УДК 813.161.4

# ПЕРСПЕКТИВИ РОЗВИТКУ НАЦІОНАЛЬНОГО КОРПУСУ РОСІЙСЬКОЇ МОВИ ДЛЯ ПЕРЕКЛАДУ СПЕЦІАЛІЗОВАНИХ ТЕКСТІВ

**Дем'янчук Ю.І., к. екон. н.,**
*викладач кафедри технічного перекладу*
*Львівський державний університет безпеки життєдіяльності*

У статті розглядаються автоматизовані паралельні додатки до Національного корпусу російської мови (НКРМ), які вважаються актуальними для перекладу спеціалізованих тестів НАТО, ООН, СОТ. Визначається роль статистичних методик виокремлення термінологічних колокацій з автоматизованих паралельних додатків. Здійснюється пропозиція щодо комплексного поєднання в НКРМ додатків спец-корпусу Sketch Engine, паралельних корпусів CRATER, InfoStream, Reverso Context, статистичних методі виокремлення колокації (log-likelihood, MI, t-score), за допомогою яких здійснюватиметься розробка словника-тезауруса спеціалізованих термінологічних сталих виразів НАТО, ООН та СОТ.

**Ключові слова:** НКРМ, СОТ, НАТО, ООН, колокація, паралельний переклад, спеціалізований текст, термінологічні сталі вирази, статистичний метод.

В статье рассматриваются автоматизированные параллельные приложения к Национальному корпусу русского языка (НКРЯ), которые считаются актуальными для перевода специализированных текстов НАТО, ООН, ВТО. Определяется роль статистических методик выделения терминологических коллокаций автоматизированных параллельных приложений. Предлагается комплексное сочетание в НКРЯ приложений спец-корпуса SketchEngine, параллельного корпуса CRATER, InfoStream, ReversoContext, статистических методов выделения коллокации (log-likelihood, MI, t-score), с помощью которых будет осуществляться разработка словаря-тезауруса специализированных терминологических постоянных сочетаний НАТО, ООН и ВТО.

**Ключевые слова:** НКРЯ, ВТО, НАТО, ООН, коллокация, параллельный перевод, специализированный текст, терминологические словосочетания, статистический метод.

**Demianchuk Yu.I. DEVELOPMENT PROSPECTS OF THE RUSSIAN NATIONAL CORPUS FOR TRANSLATION OF SPECIALIZED TEXTS**

The article deals with automated parallel applications to the Russian National Corpus (RNC), which are considered relevant for the translation of specialized texts of NATO, the UN, the WTO. The role of statistical methods of highlightthe terminological collocations from automated parallel applications. The offer on integrating in NKRM the applications of special-corps Sketch Engine, parallel corpuses CRATER, InfoStream, Reverso Context, statistical methods for allocation of collocation (log-likelihood, MI, t-score), by which it will be the development of a dictionary-thesaurus of specialized terminological constant expressions of NATO, the UN and the WTO.

**Key words:** RNC, the WTO, NATO, the UN, collocation, parallel translation, specialized text, terminological phrases, statistical method.

**Постановка проблеми та аналіз останніх досліджень та публікацій.** Корпусний багатомовний переклад економічної, суспільно-політичної, юридичної та військової термінології відіграє вагому роль у розвитку міжнародних відносин між державами-учасницями СОТ, ООН та НАТО, а також юридичними і фізичними особами. Спеціалізовані терміни міжнародно-правових організацій класифікуються за різними сферами діяльності: екологічною, правовою та військовою безпекою, промисловістю, сільським господарством, торгівлею, спеціалізованою освітою, зв'язком, мілітаризацією тощо. Відповідно, існує тісний тематичний взаємозв'язок між сталими виразами СОТ, НАТО, ООН. Виокремлення стійких словосполучень застосовується в багатьох сферах, серед яких – семантичні і лексикографічні дослідження (зокрема, створення електронних словників для Національних корпусів), у сфері автоматизованого перекладу та аналізу спеціалізованих термінів. Тому перспектива застосування статистичного підходу для виокремлення стійких словосполучень (за допомогою методу створюються частотні списки слів), статистичних заходів асоціації (log-likelihood, MI, t-score), які засновані на формулах, що використовують частоту спільного явища слів в колокації, частоти кожного компонента словосполучення, обсягу корпусу, спец-корпусу Sketch Engine, паралельних корпусів CRATER, InfoStream, Reverso Context – важливий фактор якісного корпусно-лінгвістичного дослідження спеціалізованих юридичних текстів та розробки термінологічного багатогалузевого слов-

ника-тезауруса до НКРМ. Перспективою розвитку НКРМ та розробкою спеціалізованих додатків займалися М.В. Хохлова [13], Е.В. Ягунова, Л.М. Пивоварова [14], А.І. Левінзон [2], І.Г. Федотова [15], Д.В. Січинава [11], Л.Л. Цінман [12], А.Г. Мустайокі [3] тощо. Проте перспектива розробки спеціалізованого додатка (словника-тезауруса) зі спеціалізованою термінологією ООН, НАТО та СОТ залишається відкритою.

**Постановка завдання.** Мета дослідження – розглянути перспективу розвитку НКРМ у контексті застосування в корпусі спеціалізованих паралельних додатків. Концептуальні завдання: визначити роль НКРМ у процесі виокремлення термінологічних колокацій СОТ, НАТО та ООН; розглянути перспективу застосування додаткових паралельних підкорпусів Sketch Engine, CRATER, InfoStream, Reverso Context та статистичних методів виокремлення колокації (log– likelihood, MI, t-score); результат дослідження співвіднести з можливою розробкою словника спеціалізованих термінів ООН, НАТО та СОТ.

**Виклад основного матеріалу дослідження.** Для перекладу спеціалізованих документів НАТО, СОТ та ООН, з метою корпусного виокремлення колокації пропонуємо застосовувати Національний корпус російської мови. НКРМ (Національний корпус російської мови) – це великий, збалансований за складом електронний корпус текстів; ядром НКРМ є російськомовні тексти. Також в НКРМ входить паралельний корпус, який складає багатомовна частина. Підрозділами НКРМ є [8]: основний корпус, синтаксичний корпус, газетний, паралельний (офіційно-ділові, юридичні, правові блоки), навчальний, діалектний, поетичний, усний, акцентологічний, мультимедійний і історичний корпус. Прямий пошук у НКРМ дає можливість точної вибірки. Більш складний і спеціалізований лексико-граматичний пошук у корпусі здійснюється за граматичним, семантичним і додатковим (зокрема, розділових знаків) рівнями. Доступний пошук за кількома словами дає можливість 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 у даному вікні. 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 [8]: WebCorp, Word Filter, IntelliText. Зокрема, «WebCorp» працює над обраною інформаційно-пошуковою системою, обробляючи список 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. У «WebCorp» є функції обробки результатів. Також можливе групування колокацій за алфавітом та за часовими ознаками. Є дві можливості сортування за часом: можна вибрати період часу з меню, що випадає (в минулому місяці, протягом останніх трьох місяців, протягом останніх шести місяців, в минулому році, більше одного, двох чи п'яти років). Функція «Intelli Text» має спеціальну функцію «Affixes», що дає змогу здійснювати пошук префіксів або суфіксів. Якщо необхідно знайти префіксоїд, то використовується пошук по префіксах.

Перспективним є підкорпус НКРМ «Економіка, бізнес, фінанси» (як приклад для перспективних розробок щодо юридично-правових підкорпусів СОТ, НАТО, ООН ) [4; 5; 6], який ґрунтується на матеріалах ЗМІ, відображає щоденні стрімкі зміни в термінології за темою. Даний підкорпус може функціонувати як загальний словник, а також словник автоматичної системи перекладу. Економічний розділ паралельних текстів представлений у НКРМ лише в російсько-англійської версії, проте розробляється російсько-німецький, а також російсько-український та українсько-російський корпус. Для вирівнювання знайдених корпусів застосовується інтерфейс користувача (GUI) та програма вирівнювання текстів HunAlign.

Для виявлення інформаційних дублікатів, представлених на різних мовах (російською та українською) як додаток до НКРМ, а також до Національного корпусу української мови, пропонуємо засто-

совувати програму CRATER [7] та контент InfoStream [9]. На основі контенту InfoStream створення паралельних корпусів офіційних документів можна розділити на дві групи [9]: традиційні і статистичні. Перспективним є підхід до створення паралельних корпусів документів, заснований на алгоритмі пошуку дублікатів документів на різних мовах. Підхід дає можливість відшукати схожі документи на різних мовах у великому масиві документів. У результаті можна переконатися в тому, що в корпус потрапили паралельні документи з різних джерел. Доцільність застосування запропонованого додатка полягає в тому, що традиційні методи побудови паралельних корпусів у НКРМ використовують паралельні дані, що робить їх у даному випадку непридатними для використання. Запропонований контент дає можливість створити двомовний українсько-російський паралельний корпус текстів для роботи з електронними архівами, документами. В інтерфейсі російською та українською мовами наявний обсяг документів понад 500 тисяч пар. Натомість точність запропонованого алгоритму становить 98%. НКРМ можна якісно інтегрувати в контент-моніторинг InfoStream, оскільки вказаний додаток враховує не лише статистичні властивості

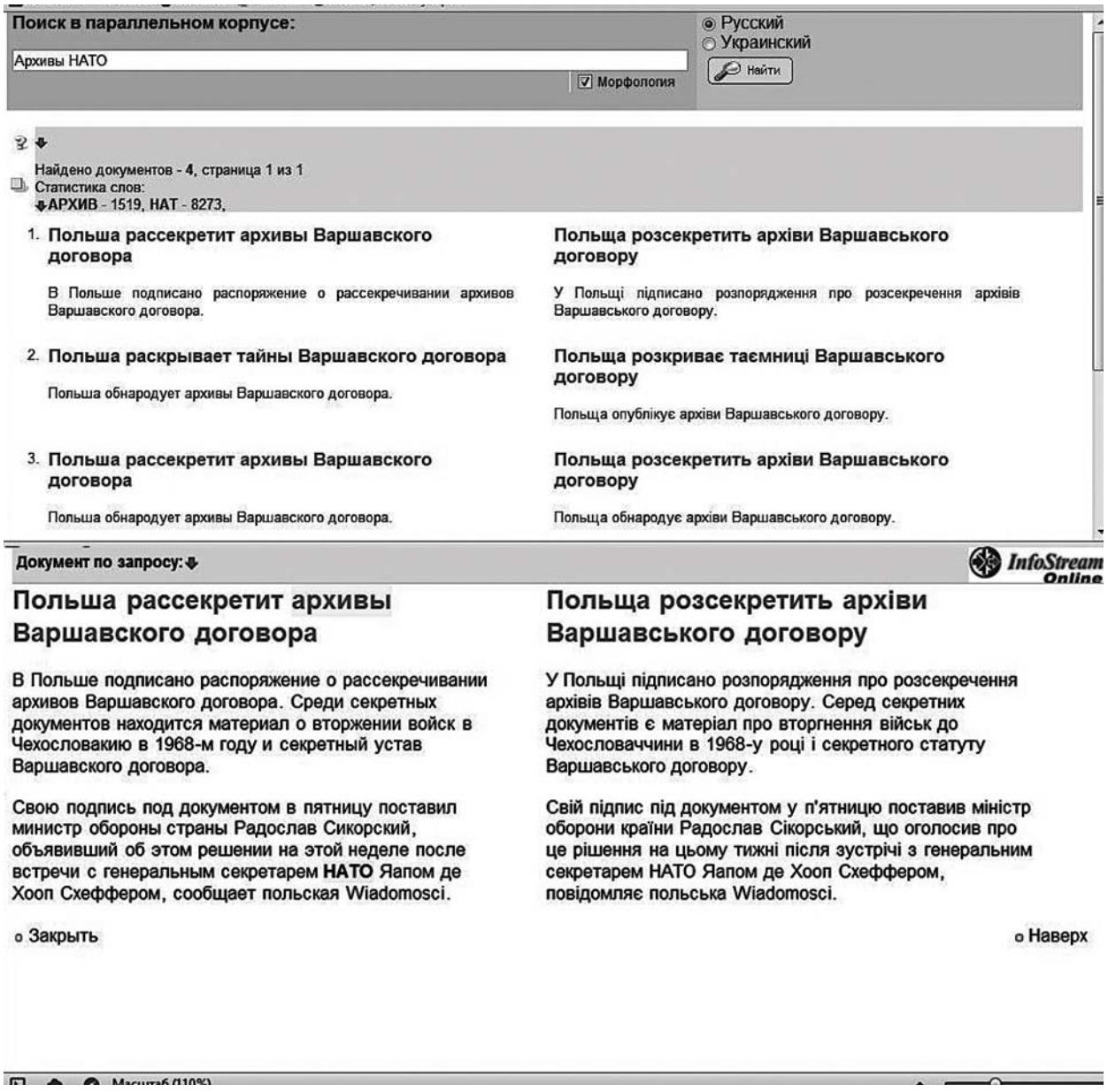


**Рисунок 1. Приклад результатів пошуку офіційно-ділових документів на основі термінів-ключів та аналізу окремих лексем в додатку InfoStream**

офіційних текстів, а й деякі морфологічні ознаки. Відповідно з цим алгоритмом побудова паралельного корпусу відбувається в кілька основних етапів:

1) створення морфологічних словників офіційних документів;
2) створення частотних морфологічних словників офіційних документів;
3) створення словників перекладів;
4) створення процедури визначення опорних слів в міжнародних документах;
5) визначення різномовних дублікатів.

Доцільним, на нашу думку, є доповнення морфологічних спеціалізованих словників неологізмами, назвами міжнародних організацій, відомими прізвищами секретарів та політичних діячів, яких не було у вихідних словниках. Доцільним є додавання та застосування електронних ресурсів публікації документів та новин (наприклад, офіційного ресурсу «НАТО», «СОТ» та «ООН») (Рисунок 1) [9].

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

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

Переклад правових, економічних термінів ООН та СОТ пропонуємо здійснювати в паралельному додатку Reverso Context [10]. Корпус має наступні характеристики: *наявний модус*, який перекладає автентичні тексти; з*а типом текстів* – досліджувані тексти написані на офіційних мовах ООН, СОТ і є офіційними документами ООН, СОТ із рівноправною юридичною та торгівельно-економічною силою; *за обсягом* – складається з підкорпусу (близько 25,5 тис. слововживань англійської та російської мови); *за мовною спеціалізацією* – належить до корпусу правової лексики; *за кількістю слововживань* – вважається об'ємним, щоб можна було робити достовірні спостереження і висновки з урахуванням поставлених перед дослідниками завдань; *за жанром (регістру)* – здебільшого юридичні тексти; *за рівнем вирівнювання мов* – це паралельний двомовний корпус (англійсько-російський); *за рівнем спільності*, *який продукує корпус* – призначений для носіїв російської та англійської мов, що займаються питаннями дотримання і захисту прав осіб, екологічної та економічної безпеки; *за рівнем маркування* – тексти рівноправно поділені та автентичні, тобто корпус паралельних текстів, в якому не розрізняється текст оригіналу і текст

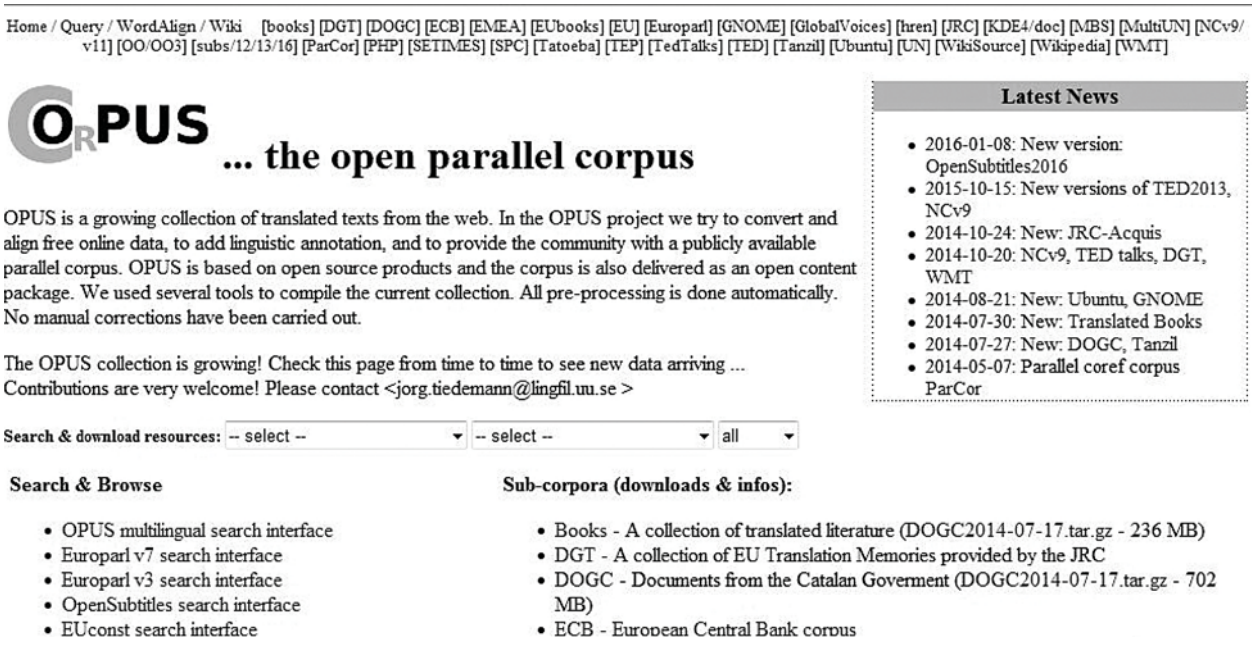


**Рисунок 2. Зразок застосування системи Sketch Engine**

перекладу; *за ступенем відкритості* – існує можливість постійного поповнення та накопичення даного підкорпусу; *за національним вирівнюванням* – це частина Національного корпусу російської мови (НКРМ), з можливою перспективою доповнення корпусу відповідними паралельними документами ООН та іншими офіційними документами (НАТО, СОТ).

Для вилучення колокацій зі стійких термінологічних виразів пропонуємо автоматизовану систему Sketch Engine (як додаток до НКРМ). Система Sketch Engine [1] розроблена англійськими і чеськими дослідниками, оперує поняттями «лексичних портретів» (word sketches), які фіксують лексичну і граматичну сполучуваність лексичних одиниць (Kilgarriff et al). На основі морфологічно розміченого корпусу дана система формує списки слів, в яких міститься інформація про їх «лінгвістичну структуру. Результат роботи програми представлений найбільш частотними (стійкими) словосполученнями, які класифікуються за типами відповідно до лексико-синтаксичних шаблонів граматики (Рисунок 2).

Sketch Engine може видавати список колокацій на потрібному лексичному рівні. Також висвітлюється список із зазначенням частоти кожної колокації в корпусі і значення зв'язку між ключовим словом і колокацією. У системі Sketch Engine є спеціальні інструменти, які визначають рівень синтагматичних та парадигматичних зв'язків на основі дистрибуції лексем в корпусі: *тезаурус* (thesaurus), *кластеризація* (clustering) і *диференціація* (differences) [1]. Робота даних інструментів ґрунтується як на статистичних критеріях, так і на розроблених багатомовних лексико-синтаксичних шаблонах. У процесі виокремлення колокацій у системі Sketch Engine висвітлюються наступні кластери [1]: 1). «джерело дослідження» – «корпус» – «словник» – «матеріал» – «система» – «база»; 2). «об'єкт дослідження» – «слово» – «одиниця» – «дієслово» – «лексема» – «іменник»; 3). «конструкції» – «словосполучення» – «термін»; 4). «термін» – «термін» – «зв'язок».

Для вилучення складових найменувань зі сталих військових, економічних, екологічних, юридично-правовоих термінологічних виразів НАТО, ООН та СОТ пропонується застосувати кілька статистичних методів. Найбільш важливим, на нашу думку, є так званий критерій сили зв'язку, який використовується для визначення сили залежності між компонентами вираження. Загальна кількість цих заходів між зв'язками обраховується біграмами. Значення заходів асоціації можна вважати показниками сили синтагматичною зв'язку між елементами словосполучень. Для опису найбільш поширених заходів найчастіше застосовуються критерії MI, t-score і log-likelihood. Окремі корпусні менеджери надають можливість обрахунку потрібних заходів. Зокрема, міра MI (mutual information) порівнює залежні контекстно-пов'язані частоти з незалежними, та словом, яке з'являлося в тексті випадково. Якщо значення MI (n, c) більше визначеного значення, тоді дане поєднання слів можна вважати статистично значущим. Міра t-score також враховує частоту спільного утворення ключового слова і його колокації, відповідаючи на запитання, наскільки невипадковою є сила асоціації (пов'язаності) між колокаціями. Також досить часто застосовується міра log-likelihood, або логарифмічна функція правдоподібності.

Загалом, застосування статистичних заходів (MI і t-score) дає можливість охарактеризувати предметну сферу і стилістику спеціалізованих текстів. Списки колокацій з окремих спеціалізованих термінологічних виразів, отриманих за допомогою MI і t-score, принципово різні. Наприклад, колокації, що виділяються за допомогою MI, дають можливість визначати назви об'єктів, терміни, складні номінації, що відображають предметну сферу, а критерій t-score спрямований на виокремлення «загальномовних стійких сполучень» (похідних службових слів, дискурсивних слів) і «стійких конструкцій», де ті та інші характеризують стилістичні особливості спеціалізованих текстів.

Важливим результатом дослідження є той факт, що в процесі експерименту були виділені сталі військові вирази, які не зафіксовані ні в одному зі словників. Тому аналіз таких поєднань показав, що відомі біграми знаходяться вгорі списку (відсортованого за зменшенням). Невідомі вирази з деякою часткою ймовірності виявляються стійкими і, відповідно, можуть бути внесені в електронний словник НКРМ.

Запропоновані інструментальні рівні можуть бути використані для подальшого лінгвістичного дослідження та розробки спеціалізованого термінологічного словника колокацій ООН, СОТ, НАТО. Головне обмеження існуючих паралельних додатків – це

їх малі розміри в порівнянні 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. Проте можна використовувати різні лексичні варіанти в системі Sketch Engine, що розширює можливості статистичного підходу, збільшує вірогідність колокаційних зв'язків із заданим ключовим словом. В окремих випадках автоматизована система співвідносить лексичні терміни та сталі термінологічні вирази. Статистичні заходи (MI і t-score) дають можливість охарактеризувати предметну сферу спеціалізованого тексту. До недоліків використання заходів t-score можна віднести те, що вона виділяє колокації з великою кількістю частотних слів-колокацій (стоп-слів). Тому для t-score необхідно задавати список стоп-слів, щоб відкинути непотрібні частотні слова. Багатозначні ж колокації характеризуються високими значеннями заходів t-score. Тому здійснене дослідження стане важливим додатком до НКРМ як головного джерела для паралельного перекладу офіційних документів НАТО.

**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. Автоматизована система Sketch Engine [Електронний ресурс]. – Режим доступу : http://www.sketchengine.co.uk
2. Левинзон А.И. Использование НКРЯ в преподавании русского языка иностранным студентам, специализирующимся в области экономики и финансов / А И. Левинзон // Национальный корпус русского языка и проблемы гуманитарного образования: материалы международной научной конференции. – М. : ГУ-ВШЭ, 2007. – № 12. – С. 127–136.
3. Мустайоки А. Роль корпусов в лингвистических исследованиях языков / А. Мустайоки // Национальный корпус русского языка и проблемы гуманитарного образования: материалы международной научной конференции. – М. : ГУ-ВШЭ, 2007. – № 12. – С. 152–166.
4. Офіційний сайт архівних документів «НАТО» [Електронний ресурс]. – Режим доступу : http://www.nato.int/cps/ru/natohq/official_texts.htm.
5. Офіційний сайт архівних документів «ООН» [Електронний ресурс]. – Режим доступу : http://www.un.org/ru/index.html.
6. Офіційний сайт документів «СОТ» [Електронний ресурс]. – Режим доступу : https://www.wto.org.
7. Сайт CRATER Multilingual Aligned AnnotatedCorpus [Електронний ресурс]. – Режим доступу : http://www.comp. lancs.ac.uk/linguistics/crater/corpus.html.
8. Сайт Национального корпуса русского языка [Електронний ресурс]. – Режим доступу : http://www.ruscorpora.ru/corpora-biblio.html.
9. Сайт інтерфейсу InfoStream [Електронний ресурс]. – Режим доступу : http://ling.infostream.ua.
10. Сайт паралельного перекладу юридичних документів «Reverso Context» [Електронний ресурс]. – Режим доступу : http://context.reverso.net/перевод/русский.
11. 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.
12. 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.
13. Хохлова М.В. Экспериментальная проверка методов выделения коллокаций / М.В. Хохлова // SlavicaHelsingiensia 34. Инструментарий русистики: Корпусные подходы / Под ред. А. Мустайоки, М.В. Копотева, Л.А. Бирюлина, Е.Ю. Протасовой. – Хельсинки, 2008. – С. 343–357.
14. Ягунова Е.В. Извлечение и классификация коллокаций на материаленаучных текстов. Предварительные наблюдения. / Е.В. Ягунова, Л.М. Пивоварова. – СПб., 2010. – 250 с.
15. Федотова И.Г. Юридические понятия и категории в английском языке / И.Г. Федотова, Г.П. Толстопятенко. – Дубна : Феникс, 2008. – 376 с.