Исследования по русскому языку. – СПб., 1885–1886. – Т. 1. – С. 131–285.

УДК 8.81.81-13

М. О. Ш в е д о в а

Київський національний лінгвістичний університет

Корпусні методи дослідження регіональних відмінностей у межах однієї мови (на матеріалі регіональних корпусів української та російської мов)

Шведова М. О. Корпусні методи дослідження регіональних відмінностей у межах однієї мови (на матеріалі регіональних корпусів української та російської мов). У статті представлено короткий огляд корпусних ресурсів, призначених для дослідження української мови, а також огляд світового досвіду у дослідженні регіолектів та національних варіантів мов на базі корпусів. Автором було розроблено два корпуси, розмічених за регіональною приналежністю текстів, –російської мови України (після 1991 р.) та української мови. У статті описані основні характеристики цих корпусів і наведено приклади конкретних лексичних досліджень на їх базі.
Ключові слова: корпусна лінгвістика, регіолект, регіоналізми, лексика, українська мова, російська мова в Україні.

Шведова М. А. Корпусные методы исследования региональных отличий в пределах одного языка (на материале региональных корпусов украинского и русского языков). В статье представлен краткий обзор корпусных ресурсов, предназначенных для изучения украинского языка, а также обзор мирового опыта в исследовании регионалектов и национальных вариантов языков на базе корпусов. Автором разработаны два корпуса, размеченных по региональной принадлежности текстов, – русского языка Украины (после 1991 г.) и украинского языка. В статье описываются основные характеристики этих корпусов и приводятся примеры конкретных лексических исследований на их базе.
Ключевые слова: корпусная лингвистика, регионалект, регионализмы, лексика, украинский язык, русский язык Украины.

Shvedova M. Corpora methods in studying regional language varieties, exemplified by geographically annotated Ukrainian and Russian corpora. The paper briefly presents existing corpora resources featuring Ukrainian texts as well as the previous experience in corpora-based studies of national language varieties (eg British vs. American English, New world Englishes, varieties of Spanish, Russian language in Belarus). The author has built two large corpora, respectively of the (written standard-oriented) Russian language in Ukraine (the so-called Ukrainian Russian) and of the Ukrainian language. Both corpora feature metatextual annotation by the regions where the texts were created (including, for Ukrainian, different countries of diaspora). The paper describes the main properties of these corpora and features examples of corpora-based lexical studies: borrowings in Ukrainian Russian (as compared to the standard of Russia proper, represented in the Russian National Corpus) and geographically distributed synonyms in Ukrainian.
Key words: corpus linguistics, regional language varieties, regional lexica, Ukrainian language, Russian language of Ukraine.

Українська корпусна лінгвістика в останні роки розвивається досить активно в різних університетах та робочих групах. Корпусні дослідження проводять вчені інституту філології КНУ ім. Т. Г. Шевченка, Києво-Могилянської академії, Донецького національного університету (м. Вінниця), Національного університету «Львівська політехніка». Триває робота над Браунським українським корпусом обсягом 1 мільйон слововживань (Корпусна група БрУК). В інтернеті українська мова представлена кількома корпусами.

1. Корпус української мови на сайті mova.info – проект лабораторії комп'ютерної лінгвістики інституту філології КНУ ім. Т. Г. Шевченка, під керівництвом Н. П. Дарчук. Корпус складається з підкорпусів художніх, наукових, поетичних, фольклорних, законодавчих та публіцистичних тестів. Загальний обсяг корпусу приблизно 60 мільйонів словоформ.
2. Корпуси текстів української мови кафедри загального та прикладного мовознавства і слов'янської філології Донецького національного університету (м. Вінниця) на сайті corpora.donnu.edu.ua – доступні для он-лайн пошуку корпуси художніх, наукових, поетичних текстів, а також корпус українських граматик. Загальний обсяг корпусів близько 4 млн. словоформ.
3. Відкриті корпуси української мови групи lang-uk, доступні для завантаження за адресою: lang.org.ua/uk/corpora/: 229 текстів з українського браунівського корпусу, корпус UberText (корпус художніх та публіцистичних текстів обсягом близько 600 млн. словоформ, призначений для комп'ютерної обробки), корпус законів та нормативно-правових актів України.
4. Український веб-корпус Лідського університету (Centre for Translation Studies, University of Leeds), доступний для пошуку на сайті corpus.leeds.ac.uk/internet.

5. Паралельні українсько-російський та російсько-український корпуси на сайті Національного корпусу російської мови: www.ruscorpora.ru/search-para-uk.html. Загальний обсяг корпусів близько 9 млн. словоформ.
6. Польсько-український паралельний корпус обсягом понад 3 мільйони словоформ (domeczek.pl/~polukr/index.php)

Ми досі не маємо корпусу української мови, який би був достатньо великим (для порівняння: Національний корпус російської мови містить понад 500 мільйонів словоформ), репрезентативним, доступним в Інтернеті із достатнім лінгвістичним інструментарієм для пошуку. Зокрема бракує такого корпусу, де можна було б формувати підкорпуси для окремих досліджень – за часом написання текстів, авторами, регіонами, до яких належать тексти.

1. Використання корпусів для дослідженні регіолектів

Регіональні та національні варіанти мов представлені в корпусній лінгвістиці. Є окремі корпуси не тільки британської англійської (наприклад, Британський національний корпус, BNC) та американської англійської (наприклад, Сучасний та Історичний корпуси американської англійської, COCA та COHA), є також корпуси так званих «нових світових англійських мов» (New World Englishes), які виникають у різних країнах світу і зазвичай не є рідними для тих, хто ними говорить. Ще у 1988 році з'явився проект Міжнародного корпусу англійської мови. Сучасний корпус, укладений на базі Інтернету, The corpus of Global Web-based English (GloWBE), має обсяг майже 2 мільярди слововживань. Цій проблематиці присвячені як досить давні [8], так і сучасні праці [7]. На матеріалі цих корпусів були проведені зіставні дослідження лексики та граматичних категорій у різних регіональних різновидах англійської мови, наприклад, перфекта (у збірнику [9]).

Аналогічний корпус (2 мільярди слововживань з 21 країни) укладений Марком Девісом для іспанської мови (https://www.corpusdelespanol.org/web-dial/). В його складі є зокрема й тексти, які походять з країни з іншою домінантною мовою (США).

У складі Національного корпусу російської мови є підкорпус зарубіжної преси, зокрема колекція газет Гродненщини російською та білоруською мовами, за яким також досліджували регіональні лексичні та граматичні особливості [4; 5].

Дослідження мови з точки зору регіональних відмінностей проводять також на матеріалі масиву текстів Інтернету із використанням можливостей пошуку за регіонами (зокрема по блоґах), які надають пошукові системи Яндекс та Гугл [1; 2; 3]. Ці пошукові системи не є спеціально лінгвістичним інструментом, не мають засобів для лінгвістичного пошуку та не дають точної інформації щодо кількості даних, яка потрібна для статистичних досліджень. Проте використання Інтернету для пошуку регіонально маркованих мовних одиниць має також очевидні переваги: великий обсяг текстів, широке географічне, стилістичне охоплення, швидкість та легкість пошуку даних, визначення регіону, датування, відомостей про автора тощо.

2. Корпус російської мови в Україні

Існує чимало досліджень специфіки російської мови на території України, здебільшого з точки зору інтерференції та культури мови (роботи Г. П. Їжакевич, Т. К. Чорторизької, Н. Г. Озерової, В. І. Кононенка, Р. В. Болдирєва та ін.). За допомогою сучасних корпусних технологій можна продовжити та розширити ці дослідження, описати російську мову України не тільки як таку, що має певні характерні «неправильності» та «помилки», але і з позицій сучасної лінгвістики: як регіональний варіант мови.

Для вирішення цього завдання нами було укладено корпус сучасної російської мови України обсягом близько 27 мільйонів словоформ (понад 3600 текстів 320 авторів з різних областей України). Корпус складається з текстів, які були написані після 1991 року: художніх, публіцистичних, текстів блогів. Цей корпус ми порівняли з Національним корпусом російської мови (http://ruscorpora.ru/), де обрали для порівняння підкорпус текстів того ж періоду (після 1991 року) обсягом 102 мільйони слів.

За цим корпусом було підтверджено висновки попередніх робіт (щодо більшої частотності в російській мові України порівняно з Росією демінутивів, конструкцій з прийменником *про*), а також зроблено власні спостереження. Зокрема знайдено, що для російської мови України характерна більша частотність слів іншомовного походження (відносно суто російських синонімів, слов'янських за походженням).

	Національний корпус російської мови (тексти після 1991 року)	Корпус сучасної російської мови України
удача / фортуна	15/1	7/1
законный / легитимный	15/1	4/1
подделка / фальсификат	21/1	8/1

подлинный / аутентичный	31/1	10/1
маленький / мизерный	99/1	20/1
показать / презентовать	123/1	23/1
область / сфера	3/1	1/1
нарушение / аберрация	99/1	35/1
удобный / комфортный	6/1	3/1
избиратель / электорат	4/1	1/1
одаренность / харизма	1,5/1	1/3
предприятие / бизнес	1,5/1	1/1,5
мышление / менталитет	5/1	3/1
облик / имидж	2/1	1/1
руководитель / менеджер	5/1	4/1
встреча / саммит	22/1	6/1
соглашение / консенсус	19/1	4/1
настоящий / реальный	2/1	1/1
совершенно / абсолютно	3/1	1/1
вольный / фамильярный	25/1	8/1
выпуклый / рельефный	3/1	2/1

З іншого боку, деякі іншомовні запозичення, добре засвоєні російською мовою, можуть мати рівне співвідношення по частотності з російським синонімом в обох корпусах, такі, наприклад, *неожиданность – сюрприз, противоречие – антагонизм, приспособленец – конъюнктурщик, материк – континент, платок – шаль, хлопать – аплодировать, наибольший – максимальный, выделяться –контрастировать* [6].

3. Корпус української мови, розмічений за регіонами

Зараз у стані розробки корпус української мови, розмічений за регіонами. Нами зібрано 6255 текстів загальним обсягом понад 100 мільйонів словоформ. Кожному тексту приписано таку інформацію: ім'я автора, назву, жанр, рік написання (для перекладів – рік створення українського перекладу), якщо текст перекладний, також ім'я перекладача та мову оригіналу. Кожному автору приписано один або кілька регіонів, де він народився, вчився або жив тривалий час. Перекладному тексту в корпусі відповідають регіони, які приписано перекладачу. Частка перекладів в корпусі – приблизно 30 %. Корпус містить переклади з 38 мов, найбільше – з англійської (8,5 %) та російської (7,4 %).

За жанром найбільша частка – це художні тексти (майже 60 %), решта (приблизно 40 %) – це публіцистичні, наукові та навчальні тексти (з історії, права, філософії тощо).

За часом написання текстів корпус охоплює майже 200 років – від 1819 до 2017 р. Кількість і обсяг текстів за різні періоди такі:

1819–1913	497	40735932	4 %
1914–1940	315	59384405	6 %
1941–1990	2064	456825001	43 %
1991–2017	2199	500491613	47 %

В основу розмітки корпусу за регіонами покладено сучасний адміністративний поділ України. За областями тексти розподіляються нерівномірно. Один текст може належати одночасно до кількох регіональних підкорпусів: якщо автор мешкав тривалий час у різних місцях, йому приписували два чи більше регіонів. Багато людей переїжджали з різних місць до Києва, тому київський підкорпус найбільший – 60 % від загального обсягу текстів, але він найменш інформативний з точки зору регіональної мовної інформації. Серед інших підкорпусів найбільший за обсягом підкорпус текстів Львівської області – майже 20 %. 9 % Полтавської. Приблизно по 5 % – Івано-Франківської, Житомирської, Волинської, Київської, Дніпропетровської, Тернопільської та Сумської областей і менша частка інших. У корпусі представлені тексти з усіх областей України і з Криму.

У випадках, коли автор довгий час жив за кордоном або емігрував, йому приписували також іншу країну. Таким чином, маємо чималий

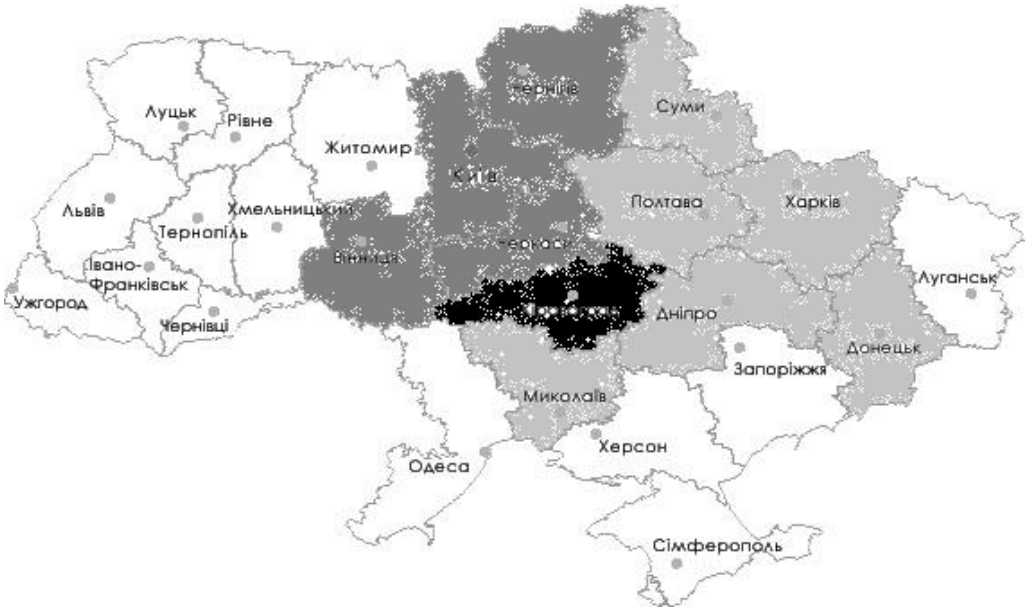
підкорпус текстів авторів діаспори. Тексти США, Канади, Польщі, Німеччини, Великої Британії, Франції здебільшого належать емігрантам 40-х років, менша частка – емігрантам 20-х рр. Обсяг корпусу української мови діаспори – понад 10 млн. словоформ.

Для пошуку за корпусом використовується пошукова система diskMETA-Personal. Це програма для локального пошуку, яка шукає за лексемою, словоформою та точною фразою. За результатами такого пошуку можна рахувати частотність слів, словоформ та сталих виразів, порівнювати таку частотність у різних областях.

Було порівняно кількість вживання *спасибі* і *дякую* у різних українських регіонах. Обидві одиниці належать до літературної мови і вживаються по всій території України. У текстах корпусу *дякую* було вжито 2112 раз, а *спасибі* – 1121 раз. Частка *дякую* становить 65 % від загальної кількості вживання обох форм, а *спасибі* – 35%. У деяких областях це співвідношення близьке до середнього, але в деяких воно помітно відхиляється від нього. В Івано-Франківській, Львівській, Хмельницькій областях частка *дякую* становить понад 80 %. В Київській, Черкаській, Вінницькій, Чернігівській областях кількість *спасибі* наближається до *дякую*. А у Кіровоградській області переважає *спасибі*.

Область	Кількість *дякую*		Кількість *спасибі*	
Івано-Франківська	85	87%	13	13%
Львівська	216	81%	51	19%
Хмельницька	16	80%	4	20%
Крим	7	78%	2	22%
Запорізька	25	74%	9	26%
Чернівецька	55	74%	19	26%
Волинська	65	73%	24	27%
Одеська	18	72%	7	28%
Херсонська	25	69%	11	31%
Тернопільська	43	68%	20	32%
Житомирська	44	67%	22	33%
Луганська	8	67%	4	33%
Рівненська	16	67%	8	33%
Закарпатська	13	65%	7	35%
Донецька	73	60%	49	40%
Миколаївська	42	60%	28	40%
Полтавська	157	60%	103	40%
Сумська	50	60%	33	40%
Дніпропетровська	94	59%	66	41%
Харківська	173	59%	120	41%
Київська	50	56%	40	44%
Черкаська	72	56%	57	44%
Вінницька	72	55%	59	45%
Чернігівська	101	53%	89	47%
Кіровоградська	45	45%	55	55%

Якщо зобразити це на карті, добре видно що області, де частотність *спасибі* вища за середню, це територія в центрі України.



Області з найвищою частотністю *дякую* – на заході.

Корпус, розмічений за регіонами, має стати основою дослідження регіональних відмінностей в українській мові, а в перспективі – системного опису українських регіональних стилів.

Висновки

Корпусна лінгвістика в Україні активно розвивається, частина корпусів української мови доступна в Інтернеті для пошуку або завантаження. Проте досі існує потреба у великому, репрезентативному корпусі, який був би доступний в Інтернеті і мав розгалужений інструментарій для лінгвістичного пошуку, можливість укладання різноманітних підкорпусів, побудови графіків тощо.

Для системного вивчення регіональних варіантів мов необхідно мати корпуси достатньо великого обсягу, розмічені за регіонами. За такими корпусами можна рахувати і порівнювати частотність слів і граматичних одиниць у різних регіонах. Вивчення регіональних стилів необхідно для глибокого розуміння функціонування мови, а також пошуку нових підходів до її лексикографічного та граматичного опису.

Література

1. Ахметова М. В. Лексические регионализмы и локализмы в русскоязычном Интернете: проблема сбора материала / М. В. Ахметова // Русский язык и новые технологии. – М., 2014. – С. 155–171.

2. Беликов В. И. К методике корпусного исследования лексики / В. И. Беликов // Русский язык и новые технологии. – М., 2014. – С. 99–130.

3. Романий Г. И. Региональная городская лексика в русском языке: происхождение, типология и ареализация по данным узуса в сети Интернет / Г. И. Романий // Русский язык и новые технологии. – М., 2014. – С. 172–186.

4. Рычкова Л. В. Многоцелевой лингвистический корпус региональных СМИ / Л. В. Рычкова, А. Ю. Станкевич // Информационно-коммуникативные технологии в лингвистике, лингводидактике и межкультурной коммуникации. – Вып. 6. – М., 2014. – С. 488–497.

5. Савчук С. О. Корпус как инструмент для исследования особенностей функционирования русского языка в региональной прессе / С. О. Савчук // Труды института русского языка им. В. В. Виноградова. – М., 2015. – С. 333–365.

6. Шведова М. А. Иноязычные влияния в русской речи Украины (Лексика) / М. А. Шведова // Мовознавчий вісник : Збірник наукових праць / Ред. кол. : Г. І. Мартинова та ін. – Черкаси, 2016. – Вип. 21. – С. 86–92.

7. Mair C. World Englishes and Corpora / C. Mair // M. Filppula et al. (eds.) The Oxford Handbook of World Englishes. Oxford: Oxford University press, 2017.

8. Schmied J. Corpus linguistics and non-native varieties of English / J. Schmied // World Englishes, 1990 – Vol. 9 – No. 3. – pp. 255–268.

9. Werner V. Re-Assessing the Present Perfect: Corpus Studies and Beyond / V. Werner, E. Seoane (eds.) – Berlin : Mouton de Gruyter, 2016.