

Юлія Цигвинцева

**Проблематика Міжнародного семінару
«Інструменти комп'ютерної лінгвістики та ефективні методики
опрацювання природної української мови»
(Єнський університет імені Фрідріха Шиллера,
Німеччина, 5 грудня 2022 р.)**

5 грудня 2022 року на базі Єнського університету імені Фрідріха Шиллера (Friedrich-Schiller-Universität Jena) відбувся Міжнародний семінар «Інструменти комп'ютерної лінгвістики та ефективні методики опрацювання природної української мови», в якому взяли участь науковці-фахівці з опрацювання природної мови, прикладної та квантитативної лінгвістики. Цей науковий захід мав на меті сприяти обміну досвідом і найкращими здобутками у викладанні комп'ютерної лінгвістики, навчанні програмування студентів філологічних спеціальностей, у взаємодії технічних та філологічних дисциплін у навчальних планах українських вишів, навчанні із залученням корпусів та використанні додаткових програмних засобів у викладанні мови. Семінар проведено аудиторно в Єні та одночасно у форматі відеоконференції на платформі Zoom.

Відкрив захід керівник українознавчих студій у Єнському університеті *проф. д-р Рупрехт фон Вальденфельс*, який привітав учасників і розповів про ініціативи на підтримку українських академічних досліджень та створення при університеті мережі українських і західних науковців, які працюють над спільними проектами, що дає можливість українським ученим в Україні та за кордоном продовжувати й розвивати свої дослідження, налагоджувати партнерство, яке триватиме після закінчення нинішньої російсько-української війни, зміцнювати позиції україністики в світі у важкі для України і української науки часи. Професор ознайомив присутніх із сайтом українознавчих студій у Єні (<https://www.ukr.uni-jena.de/>), звернув увагу на програми онлайн-навчання, орієнтовані на студентів з України, стипендіальні пропозиції та інші можливості, розповів про команду, яка працює над проектами й окреслив порядок роботи нинішнього семінару.

Модерували семінар Р. фон Вальденфельс та учасники програми О. Каніщева (Національний технічний університет «Харківський політехнічний інститут») і Н. Чейлитко (Єнський університет). Усього було виголошено та обговорено 11 доповідей.

С. Бук у доповіді «Курс «*Основи статистичної лінгвістики*» у системі філологічної освіти прикладного лінгвіста» розповіла про свій навчальний посібник «Основи статистичної лінгвістики» для студентів філологічного факультету Львівського національного університету імені Івана Франка, зауважила, що в лінгвістиці зламу ХХ – ХХІ століть особливо актуальні міждисциплінарні дослідження мови, схарактеризувала зв'язки статистичної лінгвістики з іншими мовознавчими дисциплінами, презентувала структуру посібника та його змістове й методичне наповнення. Соломія Несторівна звернула увагу на актуальність статистичних досліджень, використа-

вши цитату німецького лінгвіста Г. Альтмана, про те, що «кожна достатньо розвинута наукова дисципліна рано чи пізно, принаймні, на певному етапі свого розвитку може опинитися на порозі математизації», зазначила, що саме статистичні методики з комп'ютерною підтримкою відкривають нові шляхи для дослідження мови та літератури, а також мають величезний потенціал для розв'язання багатьох теоретичних завдань лінгвістики і практичних завдань опрацювання текстових даних, бо результати, виявлені методами статистичної лінгвістики, плідно застосовують у багатьох сферах сучасної науки: судовій та кримінальній лінгвістиці, лінгводидактиці, дешифруванні зниклих писемностей, глотохронології, стенографії, стилеметрії, комп'ютерних технологіях тощо. На думку доповідачки, без такого курсу складно уявити підготовку сучасного висококваліфікованого фахівця-лінгвіста.

О. Канищева з НТУ «Харківський політехнічний інститут» як спеціалістка з комп'ютерних наук та автоматичного опрацювання природної мови в доповіді *«Викладання технічних дисциплін на гуманітарних спеціальностях: труднощі та перспективи»* поставила перед слухачами кілька проблемних питань. Що легше: навчити комп'ютерного спеціаліста лінгвістики чи філолога програмування? Коли викладати програмування? У якому обсязі та яким чином це робити? Доповідачка вбачала головну проблему українських навчальних програм із комп'ютерної лінгвістики в тому, що всі прикладні філологічні курси занадто комп'ютеризовані як для філологів, які не завжди планують вивчати математику після школи, часто мають складнощі з опануванням програмування. Для розв'язання цих труднощів Ольга Валеріївна запропонувала, по-перше, посилити профорієнтаційну роботу зі школярами, щоб студенти чітко розуміли, для чого вони вступають на спеціальності, пов'язані з комп'ютерними та технічними науками, що вивчатимуть і де працюватимуть після навчання; по-друге, починати вивчати програмування якомога раніше; по-третє, постійно застосовувати свої знання на практиці хоча б у невеликих проєктах; по-четверте, викладачам не вимагати виконання від студентів складних задач, краще давати прості, але посильні завдання, які студенти точно виконають самі; по-п'яте, якщо студент-філолог не вміє програмувати, то краще давати йому можливість використовувати у своїй діяльності готові програмні платформи й застосунки.

В. Старко в доповіді *«Реальні задачі в курсах із комп'ютерної лінгвістики»* поділився власним досвідом викладання курсів із комп'ютерної лінгвістики в Українському католицькому університеті у Львові. Під реальними задачами доповідач розумів *real-world tasks* – завдання, пов'язані з реальними проєктами (буквально – з реальним світом), які виходять за межі аудиторії, мають практичне застосування, є елементами суспільно-зорієнтованого навчання. Науковець переконаний, що найкраще працюють ті завдання, у яких студенти на практиці можуть застосовувати свої лінгвістичні знання для суспільно важливої справи. Такими проєктами Василь Феодосійович вважає розробки гурту r2u (<https://r2u.org.ua/>): Великий електронний словник української мови (ВЕСУМ), Генеральний регіонально анотований корпус української мови (ГРАК), програму перевірки орфографії, граматики й стилю «Правописник LanguageTool» та Браунський корпус української мови (БрУК). Студенти-філологи, які обирають курси з комп'ютерної лінгвістики в УКУ, щороку долучаються до розвитку цих проєктів під керівництвом викладача, котрий пропонує їм, наприклад, такі

типи завдань: подання кодів словозміни до нових слів у ВЕСУМі, розмежування омонімів, додавання нових слів в Український семантичний лексикон із семантичними тегами, підготовка текстів із метаданими для корпусів БрУК, ГРАК тощо. Доповідач дійшов висновку, що такі задачі справді покращують мотивацію студентів та сприяють глибшому засвоєнню навчального матеріалу. Окрім цього, В. Старко традиційно запросив усіх охочих до співпраці над проектами гурту r2u.

С. Фокін у доповіді *«Розробка ономасіологічних словників як академічний виклик»* розповів про власний досвід поєднання мовознавчих досліджень та програмування на прикладі проєктів, які розробляє разом зі своїми студентами в Київському національному університеті імені Тараса Шевченка. Зокрема, ішлося про програмні застосунки для вивчення романських мов, які викладає доповідач, а саме: тренажер дієслів, програми для тренування лексики, що добувають еквіваленти з паралельних текстів, паралельні конкорданси, деякі електронні тематичні словники, словники фразеологічних синонімів, ситуативний словник, термінологічний перекладознавчий тезаурус та ономасіологічний словник синтаксем. Дослідник зосередив увагу на останньому проєкті, зауваживши, що сьогодні ще не вироблено єдиного підходу до розуміння поняття ономасіологічності, спостерігається значний різнобій у термінах на позначення цього типу словників: тематичний словник, ідеологічний словник, тезаурус, ідеографічний словник, лінгвістичний атлас тощо. Науковець окреслив традицію ономасіологічного словникарства та коротко розповів про роботу над словником українських синтаксем з базовим реєстром у 200 одиниць, в основі якого лежить семантичний підхід до граматики, а теоретичним підґрунтям є «Грамматика української мови. Синтаксис» І. Р. Вихованця. Найбільшим викликом для упорядників подібних словників, на думку доповідача, є класифікація понять, від яких користувач має переходити до лексичних засобів вираження, а також незгоджений (авторський) термінологічний глосарій, відсутність функційного розміченого корпусу й переваження викладачів і студентів. Незважаючи на це, Сергій Борисович схильний до оптимізму й вважає ономасіологічні словники надзвичайно цінним знаряддям для філологів.

О. Загородня в доповіді *«Дослідження семантики мовних одиниць за допомогою комп'ютерних технологій»* представила авторський мережевий інструмент «Стимулус» (<http://stimulus.tools/uk>), розповіла про його мету та корисні функції для проведення асоціативних експериментів та опрацювання їх результатів, порівняльного аналізу та моделювання асоціативного режиму роботи мовної свідомості (стимул-реакція), дослідження денотативних і конотативних макроелементів асоціативних полів, створення масштабної бази асоціативних опитувань для вивчення динаміки показників мовної свідомості, автоматичної побудови зведених таблиць і графіків з візуалізацією результатів дослідження. Доповідачка зазначила, що «Стимулус» може автоматично підрахувати сентимент, яскравість (семантичну близькість, відношення між реакціями на певний стимул), частотність лексем та ін., показала конкретні дослідження, виконані за допомогою цього застосунку. Ольга Феліксівна підкреслила доступність інструменту, простоту в роботі з ним, полілінгвальність, наголосила на тому, що створити власне асоціативне лінгвістичне дослідження може

будь-яка зацікавлена особа, не тільки науковець. Для застосування цього корисного ресурсу треба тільки зареєструватися на сайті.

Н. Чейлитко виголосила доповідь *«Лінгвістичні корпуси для вивчення мов»*, яку підготувала спільно зі С. Бук і М. Шведовою. Вона зацентувала увагу присутніх на тому, що українська мова зараз дуже затребувана та популярна в Європі, а, отже, постає необхідність актуальних, зрозумілих та доступних інструментів вивчення української мови як іноземної. Таким зручним та ефективним інструментом доповідачка вважає навчальний корпус, орієнтований на осіб, які вивчають мову як іноземну і як рідну, котрі навчають мови або розробляють навчальні курси. Цей корпус мусив би містити адаптовані тексти, репрезентувати найпоширенішу лексику й фразеологію, граматичні конструкції, словотвірні моделі тощо, мати лінгвістичне анотування, притаманне загальномовним референтним корпусам, можливо, містити додаткові інструменти для роботи з ним. Наталія Геннадіївна звернула увагу на два типи корпусів для вивчення мов: 1) учнівські типу L2 learner corpora, які містять тексти, створені тими, хто вивчає мови; 2) класичні – адаптовані корпуси для вивчення мови з відповідним інструментарієм, наприклад корпуси на основі програмного забезпечення Sketch Engine. Навчальний корпус, зауважила доповідачка, знімає проблему багатозначності, варіативності словотвірних форм, керування дієслів та багато інших, з якими виникають труднощі в іноземців, адже відбиває реальне функціонування мови та слововжиток у всіх виявах, на відміну від словників. Наприкінці виступу дослідниця навела приклади застосування ГРАКу для вивчення української мови як іноземної, а під час обговорення доповіді присутні дійшли висновку, що укладання корпусу на базі шкільних текстів – це оптимальний варіант для навчальних словників і корпусів.

М. Шведова в доповіді *«ГРАК як база практики для студентів-філологів»* увиразнила переваги корпусу не просто як дослідницького, а і як навчального засобу. Доповідачка зауважила, що ця велика колекція текстів із розміткою та інструментарієм може стати базою різномісних завдань для студентів-філологів будь-якої спеціалізації та року навчання: коректорських і дослідницьких. На прикладі виконаних робіт студентів Національного університету «Львівська політехніка», Одеського національного університету імені І. І. Мечникова Марія Олексіївна проілюстрували різні можливості ГРАКу під час практики: вичитування та коректорська робота зі збірником «Українські приказки, прислів'я і таке інше» Номиса 1864 року, додавання в корпус журналів із метаінформацією, збирання, розпізнавання та додавання до корпусу листів з родинних архівів, створення корпусу творів Михайля Семенка (644 тексти) у межах курсу «Комп'ютерна лексикографія» для укладання словника авторських неологізмів, поповнення ВЕСУМу, анотування лексико-семантичної групи прикметників для семантичної розмітки корпусу та ін. Дослідниця підтримала тезу, яка вже звучала під час семінару: коли студенти працюють із реальними проектами, вони мотивованіші та ефективніші, бо виконують не просто навчальне завдання «для галочки», а конкретну, нехай і невелику, суспільно важливу роботу.

У. Штанденко в доповіді *«Староукраїнські тексти як складова корпусу ГРАК»* презентувала рукописну пам'ятку ділового письменства другої пол. XVII – поч. XVIII ст. «Акти Полтавського полкового суду 1668–1740 рр.» як цінні матеріа-

ли для корпусу, своєрідний історичний підкорпус, зорієнтований не тільки на широту охоплення матеріалу, а й на його глибину. У. М. Штанденко підготувала до друку пам'ятку разом із колегами І. П.Чепігою і Л. А. Москаленко, а Інститут української мови НАН України видав її у двох книгах у серії «Пам'ятки української мови» (2017). Доповідачка розповіла, як упродовж кількох років опрацьовувала тексти для друку, схарактеризувала особливості структури, змісту та графіки пам'ятки (т.зв. козацький скоропис), окреслила специфічні риси староукраїнського тексту (властиві стилю та періоду лексеми, синоніміїні ряди, сталі сполуки, синтаксичні моделі, фонетичні особливості та їх передавання на письмі), звернула увагу присутніх на актуальність дослідження живомовних елементів у ньому. Уляна Михайлівна наголосила, що «Акти Полтавського полкового суду 1668–1740 рр.» є надійним матеріалом для лінгвістичного дослідження історії української мови одного з найвпливовіших культурних центрів України, не зіпсованої російським впливом, є істинною історією України, а тому повинні стати частиною корпусу ГРАК і таким чином посилити та урізноманітнити його. Учасники семінару висловили міркування, що зацифрування пам'яток необхідне для дослідницьких цілей та створення діахронного корпусу текстів української мови, якого досі не існує, а насамперед – для збереження культурної спадщини українського народу.

В. Жуковська в доповіді «*Корпусні технології та жанрово-аналітичний підхід у курсі англійської для академічних цілей*» показала практичне застосування корпусів під час викладання лінгвістичних дисциплін, сфокусувалася на трьох основних моментах: презентації інтегрованої корпусно-базової та жанрово-аналітичної моделі розвитку наукової писемної компетенції студентів, етапах реалізації моделі в освітній компоненті «Англійська мова для академічних цілей», перевагах та труднощах імплементації інтегрованої моделі, апробованої в Житомирському державному університеті імені Івана Франка. Доповідачка ознайомила присутніх із корпусно-базовою та жанрово-аналітичною моделлю розвитку наукової писемної компетентності, продемонструвала орієнтовні завдання для студентів на базі корпусів, створених ними самими зі проблематики свого дослідження, наприклад: генерування частотних списків різного роду, визначення найчастотніших лексем чи сполук, частин мови, аналіз типів сполучуваності (ліво- чи правостороння), опис чинників, які впливають на частотність одиниць, розмежування галузевої, загальнонаукової та загальнонавчальної лексики тощо. Вікторія Вікторівна показали студентські мінікорпуси, на основі яких філологи укладають списки лексико-граматичних патернів для різних структурних частин наукової статті, а на завершальному етапі роботи створюють англо-український словник академічного мовлення, який містить прототипову лексику чи сполуки для грамотного написання та логічного структурування статей англійською мовою. Доповідачка відзначила переваги корпусно-базового навчання: більш точний опис мови, ніж у підручнику, доступ до контекстного вживання мови в реальних умовах функціонування, приклади вживання в конкретному жанрі/реєстрі мови, самостійність та індивідуальна зацікавленість темою, а також звернула увагу на труднощі: недостатність навичок роботи з корпусними технологіями, відсутність технічних умов, недостатній рівень мовної підготовки, значні затрати часу та зусиль для розроблення завдань тощо.

Л. Завальська в доповіді «*Використання національного корпусу української мови в дослідженні українського президентського дискурсу*» окреслила традицію вивчення українськими науковцями президентського дискурсу України та інших країн, специфіку, основні жанри та функції цього особливого типу мовлення. Здійснивши пошук за ключовими словами в законодавчому та публіцистичному підкорпусах корпусу на лінгвістичному порталі Mova.info, доповідачка з'ясувала концептуальні метафори, репрезентанти українського президентського дискурсу (*Україна, мир, війна*), особливості функціонування окремих лексичних одиниць та граматичних конструкцій, дослідила динаміку вживання аналізованих одиниць. Любов Володимирівна продемонструвала, що залучення корпусів для дослідження президентського дискурсу є релевантним у таких напрямках як визначення основних концептів (частотності) цього різновиду текстів, аналіз окремих їхніх жанрів або груп жанрів, виявлення в них політичних метафор, характеристика синтаксичної сполучуваності слів тощо.

Р. Махачашвілі в доповіді «*Лінгвістичний електронний інструментарій і цифрові навички під час війни та після неї*» розповіла про свою роботу та роботу колег із Київського університету імені Бориса Грінченка в проєкті програми Європейської Комісії Горизонт 2020 «Elexis» (<https://elex.is/>) у сфері цифрової гуманітаристики. Доповідачка зауважила, що, працюючи зі студентами, вона переконалася, що сьогодні навчати потрібно не тільки природної іноземної мови, а й мови, яка функціонує в комп'ютерному просторі – мультимодальної, піктографічної, постійно оновлюваної. У цьому викладачам допомагають цифрові інструменти опрацювання мовних даних, у тому числі і ті, які безкоштовно пропонує платформа «Elexis». Зокрема, доповідачка розповіла про такі інструменти навчання: служба «Elexifier» для створення автоматичних словників, яка запам'ятовує принцип анотації словникової статті і відтворює його, може витягати дані просто з корпусу; система «Lexopom» для написання та онлайн-публікації словників; платформа створення, побудови та керування корпусом «Sketch Engine», що дозволяє користувачам укладати та працювати з текстовими корпусами різними мовами та шрифтами, містить низку унікальних інструментів для аналізу корпусів. Дослідниця розповіла про деякі реалізовані проєкти: семантичний корпус на основі текстів із актуального сповіщення про ситуацію з розгортання бойових дій в Україні кількома мовами з метою виявити інформаційні лакуни в різних мовах та швидко заповнити їх перекладними матеріалами; електронний словник російсько-української війни та ін. На думку Русудан Кирилевни, цифрові інструменти дозволяють філологам не просто ефективно виконувати дослідницькі завдання, а й бути корисними для перемоги, бо ця війна є також війною слів.

Після кожної доповіді відбувалося зацікавлене та жваве обговорення, до якого мали змогу долучитися й онлайн-учасники. Семінар був насиченим новою інформацією та пізнавальним. Відеозапис заходу у доступнено на сайті спільноти українознавчих студій при Єнському університеті імені Фрідріха Шиллера (доступ: <http://surl.li/eb1bx>).