2. Анализ метаязыка словаря с помощью ЭВМ. – М.: Наука, 1982.
3. Български римен речник. – София: Наука и изкуство, 1967.
4. Гончарова Б. Звуковая организация стиха и проблемы рифмы. – М.: Наука, 1973.
5. Гурин І. Словник рим Є.Гребінки. – Миргород, 1982.
6. Жирмунский В. Рифма, ее история и теория. – Петроград, 1923.
7. Качуровський І. Фоніка. – К.: Либідь, 1994.
8. Квятковский А. Поэтический словарь. – М.: Советская энциклопедия, 1966.
9. Ковалівський В. Рима. Ритмічні засоби українського вірша. – К.: Рад. письменник, 1965.
10. Лесин В., Пулинець О. Словник літературознавчих термінів. – К.: Рад. школа, 1965.
11. Літературознавчий словник-довідник. – К.: Академія, 1997.
12. Словарь языка русской поэзии XX века. / Отв. ред. В.Григорьев. – М.: Языки слав. культ., 2003. – Т.II.
13. Украинский семантический словарь: проспект. – К.: Наук. думка, 1990.
14. Українська мова: Енциклопедія. – К.: Вид-во „Укр. енциклопедія" ім. М.П.Бажана, 2000.
15. Цывин А. К вопросу о классификации русских словарей. // Вопросы языкознания. – 1978. – № 1. – С. 100 – 108.

Балог Валентина, к.ф.н.[*]
Інститут української мови (Київ)
УДК 161.2.81'374.72'22

СУЧАСНИЙ СТАН УКРАЇНСЬКОЇ КОМП'ЮТЕРНОЇ ЛІНГВІСТИКИ

У статті розглянуто стан української комп'ютерної лінгвістики на сучасному етапі розвитку мовознавчої науки. Оглядово подано інформацію про наукові установи, де працюють у галузі комп'ютерної лінгвістики, й основні здобутки науковців у цій галузі, а також коротко окреслено перспективи її розвитку.

Для українського мовознавства тема комп'ютеризації лінгвістичних досліджень не нова. На нашу думку, початком такого роду досліджень саме в лінгвоукраїністиці можна вважати Всесоюзну конференцію з проблем створення машинного фонду для автоматизованої системи

лексикографічних досліджень (1983 рік) за участі українських науковців – працівників Інституту мовознавства ім. О.О. Потебні та Інституту кібернетики ім. В. М. Глушкова (В. С. Перебийніс, М. М. Пещак, І. П. Білецька (Севбо)), на якій було підтримано ідею створення проекту Машинного фонду російської мови як системи комплексної автоматизації лінгвістичних досліджень та розробок [2]. Цей проект мав стати типовим для розроблення машинних фондів мов інших народів СРСР. Тож робота в цьому напрямку ведеться вже давно, проте проблема автоматизації лінгвістичних досліджень, незважаючи на чималий здобуток українських науковців, не просто залишається актуальною, але й набрала нових обертів. На сьогоднішній день диктують умови не лише *внутрішньонаукова* потреба об'єктивації наукових досліджень (необхідність впорядкування локалізованої потужної бази фактичного матеріалу, зручний та швидкий пошук потрібної інформації, оптимізація та раціоналізація дослідницької роботи мовознавця, тобто створення так званого „робочого місця" лінгвіста), *зовнішньонаукова* потреба інтегрування української мовознавчої науки в загальнонауковий процес комп'ютеризації гуманітарних наук, але й *позанаукова* потреба – задовольнити зростаючий попит пересічного користувача на адаптовану для нього об'єктивну, достовірну інформацію мовознавчого характеру (а це – словники різного типу, підручники, посібники тощо в електронному вигляді), оскільки постулат „наука – для потреб людини" завжди був і залишається актуальним. Зважаючи на стрімке зростання можливостей системотехнічного, програмного забезпечення комп'ютерної техніки, роботу в галузі комп'ютерної лінгвістики можна вважати лише початковим етапом на шляху до створення комплексної автоматизованої системи лінгвістичних досліджень, оскільки постає вже не тільки питання створення чогось нового в цій галузі, але й удосконалення і поповнення вже наявних надбань.

В українській науковій думці теоретичні засади комп'ютерної лінгвістики, методики та технології лінгвістичних досліджень за допомогою комп'ютерних технологій висвітлені насамперед у ґрунтовних працях, які допомагають окреслити проблематику, основні напрямки роботи та зорієнтуватися в термінології комп'ютерної лінгвістики (В. С. Перебийніс, Т. О. Грязнухіна, Н. Ф. Клименко, Н. П. Дарчук, М. М. Пещак, Є. А. Карпіловська, Л. І. Комарова, О. М. Демська-Кульчицька, В. А. Широков та ін.). Поширенню

відповідної інформації сприяють також наукові та науково-практичні конференції (зазначимо, що тема впровадження комп'ютерних технологій у лінгвістичні дослідження є предметом обговорення майже на кожній конференції з проблем мовознавства), широкий спектр науково-теоретичних, науково-практичних, науково-популярних публікацій, особливо в мережі Internet.

Серед установ, де проводиться наукова робота з впровадження комп'ютерних технологій у царину лінгвістичних досліджень, передусім варто назвати такі: відділ структурно-математичної лінгвістики Інституту мовознавства ім. О. О. Потебні (на нашу думку, це фундатор комп'ютерної лінгвістики в українському мовознавстві) НАН України, Лабораторія комп'ютерної лінгвістики кафедри сучасної української мови Київського національного університету імені Тараса Шевченка, Національний мовно-інформаційний фонд НАН України, відділ лексикології та комп'ютерної лексикографії Інституту української мови НАН України, кафедра української мови Донецького національного університету, університет „Львівська політехніка". Тут відбувається як процес ґрунтовного теоретичного опрацювання аналізованої проблеми, адаптації до лінгвоукраїністики наявних теоретичних напрацювань зарубіжних науковців, так і власне практичне впровадження ідей та розробок. Таким чином, українська комп'ютерна лінгвістика загалом має значні здобутки.

Зокрема, у відділі структурно-математичної лінгвістики Інституту мовознавства ім. О. О. Потебні створено Морфемно-словотвірний фонд української мови. Він має розгалужену архітектуру. Складається з трьох основних підфондів: текстової бази, генерального реєстру українських слів, зорієнтованого на п'ять словників-джерел, та лінгвістичних процесорів, що виконують основну роль у процесі опрацювання фактичного матеріалу. На сьогодні повнотекстова база даних фонду містить близько 700 тис. слововживань, оснащена процедурами орфографічного контролю текстів, аналізу їхньої морфологічної, синтаксичної та семантичної структури. Генеральний реєстр слів української мови, зведений за матеріалами 5 словників (Словника української мови в 11-ти томах (К., 1970–1980 рр.), словника-довідника І. Т. Яценка у 2-х томах „Морфемний аналіз" (К., 1980–1981), „Частотного словника сучасної української художньої прози" у 2-х томах (К., 1981), „Словника іншомовних слів" за ред. О. С. Мельничука (К., 1974) та орфографічної частини „Словника-довідника з правопису

та слововживання" С. І. Головащука (К., 1989)), становить 166385 слів із відомостями про їхню морфемну будову, кількість властивих їм значень, абсолютну частоту вживання у півмільйонній текстовій вибірці сучасної української художньої прози та частиномовну належність. Для опрацювання створених баз даних розроблені текстові процесори для морфологічного, синтаксичного та логіко-семантичного аналізу тексту. За матеріалами фонду було укладено комп'ютерні „Словник символьних моделей морфемної будови слова", „Словник афіксальних морфем української мови" (це перший в Україні комп'ютерний словник, виданий у паперовому вигляді у 1998 р.; автори – Н. Ф. Клименко, Є. А. Карпіловська, В. С. Карпіловський, Т. І. Недозим), "Кореневий гніздовий словник української мови" Є. А. Карпіловської (К., 2002), Ідеографічний словник іменників української мови, укладений Н. В. Сніжко, Ідеографічний словник дієслів переміщення української мови, укладений А. Я. Середняцькою, здійснено також спроби створити комп'ютерні версії „Словника староукраїнської мови XIV – XV ст." та словників лінгвістичних термінів Є. В. Короткевича, Н. С. Родзевич і Д. І. Ганича та І. С. Олійника [3].

У Національному мовно-інформаційному фонді НАН України під керівництвом В. А. Широкова створюється Національна словникова база України. У 2001 році створено компакт-диск Інтегрованої лексикографічної системи „Словники України". Користувачеві запропоновано словник в абетковому порядку із пошуковою системою обсягом близько 152 тис. лексем, основу якого становить реєстр „Орфографічного словника української мови" (К., 2002), роботу з яким можна здійснювати у 5-ти режимах (підґрунтям для кожного є окремий словник): 1) „парадигма" – інформація про словозмінні властивості реєстрової одиниці (цей модуль функціонує на основі розробленої словозмінної класифікації української лексики, у якій виділено за певними формальними ознаками близько 1500 парадигматичних класів для всіх відмінюваних повнозначних частин мови, а з урахування акцентуації – близько 3000 класів; таким чином, повне число словоформ наближується до 3 млн.); 2) „транскрипція" – інформація про артикуляцію лексичних одиниць згідно з нормами сучасної української літературної мови (джерелом для роботи цього модуля є „Орфоепічний словник української мови" (К., 2001–2003)); 3) „фразеологія" – інформація про здатність лексичної одиниці створювати фразеологізми української мови різного типу, кількість яких становить близько 56 тис.

одиниць (джерела – два видання „Фразеологічного словника української мови" (К.,1993; 2-ге вид. – К., 1999)); 4) „синонімія" – інформація про належність лексичної одиниці до синонімічної пари/ряду; модуль словника містить близько 9200 синонімічних рядів (джерело – „Словник синонімів української мови" (К., 1999–2000)); 5) „антонімія" – інформація про наявність у лексеми антонімічного відповідника; у модулі представлено понад 2200 компонентів антонімічних пар (джерело – „Словник антонімів української мови" Л. М. Полюги (К., 1999)). Останні три модулі містять також інформацію про семантику компонентів, ілюстративний матеріал. Таким чином, ця робота є інтегрованим поєднанням у єдину систему п'яти словників, які мають як традиційний вигляд (тобто видані друком в серії „Словники України"), так і електронний. У цьому фонді започатковано також створення Українського лінгвістичного корпусу, який проектувався як універсальна система підтримки дослідницьких процесів у мовознавстві з орієнтацією на створення лексикографічних продуктів. На сьогодні ця система нараховує понад 36 млн. слововживань, які містяться в 1535 джерелах українських текстів художньої, наукової, конфесійної літератури, публіцистики, ділових та юридичних документів. Подальша робота фахівців Національного мовно-інформаційного фонду проводиться у напрямку створення нової версії тлумачного „Словника української мови", обсяг якого планується на рівні 20 томів.

Основними напрямками наукової роботи Лабораторії комп'ютерної лінгвістики кафедри сучасної української мови Київського національного університету імені Тараса Шевченка є комп'ютерна лексикографія, навчальні програми та машинний переклад. З більшою частиною проектів Лабораторії можна познайомитись в електронному вигляді. Основні здобутки такі.

1. Електронний підручник української мови з інтерактивним тестуванням, можливістю перегляду допущених помилок та збереженням результатів тестування на сервері. Комплексна навчальна система розрахована на широкий загал – учнів загальноосвітніх шкіл, абітурієнтів, студентів філологічних та інших спеціальностей вищих навчальних закладів і всіх тих, хто бажає вдосконалити та перевірити свої знання з української мови.

2. Частотні словники художньої прози та публіцистики, а також частотний словник сучасної поетичної української мови. Останній

словник репрезентує лексику з поезій кращих українських митців кінця XX століття (обсяг вибірки – 300 тисяч слововживань).

3. Граматичний словник українських дієслів подає парадигматичні, семантичні й синтаксичні особливості близько трьох тисяч дієслів української мови з їхнім перекладом італійською мовою. Цей словник є першим в українській лексикографії досвідом систематизації словозміни дієслів у зв'язку з їхнім синтаксисом і лексичною семантикою. Призначений для студентів вищих закладів освіти, науковців, викладачів української мови, перекладачів з італійської мови.

4. Програма автоматичного синтезу парадигми англійського іменника та дієслова, яку створено в рамках бакалаврської роботи "Автоматичний синтез словоформ іменника і дієслова".

5. Морфемно-словотвірна база української мови. Мета проекту – створення морфемної бази даних української мови (170 тисяч слів) та укладання морфемно-словотвірних гнізд зі словотворчими значеннями морфем. Автоматичний морфний сегментатор українського тексту представляє собою систему, на вході якої словоформи звичайного тексту, на виході – ті ж самі словоформи, заіндексовані кодами граматичної приналежності до певної частини мови та розчленовані на морфи (кореневі й афіксальні) з відповідними індексами. Також у Лабораторії досліджують принципи англо-українського та українсько-англійського машинного перекладу, розробляють перекладні й термінологічні словники, програми аналізу та синтезу людського мовлення.

Теоретична робота з проблем комп'ютерної лінгвістики, зокрема корпусної лінгвістики, ведеться також у відділі лексикології та комп'ютерної лексикографії. Започатковано роботу над створенням Національного корпусу української мови, призначеного для суто наукових завдань: збереження текстового матеріалу, забезпечення наукових досліджень лексичної та граматичної структури мови, а також простеження динаміки і якості змін у мовній системі протягом певного хронологічного періоду. Структура корпусу передбачається складною, побудованою за моделлю „генеральний корпус – система підкорпусів", з мінімальною кількістю слововживань в 1 млн. одиниць без обмежень на верхню межу. За визначенням це буде організована, систематизована, анотована сукупність текстів української мови, які є репрезентативними для всіх (як історичних, так і географічних) варіантів та форм її

існування, оформлена згідно зі Стандартами кодування корпусу[1]. Серед основних здобутків відділу – електронна картотека, яка працює в тестовому режимі, а також перебуває в стадії наповнення електронними текстами.

Львівські дослідники О. М. Коссак та С. Л. Маньковський (університет „Львівська політехніка") створили лексикографічний процесор „СЛОВО", який дозволяє на основі бази даних термінологічних одиниць створювати одно- та багатомовні термінологічні словники. Він має зручний інтерфейс, що дозволяє користувачеві працювати з базою даних в інтерактивному режимі як з інформаційно-пошуковою, навчальною, дослідницькою системою, коригувати, поповнювати її. Процесор може виконувати такі функції:

1) запис текстів до відповідного словника (за мовою представлення);
2) коригування словників;
3) пошук слів та їхніх перекладів у базі за заданими ключами;
4) вилучення слів та/або їхніх перекладів зі словників.

В основу "СЛОВА" покладено досвід укладання „Англо-українсько-російського словника з інформатики та обчислювальної техніки".

На кафедрі української мови Донецького національного університету під керівництвом Л. А. Фроляк упорядковано та видано в 2000 році на компакт-диску фонотеку „Українські говірки Донеччини". Це текстозорієнтована база даних у звуковій та графічній формах, супроводжена засобами роботи знею. До диска увійшли аудіозаписи зв'язних текстів, здійснені у 65 населених пунктах Донецької області у 1997–2000 роках. Діалектний матеріал розміщено в папках за адміністративними районами та містами. Окремими файлами подано список обстежених населених пунктів та список інформантів. Записи зроблено у форматі WAV.compress 24 KBit/sec. (24000 Hz Mono). Працюючи з комп'ютерним диском, можна швидко знайти потрібний запис, перейти від одного тексту до іншого чи від однієї частини запису до іншої, що значно полегшує роботу дослідника, дозволяє систематизувати великий за обсягом матеріал. Прослуховуючи записи (в обраному користувачем темпі звучання), можна не тільки дослідити зміст тексту, вивчити його лексичний склад і структуру, а й виявити фонетичні особливості говірки (звукова система, інтонація, наголос та ін.).

[1] Детальніше про це див. статтю О.М. Демської-Кульчицької [1]

Аналізуючи досягнення українських науковців, бачимо, що попри безумовно прогресивний розвиток майже всіх аспектів лінгвістики, особливо лексикографії, термінографії, морфології, фонології, через призму комп'ютерних технологій поки що не створено *єдиного* комплексу, або, іншими словами, так званого „робочого місця" лінгвіста з доступом до масиву фактичного матеріалу (зокрема, лексичної картотеки, тезаурусу, різного роду словників, конкордансів тощо), з можливістю використання його для досліджень, аналітичної роботи, інформації.

Хочеться особливо наголосити на глобальності та масштабності роботи в галузі комп'ютерної лінгвістики, що потребують об'єднання зусиль, наукових пошуків та власне наповнення баз даних як основи комп'ютернолінгвістичних досліджень та розробок фактичним матеріалом. Постає необхідність проведення спеціалізованих наукових конференцій з цього питання, співпраці, обміну інформацією. Надзвичайно важливим є винесення цієї теми в інформаційний простір Internet, а також використання її текстових ресурсів. Є. А. Карпіловська зазначає, що „накопичений в україністиці досвід створення лінгвістичних баз даних, формування машинних копій та версій різнотипних традиційних („паперових") словників, розроблення лінгвістичних словникових та текстових процесорів ставить на часі завдання об'єднання наявної інформації на єдиній концептуальній та методико-процедурній основі в загальнодержавний комп'ютерний фонд української мови, який виконував би всі властиві такій інституції функції: інформаційно-довідкову, дослідницьку, навчальну та редакційно-видавничу" [3, с. 101]. Успішне ж розв'язання цього завдання передбачає співробітництво усіх фахівців – лінгвістів та математиків-програмістів, а також вироблення правової основи для такої співпраці.

Література

1. Демська-Кульчицька О. Корпуси мов і один із важливих підходів до проектування корпусу текстів української мови // Лінгвістичні студії: Зб. наук. праць. – Донецьк, 2005. – Вип. 13.
2. Казакевич О. Машинные фонды языков народов СССР // Науч.-техн. Информация. – Сер. 2 – 1989. – №10
3. Карпіловська Є. Вступ до комп'ютерної лексикографії. – К., 2004.
4. Широков В. Феноменологія лексикографічних систем. – К., 2004.
5. Проблеми українізації комп'ютерів. – К., 1993.