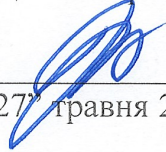


Львівський національний університет імені Івана Франка
Філологічний факультет
Катедра загального мовознавства

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


“27” травня 2025 року

РОБОЧА ПРОГРАМА

ВИРОБНИЧОЇ ПРАКТИКИ З КОМП'ЮТЕРНОЇ ТА КОРПУСНОЇ ЛІНГВІСТИКИ

галузь знань	<u>03 Гуманітарні науки</u>
спеціальність	<u>035 Філологія</u>
спеціалізації	<u>035.10 Прикладна лінгвістика</u>
факультет	<u>філологічний</u>

Львів – 2025 рік

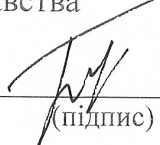
Робоча програма виробничої практики з комп'ютерної та корпусної лінгвістики для студентів 2 курсу другого (магістерського) рівня вищої освіти за спеціальністю 035 Філологія, спеціалізацією 035.10 Прикладна лінгвістика. Львів: ЛНУ імені Івана Франка, 2025.

Розробники: Мацюк Г.П., докт. філол. наук, професор;
Григорук 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 кафедри загального мовознавства.

Протокол від "30" квітня 2025 року № 8.

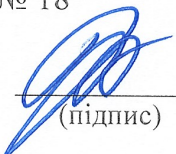
Завідувач кафедри загального мовознавства


_____ (проф. Бацевич Ф.С.)
(підпис)

"30" квітня 2025 року

Затверджено Вченою радою філологічного факультету.

Протокол від "27" травня 2025 року № 18

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

"27" травня 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> 03 Гуманітарні науки	Нормативна	
	<u>Спеціальність</u> 035 Філологія		
Модулів – 2	<u>Спеціалізація:</u> 035.10 Прикладна лінгвістика	Рік підготовки:	
Змістових модулів – 2		2-й	
Індивідуальне науково-дослідне завдання _____ (назва)		Семестр	
Загальна кількість годин - 90		3-й	
		Лекції	
		Практичні, семінарські	
Тижневих годин для денної форми навчання: аудиторних – самостійної роботи студента -	<u>Освітній ступінь:</u> магістр	Лабораторні	
		Самостійна робота	
		90 год.	
		Індивідуальні завдання:	
		Вид контролю: диференційований залік	

Практика з комп'ютерно-корпусної лінгвістики – обов'язковий компонент навчального процесу в отриманні другого (магістерського) рівня вищої освіти. Вона забезпечує поєднання теоретичної підготовки майбутніх прикладних лінгвістів з їхньою практичною діяльністю у різних установах працевлаштування, сприяє розвитку практичних навиків у використанні інформаційних технологій у текстотворенні та лінгвоекспертології.

Головна мета практики для магістрів 2024 року вступу полягає в закріпленні вже отриманих навичок, пов'язаних із застосуванням інформаційних технологій у лінгвістиці, знання про які здобувачі отримали на заняттях з корпусної лінгвістики (викладач – доц. Н. Я. Лотоцька).

Під час виробничої практики здобувачі знайомляться із діяльністю провідної установи з розробки та використання інформаційних технологій, а саме Українським мовно-інформаційним фондом НАН України, оскільки відомо, що Фонд є координатором та основним виконавцем робіт зі створення Українського національного лінгвістичного корпусу.

Керівником практики від цієї науково-дослідної інституції є кандидат технічних наук, завідувач відділу інформатики Максим Надутенко. Також зі студентами буде контактувати кандидат філологічних наук, старший науковий співробітник Маргарита Надутенко, яка є Вченим секретарем інституту.

Практика надає можливість використати інформаційні технології в текстотворенні та лінгвоекспертології. Вид практики, її тривалість і терміни проведення визначені оновленим під впливом рекомендацій та зауважень ГЕР робочим навчальним планом для слухачів 2024/2025 року вступу.

Пререквізити практики: прослухані заняття з впровадження інформаційних технологій у лінгвістиці (корпусна лінгвістика) на магістерському семінарі (викладач – доц. Лотоцька Н.Я.).

2. Мета і завдання виробничої практики

Корпусна лінгвістика формує методологію про способи використання конкретних ресурсів, що представляють великі обсяги текстових даних. Корпуси – один із найпотужніших інструментів прикладної лінгвістики. Вони створюються та використовуються у різних галузях людської діяльності на основі автоматизації процесу підбору, укладання та аналізу текстових масивів практично необмеженого обсягу.

Для кожного із магістрів-прикладників корпус і спеціальні програмні засоби роботи з цим корпусом є спеціалізованим інструментом лінгвістичного аналізу. Магістри повинні отримати навички роботи з усіма складовими корпусних завдань, а саме: створення власних корпусів, опанування різних видів корпусної розмітки,

використання Інтернету для корпусних досліджень та вживання різних статистичних методів при роботі з корпусами.

Тривалість практики магістрів в Українському мовно-інформаційному фонді – 2 тижні (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, які дають змогу висловити свої ідеї та висновки чітко й професійно. На практиці магістри працюють з теоретичною частиною своєї магістерської роботи. Головні джерела для корпусів, які використовують магістри, – наукові тексти, що є у вільному доступі в мережі Інтернет, які вони використовують для написання теоретичного розділу.

Тематика магістерських досліджень здобувачів ОПІ «Прикладна лінгвістика» 2024 року вступу:

Костельна Л. Кольористика в процесах текстотворення: концепт «ЧЕРВОНІЙ» в національно-мовній картині українців сучасних років. (проф. Т.Єщенко).

Луньо М. Лінгвістична експертиза соціальних мереж: дослідження маніпулятивних впливів у час гібридних викликів національній безпеці України (доц. Л. Гонтарук).

Мрак К. Лінгвостатистичні характеристики паралельного корпусу медійних текстів (проф. Ф. Бацевич).

Федорів Д.-М. Прагматичний потенціал текстів про кохання у формуванні лінгвістичної компетентності іноземців (на прикладі текстів Івана Франка та Лесі Українки) (проф. Ф. Бацевич, асист. О. Ясіновська).

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

Завдання практики:

1. Обрати 20 наукових статей, які входять в теоретичний розділ магістерського дослідження, для створення текстового корпусу.
2. Здійснити нормалізацію текстів: видалити зайві символи (пробіли, коди, знаки форматування); адаптувати форматування для сумісності з програмним забезпеченням Sketch Engine.

3. 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.

4. Створити словникові статті. Для цього вибрати найчастотніші 5 термінів із корпусу і укласти щодо них словникові статті за схемою: термін, визначення, приклади вживання.

4. Переглянути відеоінструкцію щодо використання капсули E-devel. Це допоможе зрозуміти основні функції платформи та способи роботи з нею.

5. Занести інформацію про себе до капсули E-devel, увійшовши до системи капсули та додавши профільну інформацію (розповідь про себе; спеціалізацію; наукові публікації; нагороди, сертифікати).

6. Занести інформацію про роботу з корпусом.

7. Занести інформацію про словник, вказавши: мету (пояснення ключових термінів); специфіку використання (для науковців, студентів, викладачів); критерії вибору термінів (релевантність і частотність).

До капсули додати звіт про роботу з корпусом: опис методики, основні функції, інструменти аналізу.

Виробнича практика спрямована на формування таких загальних 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

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.

4. Програма виробничої практики

Змістовий модуль 1 передбачає:

Знайомство з діяльністю Українського мов-інформаційного центру НАН України

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.

Змістовий модуль 2 передбачає:

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

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

Узагальнення результатів практики. Підготовка Звіту. Захист практики.

5. Структура виробничої магістерської практики					
Назви змістових модулів		Кількість годин			
		Денна форма			
	Усього го	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 на його базі словник відповідно до завдань теоретичного розділу магістерської роботи.

7. Критерії оцінювання знань і навичок

Підсумковий контроль здійснюють керівник-методист від кафедри 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 кафедри 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.	35 балів
4.	Ведення щоденника практики	5 балів
5.	Оформлення звіту про проходження практики	5 балів
6.	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

- Демська-Кульчицька О. Корпусна рецепція тексту // Наукові записки НаУКМА. Сер. Філологічні науки. 2010. т. 111. С. 3–7.
- Демська -Кульчицька О. Базові поняття корпусної лінгвістики // Українська мова. 2003. №1. С 42-47.
- Жуковська В. В. Вступ до корпусної лінгвістики: навчальний посібник. Житомир: Вид-во ЖДУ ім. І Франка. 2013.
- 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.
- 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
- 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.
- 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.
- Широков В.А. Бугаков О. В. Грязнухіна та ін. Корпусна лінгвістика. К.. Довіра, 2005.
- Широков В. А., Надутенко М. В., Стрижак О. Є., Ющенко С. С. Технологічні засади логіко-лінгвістичних досліджень законодавства //Науково-технічний журнал «Біоніка інтелекту». Т. 2. № 95 (2020). С. 3–14.
- 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 листоп. 2022 р.) / [редкол.: В. В. Черней, С. Д. Гусарєв, С. С. Чернявський та ін.]. Київ : Нац. акад. внутр. справ, 2022. С. 69-74.

10. Додатки

Про напрями діяльності Українського мовно-інформаційного фонду НАН України

«Розвиток теорії, методології, лінгвістичних технологій та програмних інструментів підтримки та супроводу текстоаналітичної та лінгвоекспертної діяльності відображено у наукових звітах Фонду за результатами НДР та в понад двох сотнях наукових праць співробітників, аспірантів, докторантів Фонду (публікації у вітчизняних та закордонних фахових виданнях, монографії, дисертаційні роботи тощо), що охоплюють усі відомі на сьогодні аспекти лінгвістичної експертизи: теоретичні засади; лінгвістичні дослідження законодавства як масиву природномовних текстів; авторознавчі дослідження та лінгвоперсонологія; дослідження на ознаки плагіату; стилістичні дослідження; семантичні дослідження; дискурс-аналіз; статистичні дослідження; термінологічні дослідження; концептографічні дослідження; онтологічні дослідження; квантитативний аналіз; класифікаційний аналіз; комп'ютерні корпусні та інтелектуальні технології.

Науковцями Фонду розроблено комп'ютерні інструменти підтримки текстоаналітичної та лінгвоекспертної діяльності – Систему аналізу текстів [2] та Когнітивну ІТ-платформу «ПОЛПЕДР»¹ (у співпраці з Національним центром «Мала академія наук України» та Інститутом кібернетики імені В. М. Глушкова НАН України), аплікабельність яких для вирішення прикладних завдань, пов'язаних із аналізом природномовних текстів у різних галузях знань та сферах діяльності, можна вважати успішно доведеною [3]. (*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 листоп. 2022 р.) / [редкол.: В. В. Черней, С. Д. Гусарев, С. С. Чернявський та ін.]. Київ : Нац. акад. внутр. справ, 2022. С 71).*

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 вважаємо за доцільне використовувати на основі вже наявних програмних продуктів як необхідну складову для створення мінікорпусів відкритого типу. Нейромережі розглядаємо як підмножину штучного інтелекту, що зосереджена переважно на проектуванні систем, які дозволяють навчатися та робити прогнози на основі певного досвіду.

Штучний інтелект розглядаємо як форму індивідуалізації систем, якій властивий мовний статус. Лексикографічний метод вважаємо за доцільне застосовувати для сучасних досліджень лексичної системи мови. Метод спрямований передусім на послідовне виділення та вибіркове 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) // Волинська 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).*

Основні категорії, які характеризують зміст практики

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 знаряддям, 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).*

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 [5]. Такий процес називаємо 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 прогалін підряд і т. ін.), уніфікацію засобів та способів форматування тощо [17].

Власний досвід доводить, що процес збирання та готування текстів до внесення їх у корпус є стилезалежним. Звичайно, багато дій є однаковими для всіх стилів тексту. Однак 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. 51).*

Система аналізу текстів (САТ)

«Система аналізу текстів (САТ) – програма, що дозволяє 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 та 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).

Операціональною основою САТ є метод N-грам, згідно з яким текст мислиться множиною комбінацій слів як деяких випадкових подій. Аналізований текст T2 є слабко синонімічним (близьким за змістом) еталонному 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.

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

«Когнітивна ІТ-платформа «ПОЛІЕДР» (КІТ «ПОЛІЕДР») – 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 України³. 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 провозастосовних процесів // International scientific conference «New approaches and current legal research», November 3–4, 2022. Riga, the Republic of Latvia).

«Призначенням Когнітивної ІТ-платформи «ПОЛІЕДР» (КІТ «ПОЛІЕДР»), операціональною 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. На Рис. 3 наведено фрагмент онтографа, побудованого Когнітивною ІТ-платформою «ПОЛІЕДР» на тексті Закону України «Про засади державної антикорупційної політики в Україні (Антикорупційна стратегія) на 2014–2017 роки» від 14 жовтня 2014 року № 1699-VII. Онтограф – це ієрархічна структура сукупності понять певної предметної області, кореневими вершинами якої є визначені експертами поняття або, як у розглянутому нижче прикладі, 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 (Рис. 3) надає користувачеві знання про засади та заходи із формування негативного ставлення суспільства до корупції – одного з ключових факторів успішної їй протидії. Згідно з розглянутим Законом, зазначений процес спирається на антикорупційний потенціал суспільства та передбачає низку заходів (внесення змін до Кримінального кодексу України; проведення інформаційних кампаній, антикорупційних та освітніх заходів; подолання пасивності тощо). Зрозуміло, що перелік заходів, наведених в окремому Законі, не є вичерпним, що має бути враховано в інших актах антикорупційного законодавства» (*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 листоп. 2022 р.) / [редкол.: В. В. Черней, С. Д. Гусарев, С. С. Чернявський та ін.]. Київ : Нац. акад. внутр. справ, 2022. С. 71-73).*