УДК 81'33=134. 2

Кундис Л. Т.,
Львівський національний університет імені Івана Франка

НАЦІОНАЛЬНІ КОРПУСИ ІСПАНСЬКОЇ МОВИ

У статті подано опис двох національних корпусів текстів іспанської мови (El banco de datos del español і El Corpus del Español), розглянуто їх основні властивості, з'ясовано переваги та недоліки, а також проаналізовано дослідні можливості кожного з них.

Ключові слова: *корпусна лінгвістика, корпус текстів, національний корпус, іспанська мова.*

В статье представлено описание двух национальных корпусов текстов испанского языка (El banco de datos del español и El Corpus del Español), рассмотрены их главные свойства, определены преимущества и недостатки, а также проанализированы исследовательские возможности каждого из них.

Ключевые слова: *корпусная лингвистика, корпус текстов, национальный корпус, испанский язык.*

The article deals with the description of two national text corpora of Spanish language (El banco de datos del español and El Corpus del Español). Were examined their main features, determined advantages and disadvantages and was carried out the analysis of research possibilities of each one of them.

Key words: *corpus linguistics, text corpus, national corpus, Spanish language.*

Корпусна лінгвістика, що виникла як наука близько півстоліття тому, останнім часом набуває все більшого поширення. Про стрімкий розвиток цієї галузі прикладної лінгвістики свідчить постійне збільшення кількості мовознавців, які здійснюють різноманітні дослідження всіх рівнів мовної системи на матеріалі одного чи декількох корпусів текстів (КТ), число яких теж постійно зростає.

На сьогодні більшість мов вже представлена електронними ресурсами такого типу. Іспанська мова, як одна з найбільш поширених мов у світі, не є винятком: світ побачила значна кількість різноманітних текстових корпусів, які дають у руки лінгвісту безцінний матеріал для опису та вивчення мови в усіх її проявах. Найвідомішими та найбільш використовуваними з усіх корпусів іспанської мови є Банк даних іспанської мови (El Banco de datos del español) та Корпус іспанської мови (El Corpus del Español).

Банк даних іспанської мови (El Banco de datos del español) [10] – текстовий корпус, створений наприкінці 1990-х років лінгвістами Королівської академії іспанської мови. За стратегією побудови та призначення цей КТ є дослідницьким, тобто матеріал до даного типу корпусу був відібраний таким чином, щоб він міг адекватно репрезентувати мовну систему в її реалізації. За способом подання мовного матеріалу Банк даних іспанської мови є динамічним, тобто не є попередньо встановленим і незмінним набором текстів, а передбачає поповнення їх множини через певні проміжки часу. За хронологічним параметром розглянутий корпус є синхронно-діахронним, тобто містить тексти, що репрезентують сучасну мову та її історичний розвиток. За предметною галуззю він є загальномовним (репрезентує національну мовну систему та її реалізацію), а за кількістю представлених мов – одномовним (конститутивною для нього є лише одна мова, а саме – іспанська). За типом охоплення тексту – повнотекстовий, тобто до його складу увійшли тексти повністю, без жодних скорочень. За розміром Банк даних іспанської мови належить до великих, або мегакорпусів, оскільки його обсяг становить понад 410 мільйонів слововживань. Щодо типу реалізації мовної системи описуваний КТ належить до ресурсів мішаного типу: він подає тексти як писемного, так і усного мовлення.

Що стосується анотації корпусу, то його текстам було присвоєно маркування, визначене міжнародним стандартом SGML (Standard Generalized Markup Language). Та на сьогодні рівень його анотації (за визначенням авторів корпусу [10], мінімальний: він містить лише зовнішнє та структурне маркування) значно програє іншим, пізнішим проектам. Крім того, Банк даних іспанської мови належить до нелематизованих КТ, що, як зазначають дослідники [8, с. 152], робить неможливою велику кількість лінгвістичних досліджень на морфологічному і синтаксичному рівнях. До того ж, оскільки іспанська мова є багатою на морфологічні форми, відсутність лематизації суттєво ускладнює роботу з даними корпусу і в десятки разів збільшує час, затрачений на пошук потрібної інформації.

Структурно Банк даних іспанської мови організований як генеральний корпус з двома субодиницями: діахронним (*CORDE*) та синхронним (*CREA*) підкорпусами, що є взаємодоповнюваними ресурсами. Їх інтегративність полягає у можливості переходу текстів з синхронної частини у діахронну, що здійснюватиметься тоді, коли дані перестануть входити до предметної галузі першої з них [10].

Діахронний корпус іспанської мови (El Corpus Diacrónico del Español, CORDE) [7] – це корпус текстів "усіх часів та з усіх країн, де коли-небудь говорили іспанською, від зародження мови до 1975 року" [10]. Основними перевагами цього КТ є досить значний обсяг репрезентованих текстів (понад 250 млн. слововживань, що роблять його найбільшим на сьогодні іспаномовним ресурсом такого типу) усіх періодів розвитку іспанської мови, а також можливість використання корпусу широким загалом. Саме ці особливості Діахронного корпусу іспанської мови посприяли популярності застосування його даних в історичних дослідженнях цієї романської мови.

Матеріалами для CORDE стали лише письмові тексти різноманітних жанрів (від лірики, драми і літературної прози до науково-технічних, історичних, релігійних текстів, а також періодики). Автори цього корпусу ставили собі за мету зібрати все "географічне, історичне і жанрове різноманіття іспанської мови задля досягнення таким зібранням достатньої репрезентативності" [10].

Хронологічно тексти описуваного КТ поділені на три великі періоди:

Середні віки (підгрупи: тексти періоду до 1250 р. ; тексти 1250 – 1492 рр.), що становлять 21 % всього корпусу;

Золотий вік (підгрупи: тексти 1493 – 1598 рр. ; тексти 1599 – 1713 рр.), який складає 28 % текстів;

– сучасна іспанська мова (підгрупи: тексти 1714 – 1812 рр. ; тексти 1813 – 1898 рр. ; тексти 1899 – 1936 рр. ; тексти 1937 – 1974 рр.), що охоплює 51 % даних усього корпусу.

Територіальний розподіл текстів Діахронного корпусу іспанської мови має такий вигляд: 74 % текстів походять з Іспанії, 25 % – з країн Латинської Америки і 1 % – іспанська мова сефарді та ін.

За формою та жанром серед текстів, представлених у CORDE, можна виокремити дві великі групи: художню (складається з поезії та прози у відсотковому співвідношенні 15 % до 85 % відповідно) та нехудожню літературу (сюди увійшли тексти наукові, дидактичні, суспільні, релігійні, історично-документальні, юридичні, рекламні, публіцистичні і т. д.).

Попри певні недоліки (як вже було зазначено, тексти Банку даних іспанської мови мають лише зовнішнє та структурне маркування, а також відсутня лематизація словоформ), Діахронний корпус іспанської мови залишається обов’язковим джерелом діахронно орієнтованих студій, стаючи все частіше в нагоді не лише дослідникам-іспаністам, але й тим, хто лише опановує іспанську мову.

Корпус сучасної іспанської мови (*El Corpus de Referencia del Español Actual, CREA*) [5] – корпус текстів, який охоплює широке коло письмових і усних текстів, створених у всіх іспаномовних країнах в період з 1975 р. по 2004 р. Як зазначають автори корпусу, він був розроблений таким чином, щоб “надавати вичерпну інформацію про мову у певний визначений період її історії та, відповідно, бути достатньо об’ємним, аби відображати всі найважливіші особливості цієї мови” [10].

Виконання поставленого завдання передбачає, насамперед, розв’язання двох питань: обсягу корпусу та його збалансованості. Що стосується об’єму описуваного КТ, він становить більше 160 млн. слововживань. Такий значний розмір ресурсу (зважаючи також на те, що він охоплює лише 30 років у розвитку мови) свідчить про справді широке коло відображеного матеріалу, який дозволить проводити найрізноманітніші дослідження сучасної іспанської мови.

Репрезентативності і збалансованості як першочергових вимог до КТ творці CREA досягли шляхом включення в його склад матеріалів, відібраних за часовим, географічним і тематичним чинниками з урахуванням походження текстів щодо форми побутування мови.

Часові рамки Корпусу сучасної іспанської мови спершу були обмежені 1975 – 1999 роками. Та у 2008 р. він був доповнений матеріалами наступних п’яти років (2000 – 2004 рр.), тож на сьогодні весь корпус поділений на шість періодів по п’ять років кожен: 1975 – 1979 рр. ; 1980 – 1984 рр. ; 1985 – 1989 рр. ; 1990 – 1994 рр. ; 1995 – 1999 рр. ; 2000 – 2004 рр.

Що стосується територіального чинника, тут варто зважати на той факт, що іспанська мова є державною не лише для Королівства Іспанія, але й ще для 21 країни світу. Таким чином, розподіл текстів у CREA (як в частині усних, так і писемних текстів) щодо країни їх походження має такий вигляд: 50 % усіх матеріалів становлять тексти, що бути створені в Іспанії, і 50 % припадає на тексти з решти країн.

Оскільки CREA є корпусом мішаного типу, тобто в ньому представлені тексти як усного, так і писемного мовлення, важливим фактором досягнення репрезентативності є співвідношення між цими двома підгрупами. Даний корпус складається з 90 % писемних та 10 % усних текстів (подібне співвідношення є досить типовим для національних корпусів мов, зокрема для класичного на сьогодні Британського національного корпусу).

Писемна частина CREA (*CREA Escrito*), у свою чергу, поділяється на книги (49 %), пресу (49 %) та інші види текстів, а саме буклети, рекламні проспекти, електронні листи, блоги та ін. (2 %). Усна частина корпусу (*CREA Oral*) складається з записів радіо- та телепередач (сюди увійшли записи новин, репортажів, інтерв’ю, дебатів, спортивних трансляцій та оглядів і т. п.), що становлять основу CREA Oral, та записів іншого типу (телефонні розмови, повідомлення на автовідповідачах, неофіційні діалоги тощо).

Підводячи підсумки, варто сказати, що попри певні недоліки, Банк даних іспанської мови має цілу низку особливостей та переваг, які роблять його надійним джерелом у лінгвістичних дослідженнях іспанської мови.

Корпус іспанської мови (El Corpus del Español) [6] – корпус текстів, створений у 2002 році Марком Дейвісом, професором Університету Бригама Янга (США). Цей КТ є дослідницьким за стратегією побудови й використання, загальномовним за предметною галуззю та одномовним за кількістю мов. Складається він з текстів без жодних скорочень, тобто належить до повнотекстових. Обсяг Корпусу іспанської мови становить понад 100 млн. слововживань із близько 14 тис. текстів, тобто ресурс є мегакорпусом. За хронологічним параметром описуваний КТ є діахронним (репрезентує розвиток іспанської мови з XIII по XX ст.). За типом реалізації мовної системи він є мішаним: до його складу увійшли тексти як писемного (95 % текстів усього корпусу), так і усного (5 % текстів) мовлення.

За рівнем кодування Корпус іспанської мови належить до лінгвістично анотованих ресурсів. Він містить структурне та лінгвістичне (морфологічне, синтаксичне і семантичне) маркування, завдяки якому можливо здійснювати різні типи пошуку. Зокрема, зазначений КТ дозволяє шукати, крім окремих слів чи фраз, ще й певні граматичні категорії, а також оточення слів, що є важливим для тих, то вивчає мову. Пошук можна проводити і за частотністю вживання слова, порівнюючи її з частотністю іншого тощо. Загалом, Корпус іспанської мови уможливлює проведення пошуку на семантичному, морфологічному та синтаксичному рівнях, демонструючи таким чином свої переваги над нелематизованими та не розміченими лінгвістично ресурсами.

Зміст Корпусу іспанської мови, як вже зазначалось, становлять тексти з усіх іспаномовних країн, створені з XIII ст. по XX ст. Розподіл ресурсу здійснено по століттях (а в межах останнього, XX ст., тексти розмежовані також за походженням) [6]:

СТОЛІТТЯ		КІЛЬКІСТЬ СЛОВОВЖИВАНЬ	ВІДСОТОК КОРПУСУ
XIII		6 905 000	6,9 %
XIV		2 820 000	2,8 %
XV		8 515 000	8,5 %
XVI		18 001 000	17,9 %
XVII		12 746 000	12,7 %
XVIII		10 263 000	10,3 %
XIX		20 465 000	20,5 %
XX	наукова література	5 138 077	5,1 %
	періодичні видання	5 144 631	5,1 %
	художня література	5 144 073	5,1 %
	усні тексти	5 113 249	5,1 %

Для ефективнішої та простішої роботи з зазначеним корпусом до нього на сайті додаються численні записи з порадами щодо здійснення кожного окремого типу пошуку, що значно спрощує ознайомлення з його роботою. Інтерфейс ресурсу зручний у використанні і теж сприяє швидкому отриманню потрібних даних із корпусу. Важливою також є можливість створення власних списків запитів для їх подальшого опрацювання (доступно після реєстрації).

Корпус іспанської мови є одним із двох національних корпусів цієї романської мови. Але проявивши свої численні переваги в порівнянні з другим, Банком даних іспанської мови, він, безперечно, стає незамінним при проведенні лінгвістичних досліджень всіх мовних рівнів, при вивченні мови та для студій з історії мови.

Література:

1. Демська О. Текстовий корпус : ідея іншої форми / О. М. Демська. – К. : ВПЦ НаУКМА, 2011. – 282 с.

2. Демська-Кульчицька О. Основи Національного корпусу української мови / О. М. Демська-Кульчицька. – К. : Наук. видання ІУМ НАН України, 2005. – 219 с.

3. Корпусна лінгвістика / [В. А. Широков, О. В. Букагов, Т. О. Грязнухіна та ін.]. – К. : Довіра, 2005. – 471 с.

4. Banco de datos del español. Manual de consulta [Електронний ресурс]. – Режим доступу : http://corpus.rae.es/ayuda_c.htm (дата візиту : 15. 12. 2011).

5. Corpus de Referencia del Español Actual (CREA) [Електронний ресурс]. – Режим доступу : http://corpus.rae.es/creanet.html (дата візиту : 15. 12. 2011).

6. Corpus del Español [Електронний ресурс]. – Режим доступу : http://www.corpusdelespanol.org (дата візиту: 28. 12. 2011).

7. Corpus Diacrónico del Español (CORDE) [Електронний ресурс]. – Режим доступу : http://corpus.rae.es/cordenet.html (дата візиту: 15. 12. 2011).

8. Davies M. Creating Useful Historical Corpora : A Comparison of CORDE, the Corpus del Español, and the Corpus do Português. En : Diacronía de las lenguas iberorromances : nuevas perspectivas desde la lingüística de corpus. Frankfurt / Madrid : Vervuert/Iberoamericana, 2010. – P. 139-168.

9. Davies M. Un corpus anotado de 100. 000. 000 palabras del español histórico y moderno. En : Proceedings of Sociedad Española para el Procesamiento del Lenguaje Natural. Valladolid, España, 2002. – P. 21-27.

10. El banco de datos [Електронний ресурс]. – Режим доступу : http://www.rae.es/rae/gestores/gespub000019.nsf/voTodosporId/DBC9D1B343D484B0C1257164003C8BFE?OpenDocument (дата візиту: 28. 12. 2011).

11. Sánchez Sánchez M., Domínguez Cintas C. El banco de datos de la Real Academia Española : CREA y CORDE. En : Per Abbat, 2 (2007). – P. 137-146.