

Львівський національний університет імені Івана Франка
Філологічний факультет
Катедра загального мовознавства

“ЗАТВЕРДЖУЮ”
Декан
філологічного факультету
доц. Р.О. Крохмальний


“29” серпня 2025 року

РОБОЧА ПРОГРАМА

ВИРОБНИЧОЇ ПРАКТИКИ З КОМП'ЮТЕРНОЇ ТА КОРПУСНОЇ ЛІНГВІСТИКИ

галузь знань	<u>Культура, мистецтво та гуманітарні науки</u>
спеціальність	<u>В11 Філологія</u>
спеціалізації	<u>В11.10 Прикладна лінгвістика</u>
факультет	<u>філологічний</u>

Львів – 2025 рік

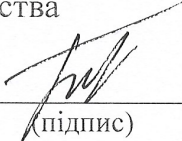
Робоча програма виробничої практики з комп'ютерної та корпусної лінгвістики для студентів 2 курсу другого (магістерського) рівня вищої освіти за спеціальністю В11 Філологія, спеціалізацією В11.10 Прикладна лінгвістика. Львів: ЛНУ імені Івана Франка, 2025.

Розробники: Мацюк Г.П., докт. філол. наук, професор;
Григоруk С.І., канд. філол. наук, доцент;
Надутенко М.В., канд. філол. наук, старший науковий співробітник Українського мовно-інформаційного фонду НАН України

Робоча програма розглянута на засіданні катедри загального мовознавства.

Протокол від "27" серпня 2025 року № 1.

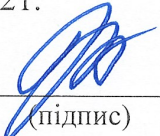
Завідувач катедри загального мовознавства


_____ (проф. Бацевич Ф.С.)
(підпис)

"27" серпня 2025 року

Затверджено Вченою радою філологічного факультету.

Протокол від "29" серпня 2025 року № 21.

Голова 
_____ (доц. Крохмальний Р.О.)
(підпис)

"29" серпня 2025 року

1. Опис практики

Найменування показників	Галузь знань, напрям підготовки, освітньо-кваліфікаційний рівень	Характеристика навчальної дисципліни	
		денна форма навчання	заочна форма навчання
Кількість кредитів – 3	<u>Галузь знань</u> В Культура, мистецтво та гуманітарні науки <u>Спеціальність</u> В11 Філологія	Нормативна	
Модулів – 2	<u>Спеціалізація:</u> В11.10 Прикладна лінгвістика	Рік підготовки:	
Змістових модулів – 2		1-й	
Індивідуальне науково-дослідне завдання _____ (назва)		Семестр	
Загальна кількість годин - 90		2-й	
		Лекції	
Тижневих годин для денної форми навчання: аудиторних – самостійної роботи студента -	<u>Освітній ступінь:</u> магістр	Практичні, семінарські	
		Лабораторні	
		Самостійна робота	
		90 год.	
		Індивідуальні завдання:	
		Вид контролю: диференційований залік	

Практика з комп'ютерної та корпусної лінгвістики – обов'язковий компонент навчального процесу. Практика забезпечує поєднання теоретичної підготовки майбутніх прикладних лінгвістів з їхньою діяльністю у різних установах працевлаштування, сприяє розвитку практичних навиків у використанні інформаційних технологій у текстотворенні та лінгвоекспертології.

Головна мета цієї практики для магістрів 2025/2026 року вступу полягає в закріпленні вже отриманих навичок, пов'язаних із застосуванням інформаційних технологій, знання про які здобувачі отримали на дисциплінах: «Комп'ютерна лінгвістика і опрацювання природної мови» (4 кредити, 120 год загальний обсяг, 16 лекцій, 16 практичних, 88 год самостійна робота – 1-й семестр) та курсу «Кількісні та корпусні підходи у прикладній лінгвістиці» (6 кредитів, 180 год загальний обсяг, 32 год лекції, 32 год практичні, 116 год самостійна робота – 2-й та 3-й семестри).

Практика сприяє розвитку та здобуттю практичних навиків у використанні інформаційних технологій у текстотворенні, лінгвоекспертології та інших галузях лінгвістики. Вид практики, її тривалість і терміни проведення визначені навчальним планом на 2025/2026 н. р. для слухачів 2025 року вступу. Зміст і послідовність проведення практики визначає наскрізна програма, розроблена згідно з оновленим навчальним планом для вступників в магістратуру в 2025/2026 навчальному році.

Пререквізити практики: прослухані дисципліни з впровадження інформаційних технологій: «Комп'ютерна лінгвістика і опрацювання природної мови» (проф. Кушнір Олег Степанович, докт. фіз.-мат. наук, професор, професор кафедри оптоелектроніки та інформаційних технологій), «Кількісні та корпусні підходи у прикладній лінгвістиці» (проф. Ровенчак Андрій Адамович, докт. фіз.-мат. наук, професор, професор кафедри теоретичної фізики імені Івана Вакарчука).

Під час виробничої практики здобувачі знайомляться із діяльністю провідної установи з розробки та використання інформаційних технологій, а саме Українським мовно-інформаційним фондом НАН України та з порталом «Mova.info: про українську мову, лінгвістику і не тільки».

2. Мета і завдання практики

Виробнича практика з комп'ютерної і корпусної лінгвістики сприяє формуванню методології використання конкретних електронних ресурсів, які відповідають практичній роботі з великими обсягами текстових даних, зокрема методиці їхнього комп'ютерного аналізу. Корпуси – один із найпотужніших інструментів прикладної лінгвістики. Їх створюють і використовують у різних галузях людської діяльності на основі автоматизації процесу підбору, укладання та аналізу текстових масивів практично необмеженого обсягу. Допоміжними засобами є методи комп'ютерного, статистичного та ін. аналізу цих даних, наприклад препроцесинг текстів, стандарти збереження текстових даних, парсинг, стемінг і

лематизація, визначення середньої довжини слів і речень, вивчення словників текстів і параметра Type-Token Ratio, фонетичний і силабічний аналіз текстів та визначення консонантного коефіцієнта, робота з базами лінгвістичних даних і побудова запитів, встановлення подібності текстів і плагіату тощо.

Для кожного із магістрів корпус і спеціальні програмні засоби роботи з цим корпусом є спеціалізованим інструментом лінгвістичного аналізу. Магістри повинні отримати навички роботи з усіма складовими комп'ютерно-корпусних завдань, зокрема попереднє опрацювання текстів, створення власних корпусів, опанування різних видів корпусної розмітки, використання Інтернету для корпусних досліджень і вживання різних статистичних методів при роботі з корпусами і/або окремими текстами з корпусів.

Тривалість практики магістрів першого курсу в Українському мовно-інформаційному фонді - два тижні (червень 2026р.). За навчальним планом виробнича практика відбувається впродовж 2 тижнів (90 годин). Тривалість робочого часу здобувачів при проходженні виробничої практики – 45 годин на тиждень.

Відомо, що написання магістерського дослідження – це процес, який вимагає не тільки знань у певній предметній області, але й навичок наукового аналізу, застосування інформаційних технологій, синтезу та аргументації, а також знання основних термінів і навичок, які дають змогу здобути наукові дані та чітко й професійно висловити свої ідеї та висновки. На практиці магістри працюють з теоретичною та методичною частинами своєї магістерської роботи. Головні джерела для корпусів, які використовують магістри, – це публічні та відкриті наукові тексти, що є у вільному доступі в Інтернеті та використовуються для написання теоретичного розділу.

Мета практики: закріплення теоретичних знань і набуття практичних навичок у використанні інформаційних технологій для створення корпусу наукових текстів для теоретичного розділу магістерського дослідження, аналізу лінгвістичних даних корпусу щодо частотності термінів, роботи зі словниковими статтями і створення словника найчастотніших термінів аналізу матеріалу магістерського дослідження.

Основні завдання практики:

1. Підготувати і провести дискусію на тему: «Комп'ютерна лексикографія як напрям сучасної комп'ютерної лінгвістики: можливості, стан і перспективи розвитку» (на матеріалі Інтегрованої лексикографічної системи «Словники України» та порталу «Mova.info: про українську мову, лінгвістику і не тільки»).
2. Виокремити 20 наукових статей з теоретичного розділу магістерського дослідження для створення текстового корпусу.

3. Здійснити нормалізацію текстів: видалити зайві символи (пробіли, коди, знаки форматування); адаптувати форматування для сумісності з програмним забезпеченням Sketch Engine і/або іншим використаним програмним забезпеченням.
 4. Завантажити нормалізовані статті в середовище Sketch Engine, створивши корпус із можливістю лінгвістичного аналізу. Перевірити роботу інструментів для пошуку ключових слів, частоти вживання та контекстного аналізу.
 5. Створити словникові статті. Для цього вибрати найчастотніші 10 термінів із корпусу і укласти щодо них словникові статті за схемою: термін, визначення, приклади вживання.
 6. Переглянути відеоінструкцію щодо використання капсули E-devel, що допомагає зрозуміти основні функції платформи та способи роботи з нею.
 7. Занести інформацію про себе до капсули E-devel, увійшовши до системи капсули та додавши профільну інформацію (розповідь про себе; спеціалізацію; наукові публікації; нагороди, сертифікати).
 8. Занести інформацію про роботу з корпусом.
 9. Занести інформацію про словник, вказавши мету (пояснення ключових термінів); специфіку використання (для науковців, студентів, викладачів тощо); критерії вибору термінів (релевантність і частотність).
- До капсули додати звіт про роботу з корпусом: опис методики, основні функції, інструменти аналізу.
10. У разі потреби та зацікавленості студентів вивченням додаткового програмного забезпечення можливі додаткові індивідуальні завдання для практикантів, які включають: визначення середньої довжини слів і речень текстів деякою мовою; вивчення словника текстів і порівняння їхніх параметрів Type-Token Ratio; фонетичний і силабічний аналіз текстів однією із східнослов'янських мов; визначення консонантного коефіцієнта; професійну роботу з базами лінгвістичних даних різних типів; побудова запитів з урахуванням лінгвістичних завдань; визначення та порівняння семантичного навантаження різних текстів корпусу; встановлення подібності текстів і наявності плагіату; роботу з програмним пакетом Grammarly із елементами штучного інтелекту для комп'ютерного редагування, машинного перекладу та перевірки текстів на плагіат.
 11. У разі технічної можливості: проаналізувати текстовий корпус, використовуючи моделі «торба слів» (Bag of Words), TF-IDF, та отримати векторизації термінів за допомогою Word2Vec. Застосувати базові можливості великих мовних моделей (Large Language Models, LLMs) за допомогою фреймворку LangChain для виконання простих завдань, таких як генерація тексту або класифікація, на основі матеріалів корпусу.

компетентностей:

- ЗК 1. Здатність спілкуватися державною мовою як усно, так і письмово.
- ЗК 3. Здатність до пошуку, опрацювання та аналізу інформації з різних джерел.
- ЗК 4. Уміння виявляти, ставити та вирішувати проблеми. ЗК
- ЗК 5. Здатність працювати в команді та автономно.
- ЗК 6. Здатність спілкуватися іноземною мовою.
- ЗК 7. Здатність до абстрактного мислення, аналізу та синтезу.
- ЗК 8. Навички використання інформаційних і комунікаційних технологій.
- ЗК 9. Здатність до адаптації та дії в новій ситуації.
- ЗК 11. Здатність проведення досліджень на належному рівні.
- ЗК 12. Здатність генерувати нові ідеї (креативність).

фахових компетентностей

ФК 4. Здатність здійснювати науковий аналіз і структурування мовного / мовленнєвого й літературного матеріалу з урахуванням класичних і новітніх методологічних принципів.

ФК 5. Усвідомлення методологічного, організаційного та правового підґрунтя, необхідного для досліджень та/або інноваційних розробок у галузі філології, презентації їх результатів професійній спільноті та захисту інтелектуальної власності на результати досліджень та інновацій.

ФК 6. Здатність застосовувати поглиблені знання з обраної філологічної спеціалізації «Прикладна лінгвістика» для вирішення професійних завдань.

ФК 7. Здатність вільно користуватися спеціальною термінологією в обраній галузі філологічних досліджень.

Виробнича практика спрямована на досягнення таких **програмних результатів навчання:**

ПРН 1. Оцінювати власну навчальну та науково-професійну діяльність, будувати і втілювати ефективну стратегію саморозвитку та професійного самовдосконалення.

ПРН 2. Упевнено володіти державною та іноземною мовами для реалізації письмової та усної комунікації, зокрема в ситуаціях професійного й наукового спілкування; презентувати результати досліджень державною та іноземною мовами.

ПРН 3. Застосовувати сучасні методики і технології, зокрема інформаційні, для успішного й ефективного здійснення професійної діяльності та забезпечення якості дослідження в конкретній філологічній галузі.

ПРН 4. Оцінювати й критично аналізувати соціально, особистісно та професійно значущі проблеми і пропонувати шляхи їх вирішення у складних і непередбачуваних умовах, що потребує застосування нових підходів та прогнозування.

ПРН 5. Знаходити оптимальні шляхи ефективної взаємодії у професійному колективі та з представниками інших професійних груп різного рівня.

ПРН 9. Характеризувати теоретичні засади (концепції, категорії, принципи, основні поняття тощо) та прикладні аспекти обраної філологічної спеціалізації «Прикладна лінгвістика».

ПРН 11. Здійснювати науковий аналіз мовного, мовленнєвого й літературного матеріалу, інтерпретувати та структурувати його з урахуванням доцільних методологічних принципів, формулювати узагальнення на основі самостійно опрацьованих даних.

ПРН 12. Дотримуватися правил академічної доброчесності

ПРН 15. Обирати оптимальні дослідницькі підходи й методи для аналізу конкретного лінгвістичного чи літературного матеріалу.

ПРН 16. Використовувати спеціалізовані концептуальні знання з обраної філологічної галузі для розв'язання складних задач і проблем, що потребує оновлення та інтеграції знань, часто в умовах неповної/недостатньої інформації та суперечливих вимог.

ПРН 17. Планувати, організовувати, здійснювати і презентувати дослідження та/або інноваційні розробки в конкретній філологічній галузі.

3. Організація проведення практики

Обов'язки керівника практики від катедри загального мовознавства: організувати наказ про практику, пояснити особливості практичного навчання відповідно до програм практики; провести інструктаж з правил техніки безпеки й охорони праці, контактувати з магістрами під час практики; перевірити матеріали практики; організувати захист практики, на якому здобувачі прозвітують про виконані завдання.

База практики: Український мовно-інформаційний фонд НАН України.

Керівником практики від Українського мовно-інформаційного фонду НАН України як науково-дослідної інституції є кандидат технічних наук, завідувач відділу інформатики Максим Надутенко; також зі студентами працює кандидат філологічних наук, старший науковий співробітник Маргарита Надутенко, Вчений секретар інституту.

Керівник практики від Українського мовно-інформаційного фонду НАН України організовує роботу над виконанням завдань практики: знайомить і контролює дотримання здобувачами-практикантами правил внутрішнього трудового розпорядку інституту; забезпечує виконання узгодженого з навчальним закладом календарного графіку етапів проходження виробничої практики; перевіряє виконану практикантами роботу, зокрема щоденники та матеріали практики.

Терміни проходження виробничої практики:

Виробнича практика триває два тижні з відривом від навчання в 2-му семестрі.

4. Програма виробничої практики

Змістовий модуль 1 передбачає:

Знайомство з діяльністю Українського мовно-інформаційного центру НАН України.

Відомості про розвиток і перспективи комп'ютерної лінгвістики і комп'ютерної лексикографії.

Поняття про Sketch Engine – програмний продукт для укладання та роботи з корпусами, якнайкраще відповідає завданням, які постають під час роботи з фаховими текстами. Він допомагає відбору активної лексики та значущої термінології та типових колокацій.

Поняття про нормалізацію текстів. Процес нормалізації постає як сукупність інформаційних процедур, які роблять текст придатним до внесення його в корпус. Приведення всіх текстів до однієї кодової таблиці, перевірка їх на пунктуаційну коректність (однакові за смыслом сутності мають бути позначені одним знаком), усунення зайвих символів (наприклад, порожні абзаци, декілька пробілів поспіль і т. ін.), уніфікація засобів та способів форматування та ін.

Поняття про словник термінів: мету як пояснення ключових термінів аналізу; словникову статтю.

Поняття про капсулу E-devel, яку практиканти наповнюють своїми матеріалами.

Змістовий модуль 2 передбачає:

Поняття про додаткове програмне забезпечення: для виконання просунутого попереднього опрацювання текстів, парсингу, стемінгу та лематизації; визначення середньої довжини слів і речень; вивчення словника текстів і параметрів TTR; фонетичного та силабічного аналізу текстів; побудови запитів з урахуванням лінгвістичних міркувань; визначення семантичного навантаження різних текстів; встановлення подібності текстів і плагіату; роботи з програмним пакетом Grammarly.

Поняття про подання текстів. Це методи, які перетворюють неструктурований текст на числові формати, придатні для обробки комп'ютером. Ці моделі враховують семантичні та синтаксичні властивості мови.

- **Bag of Words (Модель «торба слів»).** Модель, де текст розглядають як набір слів без урахування їхнього порядку, а важливість слів визначають частотою їхнього вживання.

- **TF-IDF (Term Frequency-Inverse Document Frequency).** Статистичний підхід, що вимірює важливість слова в документі щодо цілого корпусу, надаючи більшу вагу унікальним словам.
- **Word2Vec.** Нейромережева модель, що створює векторні представлення слів, де слова з подібним значенням розташовуються ближче одне до одного у векторному просторі.
- **GloVe (Global Vectors for Word Representation).** Метод, що створює векторні представлення слів, ґрунтуючись на глобальних статистиках спільного вживання слів.

Поняття про великі мовні моделі (Large Language Models, LLMs) та їх застосування. Це сучасні моделі штучного інтелекту, навчені на величезних обсягах текстових даних, які здатні розуміти, генерувати та обробляти природну мову.

Узагальнення результатів практики. Підготовка Звіту. Захист практики.

5. Структура виробничої магістерської практики					
Назви змістових модулів	Усього го	Кількість годин			
		Денна форма			
		у тому числі			
		л	п	інд.	с р с
1	2	3	4	5	6
Змістовий модуль 1.					
	45				45
Змістовий модуль 2.					
	45				45
Усього годин	90				90

6. Вимоги до звіту практикантів

Кожен звіт повинен містити докладний і належний опис плану практики та виконаної практикантом роботи. Зокрема, кожен здобувач створює корпус текстів і словник на його основі, а також виконує додатковий комп'ютерний аналіз одержаних даних відповідно до завдань теоретичного розділу магістерської роботи.

7. Критерії оцінювання знань і навичок

Підсумковий контроль здійснюють керівник-методист від катедри та керівники від Українського мовно-інформаційного фонду НАН України.

Загальна оцінка результатів проходження практики здійснюється з урахуванням оцінки за звіт та публічний захист практики і становить сумарний підсумок. При оцінюванні звіту враховується письмове оформлення звітної документації, ступінь реалізації індивідуальної програми практики, характеристики керівників від бази практики та від кафедри, додержання календарного плану та графіка індивідуально-консультативної роботи тощо. За наявності негативної характеристики керівника від катедри або бази практики загальна оцінка практики не може бути позитивною.

8. Розподіл балів, які отримують студенти

	<i>Виробничу практику магістрантів-прикладників оцінюють за видами діяльності відповідно до розробленої системи балів.</i>	<i>Максимальна кількість балів -100 б.</i>
	<i>Види роботи</i>	
1.	Ознайомлення з діяльністю і завданнями Українського мовно-інформаційного фонду НАН України.	5 балів
2.	Створення корпусу текстів і словника на його основі.	25 балів
3.	Комп'ютерний аналіз одержаних даних відповідно до завдань теоретичного розділу магістерської роботи.	15 балів
4.	Ведення щоденника практики, оформлення звіту про проходження практики.	5 балів
5.	Захист практики.	50 балів

Шкала відповідності оцінок

Шкала оцінювання: національна та ECTS

Сума балів за всі види навчальної діяльності		Оцінка ECTS	Оцінка за національною шкалою	
			для екзамену, магістерського проекту (роботи), практики	для заліку
90 – 100		A	відмінно	зараховано
81-89		B	добре	
71-80		C		
61-70		D	задовільно	
51-60		E		
21-50		FX	незадовільно з можливістю повторного складання	не зараховано з можливістю повторного складання
0-20		F	незадовільно з обов'язковим повторним вивченням дисципліни	не зараховано з обов'язковим повторним вивченням дисципліни

У встановлений деканатом філологічного факультету термін студент має змогу захистити результати практики за талоном № 2 та за талоном форми „К”.

Студент, який не виконав програми, скеровується на практику вдруге в період канікул або відраховується з навчального закладу.

9. Методичне забезпечення

1. Положення про проведення практик здобувачів вищої освіти Львівського національного університету імені Івана Франка. Львів, 2021. 22с.
2. Робоча програма виробничої практики.
3. Силабуси з відповідних лінгвістичних дисциплін.

10. Рекомендована література

- 1.Бойко М. І. Українська комп'ютерна лексикографія: суспільні запити, проблеми та перспективи // Науковий вісник Міжнародного гуманітарного університету . Серія «Філологія». 2022. №56. С.12-15. <http://www.vestnik-philology.mgu.od.ua/archive/v56/3.pdf>
- 2.Ваховська О. В. Основи комп'ютерної лінгвістики: Навчально-методичний посібник. К.: Видавничий центр КНЛУ, 2023. 112 с
- 3.Волошин В. Г. Комп'ютерна лінгвістика. Суми : Університетська книга, 2004. 382 с.
- 4.Демська-Кульчицька О. Корпусна рецепція тексту // Наукові записки НаУКМА. Сер. Філологічні науки. 2010. Т. 111. С. 3–7.
- 5.Демська-Кульчицька О. Базові поняття корпусної лінгвістики // Українська мова. 2003. №1. С. 42–47.
- 6.Жуковська В. В. Вступ до корпусної лінгвістики: навчальний посібник / Житомир: Вид-во ЖДУ ім. І. Франка, 2013.
- 7.Калимон Ю. Комп'ютерна лексикографія: виклики та перспективи // Актуальні питання іноземної філології. 2019. № 10. С. 112–118.
- 8.Карпіловська Є.А. Вступ до прикладної лінгвістики: комп'ютерна лінгвістика: Підручник. Донецьк: ТОВ «Юго-Восток, Лтд», 2006. 188 с.
- 9.Кульчицький І. Унормування тексту під час докорпусного опрацювання: досвід застосування // Вісник Національного університету «Львівська політехніка». Серія: Інформаційні системи та мережі. 2020. Т. 7. С. 51–58.
- 10.Купріянов Є. Комп'ютерна лексикографія як проблема сучасного мовознавства. 2008. URL: http://repository.kpi.kharkov.ua/bitstream/KhPI-Press/2480/1/2008_Kupriyanov_Kompiuterna%20leksykohrafiia.pdf
- 11.Кушнір О. С. Основи комп'ютерної лінгвістики (конспект лекцій). Львів: Видавн. Львів. ун-ту, 2023. 292 с.
- 12.Лінгвістично-інформаційні студії : праці Українського мовно-інформаційного фонду НАН України : у 5 т. / В. А. Широков та ін. Т. 4 : Корпусна та когнітивна лінгвістика. Київ. Український мовно-інформаційний фонд НАН України. 2018. 246 с. https://ulif.mon.gov.ua/system/files/ling_inf_studio_tom_4_umif_b5.pdf
- 13.Надутенко Маргарита, Надутенко Максим, Семенов Олена. Застосування цифрового методу у викладанні філологічних дисциплін (на прикладі віртуальної лексикографічної лабораторії) // Волинь філологічна: текст і

контекст. Вип. 34: Філологія та цифрові технології / упоряд. Т. Левчук. Луцьк: Волин. нац. ун-т ім. Лесі Українки, 2022. С. 7–26.

14.Надутенко М. В. Загальний огляд та перспективи використання національних лінгвістичних цифрових ресурсів Українського мовно-інформаційного фонду НАН України // Проблеми загального та порівняльно-історичного мовознавства : тези доповідей міжнародної наукової конференції на пошану пам'яті професора В. В. Лучика, 3 березня 2020 р. / [орг. комітет: Куранова С. І., Лучик А. А. та ін.] ; Національний університет «Києво-Могилянська академія», кафедра загального і слов'янського мовознавства. Київ : НаУКМА, 2020. С. 91–95.

15.Сидоренко О. Українська комп'ютерна лексикографія як важливий інноваційний чинник навчального процесу // Актуальні проблеми слов'янської філології. 2010. Випуск XXIII. Частина 1. С. 524–528.

16.Широков В. Комп'ютерна лексикографія : монографія. К. : Наук. думка, 2011. 351 с.

17.Широков В.А., Бугаков О. В., Грязнухіна та ін. Корпусна лінгвістика. К. : Довіра, 2005.

18.Широков В. А., Надутенко М. В., Стрижак О. Є., Ющенко С. С. Технологічні засади логіко-лінгвістичних досліджень законодавства // Науково-технічний журнал «Біоніка інтелекту». 2020. Т. 2, № 95. С. 3–14.

19.Широков В. А., Ющенко С. С. Лінгво-експертна діяльність Українського мовно-інформаційного фонду: теорія, методологія, практика, інструменти // Українська мова в юриспруденції: стан, проблеми, перспективи [Текст] : матеріали XVIII Всеукр. наук.-практ. конф. (Київ, 17.10.2022 р.) / [редкол.: В. В. Черней, С. Д. Гусарев, С. С. Чернявський та ін.]. Київ : Нац. акад. внутр. справ, 2022. С. 69–74.

11. ДОДАТКИ

Основні категорії, які розкривають зміст практики

1) Дисципліна «Комп'ютерна лінгвістика і опрацювання природної мови»

Комп'ютерна лінгвістика — «галузь мовознавства, що вивчає мову за допомогою комп'ютера, а також створює лінгвістичне забезпечення для комп'ютерних систем опрацювання інформації. Як самост. наук. напрям сформувалася в 1960-ті рр. на базі досягнень структур., матем. та приклад. лінгвістики, лінгвосеміотики, а також обчислюв. техніки, кібернетики й інформатики. Появу К. л. спричинила потреба суспільства в нових оператив. способах опрацювання мовної інформації, зокрема необхідність створення систем машин. перекладу, в основу яких покладено формал. аналоги мови. У світ. науці сформувалося широке розуміння об'єкта й предмета вивчення К. л., що розв'язує як фундам. теор., так і прикладні завдання сучас. мовознавства. До фундам. завдань належить створення моделей мовних явищ і процесів, придат. для опрацювання комп'ютером, що передбачає виявлення таких закономірностей будови та функціонування мови, які можна описати з використанням формал.-логіч. і матем. методів. Різноманітні моделі мови — статичні й динам., дедуктивні й індуктивні, аналіт. (інтерпретац.) й синтетичні (генеративні, породжувальні) — закладають основу для розв'язання практич. завдань комп'ютер. опрацювання мовної інформації: автоматич. укладання словників, створення систем машин. перекладу, інформ.-пошук., навч. та експерт. систем, корпусів мови, аналізаторів і синтезаторів усного мовлення.

Провід. напрямками сучас. К. л. є комп'ютерна лексикографія, корпусна лінгвістика, автоматич. аналіз тексту. З появою Інтернету формується новий напрям — мережна лінгвістика, або інтернет-лінгвістика (*Карпіловська Є. А.* Комп'ютерна лінгвістика // Енциклопедія сучасної України, 2014; <https://esu.com.ua/article-4396>).

Комп'ютерна лексикографія — «галузь комп'ютерної лінгвістики та лексикографії, яка спрямована на розроблення комп'ютерних технологій створення і використання словників різних типів. Саме комп'ютерна лексикографія покликана виконувати низку прикладних завдань, а саме формування: — комп'ютерних лексикографічних баз; — машинних фондів національних мов; — лексиконів як додаткової інформації для лінгвістичних процесорів комп'ютерної обробки мови; — інформаційно-пошукових систем тощо. Комп'ютерна лексикографія характеризується низкою дистинктивних ознак, а саме: — застосування палітри кольорів та візуальних ефектів; — висвітлення значного обсягу інформації за рахунок використання гіперпосилань; — розширені можливості пошуку відповідної лексеми є не лише у межах певної словникової статті, але й усього словника; — можливість використання функції виноски, що дозволяє додавати особистий коментар, власний перекладу тощо; —

використання мультимедійних елементів (ілюстрації, анімації, аудіозаписи, відеоролики); – можливість своєчасного систематичного оновлення інформації; – можливість вбудовуватися в основні офісні програми; – доступність та простота у використанні; – економія часу та матеріальних ресурсів. Відтак одним з пріоритетних завдань комп'ютерної лексикографії є створення електронних словників, які виконують функцію збереження, маніпуляції та передачі інформації. Безперечно вміння ефективно працювати з інформацією сприяє успіху в усіх галузях життєдіяльності людини. Відповідно ефективність електронних словників безсумнівна, оскільки на відміну своїх традиційних відповідників, вони характеризуються можливістю безперервного поповнення, швидким алгоритмічним пошуком слів тощо. Важливим напрямом електронної лексикографії є науково-технічна лексикографія, яка теоретично базується на галузі термінознавства. Термінологічна лексикографія (термінографія) займається безпосередньо лексикографуванням терміна. Термінографія займає проміжну позицію між лінгвістичною та енциклопедичною лексикографією, оскільки, не лише описує термінологічну одиницю як одиницю мовної системи, а й окреслює понятійне, предметне наповнення певного терміна» (Черниш О. А., Білошицька З. А., Кравченко А. О. Сучасна лексикографія: виклики та перспективи// Вчені записки Таврійського ТНУ імені В. І. Вернадського. Серія: філологія. Журналістики. Т. 33(72), №6. Ч. 2. 2022. С. 11 https://www.philol.vernadskyjournals.in.ua/journals/2022/6_2022/part_2/2.pdf).

Український мовно-інформаційний фонд НАН України

«Основні напрямки діяльності: дослідження системної будови природної мови; створення та ведення фундаментального архіву української та інших мов народів світу; розробка і створення інформаційно-лінгвістичних систем (традиційних і комп'ютерних словників, підручників, навчальних, експертних систем, систем автоматизованого опрацювання природної мови тощо, лінгвістичних баз даних та знань); дослідження лінгвістичних аспектів інтелектуальної діяльності; розробка та експлуатація інформаційних систем, баз даних і знань загальнокультурного характеру; координація робіт в Україні у галузі лінгвістичної технології, комп'ютерної та когнітивної лінгвістики і лексикографії, налагодження зв'язків з фаховими організаціями світу із зазначених галузей» В Українському мовно-інформаційному фонді створено теорію лексикографічних систем, яку за своїм логіко-лінгвістичним статусом може бути прирівняно до теорії формальних граматики. Теорія лексикографічних систем, а також її узагальнення - лексикографічні середовища, лексикографічні числення тощо визначають підґрунтя сучасної лінгвістичної технології.

В Українському мовно-інформаційному фонді розвинено теорію семантичних станів, яка становить базу для формалізованого опису широкого кола мовних

явищ, об'єднуючи граматичний та лексикографічний різновиди опису мовної системи. Український мовно-інформаційний фонд започаткував (1994 р.) і координує програму зі створення серії українськомовних словників нового покоління - "Словники України", яка вже налічує понад 70 праць. Серед них слід відзначити перший в Україні повномасштабний електронний словник української мови - Інтегровану лексикографічну систему "Словники України", одинадцять версій якої промислово випущено на лазерних дисках, починаючи з 2001 року.

В Українському мовно-інформаційному фонді створено перший в Україні лінгвістичний корпус, який має обсяг понад 200 млн. слововживань і становить сучасну експериментальну основу для проведення фундаментальних мовознавчих досліджень та укладання лексикографічних праць нового покоління. Розроблено і введено в експлуатацію інформаційно-лінгвістичну систему, що функціонує в Інтернеті - Український лінгвістичний портал (www.ulif.org.ua).

Український мовно-інформаційний фонд НАН України створив і забезпечує функціонування та розвиток єдиного в Україні лінгвістичного наукового об'єкта, який рішенням Уряду України внесено до державного реєстру наукових об'єктів, що мають статус національного надбання, а саме Національної словникової бази Українського мовно-інформаційного фонду НАН України» <https://www.ulif.org.ua/about>

Інтегрована лексикографічна система «Словники України»

«Український мовно-інформаційний фонд НАН України (УМІФ) є установою, основною науково-технічною орієнтацією якої є комп'ютерні лінгвістичні технології та лінгвістичні системи, серед яких понад 50 електронних словникових та інших мовно-інформаційних систем, а також об'єктів Національної словникової бази (<https://lcorp.ulif.org.ua/LSlist>). УМІФ започаткував (1994 р.) та координує програму зі створення серії українськомовних словників нового покоління — «Словники України», яка налічує понад 80 праць. Словникові системи створюються на основі унікального інструменту УМІФ — Віртуальної лексикографічної лабораторії, що забезпечує колективну роботу в режимі віддаленого доступу. Сучасні словники адаптуються до вимог сьогодення та активно наповнюються мультимедійним контентом.

Актуальність проблеми полягає у необхідності трансформації інформаційного цифрового простору у лінгвістичні системи нового типу. Серед таких ми виділили мультимедійні лінгвістичні технології, які формують особливу соціокультурну й технологічну можливість для створення специфічного

насиченого інформаційного поля та надають інформацію в якості нового ресурсу й активації інтересів та рівня компетентності за допомогою найрізноманітніших засобів в електронному форматі. Словникова форма як тип соціальної комунікації все більше приваблює спеціалістів найрізноманітніших предметних галузей, оскільки за повнотою, змістовністю, точністю й достовірністю інформації словники, безперечно, посідають перше місце серед інших форм, які представляють дані у стислому вигляді. Прикладом такої системи є «Мультимедійний словник з інфомедійної грамотності» (рис. 1). Для реалізації проекту «Словник української мови» було розроблено спеціальну комп'ютерну технологію, яка ґрунтується на методі віртуальних лексикографічних лабораторій (ВЛЛ). Зазначена технологія надає засоби для виконання спільних лексикографічних проєктів фахівцями різних територіально розподілених інституцій та навіть різних країн» (*Надутенко М. В. Технологія створення мультимедійних словників: етапи процесу та приклади програмних продуктів Українського мовно-інформаційного фонду / Маргарита Надутенко // Мовний простір сучасного світу : тези доповідей VII Всеукраїнської наукової конференції студентів, аспірантів і молодих учених (м. Київ, 26 травня 2023 р.) / [орг. комітет: Куранова С. І. (голова) та ін.] ; Національний університет "Києво-Могилянська академія", Кафедра загального і слов'янського мовознавства [та ін.]. Київ : НаУКМА, 2023. С. 132-133).*

Портал «Mova.info: про українську мову, лінгвістику і не тільки»
Інформаційно-довідковий портал з української мови, метою якого є здійснення довідково-інформаційної роботи стосовно української мови та української лінгвістики в мережі Інтернет. «Ключове завдання лінгвістичного порталу “Mova.info: про українську мову, лінгвістику і не тільки” – мовне консультування суспільства. На сайті розміщено чимало наукових статей, монографій, підручників та інших матеріалів наукового характеру, пов'язаних з лінгвістикою, а також надано безоплатний доступ до інформаційних ресурсів (корпус української мови, електронні словники). Досліджуючи вектори української комп'ютерної лексикографії, варто виокремити кілька лінгвістичних напрямів, які представлено на порталі “Mova.info”:

- частотні словники;
- граматичний словник української мови;
- електронний словник мови Тараса Шевченка;
- проєкт словника суржикі;
- відкритий словник;
- чотиримовний науковий словник;
- тезаурус з комп'ютерної лексикографії та тезаурус з лінгвістичної термінології;
- словник українських прийменникових колокацій та словник українських предикативних колокацій;
- семантичний словник української мови.

Проєкт створено співробітниками Інституту філології Київського університету імені Тараса Шевченка» (*Бойко М. І. Українська комп'ютерна лексикографія: суспільні запити, проблеми та перспективи // Науковий вісник Міжнародного гуманітарного університету. Серія «Філологія». 2022. №56. С.12-15. <http://www.vestnik-philology.mgu.od.ua/archive/v56/3.pdf>).*

2) Дисципліна «Кількісні та корпусні підходи у прикладній лінгвістиці»

Корпусні технології

«Їхня основа – корпус текстів, під яким розуміємо електронну збірку текстів, яку споряджено фаховою лінгвістичною інформацією у придатній для опрацювання комп'ютером формі та, за необхідності, програмним зняряддям, яке спрощує доступ до цієї інформації. Великою перевагою таких технологій є те, що дослідникам мови не доводиться покладатися на власну інтуїцію чи на інтуїцію носіїв мови або навіть на вигадані приклади. Вони можуть користуватися великою кількістю автентичних, природних лінгвальних даних, отриманих різними мовцями чи письменниками, щоб підтвердити або спростувати власні гіпотези щодо тих чи інших мовних явищ. Збирання автентичних мовних даних з корпусу дає змогу описувати мову, починаючи з доказів, а не з нав'язування певної теоретичної моделі. Позаяк в Україні корпусні технології перебувають на етапі розвитку та становлення, актуальними залишаються проблеми створення, наповнення та використання корпусів» (*Кульчицький І. Унормування тексту під час докорпусного опрацювання: досвід застосування // Вісник Національного університету «Львівська політехніка». Серія: Інформаційні системи та мережі. 2020. Т. 7. С. 51).*

Когнітивна ІТ-платформа «ПОЛІЕДР» (КІТ «ПОЛІЕДР»)

«Когнітивна ІТ-платформа «ПОЛІЕДР» (КІТ «ПОЛІЕДР») – це комп'ютерна програма, створена у співпраці науковців Українського мовно-інформаційного фонду НАН України, Національного центру «Мала академія наук України» та Інституту кібернетики імені В. М. Глушкова НАН України. КІТ «ПОЛІЕДР» призначена для підтримки процесів концептографічного аналізу великих обсягів просторово розподіленої неструктурованої інформації (Big Data), її структуризації, встановлення контекстних зв'язків, прогнозування та підтримки процесів раціонального вибору з наступним формуванням інформаційно-аналітичних WEB-орієнтованих рішень. КІТ «ПОЛІЕДР» пройшла достатню апробацію на текстах реальної складності для вирішення прикладних завдань, пов'язаних із обробкою великих масивів у різних сферах професійної діяльності (національна оборона та безпека, телекомунікаційні послуги, охорона здоров'я та ін.)» (*Широков В. А., Надутенко М. В., Ющенко С. С. Комп'ютерні інтелектуальні технології в антикорупційній експертизі законодавчих та правозастосовних процесів // International scientific conference «New approaches and current legal research», November 3–4, 2022. Riga, Latvia).*

«Призначенням когнітивної ІТ-платформи «ПОЛІЕДР» (КІТ «ПОЛІЕДР»), операційною основою якої є здатні до навчання нейронні мережі, є підтримка процесів концептографічного аналізу розподіленої неструктурованої інформації

великих обсягів (Big Data). Зазначені процеси передбачають, зокрема, структурування інформації – розпізнавання в аналізованих текстах змістовно навантажених об'єктів – концептів (термінів, понять), встановлення атрибутів цих об'єктів та контекстних зв'язків між ними із подальшим аналізом на предмет, визначений користувачем. Онтограф – це ієрархічна структура сукупності понять певної предметної області, кореневими вершинами якої є визначені експертами поняття або лексико-семантичні групи, розпізнані програмою як поняття. Дуги позначають смислові зв'язки між поняттями. Поняття кожного рівня пов'язуються із поняттями наступного (вищого), доки усі дуги онтографа не зійдуться в одній вершині (вона має назву «листова»), що надає структурі онтографа завершеності. Кількість рівнів визначається користувачем із міркувань достатнього для цілей дослідження рівня деталізації понятійного апарату предметної області, репрезентованого в тексті (текстовому масиві), що походить з цієї області. Фрагмент онтографа надає користувачеві знання про засади та заходи із формування негативного ставлення суспільства до корупції – одного з ключових факторів успішної їй протидії» (*Широков В. А., Ющенко С. С.* Лінгво-експертна діяльність Українського мовно-інформаційного фонду: теорія, методологія, практика, інструменти // Українська мова в юриспруденції: стан, проблеми, перспективи [Текст] : матеріали XVIII Всеукр. наук.-практ. конф. (Київ, 17.10.2022 р.) / [редкол.: В. В. Черней, С. Д. Гусарев, С. С. Чернявський та ін.]. Київ : Нац. акад. внутр. справ, 2022. С. 71–73).

Методи дослідження цифрової лінгвістичної інформації

«Фахівцями УМІФ НАН України розроблено теоретичні та науково-технічні засади методів дослідження цифрової лінгвістичної інформації. Статистичний метод дослідження розглядає мову як системно-структурне утворення. Методи статистичної лінгвістики використовують для лінгвістичного моніторингу функціонування мови в конкретному типі дискурсу (політичному, науковому, засобів масової інформації тощо), для контент-аналізу (виявлення стану суспільної свідомості). Метод корпусних технологій доцільно використовувати на основі вже наявних програмних продуктів як необхідну складову для створення мінікорпусів відкритого типу. Нейромережі – це підмножина систем штучного інтелекту, що зосереджена переважно на проєктуванні систем, які дозволяють навчатися та робити прогнози на основі певного досвіду. Штучний інтелект є формою індивідуалізації систем, якій властивий мовний статус. Лексикографічний метод доцільно застосовувати для сучасних досліджень лексичної системи мови. Метод спрямований передусім на послідовне виділення та вибіркоче вивчення окремих елементів та їх відношень у мовній системі (*Надутенко Маргарита, Надутенко Максим, Семенов Олена.* Застосування цифрового методу у викладанні філологічних дисциплін (на прикладі віртуальної лексикографічної лабораторії) // Волинь філологічна: текст і

контекст. Вип. 34: Філологія та цифрові технології / упоряд. Т. Левчук. Луцьк: Волин. нац. ун-т ім. Лесі Українки, 2022. С. 22–23).

Унормування тексту

«Перший етап створення корпусу – це збирання даних, що передбачає отримання текстів в електронній формі чи ручним складанням, чи розпізнаванням за допомогою програм OCR, чи як результат роботи текстового процесора, чи з PDF-файла тощо. Дані, отримані в електронному вигляді з інших джерел, майже завжди містять коди форматування та іншу інформацію, яку треба забрати або перетворити на придатну для комп'ютерного аналізу форму. Такий процес називають внормуванням тексту, під яким розуміємо сукупність інформаційних процедур, що роблять текст придатним до внесення його в корпус: приведення всіх текстів до однієї кодової таблиці, перевірку їх на пунктуаційну коректність (однакові за смыслом сутності мають бути позначені одним знаком), усунення зайвих символів (наприклад, порожні абзаци, декілька прогалин підряд і т. ін.), уніфікацію засобів та способів форматування тощо. Досвід доводить, що процес збирання та готування текстів до внесення їх у корпус є стилезалежним. Звісно, багато дій є однаковими для всіх стилів тексту, проте завжди у структурі тексту будуть особливості, що притаманні лише конкретному стилю (наприклад, поклики на джерела в науковому стилі). Такими особливостями можна нехтувати під час роботи з текстами одного стилю й обов'язково враховувати, опрацьовуючи тексти інших стилів (*Кульчицький І. Унормування тексту під час докорпусного опрацювання: досвід застосування // Вісник Національного університету «Львівська політехніка». Серія: Інформаційні системи та мережі. 2020. Т. 7. С. 52).*

Система аналізу текстів

«Система аналізу текстів (САТ) – програма, що дає змогу кількісно оцінювати семантичну близькість текстів за статистикою їхніх лексичних систем, визначати обсяг і зміст змін, внесених у вихідний текст на кожному з етапів його модифікації, й, для зручності користувача, візуалізувати ці зміни. У дослідженнях законодавства САТ може застосовуватися для визначення лінгвістичної узгодженості проєктів і чинних нормативно-правових актів із актами, яким вони мають відповідати (до таких належать, зокрема, Конституція України, Регламент Верховної Ради, міжнародні договори України, акти Європейського Союзу тощо).

Операційною основою САТ є метод N-грам, згідно з яким текст мислиться множиною комбінацій слів як деяких випадкових подій. Деякий аналізований текст T2 є слабо синонімічним (близьким за змістом) до еталонного тексту T1, якщо аналізований текст має з еталонним спільні: 1) 60 % уніграм (окремі слова)

або 2) 30% біграм (два слова, що стоять поруч) або 3) 15 % триграм (трійки слів).
Задача локалізації та підрахунку N-грам досить нескладно розв'язується
формальними методами» (*Широков В. А., Надутенко М. В., Ющенко С. С.*
Комп'ютерні інтелектуальні технології в антикорупційній експертизі
законодавчих та правозастосовних процесів // International scientific conference
«New approaches and current legal research», November 3–4, 2022. Riga, Latvia).