Людмила Алексієнко,
Наталія Дарчук
Київський національний університет імені Тараса Шевченка

ДО ДВАДЦЯТИРІЧЧЯ ЛАБОРАТОРІЇ КОМП'ЮТЕРНОЇ ЛІНГВІСТИКИ

Навчально-дослідній лабораторії комп'ютерної лінгвістики Інституту філології Київського національного університету імені Тараса Шевченка у вересні 2012 року виповнилося 20 років. Згадуючи пройдений шлях, хотілося б поділитися нашим досвідом, здобутками і планами з колегами-"прикладниками", аспірантами, студентами і висловити щиру подяку всім, хто допомагав нам на цьому шляху. І крім того, запровадження нової спеціалізації "комп'ютерна лінгвістика" на кафедрі сучасної української мови, створення під неї навчально-дослідної лабораторії, підготовка упродовж двох десятиліть бакалаврів, магістрів та аспірантів – фахівців з автоматизованого аналізу тексту – це ще одна сторінка в історії кафедри сучасної української мови Інституту філології.

Коли, у кого і чому виникла ідея готувати фахівців у галузі комп'ютерної лінгвістики, і як ми її реалізовували? Термін "комп'ютерна лінгвістика" 30 років тому на наших теренах сприймався майже як оксюморон, незважаючи на те, що вже в 60-х роках минулого століття в європейських та американських університетах активно створювалися кафедри computer sciences, які готували, зокрема, й комп'ютерних лінгвістів. Їх навчали кібернетики, інформатики та мовознавства, але в пріоритеті в цей період були інформаційно-комп'ютерні технології.

Недостатньо фахова лінгвістична параметризація тогочасних текстових корпусів, частотні словники з незнятою омонімією та недиференційованою лексичною семантикою, системи машинного перекладу недостатнього рівня семантико-граматичної глибини тощо свідчать про те, що лінгвістичному сегменту в багатьох комп'ютерних продуктах не приділялося належної уваги.

Комп'ютерна лінгвістика є новим напрямком класичної прикладної лінгвістики, яка виникла, розвивалася й розвивається паралельно з традиційною лінгвістикою. До компетенції при-

кладної лінгвістики входять: письмо (графіка), методика навчання рідної та іноземної мов, лексикографія, мовна політика – ліквідація неграмотності, вибір державної мови та її підтримка, розроблення національної термінології, національних ономастиконів тощо. Ця проблема актуальна й на сучасному етапі прикладної лінгвістики.

Разом із тим у другій половині XX століття у прикладній лінгвістиці з'явився новий вектор, спричинений активними процесами інтеграції гуманітарних, природничих, технічних і математичних наук. Результатом цього було усвідомлення і визначення спільної для багатьох предметних галузей проблеми – автоматизація оброблення, обміну і збереження різноманітної інформації, яка функціонує в суспільстві в текстовій формі. Фахівці практично всіх галузей знань користуються мовою як універсальним засобом оформлення і смислового представлення знань. Оскільки текстова інформація є природною для людини формою комунікації, лінгвістичне забезпечення інформаційних систем стає головним завданням комп'ютерної лінгвістики. У цій ситуації є необхідним розподіл компетенцій між власне лінгвістикою та інформаційно-комп'ютерними технологіями. Фаховий аналіз смислу текстів – це прерогатива лінгвістів, які глибоко розуміють систему мови в усіх її проявах. Багатовимірне впорядкування параметризованої лінгвістами текстової інформації в бази даних і бази знань, корпуси текстів, створення гіпертекстових мереж із можливістю навігації у величезних масивах тощо – це прерогатива фахівців з інформатики і кібернетики. Таким чином, комп'ютерна лінгвістика – це лінгвістика із застосуванням інформаційно-комп'ютерних ресурсів.

Нова предметна галузь потребувала підготовки відповідних фахівців, яких у 60–70-х рр. в Україні ще не було. Перший крок у цій справі було зроблено в середині 60-х років у Київському університеті. Спочатку на філологічному факультеті була створена спеціалізація зі структурно-математичної лінгвістики – групи по 8–10 студентів, яких готували як спеціалістів з автоматичної обробки текстів. Очільником спеціалізації була

проф. Феоніла Олексіївна Нікітіна, викладали спецдисципліни професори-філологи Віктор Вікторович Коптілов, Едуард Федорович Скороходько, Ірина Платонівна Севбо, с. н. с. Ірина Борисівна Штерн, професор-математик Лев Аркадійович Калужнін та програмісти з Інституту кібернетики. Через кілька років спільними зусиллями філологів та математиків на факультеті кібернетики було створено відділення і кафедру структурно-математичної лінгвістики зі щорічним набором студентів (15 осіб), які отримували диплом спеціаліста такого профілю: *автоматична обробка тексту*; *іноземна мова*; *перекладач-референт*. У підготовці кадрів із цієї спеціальності брали активну участь та надавали дієву допомогу співробітники Інституту кібернетики АН УРСР і персонально акад. Віктор Михайлович Глушков, за ініціативою якого було запроваджено наукову спеціальність: структурна, прикладна і математична лінгвістика – 10.02.21. Однак у 1985 році відділення було ліквідоване.

У цей період в Інституті мовознавства ім. О. О. Потебні АН УРСР потужно працював відділ структурно-математичної лінгвістики на чолі з проф. Валентиною Сидорівною Перебийніс. На базі української мови проводилися масштабні статистичні й стилеметричні дослідження, структурно-статистичне моделювання на різних рівнях мовної системи.

Інформаційно-комп'ютерні технології розвиваються швидкими темпами. Тому пропозиція запровадити на українському відділенні спеціалізацію "комп'ютерна лінгвістика" була підтримана кафедрою сучасної української мови, адміністрацією філологічного факультету та університету. З 1989 року на 3–5 курсах були створені групи по 7–10 студентів, які навчалися за планом спеціалізації, слухали відповідні спецкурси, працювали в спецсемінарах, писали бакалаврські, дипломні й магістерські роботи з різноманітних проблем комп'ютерної лінгвістики і отримували до диплома філолога-українєта сертифікат фахівця з автоматизованої обробки тексту.

Спецдисципліни викладали: співробітники відділу структурно-математичної лінгвістики – проф. В. Перебийніс, с. н. с. Н. Кли-

менко, Т. Грязнухіна, Н. Дарчук, Л. Орлова, Є. Карпіловська, програміст Л. Братищенко та викладачі університету – доц. Л. Алексієнко і с. н. с. І. Штерн.

З техніки на той час був один слабенький комп'ютер, але такий, без перебільшення, зірковий лекторат із самого початку зробив цю спеціалізацію рейтинговою, причому не тільки серед наших студентів, а й ширше – до нас почали звертатися за методичною допомогою викладачі-філологи з різних університетів України – Київського лінгвістичного, Донецького, Волинського, Харківського, Львівського, які почали серйозно займатися комп'ютерною лінгвістикою і підготовкою кадрів.

Фундамент спеціалізації закладався такими спецкурсами, як структурна і прикладна лінгвістика; штучний інтелект; лінгвостатистика; комп'ютерна лексикографія; лінгвістика тексту; автоматичний морфологічний аналіз; автоматичний синтаксичний аналіз; машинний переклад; основи програмування та ін. Студенти були в захваті від нового сприйняття мови, від долучення до потужних цивілізаційних процесів інформатизації, від реальної роботи в різноманітних проектах, які виконувалися в лабораторії. За 20 років було підготовлено 90 фахівців зі спеціалізації комп'ютерна лінгвістика, з них успішно захистили кандидатські дисертації дев'ять наших випускників і двоє подали дисертації до захисту. Студенти завжди тримають нас у тонусі, хочеться щиро подякувати їм за співпрацю і взаєморозуміння.

Навчальний план спеціалізації постійно вдосконалювався і впродовж 23 років увиразнився в напрямку інформаційно-комп'ютерного моделювання мови, створення автоматизованих інтелектуальних систем на базі української мови.

У 1992 році під спеціалізацію з метою вдосконалення навчального процесу було створено навчально-дослідну лабораторію комп'ютерної лінгвістики. Оптимальним способом реалізації цієї мети є співпраця викладачів і студентів у проектах зі створення автоматизованих систем аналізу текстів, баз даних, електронних словників, підручників тощо.

Першим проектом лабораторії була "Параметризована база даних українського поетичного мовлення", що планувалася як

джерело для філологічних студій функціонування української мови в літературі, зокрема дослідження ідіостилів українських поетів на різних хронологінчих зрізах. Проект викладено в Інтернет-порталі лабораторії (mova.info). На цій базі захищені три кандидатські дисертації наших випускників – Л. Гливінської, Д. Данильчука, Ю. Маковецької-Гудзь, а також понад 40 бакалаврських і магістерських робіт. Проект дістав міжнародне схвалення, зокрема грант ACLS – американської асоціації підтримки інноваційних проектів в Росії, Україні та Білорусії.

Наступний проект – "Морфемно-словотвірна база даних української мови" (≈170 тис. слів), викладена на порталі лабораторії. База є ресурсом для автоматичного укладання алфавітно-частотних словників морфем і словотвірних гнізд на матеріалі будь-яких текстів. База забезпечує високу якість, масштабність, системність та оперативність досліджень. На цій базі підготовлені кандидатські дисертації О. Тютенко і Т. Жигун, а також захищено понад 30 бакалаврських і магістерських робіт різної тематики. Ця база даних також відзначена грантом ACLS.

Важливим проектом для розвитку лабораторії став лінгвістичний портал "mova.info". На порталі розміщуються всі наші проекти, ведеться рубрика "Новини мовознавства, мовної культури і мовної політики". Проект підтриманий грантом Посольства Канади в Україні. Кількість щоденних відвідувань порталу ≈100.

Проект "Електронна граматика української мови" (для абітурієнтів та дистанційного навчання). Підручник, крім теорії, містить вправи, тести та єдину навігаційну систему користування. Як свідчать постійні звернення до порталу, підручник популярний серед абітурієнтів. Для спеціалізації він є ресурсом спецкурсу "Електронні підручники з мови".

Проект "Українсько-російсько-італійська довідково-пошукова система з питань усиновлення" (2400 юридичних термінів трьома мовами з перекладом, тлумаченням, енциклопедичною інформацією та юридичними документами трьох країн) виконувався як міжнародний, спільно із Флорентійським університетом. Одержав схвальний відгук дитячого фонду ЮНІСЕФ при ООН.

Проект на замовлення Державного комітету України з питань науки, інновацій та інформатизацій "Електронний словник лінгвістичної термінології з інформаційно-пошуковою системою (тезаурус)" – 3400 термінів з українсько-російсько-англійським перекладом. Цей проект також є ресурсом для спецкурсів та нових проектів. За його методикою магістри протягом року створили електронний тезаурус літературознавчих термінів.

Із 2010 року лабораторія почала працювати над масштабним проектом "Дослідницький корпус української мови". На сьогодні на порталі викладена його частина – корпус розмічених і параметризованих текстів обсягом понад 13 млн слововживань.

Проекти лабораторії комп'ютерної лінгвістики стали полігоном навчання, виробничих практик, наукових і методичних досліджень широкого спектру. Необхідно відзначити, що в результаті створення різноманітних електронних продуктів були одержані такі важливі комп'ютерні ресурси, як програми автоматичного морфологічного, контекстного і синтаксичного аналізу українських текстів, без яких неможлива жодна інтелектуальна інформаційна система.

Усі проекти створювалися колективом штатних співробітників лабораторії (випускників спеціалізації), науковим керівником якої є доц. Н. Дарчук. Завдяки їй та інженеру-програмісту В. Сорокіну підготовка фахівців з комп'ютерної лінгвістики і наукова робота лабораторії досягли такого рівня.

Паралельно з цими проектами в лабораторії розроблялися засади комп'ютерної морфології. Це електронний "Граматичний словник дієслів української мови" (проект, спільний із Лейпцигським університетом); електронний "Українсько-італійський граматичний словник дієслів" (проект TEMPUS-TASSIS, спільний із Флорентійським університетом).

Видані підручник "Комп'ютерна лінгвістика" та навчальний посібник "Термін у лінгвістичній інформатиці" – автор Н. Дарчук.

Видані монографії І. Козленко "Морфеміка сучасної української літературної мови" та створено колективний підручник "Морфологія української мови. Морфемологія. Словотвір. Па-

радигмологія" – автори: Л. Алексієнко, О. Зубань, І. Козленко. Ці структурно-прикладні розробки можна використовувати як у навчальному процесі, так і для створення нових автоматизованих систем на базі української мови.

Наступний етап діяльності лабораторії – запровадження спеціальності "прикладна лінгвістика", для якої розроблено навчальний план. Набутий досвід у спеціалізації "комп'ютерна лінгвістика" засвідчує, що підготовка сучасних фахівців – бакалаврів і магістрів – потребує не лише збільшення філологічного комплексу, а й уведення дисциплін математичного циклу, які читатимуться студентам упродовж всього періоду навчання. У цьому нас підтримали декан факультету кібернетики акад. А. Анісімов, який добре обізнаний з нашою предметною галуззю, а також його колеги та учні. Вони не тільки консультували навчальний план спеціальності, а й висловили готовність узяти участь у його реалізації.

Проведені консультації з викладачами різних кафедр Інституту філології та інших факультетів підтвердили не лише необхідність, а й готовність до запровадження спеціальності "прикладна лінгвістика" з 2013 року.

Для здійснення навчально-дослідної роботи на новому етапі вкрай потрібне фахове середовище, співпраця в дослідницьких проектах насамперед із кафедрами прикладної лінгвістики українських вишів. Це засвідчила нещодавно проведена кафедрою сучасної української мови традиційна конференція "Мова як світ світів". На секцію "Актуальні проблеми комп'ютерної лінгвістики" було надіслано понад 40 доповідей (Національний університет "Львівська політехніка", Східноєвропейський університет (Луцьк), Інститут російської мови РАН (Москва), Військовий інститут Київського національного університету імені Тараса Шевченка, Інститут української мови НАНУ (відділ структурно-математичної лінгвістики), Український мовно-інформаційний фонд НАНУ, Львівський університет, Київський національний лінгвістичний університет, Кіровоградський педагогічний інститут ім. Володимира Винниченка, Національний університет "Київський політехнічний інститут").

Під час Круглого столу учасники конференції висловили бажання на базі Київського національного університету імені Тараса Шевченка систематично проводити наукові семінари (раз на рік) з актуальних проблем комп'ютерної лінгвістики; започаткувати спільний онлайн-проект "Термінологія комп'ютерної лінгвістики (електронна база даних, тезаурус)"; організовувати "школи комп'ютерної лінгвістики" для студентів та аспірантів; розробляти спільні проекти (в режимі онлайн) із залученням бакалаврів, магістрів та аспірантів.

Масштабність та значущість здійснених і запропонованих проектів отримали схвальні відгуки з усієї "прикладної" України. Це дає підстави вважати лабораторію комп'ютерної лінгвістики Інституту філології Київського національного університету імені Тараса Шевченка науково-методичним центром спеціальності "прикладна лінгвістика".