УДК 81'25

СТАН ДОСЛІДЖЕННЯ ПРОБЛЕМИ АВТОМАТИЗОВАНОГО БАГАТОМОВНОГО ПАРАЛЕЛЬНОГО ПЕРЕКЛАДУ ТА ВИЛУЧЕННЯ КОЛОКАЦІЙ У СПЕЦІАЛІЗОВАНИХ ДОКУМЕНТАХ НАТО, ООН ТА СОТ

Дем'янчук Ю. І.
Львівський державний університет безпеки життєдіяльності

У статті визначено стан дослідження автоматизованого паралельного перекладу та вилучення колокацій у спеціалізованих документах НАТО, СОТ та ООН. У контексті аналізу лінгвістичних праць сучасних українських та зарубіжних науковців наголошено на необхідності подальшого дослідження проблеми автоматизованого перекладу спеціалізованих текстів. Окреслено важливість ролі зазначених досліджень для розробки багатомовного словника-тезауруса ООН, НАТО та СОТ.

Ключові слова: *колокація, лексико-синтаксичний аналіз, корпус тексту, статистичний аналіз, паралельний переклад, НКРМ.*

Демьянчук Ю. И. Состояние исследования проблемы автоматизированного многоязычного параллельного перевода и изъятия коллокаций в специализированных документах НАТО, ООН и ВТО. *В статье рассматривается состояние исследования автоматизированного параллельного перевода и изъятия коллокаций в специализированных документах НАТО, ВТО и ООН. В контексте анализа лингвистических работ современных украинских и зарубежных ученых указывается на необходимость дальнейшего исследования проблемы автоматизированного перевода специализированных текстов. Отмечается важная роль перечисленных исследований для разработки многоязычного словаря-тезауруса ООН, НАТО и ВТО.*

Ключевые слова: *коллокация, лексико-синтаксический анализ, корпус текста, статистический анализ, параллельный перевод, НКРЯ.*

Demianchuk Yu. I. Status of the investigation of the problem of automated multilingual parallel translation and the collocation exclusion in NATO, UN and WTO specialized documents. *The article deals with the state of research of automated parallel translation and the removal of collocations in the specialized NATO, WTO and UN documents. In the process of theoretical comprehension of the category of parallel texts corpus the current state of the study of the National Corpus of Languages and the problem of multilingual parallel translation of official documents becomes of particular relevance. It is pointed out that the automated process of collocation from legal texts is still unexplored because there are no software applications in parallel corpuses and work with combinations is performed by translators in manual mode or on the basis of statistical interfaces. In order to improve the quality of translation and the compiling of electronic dictionaries-glossaries, theoretical and methodical concepts of allocation of terms from collocation develop. The factors which dictate the scientific interest to the study of collocation – syntactically and semantically integral units, are considered. The stages of studying collocation include: a description of syntagmatic as a phenomenon of language; the compatibility of words is studied without detaching from the semantic features of lexical units; the syntagmatic component is distinguished in the structure of the value of a lexical unit. The article pays attention to the feasibility of using multilingual thematic corpuses of texts for the translation of the specialized UN, NATO and WTO official documents. In addition, parallel or comparative text corpuses should be used, in particular the emphasis on the Russian National Corpus. It is proposed to analyze and find possible ways of using RNC for the derivation of terminological word combinations. With the purpose of a simplified procedure for the conclusion of a vocabulary dictionary-thesaurus it is proposed to consider the possibilities of independently compiling a thematic corpus of the NATO, the UN, and the WTO official documents.*

Key words: *collocation, lexical-syntactic analysis, corpus of text, statistical analysis, parallel translation, RNC.*

Постановка проблеми та обґрунтування актуальності її розгляду. У процесі теоретичного осмислення категорії паралельного корпусу текстів особливої актуальності набуває стан дослідження як національних корпусів мови, так і проблеми багатомовного паралельного перекладу офіційних документів. Зокрема, сучасний стан дослідження паралельних корпусів у складі НКРМ та інших корпусів датовано 2009–2015 роками, коли велася активна робота з формування методики та структури багатомовного корпусу. Відтоді НКРМ (Національний корпус російської мови) почав вміщувати такі паралельні двомовні корпуси, як: англійський, вірменський, білоруський, болгарський, іспанський, італійський, латиський, німецький, польський, український, французький та естонський. Відповідно паралельні багатомовні національні корпуси мови стали важливим об'єктом дослідження. З подальшими розробками науковці на теоретичному та практичному рівнях вдосконалювали структуру корпусів, і 2012 року був розроблений багатомовний російсько-французький корпус, який вміщував до 4-х варіантів перекладу спеціалізованих текстів. До паралельних національних корпусів мови (зокрема до НКРМ) тепер включено як художні, так і юридичні тексти.

Автоматизований процес виведення колокацій із юридичних текстів на разі не досліджений, оскільки в паралельних корпусах немає програмних додатків і роботу зі сполученнями перекладачі здійснюють вручну або на основі статистичних інтерфейсів.

Формулювання мети і завдань статті. Мета статті – окреслити сучасний стан дослідження паралельного багатомовного перекладу офіційних текстів та автоматизованого виведення колокацій.

Поставлена мета передбачає розв'язання концептуальних **завдань**: розглянути найновіші праці українських науковців, які вивчали проблему виведення колокацій у спеціалізованих текстах; проаналізувати зарубіжні дослідження проблеми багатомовного паралельного перекладу юридичних текстів; вказати на практичне значення лінгвістичних розробок та їх роль у процесі укладання колокаційних словників-тезаурусів.

Виклад основного матеріалу дослідження. За останні роки з'явилася значна кількість досліджень і розробок, присвячених проблемі перекладу та виведення колокацій. Серед усього науковці обґрунтовують теоретичні аспекти статистичного та практичного методів виведення колокацій (S. Evert) [12], виділяють колокації на базі статистичних методів, що є важливим у лексикографічній практиці (B. T. Sue Atkins [13, 125], A. Kilgarriff [14, 108]). Зокрема, розглянуто статистичні автоматизовані методи MI-score, t-score і log-likelihood; сервіс пошуку біграм на сайті АОТ (автоматична обробка тексту); сервіс пошуку у Національних корпусах; система Sketch Engine тощо.

Е. І. Большакова, О. А. Митрофанова, В. П. Захаров, М. Ю. Сидорова запропонували новий підхід дослідження синтагматичних зв'язків, який передбачає опис сполучуваності за допомогою лексико-синтаксичних рівнів аналізу тексту. У праці Е. І. Большакової лексико-синтаксичний шаблон – це «структурний зразок мовної конструкції, який відображає її лексичні і поверхнево-синтаксичні властивості» [1, 45]. В. П. Захаров лексико-синтаксичний рівень тексту кваліфікує як мовну конструкцію, у якій відображено істотні граматичні характеристики лексем як елементів сталих висловів, що створені відповідно до стилю тексту [8, 96].

Розвиваються теоретичні та методичні концепції виділення термінів із колокацій. Зазвичай, це такі види критеріїв, як інформаційний, дефініційний, критерій концептуальної цілісності, критерій логічних теорем тощо. Усі вони орієнтовані на покращення якості перекладу та укладання електронних словників-глосаріїв.

Науковий інтерес до вивчення колокацій (синтаксично і семантично цілісних одиниць) зумовлений різними факторами: по-перше, проводиться опис синтагматики як явища мови; по-друге, сполучуваність слів вивчається у взаємозв'язку із семантичними особливостями лексичних одиниць; по-третє, у структурі значення лексичної одиниці виокремлюється синтагматичний компонент.

Серед сучасних українських досліджень відомими є праці Т. Бобкової, присвячені еволюції корпусної лінгвістики, а також вилученню колокацій із окремих текстових корпусів (монографія «Вилучення і класифікація колокацій: корпусний підхід»). Дослідниця пов'язує історичний розвиток корпусів із необхідністю вивчення англійської мови як іноземної та можливістю швидкого перекладу декількох текстових блоків одночасно [2, 16].

Стосовно важливості досліджень багатомовних корпусів дослідниця висновує, що для таких корпусів характерне застосування статистичних методів, комп'ютерних технологій і програмного забезпечення (це відповідно дає можливість глибоко розглядати різні види корпусів тексту з їхніми стилістичними особливостями) і вони опертні на визнання тексту основним об'єктом лінгвістичного дослідження й на науковий характер власної методології дослідження (враховується лексикографічний, морфологічний та стилістичний рівні). Водночас корпусні дослідження ґрунтуються на певній дослідницькій філософії.

Серед найвідоміших методологічних розробок є:

1. Автоматизація лінгвістичного аналізу – дослідження лабораторії лексикографічного аналізу (Безансон, 1957 р.) і лабораторії Центру з автоматизації філологічного аналізу (Ґалараті, 1953 р.).

2. Розробка інструментарію машиночитаних корпусів – комп'ютерний конкордансер (1949–1967 рр.) Р. Буза при підтримці IBM; механолінгвістика й методологія формування вибірки (1956–1970 рр.) А. Джіланда.

Дослідниця В. В. Жуковська у праці «Корпусна лінгвістика: історична перспектива та сучасний стан» висвітлює системи розвитку двомовних та багатомовних корпусів. Особливу увагу авторка приділяє електронному корпусу The Survey of English Usage, укладеному Рендольфом Квірком 1959 року в University College London. Цей проект став перехідним етапом становлення корпусної лінгвістики, оскільки стосувався лінгвістичних особливостей щоденного спілкування (як писемного, так і усного) звичайних громадян і не передбачав збереження даних в електронному форматі [7, 36].

Описуючи сучасний стан корпусної лінгвістики, О. М. Демська-Кульчицька у монографії «Основи національного корпусу української мови» зазначає, що цей напрям досить розгалужений і передбачає стадії: по-перше, загальної теорії корпусної лінгвістики, над якою працюють Д. Байбер, Дж. Синклер, В. Тойберт; по-друге, кореляції корпусної лінгвістики та інших лінгвістичних дисциплін; по-третє, типології корпусів та методики інтерпретації корпусних даних; по-четверте, розроблення загальних засад створення природних мов тощо (праці Б. Алтенберга, М. Баньки, У. Френсиса, Г. Кеннеді, Г. Ліча, А. Баранова, М. Михайлова, Р. Рикова, Л. Ричкової, С. Шарова та ін.) [6, 74].

Інший дослідник Г. Г. Лук'янець наголошував, що основою корпусної лінгвістики є розроблення теоретичних засад і практичних прийомів побудови машинного опрацювання, експлуатації та аналізу мовних даних, оформлених як корпус текстів. У дослідженні «Основні напрямки сучасних корпусних досліджень мови та перспективи їх подальшого розвитку» автор наголошує, що багатомовні корпуси тексту – це осо-

блива інформаційно-довідкова система, яка слугує базою дослідження одиниць та явищ різних мовних рівнів (фонетичного, морфологічного, лексико-семантичного та синтаксичного) з метою вивчення особливостей використання природної людської мови у формах усного та писемного мовлення та для визначення специфіки функціонування мови в різних стилях (художньому, офіційно-діловому, публіцистичному, науковому, розмовному) [9, 128].

Термінологічне словосполучення не раз ставало об'єктом лінгвістичних описів і досліджувалося в багатьох терміносистемах: біології, математиці, медицині, військовій термінології, економіці тощо. Зацікавленість пошуком з використанням спеціалізованих термінів у своїх роботах виявляють Д. В. Джоханссон і Ю.-Х. Лью. Зокрема, у роботі Д. В. Джоханссон терміни використовуються для збільшення якості коефіцієнтів слів. На противагу цьому дослідження Ю.-Х. Лью були обмежені використанням індексу словника (MeSH), у який вручну вносилася інформація зі спеціалізованих текстів. Як стверджують науковці, для кожної статті індексу встановлені зв'язки з термінами словника, що дало можливість відмовитися від процедури автоматичного вилучення термінологічних словосполучень.

Не оминають увагою науковці і застосування термінів для розв'язання завдань інформаційного пошуку: розширення запитів, міжмовного пошуку, класифікації текстів, витягу ключових фраз. Численні праці А. Хотхо, Ю. Ву, К. Лі, Р. С. Бота, Дж. Грінберга вказують на високу актуальність досліджень у цій сфері.

K. Frantzi, S. Ananiadou, H. Mima у студії «Automatic Recognition of Multi-Word Terms: the C-value/NC-value Method» розглядають процедуру виділення термінологічних словосполучень зі спеціалізованих текстів. Серед них три базові підходи: 1) лінгвістичний, 2) статистичний і 3) змішаний, заснований на застосуванні статистичної і лінгвістичної інформації [15, 130].

Крім цих методів, які зазвичай застосовуються для пошуку термінів в тексті, раціонально розглянути також підходи, вироблені для вилучення стійких словосполучень (колокацій). У пропонованому дослідженні акцентовано на теорії складених найменувань (multiword expression theory), розробленій у Стенфордському університеті (США). Для вилучення складових найменувань запропоновано кілька статистичних методів; найцікавішим є так званий критерій сили зв'язку, який використовують для визначення сили залежності та зв'язку між компонентами.

Продовжує працювати над проблемою статистичного методу виділення термінологічних спеціалізованих сполучень М. В. Хохлова. У праці «Исследование лексико-синтаксической сочетаемости в русском языке с помощью статистических методов» авторка докладно описує формальні механізми лексико-синтаксичної сполучуваності та виявлення колокацій [11, 85]. Впровадження цих методів у вивчення сполучуваності має як теоретичне, так і практичне значення. Праця М. В. Хохлової є певною мірою об'єднанням трьох підходів до вивчення сполучуваності: лексичного (семантичного), синтаксичного і статистичного. Застосування статистичних методів дослідниці на базі великих корпусів текстів стало підґрунтям створення словників і граматик нового типу, серед усього і словників стійких словосполучень.

Цікаве бачення порушуваної проблеми викладено в монографії М. В. Каменського, Т. Н. Ломтєва, Н. С. Кабилкіна, С. А. Бутенко, І. Ю. Тріско, Д. А. Шереметова, Е. Р. Федіна «Формалізація дослідних процедур. Аналіз семантики мовних одиниць» [10]. Тут узагальнено дослідження лексичної семантики із застосуванням елементів автоматизованої обробки електронних корпусів текстів різними мовами. Простежено можливості застосування спеціалізованих мов предметної галузі (DSL – Domain Specific Language) для опису семантики мовних одиниць з метою укладання багатомовного словника [10, 87].

Водночас проведено чимало наукових експериментів, мета яких – проаналізувати процес писемного перекладу. Частину з них виконано на рівні корпусної лінгвістики. Важливими в цій сфері є роботи М. Уілкінсона, В. Н. Шевчука, Н. В. Владимова і Р. К. Кошкіна, де корпус електронних текстів постає засобом виявлення і усунення чинників, які призводять до похибок перекладу. Дослідники А. А. Віланденбегр, Л. Н. Бєляєва, В. П. Захаров, С. А. Коваль, Ю. М. Марчук, В. Ш. Рубашкін, В. В. Риков, Л. В. Ричкова, А. Я. Шайкевич особливу увагу приділяли спеціальному корпусу текстів. Такий корпус є багаторівневою системою кількаразового використання, що дає змогу здійснювати різні операції під час розв'язання конкретних дослідницьких завдань.

У дисертації «Принципы и методы гармонизации терминологии на основе корпуса специальных параллельных текстов (на материале документов ООН)» А. А. Віланденберг корпус спеціальних паралельних текстів розглядає як адекватну базу для гармонізації термінології відповідної предметної сфери, зокрема офіційних документів [3, 116]. Дослідниця доходить важливих висновків щодо стилістичних особливостей перекладу юридичних документів:

1. Для забезпечення семантичної однозначності правового документа і точності його юридичних формулювань у спеціальному двомовному та багатомовному словнику підлягають фіксації стійкі багатокомпонентні термінологічні словосполучення.

2. Функціональна класифікація термінів права детермінує виокремлення номінацій уточнювальних і регулювальних термінів.

3. На відміну від наукових і науково-технічних текстів у правовому тексті особливу роль відіграють терміни – модальні дієслівні терміносистеми, що є засобом вираження прагматичної спрямованості тексту, а також уточнювальні терміни – прийменникові терміносполучення як засіб необхідної конкретизації правової норми.

Варте уваги дослідження Д. Ю. Груздєва «Электронный корпус текстов как эффективный инструмент переводчика» [5, 184], де автор простежує потенційні можливості використання електронного фахового корпусу текстів при перекладі

військово-технічних текстів НАТО. На його думку, помилки найчастіше трапляються у вживанні лексичних конструкцій. До них належать: проблеми лексико-граматичної сполучуваності, орфографічні, пунктуаційні труднощі й проблеми, пов'язані з вибором оптимальної граматичної конструкції. Інший лінгвіст С. Є. Вадяєв у дисертації «Лингвистические принципы построения и использование корпуса текстов для исследования официально-делового стиля современного немецкого языка: На материале электронного корпуса "DER" [4, 12] вперше детально описав лінгвістичну концепцію і на цих засадах розробив технологію побудови німецько-російського паралельного корпусу текстів, а також наочно репрезентував можливості використання німецькомовного субкорпусу ПКТ для досліджень у стилістичній сфері. В межах роботи детально представлено методику автоматичної побудови паралельних текстів німецькою та російською мовами, а також здійснено практичне застосування німецького субкорпусу ПКТ "DER" для виявлення специфіки вживання модальних дієслів у текстах офіційно-ділового стилю сучасної німецької мови (на прикладі документів ООН).

Таким чином, на розробки багатомовних корпусів тексту суттєво впливало вироблення перекладацької тактики з офіційними документами. Оскільки більшість корпусів були створені в 90-х рр. XX століття, то їх стилістична якість доволі низька. Що стосується перекладу сучасних міжнародних документів, то потрібно враховувати морфологічні, лексикографічні та стилістичні елементи, які доповнюють сучасні багатомовні корпуси.

Розглянувши каталог базових дисертацій та монографій, актуальним для застосування (серед багатомовних спеціалізованих корпусів тексту) можемо вважати НКРМ, на основі якого перекладацька робота з юридичними та офіційними документами дає можливість виявити позитивні мовні показники та алгоритми.

Висновки та перспективи подальших досліджень у цьому напрямі. Загалом теоретичний аналіз спеціалізованих корпусів текстів здійснено на основі загального розгляду зарубіжних та українських досліджень. Уважаємо доцільним застосовувати багатомовні тематичні корпуси текстів для перекладу спеціалізованих офіційно-ділових документів ООН, НАТО та СОТ. Необхідно переглянути підходи роботи з корпусом текстів на предмет врахування особливостей термінології та колокацій міжнародних документів. Додатково варто застосовувати паралельні або порівняльні корпуси тексту. Зокрема, робимо акцент на Національному корпусі російської мови. Пропонуємо: 1) проаналізувати і знайти можливі шляхи застосування НКРМ для виведення термінологічних словосполучень; 2) розглянути можливості самостійного складання тематичного корпусу офіційно-ділових документів НАТО, ООН, СОТ з метою спрощеної процедури укладання колокаційного словника-тезауруса.

ЛІТЕРАТУРА

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : [учеб. пособие] / Е. И. Большакова, Э. С. Клышинский, Д. В. Ландэ и др. – М. : МИЭМ, 2011. – 272 с.

2. Бобкова Т. В. Вилучення і класифікація колокацій : корпусний підхід / Т. В. Бобкова. – К. : Освіта України, 2016. – 350 с.

3. Виландеберк А. А. Принципы и методы гармонизации терминологии на основе корпуса специальных параллельных текстов : На материале документов ООН : дисс. … канд. филол. наук : спец. 10.02.21 «Прикладная и математическая лингвистика» / Анна Арнольдовна Виландеберк. – СПб., 2005. – 266 с.

4. Вадяев С. Е. Лингвистические принципы построения и использования корпуса текстов для исследования официально-делового стиля современного немецкого языка: На материале электронного корпуса DER : автореф. дисс. … канд. филол. наук : спец. 10.02.04 «Германские языки» / С. Е. Вадяев. – Нижний Новгород, 2005. – 22 с.

5. Груздев Д. Ю. Электронный корпус текстов как эффективный инструмент переводчика : дисс. … канд. филол. наук : спец. 10.02.19 «Теория языка» / Дмитрий Юрьевич Груздев. – М., 2013. – 188 с.

6. Демська-Кульчицька О. М. Основи Національного корпусу української мови : [монографія] / О. М. Демська-Кульчицька. – К. : Інститут української мови Національної академії наук України, 2005. – 219 с.

7. Жуковська В. В. Корпусна лінгвістика : історична перспектива та сучасний стан / В. В. Жуковська // Ключови въпроси в съвременната наука : [материали за 8-а международна научна практична конференция]. – София. «Бял ГРАД-БГ» ООД, 2012. – Т. 18 : Филологични науки. – 72 с.

8. Захаров В. П. Тезаурус по корпусной лингвистике / В. П. Захаров // Информационные технологии и письменное наследие. El'Manuscript-10 : [материалы Международной научной конференции]. – Уфа, 2010. – С. 95–98.

9. Лук'янець Г. Г. Основні напрямки сучасних корпусних досліджень мови та перспективи їх подальшого розвитку / Г. Г. Лук'янець // Наукові праці Національного університету харчових технологій. – 2012. – № 44. – С. 127–133.

10. Формализация исследовательских процедур анализа семантики языковых единиц : [коллективная монография] / М. В. Каменский, Т. Н. Ломтева, Н. С. Кабылкинаи др. – СКФУ, 2016. – 170 с.

11. Хохлова М. В. Исследование лексико-синтаксической сочетаемости в русском языке с помощью статистических методов (на базе корпусов текстов) / М. В. Хохлова. – СПб., 2011. – 223 с.

12. Evert S. Computational Approaches to Collocations [Електронний ресурс] / Stefan Evert. – Режим доступу: http://www.collocations.de/EK/Articles/MathAM.4up.pdf.

13. Atkins B. T. Sue. The Oxford guide to Practical Lexicography / B. T. Sue Atkins and Michael Rundell. – Oxford : Oxford University Press, 2008. – 540 p.

14. Kilgarriff A. The Sketch Engine / Adam Kilgarriff, Pavel Rychly, Pavel Smrž, David Tugwell // In Proceedings of the Eleventh EURALEX International Congress. – Lorient, France: Universite de Bretagne-Sud, 2004. – P. 105–116.

15. Frantzi K. Automatic recognition of multi-word terms: the C-value/NC-value method / Katerina Frantzi, Sophia Ananiadou, Hideki Mima // International Journal on Digital Libraries. – August 2000. – Vol. 3. – Issue 2. – P. 115–130.