Юлія ДЕМ'ЯНЧУК

ОСОБЛИВОСТІ БУДОВИ НАЦІОНАЛЬНОГО КОРПУСУ РОСІЙСЬКОЇ МОВИ: ЛЕКСИКО-ГРАМАТИЧНІ ТА СТИЛІСТИЧНІ ОСОБЛИВОСТІ КОРПУСУ

Науковий вісник Ужгородського університету. Серія: Філологія.
Випуск 1(35) 2016
УДК 811.161.1
Дем'янчук Ю. Особливості будови національного корпусу російської мови: лексико-граматичні та стилістичні особливості корпусу; 14 стор.; бібліографічних джерел – 12; мова – українська.

Анотація. У статті розглядається паралельний багатомовний національний корпус російської мови, його структура, можливість збереження інформації, конотаційні особливості. Здійснюється лінгвістичне порівняння НКРМ з іншими відомими корпусами та словниками (Словацьким національним корпусом,Американським національним корпусом (ANC), Британським національним корпусом (BNC), Українським національним лінгвістичним корпусом тощо). На основі аналізу стилістичних можливостей НКРМ пропонується застосування корпусу для багатомовного перекладу міжнародних офіційно-ділових документів.

Ключові слова: національний корпус, НКРМ, багатомовний корпус тексту, стилістичні особливості, офіційно-ділові документи.

Актуальність дослідження. З розвитком комп'ютерних технологій особливого значення в галузі корпусної лінгвістики почали відігравати паралельні багатомовні корпуси текстів в електронному форматі. Зокрема, практика створення сучасних корпусів передбачає розмітку даних на рівні слова (наприклад, розмітку лем, частин мови, граматичних ознак тощо), а також більших мовних одиниць (наприклад, розмітку синтаксичних груп, комунікативного членування пропозиції, стилістичних особливостей). Саме національний корпус російської мови (НКРМ) призначений для забезпечення наукових досліджень лексики і граматики мови, а також безперервних процесів лексичних змін. Проте найважливіше завдання корпусу – надання потрібних довідок щодо певної мовознавчої галузі (лексики, граматика, стилістики, акцентології, історії мови). Дослідженням корпусів займаються такі сучасні дослідники, як: Бобкова Т.В., Данилюк І.Н., Дарчук Н.П., Савчук С.О., Burnard L., McEnery T., Wittenburg P. та інші. Проте розробки стосуються загальних мовознавчих (маркувальних, лексико-граматичних, конотаційних) ознак корпусів, без врахування функціональних властивостей та можливостей перекладу тематичних текстів (наприклад, міжнародних офіційних документів). Водночас, відсутні наукові підтвердження про важливість застосування НКРМ у багатомовному вимірі, та перспективи створення нових текстових блоків, що стане актуальним для української корпусної лінгвістики.

Метою статті є дослідження структури НКРМ, його лексико-граматичних та стилістичних особливостей, що є важливим для перекладу спеціалізованих текстів. Концептуальні **завдання:** з'ясувати роль НКРМ у системі корпусної лінгвістики; здійснити аналіз будови НКРМ, його структури, розміток, факторів вирівнювання; порівняти НКРМ з іншими корпусами тексту, з метою визначення перспективи його застосування.

Виклад основного матеріалу, результати дослідження. Багатомовні паралельні корпуси стали важливою лінгвістичною базою сучасного перекладознавства. Зокрема, паралельний корпус (Parallel Corpora) – це електронний аналог паралельних перекладних текстів, який складається із декількох блоків «тексту-оригіналуі декількох його перекладів» (Плунгян В.А.) [5, с.13]. Електронні тексти в корпусі можуть бути цілісною мовною конструкцією або будь-якою його частиною (окремим блоком).

Найвідомішими міжнародними багатомовними корпусами є: 1. EUROPARL (автор Philipp Koehn, наявно 20 млн. слововживань, відкритий корпус Європарламенту на 11 мовах); 2. GERMAN-ENGLISH TRANSLATION CORPUS (один мільйон слововживань, тексти – академічні, політичні, туристичні); 3. KACENKA «Korpus anglicko-česky; Czech» (три мільйони слововживань); 4. OPUS «an open source parallel corpus» (інтернет-корпус, наявні п'ять мов, можна вирівняти і розмітити корпус, додати лінгвістичну інформацію); 5.Lancaster's ITU (вміщує англійську, французьку та іспанську мови); 6. Англо-норвезький паралельний корпус, англо-китайський паралельний корпус HKUST; 7. Східний багатомовний корпус (наявні болгарська, чеська, естонська, угорська, румунська та словенська мови); 8. Корпус «Agenda 21» (наявні датська, англійська, французька та німецька мови); 9. INTERSECT (англо-німецький паралельний корпус) [4, с. 157].

Зазвичай, кількість слів корпусу складає 100 млн. слів. Загальновизнаним зразком є, зокрема, Британський національний корпус (BNC), на який орієнтовано багато інших сучасних корпусів. Також, вдосконалюється Американський національний корпус (ANC). Серед корпусів

слов'янських мов виділяється: Чеський національний корпус, створений в Карловому університеті Праги.; Угорський національний корпус нараховує 180 млн. слів. В Україні сформовано один український національний лінгвістичний корпус (42 мільйони слововживань). Цю роботу виконав Український мовно-інформаційний фонд – науково-дослідний інститут Національної академії наук України.

Як наголошують Киркунова Л. Г. та Ширінкіна М. А., визначення параметрів збалансованості текстів – одне з найскладніших завдань, яке вирішується застосуванням ряду методик, в тому числі з опорою на результати соціолінгвістичного анкетування [1, с. 32]. Наявні дані про співвідношення типів текстів в корпусах, які визначаються як збалансовані, свідчать про те, що «літературоцентризм» зменшується за рахунок збільшення кількості публіцистики в його складі; художні тексти при цьому займають друге місце в загальному обсязі корпусу, третє – так звані спеціалізовані тексти із внутрішньою додатковою спеціалізацією. Наприклад, Словацький національний корпус, що належить до збалансованих корпусів, вміщує тексти у такій відсотковій пропорції: публіцистика (60,6%), художня література (17,5%), спеціалізовані тексти (11,6%), інше (10, 3%). Два варіанти Словенського національного корпусу БШЛ і FidaPLUS вміщують тексти в таких пропорціях: художні тексти (63,47%), наукові (10%), інші (86,34%); книги (8,74%), газети (65,26%), журнали (23,26%), тексти з Інтернету (електронні тексти) (1,24%), інше (в тому числі незначна частка усного мовлення – стенограми парламентських слухань) (1,5%) .

Натомість, в Українському національному лінгвістичному корпусі, який не належить до збалансованих, можна визначити такий відсотковий баланс: художня література (43%), есеїстика (29,6%), публіцистика (16,9%), книги для дітей (6 %), релігійні тексти (2,8%), юридичні тексти (1,5%), фольклор (0,2%).

Проте для перекладу галузевих міжнародних текстів, актуальним залишається Національний корпус російської мови. НКРМ вміщує паралельні корпуси, в яких можна знайти переклади для певного слова або словосполучення на російську мову або з російської мови та іншу. В даний час для пошуку доступні: англо-російський, російсько-англійський, німецько-російський, українсько-російський, російсько-український, українсько-російський, російсько-білоруський і багатомовні паралельні корпуси.

Можна погодитися з Левінзон А.І., що діючі системи машинного перекладу орієнтовані на конкретні пари мов і використовують, як правило, перекладні відповідності або на поверхневому рівні, або на проміжному рівні між вхідною та вихідною мовою [3, с. 132]. Якість машинного перекладу залежить від обсягу словника, обсягу інформації, що додаються до лексичних одиниць. Сучасні апаратні і програмні засоби дають можливість застосовувати словники великого обсягу, що вміщують детальну граматичну інформацію. Інформація може бути представлена як в декларативній (описовій), так і в процедурній (враховує потреби алгоритму) формі.

У даному випадку, національний корпус російської мови – це великий, збалансований за складом електронний корпус текстів; ядром НКРМ є російськомовні тексти. Також в НКРМ входить паралельний корпус, який складається з багатомовної частини.

Підрозділами НКРМ є: основний корпус, синтаксичний корпус, газетний, паралельний (офіційно-ділові, юридичні, правові блоки), навчальний, діалектний, поетичний, усний, акцентологічний, мультимедійний і історичний корпуси [9]. Прямий пошук в НКРМ дає можливість точної вибірки. Більш складний і спеціалізований лексико-граматичний пошук в корпусі здійснюється за граматичними, семантичним і додатковим (зокрема, розділових знаків) рівнями. Доступний пошук за кількома словами з можливістю задати відстань між ними. Створення свого підкорпусу для пошуку передбачає звуження метатекстових ознак (автор і назва тексту, час створення тексту, жанрові характеристики тощо).

Словотвірна розмітка в НКРМ розглядається в двох варіантах, перший з яких – реалізація в складі семантичної розмітки; визначення параметрів словотвірної розмітки в цьому випадку проводиться вибором у формі «лексико-граматичний пошук» вікна «семантичні ознаки» і далі – вибором параметрів групи «словотвір», доступних в даному вікні. У цьому виді розмітки набір словотворчих параметрів відповідає наступним типам характеристик: морфологічно-семантичні словотвірні ознаки; розряд, який створює слова; лексико-семантичний (таксономічний) тип, що створює слова; звичайний морфологічний тип словотворення [10]. Даний варіант словотворчої розмітки доступний лише в семантично розмічених корпусах НКРМ: основному, газетному, паралельному, поетичному, усному, акцентологічному, мультимедійному.

Опціями, що забезпечують паралельні багатомовні корпуси НКРМ є: WebCorp, Word Filter, IntelliText [11; 12].

«WebCorp» працює над обраною інформаційно-пошуковою системою, обробляючи список повернутий нею URL, виймаючи зі знайдених сторінок рядки конкордансу за запитом.За допомогою оператора можна здійснити одночасний пошук за кількома словами. Квадратні дужки використовуються для згрупування елементів запиту. Опція «Word Filter» дає змогу приєднати додаткові слова, які повинні або не повинні з'являтися в лініях конкордансу, які зберігаються за пошуковим запитом. В полі «Site» можна визначити галузь пошуку через набір доменних зон або фрагментів URL.

Також можна вказати домени, які не повинні бути включені в результати пошуку, написавши їх зі знаком мінус.

У «WebCorp» є функції обробки результатів. Коли пошук завершився, на сторінці результатів надається можливість аналізувати колокації пошукового терміна, тобто слів, які найчастіше з'являються в його оточенні. Також можливе групування колокацій по алфавіту та за часовими ознаками. Є дві можливості сортування за часом: можна вибрати період часу з меню, що випадає (у минулому місяці, протягом останніх трьох місяців, протягом останніх шести місяців, в минулому році, більше одного, двох чи п'яти років). Функція «IntelliText» має спеціальну функцію «Affixes», що дає змогу здійснювати пошук префіксів або суфіксів. Якщо необхідно знайти префіксоїд, то використовується пошук по префіксах.

Вартий уваги сучасний підкорпус НКРМ «Економіка, бізнес, фінанси» (як приклад для перспективних розробок щодо юридично-правових підкорпусів), який ґрунтується на матеріалах ЗМІ, відображає щоденні стрімкі зміни в термінології по темі. Даний підкорпус може функціонувати як загальний словник, а також словник автоматичної системи перекладу. Економічний розділ паралельних текстів представлений в НКРМ лише в російсько-англійської версії, проте розробляється російсько-німецький, а також російсько-український та українсько-російський корпус.

Порівнюючи НКРМ, дослідник А. Мустайокі констатує, що національний корпус характеризується показністю, або збалансованим складом текстів [4, с. 158]. Це означає, що корпус містить по можливості всі типи письмових і усних текстів, представлених в даній мові (художні твори різних жанрів, газетні та журнальні статті різної тематики, рекламу, спеціальні тексти, щоденники, переписку), і що всі ці тексти входять в корпус по можливості пропорційно їх частці у мові відповідного періоду.

Укладачі НКРМ диференціюють тексти корпусу в такий спосіб: сучасна художня проза різних жанрів і напрямків, сучасна драматургія, мемуарно-біографічна література, журнальна публіцистика і літературна критика, газетна публіцистика і новини, наукові, науково-популярні та навчальні тексти, релігійні та релігійно-філософські тексти, виробничо-технічні тексти, офіційно-ділові та юридичні тексти, побутові тексти (в тому числі тексти, не призначені для публікації: особисте листування, щоденники тощо). Водночас, тексти НКРМ представлені в певній пропорції, що відбиває їх частку в загальному масиві сучасних текстів. Так, частка художніх текстів (включаючи драматургію і мемуари) становить не більше 40% і як підкреслює Плунгян В.А., всі ці тексти «входять в корпус по можливості пропорційно їх частці в мові відповідного періоду» [8]. Загалом, дані корпусу репрезентативно представляють письмові тексти, включаючи транскрипти усного мовлення, що відносяться тільки до інституціонального спілкування, до публічних жанрів усної офіційної комунікації. Усна комунікація може бути включена до складу НКРМ в статусі самостійного підкорпусу.

Натомість, лінгвістичні дослідження, що базуються на матеріалі корпусів і зіставленні отриманих даних з даними Національного корпусу російської мови, дають можливість з'ясувати природу помилок і сфери формування нових тенденцій, пов'язаних із розвитком лексико-граматичної системи сучасної російської мови, її зв'язків з іншими мовами (українською, англійською, німецькою).

В НКРМ здійснюється оцінка максимальної довжини N-грамів з точки зору інтересів користувачів і продуктивності пошукової системи. Прораховуються заходи стійкості колокацій, абсолютна частота входжень, кількість документів, в яких зустрілася одиниця. Типологія пропонованої розмітки включає лематизації, частини мови, граматичну розмітку, розмітку додаткових параметрів (наявність пунктуації, капіталізацію). Користувач отримує попередній аналіз видачі по корпусу (кластеризацію контекстів), оцінки стійкості колокацій, оцінки ймовірності появи мовних одиниць (леми, частини мови, форми певного відмінка) в найближчому контексті. Функціонал включає сортування за статистичними заходам, вивантаження даних он-лайн і перехід в НКРМ (видачу прикладів, які відповідають обраним критеріям). Ресурс забезпечує розвиток квантитативних корпусних досліджень і стає базою для фундаментальних досліджень в галузі російської граматики. Також, у 2010 році в складі Національного корпусу російської мови був відкритий пілотний варіант Мультимедійного російського корпусу (Мурка).

Через наявність у більшості корпусів алгоритмічних похибок, необхідно застосовувати практику ручної перевірки результатів. Для оптимізації і прискорення процесу ручної перевірки результатів у НКРМ створена спеціальна комп'ютерна програма – робоче місце постредактора. Загальний список лем розбитий на рівні частини, кожна з яких перевіряється окремо різними учасниками проекту. Після первинної перевірки окремі відредаговані частини збираються в єдиний масив для повторної перевірки на предмет однозначності прийнятих щодо спірних випадків рішень. Програма оновлюється кожні три роки, натомість, кожен рік дослідник в галузі корпусної лінгвістики, аналізуючи списки лем, вирішують складні теоретичні і практичні завдання. Наприклад, на низькому рівні перебуває розробка автоматизації розмітки суфіксів і кореневих частин. До того ж, одним з раніше актуальних напрямків роботи є удосконалення програми по відділенню окремих тем (наприклад, юридичних блоків, мови

офіційних документів ЄС тощо) та стилістичної приналежності.

Висновки. Перспективи розвитку НКРМ та інших національних корпусів, пов'язані з подальшою розробкою і поглибленням теорії і практики перекладу. Для розвитку теорії важливі результати зіставного мовознавства, загальної теорії перекладу, корпусних розробок, оптимізації і вдосконалення лінгвістичних алгоритмів. Нові та більш ефективні корпуси, які б опрацьовували тематичні офіційно-ділові документи з необхідною словниковою інформацією, термінологізацією лексики, допоможуть підвищити якість перекладу лексичних одиниць. Формальні граматики, орієнтовані на переклад, дадуть можливість оптимізувати алгоритми перекладацьких відповідників офіційно-ділових текстів. Водночас, нові можливості програмування також будуть корисними для вдосконалення і подальшого розвитку додаткових паралельних блоків Національного корпусу російської мови.

Література

1. Кыркунова Л. Г., Ширинкина М. А. Использование НКРЯ в преподавании речеведческих дисциплин в вузе // Национальный корпус русского языка и проблемы гуманитарного образования: материалы международной научной конференции. – М.: ГУ-ВШЭ, 2007., №12. – С.28-34.
2. Компьютерная лингвистика: научное направление и учебная дисциплина : сборник научных статей. Вып. 2 / В. И. Коваль (ответств. ред.) [и др.]; М-во образования РБ, Гомельский гос. ун-т им. Ф. Скорины. – Гомель: ГГУ им. Ф. Скорины, 2012. – 136 с.
3. Левинзон А. И. Использование НКРЯ в преподавании русского языка иностранным студентам, специализирующимся в области экономики и финансов // Национальный корпус русского языка и проблемы гуманитарного образования: материалы международной научной конференции. – М.: ГУ-ВШЭ, 2007., №12. – С.127-136.
4. Мустайоки А. Роль корпусов в лингвистических исследованиях языков // Национальный корпус русского языка и проблемы гуманитарного образования: материалы международной научной конференции. – М.: ГУ-ВШЭ, 2007., №12. – С.152-166.
5. Плунгян В. А. Корпус как инструмент и как идеология // Национальный корпус русского языка и проблемы гуманитарного образования: материалы международной научной конференции. М.: ГУ-ВШЭ, 19-20 апреля 2007. – С.11-21.
6. Плисецкая А. Д. Национальный корпус русского языка как один из инструментов анализа фразеологических сочетаний // В кн.: Корпусная лингвистика – 2013: Труды международной научной конференции. – СПб. : Санкт-Петербургский государственный университет, 2013. – С. 387-396.
7. Daniel M. The Second Genitive in Russian, in: Partitive cases and related categories. – Berlin, NY : Mouton de Gruyter, 2014. Ch. 9. – P. 347-377.
8. Плунгян В. А. Зачем нужен Национальный корпус русского языка: неформальное введение [Электронный ресурс] //Национальный корпус русского языка: 2003-2005. – М.: Индрик, 2005. – Режим доступу: http: // ruscorpora.ru / sbornik2005 / 02plu.pdf
9. Национальный корпус русского языка [Електронний ресурс] / Национальный корпус русского языка; Яндекс. – Режим доступу: http://www.ruscorpora.ru.
10. Тагабилева, М.Г. Словообразовательная разметка Национального Корпуса русского языка: задачи и методы [Электронный ресурс] / М. Г. Тагабилева, Ю. Н. Березуцкая. – Режим доступу: http://www.dialog-21.ru/digests/dialog2010/materials/pdf/73.pdf.
11. WebCorp [Electronic resource] / Research and Development Unit for English Studies, Birmingham City University. – Mode of access: http://www.webcorp.org.uk/live/.
12. IntelliText [Electronic resource] / University of Leeds: Centre for Translation Studies (CTS). – Mode of access: http://corpus.leeds.ac.uk/itweb/htdocs/Query.html.

Юлия Демьянчук

ОСОБЕННОСТИ УСТРОЙСТВА НАЦИОНАЛЬНОГО КОРПУСА РУССКОГО ЯЗЫКА: ЛЕКСИКО-ГРАММАТИЧЕСКИЕ И СТИЛИСТИЧЕСКИЕ ОСОБЕННОСТИ КОРПУСА

Аннотация. В статье рассматривается параллельный многоязычный национальный корпус русского языка, его структура, возможность сохранения информации, коннотационные особенности. Осуществляется лингвистическое сравнение НКРМ с другими известными корпусами и словарями (Словацким национальным корпусом, Американским национальным корпусом (ANC), Британским национальным корпусом (BNC), Украинским национальным лингвистическим корпусом и т.п.). Автор выясняет роль НКРМ в системе корпусной лингвистики; осуществляет анализ строения НКРМ, его структуры, разметок, факторов выравнивания; сравнивает НКРМ с другими корпусами текста, с целью определения

перспективы его применения. Однако для перевода отраслевых международных текстов актуальным остается Национальный корпус русского языка. На основе анализа стилистических возможностей НКРМ предлагается применение корпуса для многоязычного перевода международных официально-деловых документов.

Ключевые слова*:* национальный корпус, НКРМ, многоязычный корпус текста, стилистические особенности, официально-деловые документы.

Demyanchuk Yuliya

THE STRUCTURAL FEATURES OF THE NATIONAL CORPUS OF THE RUSSIAN LANGUAGE: A LEXICO-GRAMMATICAL AND STYLISTIC PECULIARITIES

Annotation. The article deals with a parallel multilingual National Corpus of the Russian language, its structure, the ability to store information, the connotation features. It is carried out linguistic NKRM comparison with other well-known buildings and dictionaries (Slovak National Corpus, the American national body (the ANC), the British National Corpus (BNC), Ukrainian National Linguistic housing, etc.). Author finds NKRM role in the system of corpus linguistics; will fulfill NKRM structure analysis, its structure, layouts, alignment factors; NKRM compares with other bodies of text, in order to determine the prospects of its application. However, for the translation industry international texts, remains urgent Russian National Corpus. NKRM compared the author of the international multilingual buildings: EUROPARL, GERMAN-ENGLISH TRANSLATION CORPUS, KACENKA, OPUS, Lancaster's ITU, Anglo-Norwegian parallel body, Anglo-Chinese parallel body HKUST, East multi-hull, the British national body (BNC), the US National Corps (ANC), the Slovak national body, Ukrainian national linguistic body.Based on the analysis of stylistic features NKRM, provided the use of the body for multilingual translation of international official and business documents. However, the new programming will also be useful for the improvement and further development of additional parallel unit of the National Corps of Russian.

Key words: national housing NKRM multilingual body of the text, stylistic features, official-business documents.

Стаття надійшла до редакції 10.06.2016 р.

Дем'янчук Юлія Ігорівна – кандидат економічних наук, викладач кафедри технічного перекладу Львівського державного університету безпеки життєдіяльності.