використання наявних програмних розробок.

Актуальність проблеми визначається необхідністю комплексного аналізу лінгвістичного матеріалу на всіх етапах роботи з текстом: цифровізація контенту, автоматична обробка, систематизація, класифікація та відповідна презентація великих масивів даних із забезпеченням доступності набутих знань для широкого кола користувачів.

Наукова новизна дослідження полягає у розробці комплексу методів для дослідження великих масивів цифрової лінгвістичної інформації.

Мета статті – презентація основних методів дослідження цифрової лінгвістичної інформації.

Технології обробки текстів – найбільш поширені засоби обробки текстової інформації *різного формату*. ***Текст*** – будь-яка послідовність символів: літери, пробіл, розділові знаки, цифри, знаки арифметичних операцій тощо. **Обробка текстової інформації** – сукупність операцій: **збирання, введення, записування, перетворення, зчитування, зберігання, знищення,**

реєстрація, що здійснюються за допомогою технічних та програмних засобів (рис. 1).

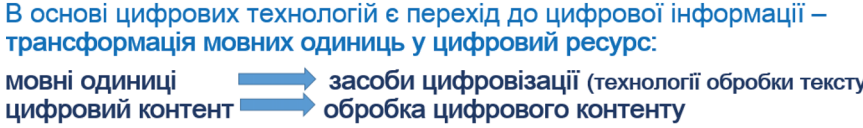

Рис. 1. Перехід до цифрової інформації

Виділено основні методи дослідження цифрової лінгвістичної інформації, які є основою сучасних лінгвістичних досліджень:

1) **статистичний метод обробки інформації** (математичний);
2) **метод корпусних технологій**;
3) **лексикографічний**;
4) **штучний інтелект**: машинне навчання, глибоке навчання, нейромережі.

Класифікація методів дослідження формується відповідно до мети їх використання та має враховувати специфіку джерельної бази дослідження. Запропоновані методи дослідження не суперечать вже існуючим класичним методам дослідження та мають бути використані у комплексі.

Статистичний метод дослідження розглядає мову як системно-структурне утворення, що має окремі підсистеми – рівні, які представлені відповідними одиницями: фонемами, морфемами, лексемами, синтагмами. Таким чином, мова характеризується на лише якісними, а й кількісними показниками. Кількісна методика стала більш ефективною з появою відповідного програмного забезпечення відкритого доступу. Приклад виявлення та порівняння *синонімічних рядів* у лексемах: *україна* (*заст.*) та сучасного багатозначного відповідника *край* (рис. 2).

Лексикографічний метод вважаємо доцільним застосовувати для сучасних досліджень лексичної системи мови, який на-

Рис. 2. Приклад виявлення та порівняння синонімічних рядів у лексемах: україна (заст.) та сучасного багатозначного відповідника край (Інтегрована лексикографічна система «Словникǵ України online»)

правлений, перш за все, на послідовне виділення та вибіркове вивчення окремих елементів та їх відношень у мовній системі. Відповідно до аспектів дослідження сучасна лінгвістика пропонує цілий ряд методик та прийомів аналізу одиниць лексичної системи. Наразі триває активний пошук нових прийомів, які використовують лексикографічний метод дослідження у поєднанні з автоматичними лексикографічними системами.

Приклади лексикографічних методів:

1) **класифікація лексики з метою створення мультимедійних словників нових типів** – створення Мультимедійного словника з інфомедійної грамотності з метою групування тематичного ряду медіатермінів (http://lcorp.ulif.org.ua/InfoMediaVLL/);
2) **аналіз зіставлень та порівнянь на межі декількох мов** здійснюється з метою дослідження встановлення частотних характеристик відповідних лексем – таблиця порівняння частотності вживання лексем:
3) **відкритий вільний асоціативний експеримент** із використанням автоматичних лексикографічних систем;

Українська мова **Український національний лінгвістичний корпус (УНЛК)** https://svc.ulif.org.ua/UNLC/virt_unlc_4.5/	***Болгарська мова*** **Български национален корпус/ Bulgarian National Corpus (BNC)** http://search.dcl.bas.bg/
Україна 669 441	*Украина* 12
український народ 12 968	*украински народ* 1
українець 56 432	*украински* 551
відродження 11 506	*възраждане* 1598
громадянин 67 104	*гражданин* 19 466
держава 157 059	*държава* 46 866
Радянський Союз 6857	*Съветския Съюз* 2536
Болгарія 5768	*България* 37 652
болгарський народ 80	*български народ* 1482
Європа 61 464	*Европа* 20 651
влада 147 282	*власт* 42 531
національні інтереси 2480	*национални интереси* 366

4) **дистрибутивний аналіз семантем** для визначення відмінності в семемах окремої лексеми у хронологічному зрізі.

Метод корпусних технологій на сучасному етапі мовознавчих досліджень вважаємо доцільним використовувати на основі вже існуючих програмних продуктів як необхідну складову для створення мінікорпусів відкритого типу.

Штучний інтелект: машинне навчання, глибоке навчання. **Нейромережі** розглядаємо як підмножину штучного інтелекту, що зосереджена переважно на проєктуванні систем, які дозволяють їм навчатися та робити прогнози на основі певного досвіду. Штучний інтелект (ШІ) розглядаємо як форму індивідуалізації систем, якій притаманний мовний статус. Під ШІ зазвичай мається на увазі імітація чи реплікація когнітивних рис людини машинами – комп'ютерними програмами. Загальне

поняття ШІ включає кілька напрямів: експертні системи, системи для характеристичного аналізу і робототехніка. Здатність інженерної системи обробляти, застосовувати та вдосконалювати здобуті знання. Це система, що має певні ознаки інтелекту, тобто здатна: розпізнавати та розуміти; знаходити спосіб досягнення результату та приймати рішення; вчитися. Наприклад, для визначення семантичної подібності між реченнями та великими контекстами, необхідної для індексування, використовуються метод семантичного хешування, який не потребує значних обчислювальних потужностей на попередню обробку текстів, як того вимагають методи на основі нейронних мереж. Для семантичного хешування використовується метод розбиття тексту на N-грами слів. Далі для побудованих наборів N-грам розраховується набір із 1000 хеш-значень на основі комбінації таких некриптографічних хеш функцій, як CityHash, spookyHash та xxHash, що мають високу колізійну якість та якість псевдовипадковості. Також для визначення семантичної подібності речень нами використовується набір алгоритмів типу Word2Vec, але, зважаючи на недоліки, зазначені вище, для побудови семантичного вектора контексту використовується набір кластерів, що відображають тематику тексту. Тематики визначаються автоматично кожен раз при індексації на основі кластеризації. Якщо слово є реєстровим, у якості початкових значень може використовуватися семантична ремарка зі Словника української мови у 20 томах (https://services.ulif.org.ua/expl/).

Машинне навчання здійснюється без учителя на основі контекстуального представлення Українського національного лінгвістичного корпусу. Кожне слово при цьому використовує контекстний зв'зок як із наступним, так і з попереднім словом. Відповідно **нечіткий пошук за реченнями** нами реалізовано на основі трьох методів: вибору оптимально близьких за наборами хешів речень, речень, ранжованих за найменшою косинусною відстанню між їхніми семантичними векторами розмірністю

3000, та речень, отриманих за допомогою векторного представлення на основі алгоритму BERT.

Українським мовно-інформаційним фондом НАН України розроблено теоретичні та науково-технічні засади зазначених методів дослідження, які довели необхідність та доцільність для подальшого використання. Їхнє практичне застосування продемонструвало високу ефективність впроваджених засобів у формі віртуальних лексикографічних лабораторій та згрупованих лінгвістичних платформ, засвідчило високу якість кінцевих мультимедійних словникових продуктів, які активно впроваджуються у загальноукраїнський та європейський простір.

Розроблене програмне забезпечення може стати основою для подальшого впровадження зазначених методів, які стануть важливою складовою розвитку лексикографічних технологій у галузі мовознавста та забезпечать ефективну фахову взаємодію у загальноукраїнському та європейському просторі.

Література

1. Nadutenko M., Prykhodniuk V., Shyrokov V., Stryzhak O. Ontology-Driven Lexicographic Systems. Advances in Information and Communication. FICC 2022. Lecture Notes in Networks and Systems. Cham : Springer. 2022. Pp. 204–215. https://doi.org/10.1007/978-3-030-98012-2_16
2. Mintser O. P., Babintseva L. Yu., Zaliskyi V. M., Nadutenko M. V., Kharchenko N. V., Ladychuk O. K. Theoretical approaches to the creation of systemic biomedicine (based on the materials of the report on SRW "System-Biological And System-Medical Regularities Of Development And Course Of Ischemic Heart Disease"). *Medical Informatics and Engineering*. 2021. Vol. 4. Pp. 16-72. https://doi.org/10.11603/mie.1996-1960.2020.4.11889
3. Semenog O. M., Nadutenko M. V., Nadutenko M. V. VIRTUAL LABORATORY "MULTIMEDIA DICTIONARY OF INFOMEDIA LITERACY". International scientific and practical conference "Philological sciences, intercultural communication and translation studies: an experience and challenges" : conference proceedings, April 23–24, 2021. Vol. 1. Czestochowa : "Baltija Publishing"; Київ : ТОВ «КАЛЕНДАР ТМ», 2021. 300 с. URL: http://baltijapublishing.lv/omp/index.php/bp/catalog/view/128/3621/7733-1 (дата звернення: 10.11.2022).

4. Stryzhak O., Prykhodniuk V., Popova M., Nadutenko M., Haiko S., Chepkov R. Development of an Oceanographic Databank Based on Ontological Interactive Documents. Lecture Notes in Networks and Systems. Cham : Springer, 2021. Pp. 97–114. https://doi.org/10.1007/978-3-030-80126-7_8
5. Борисенко Н. Д. Методика проведення наукових досліджень: навчально-методичний посібник. Житомир : Вид-во ЖДУ, 2010. 64 с.
6. Български национален корпус/Bulgarian National Corpus (BNC). URL: http://search.dcl.bas.bg/ (дата звернення: 10.11.2022).
7. Інтегрована лексикографічна система «Словникѝ України online». URL: https://svc2.ulif.org.ua/dictua/ (дата звернення: 10.11.2022).
8. Концепт освіта в українському та польському дискурсах крізь призму національних лінгвістичних корпусів / М. В. Надутенко, О. М. Семеног. *Innovative Pathway for the Development of Modern Philological Sciences in Ukraine and EU Countries* : collective monograph. Riga, Latvia : "Baltija Publishing", 2021. Vol. 2. Pp. 66–81. URL: http://baltijapublishing.lv/omp/index.php/bp/catalog/view/100/2541/5453-1 (дата звернення: 10.11.2022). https://doi.org/10.30525/978-9934-26-031-5-27
9. Лінгвістично-інформаційні студії : праці Українського мовно-інформаційного фонду НАН України : у 5 т. / В. А. Широков, М. В. Надутенко та ін. Т. 4 : Корпусна лінгвістика. Київ : УМІФ НАН України. 2018. 289 с.
10. Лінгвістично-інформаційні студії : праці Українського мовно-інформаційного фонду НАН України : у 5 т. / В. А. Широков, М. В. Надутенко та ін. Т. 5 : Віртуалізація лінгвістичних технологій. Київ : УМІФ НАН України. 2018. 289 с.
11. Надутенко Маргарита. Методи ономастичного дослідження з використанням сучасних лінгвістичних технологій : тези доповідей Міжнародної наукової конференції «АКТУАЛЬНІ ПИТАННЯ СУЧАСНОЇ ЛІНГВІСТИКИ», 6 березня 2021 р. / [орг. комітет: Куранова С. І., Лучик А. А. та ін.] ; Національний університет «Києво-Могилянська академія», Кафедра загального і слов'янського мовознавства. Київ : НаУКМА, 2021. С. 65–72.
12. Словник української мови : у двадцяти томах : Т. 1-12 / наук. кер. Широков В. А., уклад. : Шевченко І. В., Загнітко А. П., Заїка Н. М., Надутенко М. В., Оліфіренко Л. В., Сивокозова В. В., Симоненко Л. О., Томіленко Л. М., Ярун Г. М., Чумак В. В., Шевченко Л. Л., Брітікова К. В., Білоноженко В. М., Винник В. О. Київ : УМІФ НАНУ, 2010–1021. (Словники України). URL: https://sum20ua.com/ (дата звернення: 10.11.2022).
13. Український національний лінгвістичний корпус (УНЛК). URL: https://svc.ulif.org.ua/UNLC/virt_unlc_4.5/ (дата звернення: 10.11.2022).
14. Формування культуромовної особистості фахівця в умовах неперервної освіти : монографія / авт. кол. О. Семеног, В. Герман, Н. Громова,

І. Левенок, Н. Пономаренко, М. Надутенко, К. Діхнич ; за заг. ред. О. Семеног. Суми : Вид-во СумДПУ імені А. С. Макаренка, 2020. 212 с. URL: https://repository.sspu.edu.ua/bitstream/123456789/9467/3/Formuvannia%20kulturomovnoi%20osobystosti%20fakhivtsia.pdf (дата звернення: 10.11.2022).

15. Широков В. А., Загнітко А. П., Надутенко М. В., Надутенко М. В. та ін. Віртуальна лексикографічна лабораторія «Мультимедійний словник з інфомедійної грамотності». *Український мовно-інформаційний фонд НАН України*, створено в рамках грантової програми «МЕДІА& ВЧИТЕЛЬський кампус» проєкту «Вивчай та розрізняй: інфомедійна грамотність», IREX (Рада наукових досліджень та обмінів) за підтримки Посольств Великої Британії та США у партнерстві з Міністерством освіти і науки України та Академією української преси, 2020–2021. URL: https://lcorp.ulif.org.ua/InfoMediaVLL/ (дата звернення: 10.11.2022).
16. Широков В. А., Лучик А. А. Парадигмальні засади лінгвістики першої половини XXI століття. *Мовознавство*. 2021. № 5. С. 3–163.