Оксана Тищенко-Монастирська
Марія Шведова
Дмитро Січінава

Паралельні українсько-російський та російсько-український корпуси[1]

Паралельний корпус (тобто електронний анотований корпус, до якого крім оригінальних текстів залучено їхні переклади тією чи іншою мовою, які вирівняні за реченнями або абзацами) – новий напрямок корпусної лінгвістики, що стрімко поширюється нині в Європі та світі. Можливості застосування паралельних корпусів різноманітні. Вони, зокрема, використовуються в перекладній лексикографії (укладання загальномовних та термінологічних словників, добір ілюстративного матеріалу), порівняльних лексичних та граматичних дослідженнях (контрастивна лінгвістика), вивченні теорії та практики перекладу, розробці систем автоматичного перекладу, викладанні мови.

Особливу роль у європейській корпусній лінгвістиці відіграє створення паралельних корпусів кількох слов'янських мов. Окрім проекту паралельних українсько-російських та російсько-українських корпусів, пов'язаного з Національним корпусом російської мови (за участю НКРМ розвиваються також білорусько-російський та польсько-російський паралельні корпуси), динамічно розгортаються такі великі проекти, як ASPAC (Амстердамський слов'янський паралельний корпус А.Барентсена, http://home.medewerker.uva.nl/a.a.barentsen/), ParaSOL («Паралельний корпус слов'янських та інших мов», Берн–Регенсбург, [Waldenfels 2006]), чеський InterCorp [Vavřín, Rosen 2009]. Колективи розробників регулярно представляють паралельні слов'янські корпуси на конференціях з корпусної лінгвістики, де обговорюються питання обміну матеріалом та досвідом. З 2010 р. корпусна комісія міжнародного комітету славістів організовує конференції Slavicorp, а у 2012 р. на Міжнародному з'їзді славістів у Мінську слов'янським паралельним корпусам буде присвячено спеціальний круглий стіл.

Проект з розвитку паралельних українсько-російських та російсько-українських корпусів розробляється з 2010 р. спільним колективом українських і

[1]Статтю написано за підтримки програми Президії РАН «Корпусна лінгвістика».

УДК 811.2'374.72'22

російських учасників, які представляють різні науково-дослідні організації (Інститут української мови та Інститут мовознавства НАНУ, Київський лінгвістичний університет, Інститут російської мови та Інститут мовознавства РАН). І україністи, й русисти мають уже достатній досвід з теоретичної та практичної розробки одномовних корпусів. Для російської мови це передусім традиція, що іде від проекту Машинного фонду російської мови до Національного корпусу (http://ruscorpora.ru; див. на цьому сайті список публікацій), для української – Мовно-інформаційний фонд НАНУ, корпус (точніше, низка корпусів) лабораторії комп'ютерної лінгвістики КНУ (http://mova.info), проект Національного корпусу [Демська-Кульчицька 2005]. Є також досвід створення паралельних корпусів: крім уже згаданих багатомовних, що містять російську й українську мови, існують проекти польсько-українського корпусу [Kotsyba 2010], українсько-російського корпусу новин ElVisti, що створюється автоматичним шляхом [Ландэ, Жигало 2010], російсько-іншомовних корпусів (не лише слов'янських) у складі Національного корпусу російської мови [Шведова, Сичинава 2010]. Досвід розробки усіх цих проектів всебічно враховується в підготовці українсько-російського та російсько-українського корпусів.

Паралельний українсько-російський корпус відкритий для пошуку у відповідному розділі сайту НКРМ і станом на липень 2011 р. містить 131 текст обсягом 2 млн словоформ (за кількістю текстів це найбільший паралельний корпус серед тих, що були розроблені за участю НКРМ). Російсько-український корпус поки що складається з 25 текстів загальним обсягом 1 млн словоформ. Розширення його обсягу триває. Для паралельного корпусу цей параметр є не менш значущим, ніж для одномовного: чим більший за обсягом корпус, тим вища надійність невипадковості повторення тих чи інших моделей перекладу.

Обидва корпуси достатньо репрезентативні з хронологічного погляду: вони охоплюють період від творчості засновників літературних мов – Котляревського й Пушкіна – до сьогодення. Цікавою теоретичною проблемою є той факт, що під час розробки паралельних корпусів завдання жанрової репрезентативності раніше практично не ставилося. У більшості паралельних корпусів зібрано тексти, що представляють лише один жанр; зазвичай це тільки художні тексти, іноді – тільки офіційно-ділові (наприклад, у корпусі документів Європейського союзу, http://corpus.leeds.ac.uk/paraquery.html). Звісно, це пояснюється об'єктивними чинниками функціонування перекладних текстів: тиражні тексти, які часто перевидаються (особливо художні), перекладають частіше, ніж, скажімо, газетну публіцистику або тим більше приватне листування. Проте до українсько-російського та російсько-українського кор-

пусів залучено (або планується залучити) також наукові тексти, публіцистику, навіть особисті листи письменників, які увійшли до перекладних видань їхніх творів. До корпусу також залучено перекладений російською мовою український фольклор (казки та легенди). Прецедентів включення до паралельних корпусів такого матеріалу раніше не було. Ми намагаємося якомога репрезентативніше відобразити жанри та різновиди текстів, які були перекладені в різні часи з української мови на російську та навпаки.

Переклади з української на російську та з російської на українську становлять значний інтерес для вивчення обох мов. Адже йдеться про близькоспоріднені мови, різниця у лексичній та граматичній будові між якими часто є досить неочевидною і складною. Слід також враховувати, що в описах та словниках радянського часу норми цих мов з ідеологічних міркувань нерідко штучно зближувалися. Корпусне дослідження (рівносильне повному «розписуванню» величезного обсягу текстів на картки в традиційній лексикографії) надає можливість виявити всю повноту і складність картини міжмовних відповідностей у кожному конкретному випадку. Перекладні тексти тут становлять не менш цінний матеріал, ніж оригінальні. Українська школа художнього перекладу завжди цінувалася з огляду на мовне та стилістичне багатство перекладених текстів, мова українських перекладачів уже аналізувалася з лінгвістичного, зокрема з лексикографічного, погляду [Скопненко, Цимбалюк 2003]. Перекладачами виступали й відомі українські та російські письменники. Серед перекладачів з російської на українську – Леся Українка, Максим Рильський та Борис Антоненко-Давидович (крім того, деякі письменники самі перекладали свої твори російською мовою), серед перекладачів з української на російську – Корній Чуковський, Павло Антокольський і Всеволод Рождественський. Навіть у пересічних перекладах радянського часу, в яких необхідно зважати на цензурні та редакційні зміни тексту з ідеологічних причин, а також у сучасних масових перекладах публіцистики та наукових текстів відображаються динамічні зміни мови у XX столітті, не кажучи вже про історичну та культурологічну інформацію.

Тексти корпусу напівавтоматично, за реченнями, вирівнювали М. О. Шведова, О. О. Тищенко-Монастирська і Г. Г. Кривенко за допомогою безкоштовної програми LeoBilingua, доступної за адресою www.hot.ee/b/bclogic/. Використовувалися також уже вирівняні, люб'язно надані А. Барентсеном та Р. фон Вальденфельсом тексти з проектів ASPAC і ParaSOL (їхня розмітка при цьому дещо коригувалася, зокрема вирівнювання за абзацами у тих випадках, де це можливо зробити автоматично, було замінене на вирівнювання за реченнями). Надалі планується використовувати власну розробку (оболонку

до програми HunAlign, яка застосовується, зокрема, у корпусі ParaSOL). Вирівняні речення представлено у форматі XML. Всі тексти отримують автоматичну морфологічну розмітку (із незнятою омонімією; планується також зняття омонімії в частині текстів). Для пошуку доступні будь-які сполучення словоформ, лексем та граматичних характеристик; можливий також пошук з урахуванням розділових знаків. В інтернет-інтерфейсі корпусу можна задати хронологічний і жанровий підкорпус текстів, сортування пошукової видачі.

Нагальне для цілої низки потреб (зокрема лексикографічних) завдання створення великих загальнодоступних корпусів української мови вирішене ще не повністю, і розробка паралельних українсько-російських та російсько-українських корпусів є лише одним із кроків у цьому напрямку.

Література

1. Демська-Кульчицька О. Основи Національного корпусу української мови. – К., 2005.

2. Ландэ Д. В., Жигало В. В. О создании параллельного двуязычного корпуса веб-публикаций [Електронний ресурс] // http://infostream.ua/ling/ml-small-end.pdf

3. Скопненко О., Цимбалюк Т. Фразеологія перекладів Миколи Лукаша, Словник-довідник. – К., Довіра, 2003. – 735 с.

4. Сичинава Д. В., Шведова М. А. Параллельные корпуса в составе национального корпуса русского языка: технологии и решаемые задачи // Компьютерная лингвистика: научное направление и учебная дисциплина: сборник научных статей. Вып. 1 / Отв. ред. В. И. Коваль. – Гомель: ГГУ им. Ф. Скорины, 2010. – С. 31–35.

5. Kotsyba N. PolUKR (a Polish-Ukrainian parallel corpus) as a testbed for a parallel corpora toolbox. (submitted for publication to the proceedings of the international conference «SlaviCorp», 22–24 November 2010, Warsaw [Електронний ресурс] // http://domeczek.pl/~natko/papers/NKotsyba_SlaviCorp2010_paper.pdf.

6. Vavřín M., Rosen A. Korpus InterCorp [Електронний ресурс] // http://korpus.cz/intercorp-info.php.

7. Waldenfels R. Compiling a parallel corpus of slavic languages. Text strategies, tools and the question of lemmatization in alignment. In: Brehmer, B., Zdanova, V., Zimny, R. (Hrsg.); Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV) 9. München, 123-138, 2006 [Електронний ресурс] // (http://www-nw.uni- regensburg.de/%7E.war05297.slavistik. sprachlit.uni-regensburg.de/pub/ WaldenfelsParallel Corpora2006. pdf.