УДК 81’25(651.926)

Ю. І. Дем’янчук

Методика створення паралельних багатомовних корпусів для перекладу юридичних та офіційних документів

У статті пропонується розробка багатомовного паралельного корпусу (англо-російсько-українського) для перекладу юридичних та ділових текстів. На основі аналізу морфемних змін здійснюється вибірка та вибір кращого словникового варіанта. Запропонований паралельний корпус (на основі програм CRATER, контенту InfoStream та інших) визначається як вагомий додаток до Національного корпусу російської мови (НКРМ).
Ключові слова: багатомовний корпус, офіційні документи, алгоритм, вибірка, електронний словник, НКРМ.

В статье предлагается разработка многоязычного параллельного корпуса (англо-русско-украинского) для перевода юридических и деловых текстов. На основе анализа морфемных изменений, осуществляется выборка и выбор лучшего словарного варианта. Предложенный параллельный корпус (на основе программ CRATER, контента InfoStream и других), определяется как весомый дополнение к Национальному корпусу русского языка (НКРМ).
Ключевые слова: многоязычный корпус, официальные документы, алгоритм, выборка, электронный словарь, НКРМ.

The article proposes the development of a multilingual parallel corpus (English-Russian-Ukrainian) for translation of legal and business texts. On the basis of morphemic and algorithmic changes, carried out sampling and selection of the best variant of the vocabulary. The proposed parallel body (based on the CRATER program content InfoStream and others), defined as a significant addition to the Russian National Corpus (NKRM).
Key words: multilingual body, official documents, the algorithm, sampling, electronic dictionary, NKRM.

У сучасній корпусній лінгвістиці особливої актуальності набуває створення додаткових паралельних корпусів для швидкого перекладу міжнародних офіційних документів (наприклад, архівів та доповідей НАТО). Складність офіційно-ділового перекладу уповільнює розробку алгоритмів та програм для національних корпусів (зокрема, багатомовного Національного корпусу російської мови). Порівняно з

художніми текстами тексти офіційних документів є глибоко стандартизованими, це стосується як структури всього тексту, так і організації окремих параграфів. Основною рисою мови міжнародної ділової кореспонденції є збереження структурних форм і використання певних синтаксичних конструкцій (заголовка, дати, вступного звернення, основного тексту, заключного слова і підпису), все це регулюється як лексично, так і синтаксично. Саме повнота інформації, точність і лаконічність формулювань, відсутність емоційності, використання нейтрального тону, безособовість лексичних конструкцій офіційного стилю створюють додаткове навантаження для лінгвістів-розробників, адже вимагає високого рівня професійності. Паралельний багатомовний корпус як блокову частину головних національних корпусів досліджували Д. Добровольський [2], Д. Сичинава [4], В. Широков [6], Л. Цінман [5] та інші. Проте проблема перекладу міжнародних офіційних документів у додаткових паралельних компонентах національного корпусу залишається нерозглянутою.

Мета дослідження – здійснити розробку та аналіз загальних компонентів паралельного корпусу для НКРМ на прикладі додатку InfoStream. Концептуальні завдання: розглянути систему розробки паралельного багатомовного корпусу та методи його застосування; висвітлити систему пошуку офіційно-ділових текстових блоків за ключовим словом в паралельному корпусі НКРМ; запропонувати застосування онлайн-сервісу InfoStream як якісного двомовного паралельного (російсько-українського) додатка до НКРМ.

Для перекладу міжнародних ділових документів результативним можемо вважати багатомовний Національний корпус російської мови (НКРМ). Зокрема, розробка паралельного корпусу російсько-англійських і російсько-німецьких текстів на базі технологій Національного корпусу російської мови тривала до 2015 року, до сьогодні база оновлюється. Наприклад, російсько-британська частина корпусу двомовна із можливістю переходу на іншу мову.

Українсько-російський і російсько-український компоненти паралельного корпусу досягає десяти мільйонів слововживань і перевершує найбільший Національний лінгвістичний корпус української мови.

Вирівнювання текстів англійською та російською, а також німецькою та російською мовами програмними засобами з подальшим редагуванням є технологічно складним. Зокрема, для корпусу були розроблені спеціальні програмні засоби вирівнювання та створено систему управління корпусом для полегшення роботи із корпусом

(зокрема, програма ПАРТЕКС, що має на вході два паралельні тексти (оригінал і переклад)).

Крім того, в різних мовах (а іноді й у різних виданнях) реалізовуються різні способи графічного оформлення, що іноді ускладнює визначення меж пропозиції в автоматичному режимі. Наприклад, існують різні способи оформлення переходів від прямої мови персонажів до авторських ремарок. У таких випадках результати автоматичного вирівнювання потребують корекції, що здійснюється вручну.

Нами була виявлена проблема пошуку українських та російських офіційно-ділових термінів (наприклад, мовних конструкцій з документів НАТО). Даний пошук можливо реалізовувати лише англійською мовою у паралельному корпусі, далі вирівнювання та пошук алгоритмів має здійснюватися вручну (рис. 1, 2) [8].

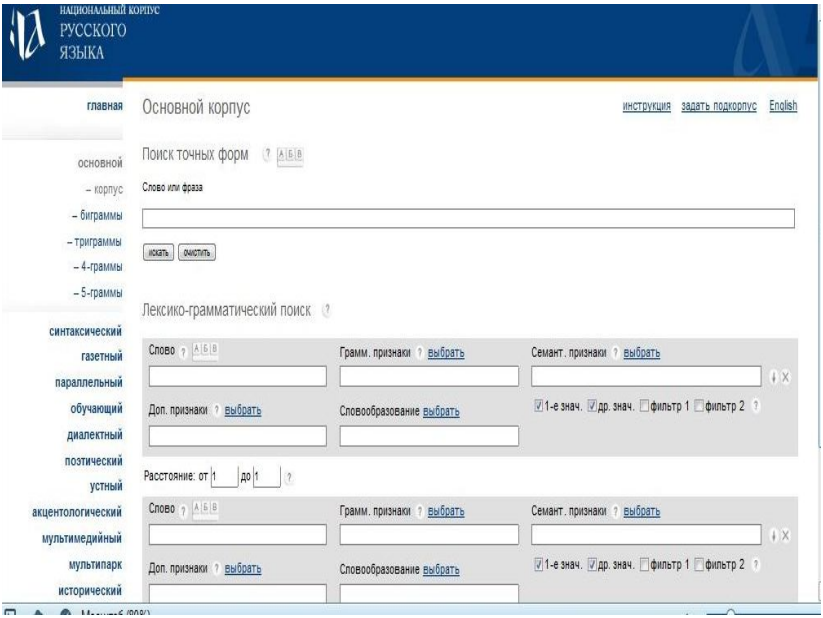

Рис. 1. ***Пошукове вікно паралельного корпусу текстів НКРМ***

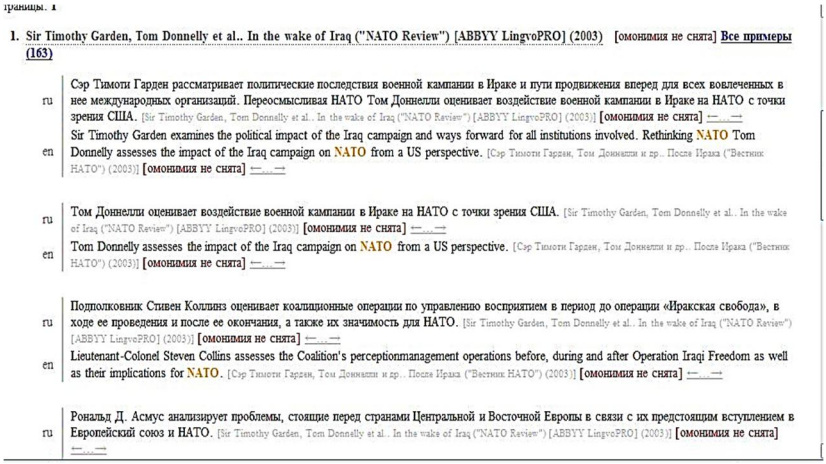

Рис. 2. ***Приклад пошуку корпусів офіційних документів за ключовими словами-термінами (НКРМ)***

Для вирівнювання знайдених корпусів застосовується інтерфейс користувача (GUI) та програма вирівнювання текстів HunAlign. Для пошуку офіційної термінології передбачені наступні автоматизовані операції (у вигляді графічних кнопок: 1) додавання порожнього рядка; 2) видалення порожнього рядка; 3) перенесення пропозиції в сусідній рядок; 4) перенесення частини в сусідній рядок; 5) додавання вручну вирізаних частин до пропозиції.

При збереженні вирівняного тексту вказується мова оригіналу і перекладу. В даному разі мовою оригіналу можуть бути лише англомовні блоки, мова перекладу – російська. Тому виникають труднощі концептуального перекладу з базової російської на українську окремих текстових блоків.

Пропонується метод, за допомогою якого реалізується виявлення інформаційних дублікатів, поданих різними мовами (російською та українською), як додаток до НКРМ, а також до Національного корпусу української мови. Програма CRATER та контент InfoStream оснащені деякими автоматично сформованими тегами і перекладами виділених лексем на двох мовах [7]. Вирівнювання даного корпусу за пропозиціями або словами, а також морфологічна розмітка корпусу має перспективи подальшого розвитку.

На основі контенту InfoStream створення паралельних корпусів офіційних документів можна розділити на дві групи [9]: традиційні і

статистичні. Актуальним у цьому разі є підхід до створення паралельних корпусів документів, заснований на алгоритмі пошуку дублікатів документів різними мовами. Підхід дає можливість відшукати схожі документи різними мовами у великому масиві документів. В результаті можна переконатися в тому, що до корпусу потрапили паралельні документи з різних джерел (рис. 3, 4).

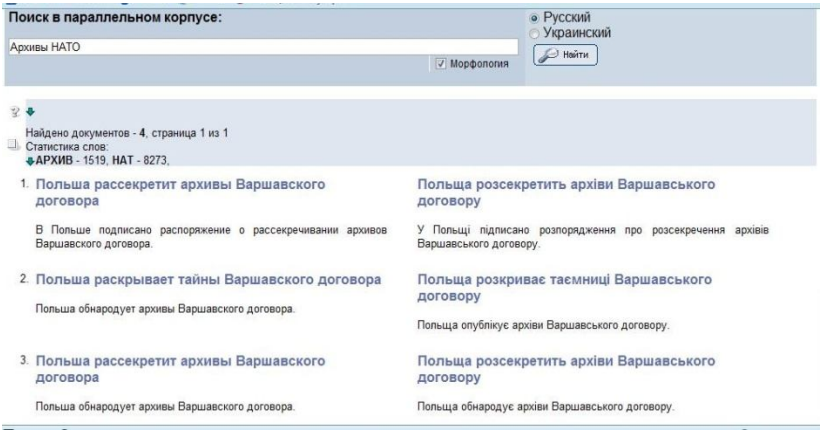

Рис. 3. ***Приклад пошуку офіційно-ділових документів на основі термінів-ключів (контент InfoStream)***

Рис. 4. ***Результати пошуку та аналізу окремих лексем у додатку InfoStream***

Доцільність застосування запропонованого додатка полягає в тому, що традиційні методи побудови паралельних корпусів в НКРМ використовують паралельні дані, що робить їх непридатними для використання. Запропонований контент дає можливість створити двомовний українсько-російський паралельний корпус текстів для роботи з електронними архівами, документами. В інтерфейсі наявний російською та українською мовами обсяг документів понад 500 тисяч пар. Натомість точність запропонованого алгоритму становить 98 %.

За допомогою додатків можливе створення частотного словника на основі морфологічного словника (МС) з використанням текстового корпусу документів, побудови алгоритму обчислення опорних слів з використанням частотного МС і модифікації загальновідомого підходу TF IDF, а також статистичного підходу для виключення похибок. НКРМ можна якісно інтегрувати в контент-моніторингу InfoStream, оскільки вказаний додаток враховує не лише статистичні властивості офіційних текстів, а й деякі морфологічні ознаки. Відповідно до цього алгоритму побудова паралельного корпусу відбувається кількома основними етапами: створення морфологічних словників офіційних документів; створення частотних морфологічних словників офіційних документів; створення словників перекладів; створення процедури визначення опорних слів у міжнародних документах; визначення різномовних дублікатів.

Доцільно, на нашу думку, доповнити морфологічні словники неологізмами, назвами міжнародних організацій, відомими прізвищами секретарів та політичних діячів, яких не було у вихідних словниках. А також додавання та застосування електронних ресурсів публікації документів та новин (наприклад, офіційного ресурсу "НАТО" [10], якщо досліджуються архівні документи організації) (рис. 5).

Із запропонованого ресурсу можна створити файл потрібних словоформ, сортувати леми, після чого проаналізувати кількість входжень кожної словоформи і кількість документів, у яких вона зустрілася. Знайдені частоти записуються в частотний словник, на підставі якого визначається ймовірна нормальна форма кожного слова (аналіз здійснюється через InfoStream в НКРМ).

Після аналізу термінологічних словоформ в InfoStream в НКРМ здійснюється другий етап аналізу лем – виявлення омонімії. Зокрема, в вихідний файл записуються потрібні відповідники, далі зберігаються підраховані частоти з усіма знайденими нормальними

формами. Останній етап – загальний підрахунок кількості омонімічних форм і збереження їх результатів в частотний словник.

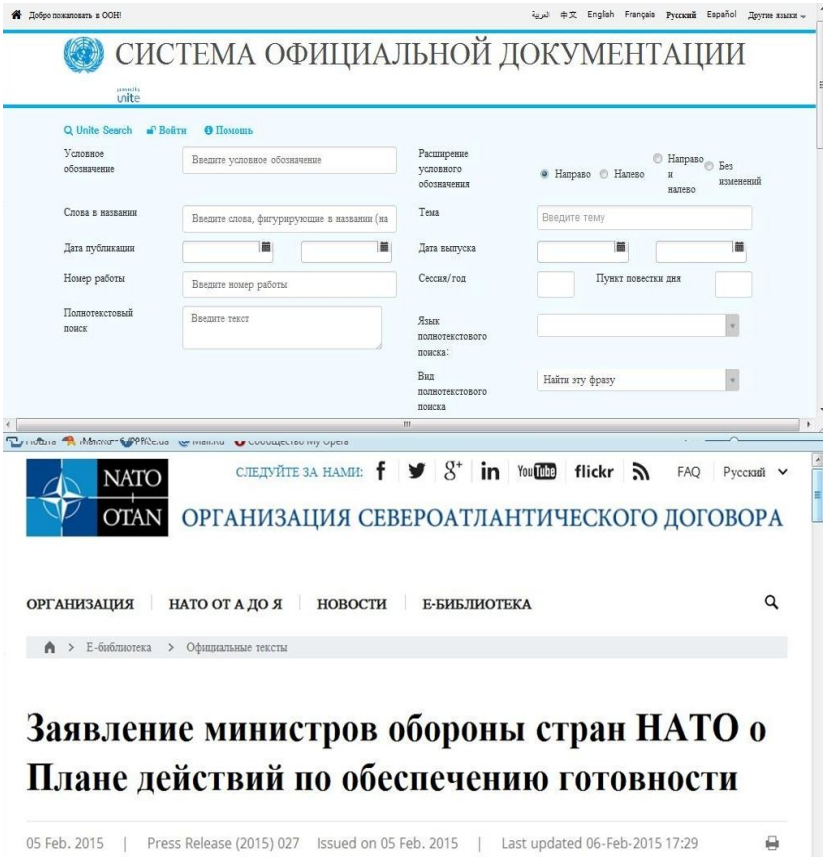

Рис. 5. ***Зразок пошукового офіційного ресурсу "НАТО" для аналізу офіційних документів організації***

На морфологічному рівні для розв'язання завдання побудови паралельних текстових корпусів у словники відбираються всі словоформи іменників.

Третій етап – синтаксичний аналіз вибраного офіційного тексту. Відповідно до інтерфейсу InfoStream можна застосувати декілька етапів аналізу: визначити синтаксичні зв'язки між формами слів у

реченні; контекстуальний аналіз; побудувати формальну синтаксичну структуру речень; побудувати формальну структуру складних синтаксичних одиниць. Усі запропоновані синтаксичні паралелі прикладаються до паралельного українсько-російського паралельного корпусу НКРМ.

Запропоновані інструментальні рівні можуть бути використані для подальшого лінгвістичного дослідження. Водночас лінгвістичні бази даних можуть бути інтегровані не лише в НКРМ, а й в україномовний, англомовний та німецькомовний корпуси з різною поточною обробкою природної мовної системи.

Таким чином, ідея проектної пропозиції полягає в об'єднанні передової інтернет-бази з традиційним акцентом на даних контрастивної лінгвістики. Головне обмеження запропонованого додатка до паралельного корпусу НКРМ – це його менші розміри порівняно з одномовними корпусами. Причина цього – алгоритмічна невідповідність. Саме тому робота над розробкою електронних словників з офіційно-діловою термінологією має продовжуватися, оскільки це дасть змогу розширити лексичну галузеву структуру національних корпусів.

Література

1. Баглей С. Г. Вероятностный подход к задаче разрешения омонимии слов и словарных пар / С. Г. Баглей, А. В. Антонов, В. С. Мешков и др. // Труды межд. конф. Диалог'2005. – М. : Наука, 2007. – С. 23–28.

2. Добровольский Д. О. Использование корпусов текстов в двуязычной лексикографии / Д. О. Добровольский // Среди нехоженых путей : сб. науч. статей к юбилею А. А. Кретова. – Воронеж : НАУКА-ЮНИПРЕСС, 2012. – С. 14–25.

3. Зинькина Ю. В. Разрешение функциональной омонимии в русском языке на основе контекстных правил / Ю. В. Зинькина, Н. В. Пяткин, О. А. Невзорова // Труды межд. конф. Диалог'2005. – М. : Наука, 2005. – С. 198–202.

4. Сичинава Д. В. Параллельные корпуса Национального корпуса русского языка как инструмент лексической типологии / Д. В. Сичинава // Труды симпозиума по лексической типологии LEXT-III. – Гранада, 2012. – С. 11–24.

5. Цинман Л. Л. Лингвистический процессор ЭТАП: дескрипторное соответствие и обработка метафор / Л. Л. Цинман, В. Г. Сизов // Труды межд. Семинара. Диалог'2000. – М. : Изд-во РГГУ, 2000. – С. 366–369.

6. Широков В. А. Корпусна лінгвістика / В. А. Широков, О. В. Бугаков, Т. О. Грязнухіна. – К. : Довіра, 2005. – 471 с.

7. Сайт CRATER Multilingual Aligned Annotated Corpus [Електронний ресурс]. – Режим доступу:
http://www.comp.lancs.ac.uk/linguistics/crater/corpus.html. – Назва з екрана.
8. Сайт Национального корпуса русского язика [Електронний ресурс]. – Режим доступу:
http://www.ruscorpora.ru/corpora-biblio.html. – Назва з екрана.
9. Сайт інтерфейсу InfoStream [Електронний ресурс]. – Режим доступу:
http://ling.infostream.ua. – Назва з екрана.
10. Офіційний сайт архівних документів "НАТО" [Електронний ресурс]. – Режим доступу:
http://www.nato.int/cps/ru/natohq/official_texts.htm. – Назва з екрана.