3. Изард К.Э. Психология эмоций / Перев. с англ. – СПб.: Издательство “Питер”, 2000. – 464 с.

4. Маслова В.А. Лингвокультурология: Учеб. пособие для студ. высш. учеб. заведений. – М.: Издательский центр “Академия”, 2001. – 208 с.

5. Прадід Ю.Ф. Фразеологічна ідеографія (проблематика досліджень). – К., Сімферополь, 1997. – 252 с.

6. Фразеологічний словник української мови / Уклад. В.М. Білоноженко та ін. – К.: Наукова думка, 1999. – 984 с.

7. Цимбалюк Ю.В., Краковецька Г.О. Крилаті латинські вислови. – К.: Вища школа, 1976. – 192 с.

УДК 81’373.72

Н.А. Кошка,
Національний університет водного господарства та природокористування, м. Рівне

ОГЛЯД РОЗВИТКУ ТА СТАНУ КОРПУСНОЇ ЛІНГВІСТИКИ В ІСТОРИЧНОМУ АСПЕКТІ

У статті розглядається історія та основні принципи створення корпусної лінгвістики. Наведено приклад метарозмітки Національного корпусу російської мови та висвітлено актуальність подальшого створення Національного корпусу української мови.

The article is focused on the history and the main principles of Corpus linguistic creation. It gives the example of Russian National Corpus meta-marking and elucidates the topicality of Ukrainian National Corpus creation.

Мова – це універсальне надбання людства й універсальна реальність суспільного існування. Це, за висловом німецького філософа Мартіна Гайдеггера, оселя людського духу [1, с. 7]. І не дивно, що люди ще в давні часи зацікавились мовою і створили про неї науку – мовознавство (лінгвістику). Проблеми сутності мови, її функції, структури і розвитку є дуже важливими, оскільки мова є необхідною умовою мислення та існування суспільства. Через пізнання мови пролягає шлях до пізнання людини.

Мовознавство – одна із найрозгалуженіших наук. Усі дослідження розділяють між двома підрозділами цієї науки – загальним та конкретним мовознавством. Проте, на даний час, в умовах стрімкого

розвитку сучасних комп’ютерних технологій все більшу увагу привертає корпусна лінгвістика як галузь конкретного мовознавства.

Актуальність даної проблеми полягає в тому, що не так багато досліджень зроблено в цій галузі. Тому ми поставили за мету розглянути основні проблеми створення, принципи та теоретичні засади корпусної лінгвістики.

Основою корпусної лінгвістики є розроблення теоретичних засад і практичних прийомів побудови, машинного опрацювання та експлуатації мовних даних, оформлених як корпус текстів. Такі корпуси становлять машинописне, стандартно організоване зібрання репрезентативних для певної мови, діалекту або іншої підмножини мов писемних або усних текстів в електронній формі, призначених для лінгвістичного аналізу й опису. Це, свого роду, інформаційно-довідкова система.

Термін “корпусна лінгвістика” почали застосовувати ще в 90-х роках ХХ століття у зв’язку з розвитком практики створення корпусів, якому сприяв, як вже зазначалось, розвиток обчислювальної техніки. Першим таким великим комп’ютерним корпусом вважається Браунівський корпус (англ. Brown Corpus), що був створений в університеті Брауна в 60-ті роки минулого століття [2]. Його автори – У. Френсис та Г. Кучера. Тексти належали п’ятнадцятьом найбільш масовим жанрам англомовної друкованої прози США. Його автори вперше вжили слово “корпус” як “сукупність текстів, яка є визначальною для даної мови і створена для лінгвістичного аналізу” [2]. Корпус супроводжувався не лише детальним описом, але і великою кількістю зібраних матеріалів. У. Френсис та Г. Кучера поставили мету представити корпус текстів, які відповідали б наступним критеріям відбору:

1. Походження і склад тексту (автор повинен бути носієм американського варіанта англійської мови, а якщо є діалог, то він повинен займати менше половини об’єму тексту);

2. Синхронізація (в корпус входили тексти, що вперше були надруковані в 1961 році);

3. Продумане співвідношення різноманітних жанрів та відбір окремих текстів;

4. Доступність для комп’ютерної обробки (спеціальні посилання для передачі графічних особливостей тексту) [2].

Згодом визначили і сформулювали основні вимоги щодо відбору текстів – об’єм окремого тексту та склад і співвідношення жанрів повинні відображати свої стильові особливості. Незабаром Браунівський корпус став стандартом для створення інших аналогічних корпусів, тим паче всі підрахунки вже можна було швидко виконувати на комп’ютері. Таким чином, корпус текстів можна назвати багаторівневою динамічною системою. В процесі використання Брау-

нівського корпуса стало зрозумілим, що зробити певні лінгвістичні порівняння можна лише використовуючи тільки великий об'єм матеріалу, який певним чином організований у корпус. Саме такий підхід став основою корпусної лінгвістики.

Незабаром, дотримуючись принципів, покладених в основу створення Браунівського корпусу, був створений британський аналог Браунівському корпусу – LOB (Lancaster – Oslo/Bergen) [2]. Це зробило можливим порівнювати два варіанти англійської мови – американський та британський. Тим паче, що програми лінгвістичної обробки відповідних текстів на комп'ютері можна було застосовувати без будь-яких змін до текстів того чи іншого корпуса.

У 1992 році була створена організація "Європейська корпусна ініціатива", яка створила близько 50 корпусів текстів об'ємом від 12 тисяч до 5 мільйонів слів кожен на європейських мовах. Метою цієї організації було створення корпусів так званих "паралельних" текстів (текст на одній мові разом з перекладом на іншій), спочатку на англійській, німецькій, французькій, іспанській мовах. Паралельні тексти, або бітексти створюються за допомогою комп'ютерних програм, які називаються "інструментами для вирівнювань" (alignment tools) або "інструментами для бітекстів" (bitext tools), що дозволяють автоматично порівнювати оригінальну версію тексту з його перекладом. Подібні програми, як правило, узгоджують два тексти (оригінал та переклад) пореченнєво. Зібрання бітекстів називається "двомовним корпусом" та може використовуватись як довідник для пошуку потрібних словосполучень [4].

Легкість доступу до великих масивів різноманітного лінгвістичного матеріалу за допомогою комп'ютера призвела до якісно нових результатів. Перш за все це стосується лексикографії. Варто зазначити, що з'явився новий тип словників COBUILD, який завоював прихильність масового користувача. Ці словники були створені на "корпусному" комп'ютерному матеріалі за допомогою комп'ютера. Більше того, так був створений частотний словник російської мови Л.Н. Засоріної об'ємом в 1 мільйон слів, що містить приблизно в рівних пропорціях суспільно-політичні, художні, наукові та науково-популярні тексти з різних галузей, що є не менш важливим. Частотні словники дають можливість порівнювати два корпуси, щоб визначити, які слова є найбільш характерними для кожного з них. Наявність великої кількості текстів в електронному вигляді значно полегшила роботу створення корпусів розміром в десятки і, навіть, сотні мільйонів слів. Такі корпуси існують (чи вже розробляються) для німецької, китайської, японської та інших мов. Для прикладу розглянемо основні параметри створення Національного корпусу російської мови.

Суттєвою частиною пошукової системи Корпусу є так звана метарозмітка текстів, що входять до нього. Під метарозміткою розумі-

ємо приписування до тексту атрибутів, що характеризують обставини його створення, автора, тематику, жанрові особливості та ін. Метарозмітка необхідна перш за все для того, щоб дослідник, що користується Корпусом, міг складати за своїм бажанням довільні вибірки текстів задані по зовнішнім параметрам: наприклад, тексти мемуарного характеру, тексти, написані авторами, що народилися між 1940 – 1960рр., тексти автобіографій, тексти романів та ін. Враховуючи об’єм та різновидність текстів корпусу, така диференціація є необхідною: більшість дослідників працюватиме не з корпусом в цілому, а з якимись підмасивами текстів (художніми, публіцистичними, діловими та ін.). Структура метарозмітки корпусу – порівняно проста, призначена не для фахівців з корпусної лінгвістики, а для пересічного користувача (у тому числі і для лінгвіста, що не знайомий з термінологією корпусних досліджень). Саме такий тип метарозмітки безпосередньо відображається в інтерфейсі. Інтерфейс для спрощеного метатекстового пошуку влаштований так, що параметри тексту об’єднуються в декілька блоків:

I. “Паспорт тексту” (автор тексту, назва тексту, час створення тексту, об’єм тексту).

Блок II складається з трьох пошукових масивів: “нехудожня проза”, “художня проза”, “драматургія”. Перші два масиви мають дещо різні структури параметрів, тому оформляються окремо.

II. Художні тексти (жанр тексту, тип тексту, хронотоп тексту).

III. Нехудожні тексти (сфера функціонування тексту, тип тексту, тематика тексту).

При розробці параметрів метатекстової розмітки укладачами корпусу був врахований світовий досвід, перш за все досвід укладачів Британського національного корпусу. В англомовній літературі існує цілий ряд пропозицій по класифікації текстів для створення представницьких корпусів, але було ухвалено рішення спиратися в основному на рекомендації Дж. Синклера (так званий стандарт EAGLES (European Advisory Group on Language Engineering Standards), прийнятий в багатьох сучасних системах автоматичної обробки текстів). Ці рекомендації були адаптовані на російському матеріалі та склали перший варіант метарозмітки (умовно – “міжнародний” варіант, або варіант Синклера-Шарова).

На даний час ведеться робота по внесенню цієї інформації в справжній корпус. Їх застосування в корпусі полегшить зіставлення результатів метарозмітки в російському та інших Національних корпусах та буде зручнішим для фахівців з корпусної лінгвістики різних країн. Суть класифікації Синклера-Шарова полягає в тому, що вона заснована переважно на логічних властивостях комунікації, тому її можна застосовувати при описі дискурсу на будь-якій мові. Розрізняють два класи чинників, що впливають на вибір текстів в

корпусі: *зовнішні* (Е), тобто позамовні чинники, які можуть вплинути на структуру або зміст тексту, та *внутрішні* (І) – чинники, що відображають властивості мови, що використовуються в тексті. Виділяються три групи Е-факторів:

Е1 (origin) – чинники, що відносяться до створення тексту автором;

Е2 (state) – чинники, що відносяться до зовнішніх ознак тексту;

Е3 (aims) – чинники, що відносяться до цілей створення тексту та його впливу на аудиторію.

Виділяють два основних І-фактора:

І1 (topic) – предметна область тексту;

І2 (style) – стилістичні особливості (частково залежні від Е-факторов).

У зв'язку з тим, що величезна кількість текстів представлена в електронній формі і доступно знаходиться в Інтернеті, то найбільшим корпусом можна вважати і сам Інтернет, а засобами доступу до нього є пошукові системи та інтерпретація результатів або за кількістю знайдених сторінок, або за першими посиланнями. Відомо, що якість пошуку можна значно підвищити шляхом застосування найпростішого тезауруса, що індексує запити користувачів, а також документів, які можна продивлятись при пошуку. Проте тексти в Інтернеті хаотичні і лінгвістично цікавий запит зазвичай складно, або навіть і неможливо сформулювати за допомогою запитів пошукової машини. Тому за результатами пошуку не можна об'єктивно оцінити вибірку: або потрібних текстів немає в Інтернеті, або не були знайдені даною пошуковою системою. Ось чому корпусна лінгвістика використовує одно– чи багатомовні корпуси текстів, які анотовані лінгвістично значущою інформацією, наприклад, частини мови, леми, морфологічні ознаки, синтаксична структура, семантична інтерпретація. В англійській мові така методологія отримала назву англ. Googleology. Для російської мови це може бути Яндексологія, яка є високоефективною, повністю реалізує контекстний пошук, містить потужну мову запитів, допускає гнучку настройку індексатора для нестандартних типів текстів. Яндекс-сервер виконує складні запити у великому масиві текстів корпусу за лічені секунди, причому об'єм корпусу ніяк не впливає на швидкість пошуку. Для української мови все більшого поширення набуває meta. Але проблема полягає в іншому. Для національного корпусу, який хоча б певною мірою достовірно описував дану мову, достатнім розміром вважається 100 мільйонів слів. Але, на жаль, для української мови поки що такого корпусу не існує. На даний час Українським мовно-інформаційним фондом створений морфологічно анотований корпус на 36 мільйонів слововживань. Причому масиви цього корпусу використані при створенні нового 20-томного тлумачного словника української мови та ряду інших словників. Тому створення Національного корпусу української мови є однією з актуальних тем корпусної лінгвістики.

Література

1. Баранов А.Н. Проблема репрезентативности корпуса текстов // Труды Международного семинара Диалог-2001 по компьютерной лингвистике и ее приложениям. – http://www.dilog-21.ru, 2001.
2. Демська-Кульчицька О.М. Базові поняття корпусної лінгвістики // Українська мова. – 2003. – №1. – С. 42-47.
3. Кочерган М.П. Загальне мовознавство: Підручник: – К.: Видавничий центр "Академія", 2003. – 464 с. (Альма-матер)
4. Рыков В.В. Корпусная лингвистика. – http://rykov-cl.narod.ru/lekcii.doc, 2001.
5. Національний корпус російської мови. – http://ruscorpora.ru
6. Adam Kilgarriff, Putting Frequencies in the Dictionary, // International Journal of Lexicography, 10(2). 1997. C. 135–155 [1].
7. Ide N. Corpus Encoding Standard. – http://lpl.univ.-aix.fr/projects/multext/CES, 2000.

УДК 81'282.162.1

Ю.К. Кратюк,
Національний університет "Острозька академія",
м. Острог

СТАН ДОСЛІДЖЕНЬ ПІВДЕННОПОГРАНИЧНОЇ ДІАЛЕКТНОЇ ПОЛЬСЬКОЇ МОВИ

У статті розглянуто стан досліджень польського південнопограничного діалекту, що функціонує за межами Польщі, у період від XV ст. до наших днів. Також представлено найголовніші мериторично-термінологічні питання, пов'язані із вказаним діалектом, з метою уточнення його природи і встановлення факту існування.

The article under consideration overlooks the state of research of the polish south border dialect starting from the 15 century until nowadays functioning outside Poland. The main merithorical-terminological problems connected with the above mentioned dialects are introduced in order to specify the dialect phenomenon and to acknow ledge the fact of its existence.

Польським діалектам, що функціонують на території сучасної України, присвячено велику кількість наукових праць, проте зацікавленість цією темою не послаблюється, а навпаки постійно зростає. Мова однією із перших реагує на політичні та соціально-економічні зрушення у житті суспільства. Оскільки такі зміни є значними і ди-