КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ТАРАСА ШЕВЧЕНКА

ISSN 0320-3077

УКРАЇНСЬКЕ МОВОЗНАВСТВО

Міжвідомчий науковий збірник

Засновано у 1973 р.

Випуск 43

КИЇВСЬКИЙ ВТЦ УНІВЕРСИТЕТ 2013

У збірнику вміщені дослідження з актуальних проблем комп'ютерної лінгвістики. Для викладачів, науковців, учителів, студентів.

This volume presents investigations on topical issues in Computational Linguistics. It is intended for lecturers, researchers, teachers, and students.

Людмила Алексієнко,
Наталія Дарчук
Київський національний університет імені Тараса Шевченка

ДО ДВАДЦЯТИРІЧЧЯ ЛАБОРАТОРІЇ КОМП'ЮТЕРНОЇ ЛІНГВІСТИКИ

Навчально-дослідній лабораторії комп'ютерної лінгвістики Інституту філології Київського національного університету імені Тараса Шевченка у вересні 2012 року виповнилося 20 років. Згадуючи пройдений шлях, хотілося б поділитися нашим досвідом, здобутками і планами з колегами-"прикладниками", аспірантами, студентами і висловити щиру подяку всім, хто допомагав нам на цьому шляху. І крім того, запровадження нової спеціалізації "комп'ютерна лінгвістика" на кафедрі сучасної української мови, створення під неї навчально-дослідної лабораторії, підготовка упродовж двох десятиліть бакалаврів, магістрів та аспірантів – фахівців з автоматизованого аналізу тексту – це ще одна сторінка в історії кафедри сучасної української мови Інституту філології.

Коли, у кого і чому виникла ідея готувати фахівців у галузі комп'ютерної лінгвістики, і як ми її реалізовували? Термін "комп'ютерна лінгвістика" 30 років тому на наших теренах сприймався майже як оксюморон, незважаючи на те, що вже в 60-х роках минулого століття в європейських та американських університетах активно створювалися кафедри computer sciences, які готували, зокрема, й комп'ютерних лінгвістів. Їх навчали кібернетики, інформатики та мовознавства, але в пріоритеті в цей період були інформаційно-комп'ютерні технології.

Недостатньо фахова лінгвістична параметризація тогочасних текстових корпусів, частотні словники з незнятою омонімією та недиференційованою лексичною семантикою, системи машинного перекладу недостатнього рівня семантико-граматичної глибини тощо свідчать про те, що лінгвістичному сегменту в багатьох комп'ютерних продуктах не приділялося належної уваги.

Комп'ютерна лінгвістика є новим напрямком класичної прикладної лінгвістики, яка виникла, розвивалася й розвивається паралельно з традиційною лінгвістикою. До компетенції при-

кладної лінгвістики входять: письмо (графіка), методика навчання рідної та іноземної мов, лексикографія, мовна політика – ліквідація неграмотності, вибір державної мови та її підтримка, розроблення національної термінології, національних ономастиконів тощо. Ця проблема актуальна й на сучасному етапі прикладної лінгвістики.

Разом із тим у другій половині XX століття у прикладній лінгвістиці з'явився новий вектор, спричинений активними процесами інтеграції гуманітарних, природничих, технічних і математичних наук. Результатом цього було усвідомлення і визначення спільної для багатьох предметних галузей проблеми – автоматизація оброблення, обміну і збереження різноманітної інформації, яка функціонує в суспільстві в текстовій формі. Фахівці практично всіх галузей знань користуються мовою як універсальним засобом оформлення і смислового представлення знань. Оскільки текстова інформація є природною для людини формою комунікації, лінгвістичне забезпечення інформаційних систем стає головним завданням комп'ютерної лінгвістики. У цій ситуації є необхідним розподіл компетенцій між власне лінгвістикою та інформаційно-комп'ютерними технологіями. Фаховий аналіз смислу текстів – це прерогатива лінгвістів, які глибоко розуміють систему мови в усіх її проявах. Багатовимірне впорядкування параметризованої лінгвістами текстової інформації в бази даних і бази знань, корпуси текстів, створення гіпертекстових мереж із можливістю навігації у величезних масивах тощо – це прерогатива фахівців з інформатики і кібернетики. Таким чином, комп'ютерна лінгвістика – це лінгвістика із застосуванням інформаційно-комп'ютерних ресурсів.

Нова предметна галузь потребувала підготовки відповідних фахівців, яких у 60–70-х рр. в Україні ще не було. Перший крок у цій справі було зроблено в середині 60-х років у Київському університеті. Спочатку на філологічному факультеті була створена спеціалізація зі структурно-математичної лінгвістики – групи по 8–10 студентів, яких готували як спеціалістів з автоматичної обробки текстів. Очільником спеціалізації була

проф. Феоніла Олексіївна Нікітіна, викладали спецдисципліни професори-філологи Віктор Вікторович Коптілов, Едуард Федорович Скороходько, Ірина Платонівна Севбо, с. н. с. Ірина Борисівна Штерн, професор-математик Лев Аркадійович Калужнін та програмісти з Інституту кібернетики. Через кілька років спільними зусиллями філологів та математиків на факультеті кібернетики було створено відділення і кафедру структурно-математичної лінгвістики зі щорічним набором студентів (15 осіб), які отримували диплом спеціаліста такого профілю: *автоматична обробка тексту*; *іноземна мова*; *перекладач-референт*. У підготовці кадрів із цієї спеціальності брали активну участь та надавали дієву допомогу співробітники Інституту кібернетики АН УРСР і персонально акад. Віктор Михайлович Глушков, за ініціативою якого було запроваджено наукову спеціальність: структурна, прикладна і математична лінгвістика – 10.02.21. Однак у 1985 році відділення було ліквідоване.

У цей період в Інституті мовознавства ім. О. О. Потебні АН УРСР потужно працював відділ структурно-математичної лінгвістики на чолі з проф. Валентиною Сидорівною Перебийніс. На базі української мови проводилися масштабні статистичні й стилеметричні дослідження, структурно-статистичне моделювання на різних рівнях мовної системи.

Інформаційно-комп'ютерні технології розвиваються швидкими темпами. Тому пропозиція запровадити на українському відділенні спеціалізацію "комп'ютерна лінгвістика" була підтримана кафедрою сучасної української мови, адміністрацією філологічного факультету та університету. З 1989 року на 3–5 курсах були створені групи по 7–10 студентів, які навчалися за планом спеціалізації, слухали відповідні спецкурси, працювали в спецсемінарах, писали бакалаврські, дипломні й магістерські роботи з різноманітних проблем комп'ютерної лінгвістики і отримували до диплома філолога-українєта сертифікат фахівця з автоматизованої обробки тексту.

Спецдисципліни викладали: співробітники відділу структурно-математичної лінгвістики – проф. В. Перебийніс, с. н. с. Н. Кли-

менко, Т. Грязнухіна, Н. Дарчук, Л. Орлова, Є. Карпіловська, програміст Л. Братищенко та викладачі університету – доц. Л. Алексієнко і с. н. с. І. Штерн.

З техніки на той час був один слабенький комп'ютер, але такий, без перебільшення, зірковий лекторат із самого початку зробив цю спеціалізацію рейтинговою, причому не тільки серед наших студентів, а й ширше – до нас почали звертатися за методичною допомогою викладачі-філологи з різних університетів України – Київського лінгвістичного, Донецького, Волинського, Харківського, Львівського, які почали серйозно займатися комп'ютерною лінгвістикою і підготовкою кадрів.

Фундамент спеціалізації закладався такими спецкурсами, як структурна і прикладна лінгвістика; штучний інтелект; лінгвостатистика; комп'ютерна лексикографія; лінгвістика тексту; автоматичний морфологічний аналіз; автоматичний синтаксичний аналіз; машинний переклад; основи програмування та ін. Студенти були в захваті від нового сприйняття мови, від долучення до потужних цивілізаційних процесів інформатизації, від реальної роботи в різноманітних проектах, які виконувалися в лабораторії. За 20 років було підготовлено 90 фахівців зі спеціалізації комп'ютерна лінгвістика, з них успішно захистили кандидатські дисертації дев'ять наших випускників і двоє подали дисертації до захисту. Студенти завжди тримають нас у тонусі, хочеться щиро подякувати їм за співпрацю і взаєморозуміння.

Навчальний план спеціалізації постійно вдосконалювався і впродовж 23 років увиразнився в напрямку інформаційно-комп'ютерного моделювання мови, створення автоматизованих інтелектуальних систем на базі української мови.

У 1992 році під спеціалізацію з метою вдосконалення навчального процесу було створено навчально-дослідну лабораторію комп'ютерної лінгвістики. Оптимальним способом реалізації цієї мети є співпраця викладачів і студентів у проектах зі створення автоматизованих систем аналізу текстів, баз даних, електронних словників, підручників тощо.

Першим проектом лабораторії була "Параметризована база даних українського поетичного мовлення", що планувалася як

джерело для філологічних студій функціонування української мови в літературі, зокрема дослідження ідіостилів українських поетів на різних хронологінчих зрізах. Проект викладено в Інтернет-порталі лабораторії (mova.info). На цій базі захищені три кандидатські дисертації наших випускників – Л. Гливінської, Д. Данильчука, Ю. Маковецької-Гудзь, а також понад 40 бакалаврських і магістерських робіт. Проект дістав міжнародне схвалення, зокрема грант ACLS – американської асоціації підтримки інноваційних проектів в Росії, Україні та Білорусії.

Наступний проект – "Морфемно-словотвірна база даних української мови" (≈170 тис. слів), викладена на порталі лабораторії. База є ресурсом для автоматичного укладання алфавітно-частотних словників морфем і словотвірних гнізд на матеріалі будь-яких текстів. База забезпечує високу якість, масштабність, системність та оперативність досліджень. На цій базі підготовлені кандидатські дисертації О. Тютенко і Т. Жигун, а також захищено понад 30 бакалаврських і магістерських робіт різної тематики. Ця база даних також відзначена грантом ACLS.

Важливим проектом для розвитку лабораторії став лінгвістичний портал "mova.info". На порталі розміщуються всі наші проекти, ведеться рубрика "Новини мовознавства, мовної культури і мовної політики". Проект підтриманий грантом Посольства Канади в Україні. Кількість щоденних відвідувань порталу ≈100.

Проект "Електронна граматика української мови" (для абітурієнтів та дистанційного навчання). Підручник, крім теорії, містить вправи, тести та єдину навігаційну систему користування. Як свідчать постійні звернення до порталу, підручник популярний серед абітурієнтів. Для спеціалізації він є ресурсом спецкурсу "Електронні підручники з мови".

Проект "Українсько-російсько-італійська довідково-пошукова система з питань усиновлення" (2400 юридичних термінів трьома мовами з перекладом, тлумаченням, енциклопедичною інформацією та юридичними документами трьох країн) виконувався як міжнародний, спільно із Флорентійським університетом. Одержав схвальний відгук дитячого фонду ЮНІСЕФ при ООН.

Проект на замовлення Державного комітету України з питань науки, інновацій та інформатизацій "Електронний словник лінгвістичної термінології з інформаційно-пошуковою системою (тезаурус)" – 3400 термінів з українсько-російсько-англійським перекладом. Цей проект також є ресурсом для спецкурсів та нових проектів. За його методикою магістри протягом року створили електронний тезаурус літературознавчих термінів.

Із 2010 року лабораторія почала працювати над масштабним проектом "Дослідницький корпус української мови". На сьогодні на порталі викладена його частина – корпус розмічених і параметризованих текстів обсягом понад 13 млн слововживань.

Проекти лабораторії комп'ютерної лінгвістики стали полігоном навчання, виробничих практик, наукових і методичних досліджень широкого спектру. Необхідно відзначити, що в результаті створення різноманітних електронних продуктів були одержані такі важливі комп'ютерні ресурси, як програми автоматичного морфологічного, контекстного і синтаксичного аналізу українських текстів, без яких неможлива жодна інтелектуальна інформаційна система.

Усі проекти створювалися колективом штатних співробітників лабораторії (випускників спеціалізації), науковим керівником якої є доц. Н. Дарчук. Завдяки їй та інженеру-програмісту В. Сорокіну підготовка фахівців з комп'ютерної лінгвістики і наукова робота лабораторії досягли такого рівня.

Паралельно з цими проектами в лабораторії розроблялися засади комп'ютерної морфології. Це електронний "Граматичний словник дієслів української мови" (проект, спільний із Лейпцигським університетом); електронний "Українсько-італійський граматичний словник дієслів" (проект TEMPUS-TASSIS, спільний із Флорентійським університетом).

Видані підручник "Комп'ютерна лінгвістика" та навчальний посібник "Термін у лінгвістичній інформатиці" – автор Н. Дарчук.

Видані монографії І. Козленко "Морфеміка сучасної української літературної мови" та створено колективний підручник "Морфологія української мови. Морфемологія. Словотвір. Па-

радигмологія" – автори: Л. Алексієнко, О. Зубань, І. Козленко. Ці структурно-прикладні розробки можна використовувати як у навчальному процесі, так і для створення нових автоматизованих систем на базі української мови.

Наступний етап діяльності лабораторії – запровадження спеціальності "прикладна лінгвістика", для якої розроблено навчальний план. Набутий досвід у спеціалізації "комп'ютерна лінгвістика" засвідчує, що підготовка сучасних фахівців – бакалаврів і магістрів – потребує не лише збільшення філологічного комплексу, а й уведення дисциплін математичного циклу, які читатимуться студентам упродовж всього періоду навчання. У цьому нас підтримали декан факультету кібернетики акад. А. Анісімов, який добре обізнаний з нашою предметною галуззю, а також його колеги та учні. Вони не тільки консультували навчальний план спеціальності, а й висловили готовність узяти участь у його реалізації.

Проведені консультації з викладачами різних кафедр Інституту філології та інших факультетів підтвердили не лише необхідність, а й готовність до запровадження спеціальності "прикладна лінгвістика" з 2013 року.

Для здійснення навчально-дослідної роботи на новому етапі вкрай потрібне фахове середовище, співпраця в дослідницьких проектах насамперед із кафедрами прикладної лінгвістики українських вишів. Це засвідчила нещодавно проведена кафедрою сучасної української мови традиційна конференція "Мова як світ світів". На секцію "Актуальні проблеми комп'ютерної лінгвістики" було надіслано понад 40 доповідей (Національний університет "Львівська політехніка", Східноєвропейський університет (Луцьк), Інститут російської мови РАН (Москва), Військовий інститут Київського національного університету імені Тараса Шевченка, Інститут української мови НАНУ (відділ структурно-математичної лінгвістики), Український мовно-інформаційний фонд НАНУ, Львівський університет, Київський національний лінгвістичний університет, Кіровоградський педагогічний інститут ім. Володимира Винниченка, Національний університет "Київський політехнічний інститут").

Під час Круглого столу учасники конференції висловили бажання на базі Київського національного університету імені Тараса Шевченка систематично проводити наукові семінари (раз на рік) з актуальних проблем комп'ютерної лінгвістики; започаткувати спільний онлайн-проект "Термінологія комп'ютерної лінгвістики (електронна база даних, тезаурус)"; організовувати "школи комп'ютерної лінгвістики" для студентів та аспірантів; розробляти спільні проекти (в режимі онлайн) із залученням бакалаврів, магістрів та аспірантів.

Масштабність та значущість здійснених і запропонованих проектів отримали схвальні відгуки з усієї "прикладної" України. Це дає підстави вважати лабораторію комп'ютерної лінгвістики Інституту філології Київського національного університету імені Тараса Шевченка науково-методичним центром спеціальності "прикладна лінгвістика".

УДК 81’322

Наталія Дарчук
Київський національний університет імені Тараса Шевченка

АВТОМАТИЧНИЙ СИНТАКСИЧНИЙ АНАЛІЗ ТЕКСТІВ КОРПУСУ УКРАЇНСЬКОЇ МОВИ

Розглядається автоматичне представлення синтаксичної структури речення на рівні словосполучення: автоматичне виокремлення словосполучення, приписування йому типу синтаксичного зв’язку (підрядного, сурядного, предикативного).

Ключові слова: автоматичний синтаксичний аналіз, словосполучення, синтаксичний зв’язок, підрядний зв’язок, сурядний зв’язок, предикативний зв’язок, ядровий підрядний зв’язок, ад’юнктний підрядний зв’язок.

Автоматичний синтаксичний аналіз (АСА) – проект, над вирішенням якого працюють розробники Корпусу української мови, співробітники лабораторії комп’ютерної лінгвістики Інституту філології Київського національного університету імені Тараса Шевченка.

На рівні словосполучення АСА передбачає автоматичне виокремлення словосполучень, приписування їм типу синтаксичного зв’язку та автоматичне укладання словників словосполучень (дієслівних, іменних, ад’єктивних). На рівні речення має здійснюватися повний синтаксичний аналіз у вигляді дерев залежностей.

За результатами роботи над проектом передбачається створити електронний алфавітно-частотний словник сполучуваностей української мови, який по завершенні дослідницької роботи буде викладений в Інтернеті на мовному порталі www.mova.info для загального користування.

Для української мови подібне лінгвістичне та програмне забезпечення розробляється вперше, тобто цей проект має унікальний характер. Це єдиний лінгвістичний ресурс, що містить синтаксичне розмічування текстів української мови, яке здійснюється тільки автоматично на базі повного автоматичного морфологічного аналізу зі знятою омонімією (про перші спроби АСА див. [Дарчук 1990; Дарчук 1999]).

У теоретичному плані вирізнення словосполучення з реченнєвої структури на великих різностильових масивах текстів, які входять до Корпусу української мови, дає можливість дослідникам української мови визначити синтаксичну і семантичну ємність цієї синтаксичної одиниці.

Необхідність вивчення сполучуваності лексичних одиниць зумовлена нерозробленістю широкого кола як теоретичних, так і прикладних проблем. Теоретичні аспекти, які потребують вивчення, – це, зокрема, граматична і лексична валентність слів, типова сполучуваність, синонімія словосполучень різних структурних типів, лексична і граматична валентність як критерій синонімічності, закони комбінаторики словосполучень різних типів і розрядів, лексична валентність як критерій розмежування вільних і фразеологічних словосполучень, взаємодія стійкості й ідіоматичності тощо. До прикладних проблем можна віднести автоматизацію лінгвістичних досліджень, автоматичне визначення меж словосполучень, установлення критеріїв членування фрази на синтагми, автоматичний синтаксичний аналіз речення, автоматичне реферування й анотування тексту на основі сполучувальнісних критеріїв тощо.

Базою для розроблення, впровадження і застосування АСА є Корпус української мови, який постійно зростає. На початок листопада 2012 р. його обсяг сягав близько 13 млн. слововживань, або близько 650 тис. речень. Отже, стала потреба у створенні потужного механізму автоматичного опрацювання українського тексту на рівні синтаксису і, відповідно, у розробленні лінгвістичного та програмного забезпечення цього ресурсу. Постали завдання створення такого типу АСА, за допомогою якого можна одержати різноманітну інформацію про функціонування граматичних синтаксичних одиниць та їх категорій. При цьому виникає дилема обсягу матеріалу і точності його опрацювання. Створення аналізатора, який би абсолютно безпомилково здійснював аналіз українського тексту, неможливе, тому якісне анотування тексту завжди пов’язане з ручним доопрацюванням.

У цьому сенсі за умови відносної обмеженості організаційних можливостей розробники Корпусу опинилися перед вибором: створення порівняно невеликого, але вивіреного корпусу чи значного за обсягом, але анотованого автоматично. Обидва підходи мають право на існування. Ми обрали другий з них: розроблення лінгвістичного і програмного забезпечення, за допомогою якого з будь-якого тексту Корпусу автоматично виділяються словосполучення з подальшою можливістю редагування одержаних даних. На цьому матеріалі, так само в автоматичному режимі, будуються словники сполучуваності для різних частин мови: окремо для слів, що виступають у ролі ядрових (“хазяїн”) і в ролі ад’юнктних (“слуга”).

Не вдаючись до теоретичних дискусій щодо деяких питань синтаксису, зазначимо, що в основі АСА лежить формально-синтаксичний аспект вивчення речення. Ані семантико-синтаксичний і функціональний, ані комунікативний підхід до розгляду речення не можуть стати основою автоматизації. Тоді як дослідження формально-синтаксичної будови речення дає можливість створити словник синтаксем, для якого попередньо слід укласти таксономічну класифікацію лексики, що у майбутньому уможливить автоматичне визначення синтаксичних відношень між членами словосполучення. Формальна граматика, адаптована для потреб автоматизації, базуватиметься на гіпотаксисі як провідному аспекті синтаксичного ладу мови; а паратаксис буде додатковим аспектом, оскільки виокремлення сурядних словосполучень з погляду автоматизації не становить суттєвих труднощів.

У ході автоматичного синтаксичного аналізу речення насамперед має здійснюватися автоматичний пошук зв’язків слів у реченні. Ознаки таких зв’язків наявні, зокрема, у словозмінних характеристиках слів. У реченні послідовно розгортається підпорядкування слів одне одному: одне слово (залежне) змінює форму, щоб адаптуватися до вимог іншого слова (головного).

Таким чином, машина має виокремлювати пари слів, пов’язані граматичним зв’язком, позначаючи напрямок залежності.

Наприклад, для речення: *Широко* (1) *обговорюються* (2) *проблеми* (3) *життя* (4) *українського* (5) *суспільства* (6). – виокремлюються такі пари слів:

(5) ← (6) українського суспільства
(1) ← (2) широко обговорюються
(3) → (4) проблеми життя
(4) → (6) життя суспільства
(2) ↔ (3) обговорюються проблеми

Цим діям можна надати алгоритмічного вигляду. Врешті отримуємо список пар залежностей. Жодних даних семантичного характеру у цьому аналізі не використовується. Єдине, що можна визначити при такому підході, – це залежність слів одне від одного і порядок їх розташування. Це і є прикладом формально-синтаксичного підходу до аналізу речення.

Словосполучення – це смислове та граматичне поєднання двох або більшої кількості слів на основі підрядного, сурядного або предикативного зв'язку [Загнітко]. Ці типи зв'язків відповідають відтворенню загальної системи відношень між компонентами описуваної ситуації у реченні. Віднесення до словосполучень тільки тих, які сполучаються підрядним прислівним зв'язком, не є вичерпним з точки зору складників речення [Вихованець].

Ми вважаємо, що словосполучення – відносно самостійна одиниця мови, що виділяється у межах речення, будується за законами поєднання слів, виявляє у мовленні валентні властивості головного слова, має мовні моделі, відтворювані у мовленні. Не є словосполученнями: складені аналітичні поєднання слів, зокрема сполуки іменника з прийменником (*через міст, в інституті*); складені аналітичні форми слів (*буду читати, більш досвідчений*); фразеологізми (*ні пари з вуст, бити байдики*).

Завданням АСА є виявлення всіх різновидів сполучуваності – предикативної, підрядної і сурядної – кожного слова з текстів. Граматичні характеристики словосполучення безпосередньо залежать від того, до якої частини мови належить слово-"хазяїн", тому що лексико-граматична природа слова визначає його здат-

ність сполучатися з іншими словами. Відповідно до цього словосполучення поділяють на іменникові, прикметникові, займенникові, числівникові, дієслівні та прислівникові. За виробленою концепцією АСА при виокремленні словосполучень було передбачено попередній етап створення **словника валентностей** для дієслова (31 206 правил), іменника (40 023), ад'єктива (6205), а також словника фразеологізмів (близько 3000 одиниць).

За складом словосполучення поділяють на прості, складні та комбіновані. Ми виділяємо тільки прості бінарні словосполучення, які можуть бути поширені у складні або комбіновані автоматизовано, оскільки при визначенні їх складу потрібен аналіз смислової структури.

Якщо у сполуках у головній позиції слово інформативно недостатнє, а залежне слово цю недостатність заповнює (так звані доповнювальні, або комплетивні відношення), то вони розглядаються як словосполучення, що виконують функцію одного члена речення, наприклад: *дехто з присутніх, четверо з них, почав працювати* і под.

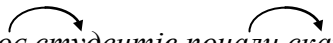
Ми відмовилися від традиційного поділу підрядного зв'язку на підвиди – узгодження, керування та прилягання. Ґрунтуючись на широкому розумінні поняття синтаксичного зв'язку, всі вони є випадками приєднання до головного слова відмінкової форми іменника або субстантива. При узгодженні залежне слово уподібнюється до головного в усіх його граматичних формах, а при приляганні воно, не маючи форм словозміни, приєднується до головного за змістом. Крім того, останнім часом у деяких роботах випадки поєднання з головним словом залежної форми іменника з атрибутивним чи обставинним значенням трактуються як прилягання [Русская граматика : 21]: *працювати лікарем* – зв'язок керування, а *прогулюватися парком* – прилягання; *допомога матері* – керування; *пам'ятник поетові* – прилягання.

Причиною такої відмови є ще й неможливість у деяких випадках автоматично, не вдаючись до аналізу значення кожного з членів словосполучення, визначити його тип. Алгоритм АСА має спиратися виключно на морфологічні форми слів (орудний відмінок залежного слова у першій парі; давальний у другій).

Підрядні зв'язки поділяються нами на ядрові і неядрові. Ядровим називаємо такий зв'язок, при якому аналізоване слово є керувальним, головним. Наприклад, у реченні – *Від економічної кризи сильно постраждали майже всі європейські держави*. – спостерігаємо такі ядрові зв'язки: ***кризи*** домінує над *економічної*; ***постраждали*** домінує над *від*; ***від*** домінує над *кризи*; ***постраждали*** домінує над *сильно;* ***всі*** домінує над *майже*; ***держави*** домінує над *всі*; ***держави*** домінує над *європейські*. Неядровий зв'язок – це такий зв'язок, при якому аналізоване слово є залежним, керованим. У попередньому прикладі неядрові зв'язки є у словах *економічної* (залежить від *кризи*), *сильно* (залежить від *постраждали*), *європейські* (залежить від *держави*) тощо.

Предикативний зв'язок – це зв'язок між основними компонентами речення "підмет – присудок", який ґрунтується на їхній двобічній залежності. У предикативній парі жодне зі слів не можна вважати домінованим, вони обидва є однаково домінувальними.

Якщо підмет або присудок виражений складеним словосполученням, то визначається підрядний зв'язок для аналізованого слова, наприклад:



Предикативний зв'язок установлюємо між *двоє студентів* і *почали скандувати*. В межах словосполучення *двоє студентів* ядровим буде ***двоє***, яке домінує над *студентів* (*студентів*, відповідно, має неядровий зв'язок); у словосполученні ***почали*** *скандувати* ядровим буде ***почали***, яке домінує над *скандувати*. Те саме стосується іменного складеного присудка:

Ядровий зв'язок встановлюємо між підметом *він* і присудком *став студентом*. У межах іменного складеного присудка ***став*** *студентом* ядровим буде допоміжне дієслово ***став***, тому що

воно домінує над іменною частиною, вираженою іменником в орудному відмінку *студентом*.

Сурядний зв'язок – це зв'язок, при якому жодне із взаємопов'язаних слів не є ані домінувальним, ані домінованим. Вважається, що два слова знаходяться в сурядному зв'язку, якщо кожне з них підпорядковане одному й тому ж третьому слову, якщо вони пов'язані через сполучник між собою або відокремлені одне від одного комою. При цьому ми дотримуємося таких умов, що сурядний зв'язок установлюється між словами, а не між словом і зворотом або синтаксичною конструкцією. Наприклад, сурядний зв'язок є між словами *чесними і прозорими* (*Вибори були чесними і прозорими*) і його немає у такому прикладі: *Президент був задоволений і в гуморі*.

Щодо ад'єктивів (прикметників, дієприкметників, займенників), які виконують одну й ту саму функцію, прийняті такі домовленості:

- якщо між ними є сполучник, то напрямок зв'язку такий: від головного слова до кожного з прикметників, а потім прикметники із сполучником, наприклад:

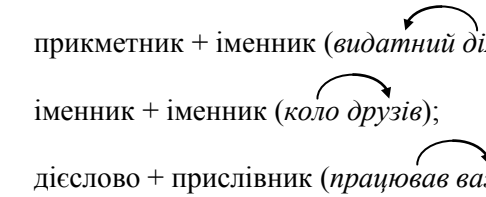
(порядні **банкіри** [ІС/ПЯ]), достойні **банкіри** [ІС/ПЯ], **порядні і достойні** [СУ]);

- якщо між ними безсполучниковий зв'язок, то сурядні зв'язки встановлюються з кожним із ад'єктивів та іменником (у будь-якій формі), а потім між самими ад'єктивами, наприклад:

порядні, достойні банкіри

(порядні **банкіри** [ІС/ПЯ]), достойні **банкіри** [ІС/ПЯ], **порядні**, **достойні** [СУ]).

Кожний тип словосполучення відображається у певному виді моделі. Модель словосполучення – це двоелементна формула, що відбиває один із типів зв'язку аналізованого слова з певним повнозначним словом, наприклад:



У тих випадках, коли прийменник (або сполучник) служить лише засобом зв'язку між двома повнозначними словами, він не вважається самостійним членом моделі. Таким чином, модель "дієслово + прийменник + іменник" (*працювати в уряді*) залишається двочленною, хоча складається з трьох слів. У проектованому словнику подаються чотири типи моделей: ядрові; неядрові (ад'юнктні, які відображають підрядні зв'язки); сурядні; предикативні.

АСА здійснюється за правилами – моделями, представленими у таблиці SyntaxRules у програмному середовищі Access, яка виконує роль диспетчера. Кожній моделі згідно з таблицею автоматично приписується певний код. Зокрема, за цією таблицею здійснюється "збирання" в один вузол складених морфологічних та синтаксичних явищ, наприклад: ГБ – аналітичний майбутній час (*буду читати*); ГЗ – аналітичний наказовий спосіб (*хай читає*); ГЧ – умовний спосіб (*читав би*); СЧ – складений числівник (*сорок три*); ПМ – складений підмет (*один з них*); ПС – складений присудок (*почав працювати*). Простим присудкам і безособовим формам дієслова приписуються коди: ПР та ГЧ відповідно. Причому в першому випадку далі ведеться пошук підмета, а в другому пошук продовжується за таблицею дієслівної валентності.

Типи синтаксичних словосполучень кодуються за частиномовною належністю: ІС – іменникове; АС – прикметникове; ДС – дієслівне; ЧС – числівникове; РС – прислівникове; ЗС – займенникове. За цією ж таблицею кодуються види синтаксичних зв'язків (vpr): КЗ – координація; ПЯ – підрядний ядровий; ПА – підрядний ад'юнктний; СУ – сурядний. Напрямки перевірки (праворуч / ліворуч) строго регламентуються набором правил для конкретного рядка, якими визначається і пріоритет у роботі групи правил.

Як свідчить тестування результатів роботи АСА, автоматично виділялися такі словосполучення і такі типи зв'язків, які повністю відповідають інтуїтивним уявленням носіїв української мови та експертів-лінгвістів. Отже, синтаксичну структуру речення автомат "зрозумів" правильно. І це відкриває широкі перспективи, зокрема можливість укладання частотного словника сполучуваностей української мови та здійснення автоматичного синтаксичного аналізу цілого речення. У свою чергу, правильний синтаксичний аналіз є запорукою створення автоматичного семантичного аналізу тексту.

1. *Вихованець І. Р.* Граматика української мови. Синтаксис / І. Р. Вихованець. – К., 1993. 2. *Дарчук Н. П.* ЭВМ в синтаксических исследованиях / Н. П. Дарчук // Использование ЭВМ в лингвистических исследованиях. – Киев, 1990. – С. 113–129. 3. *Дарчук Н. П.* Сегментация сложного предложения на составляющие предикативные части / Н. П. Дарчук // Синтаксический анализ научного текста на ЭВМ. – Киев, 1999. – С. 40–127. 4. *Загнітко А. П.* Основи українського теоретичного синтаксису. Частина 1 / А. П. Загнітко. – Горлівка, 2004. 5. Сучасна українська літературна мова. Морфологія. Синаксис. – К., 2010. 6. Русская граматика: в 2 т. Т. 2. Синтаксис. – М., 1980. – С. 21.

Рассматривается автоматическое представление синтаксической структуры предложения на уровне словосочетания: автоматическое выделение словосочетания, приписывание ему типа синтаксической связи (подчинительной, сочинительной, предикативной).

Ключевые слова: автоматический синтаксический анализ, словосочетание, синтаксическая связь, подчинительная связь, сочинительная связь, предикативная связь, ядерная подчинительная связь, адъюнктная подчинительная связь.

The article provides automated representation of syntactic analysis of sentence on the level of word-combination: automated marking of word-combination, characterising with a kind of syntactic relations (subordinating, coordinating conjunction, predicative relation).

Keywords: automated syntactic analysis, word-combination, syntactic relation, subordinating conjunction, coordinating conjunction and predicative relation.

Стаття надійшла до редакції 10.09.2012

УДК 81’322

Вікторія Завадська
Національний технічний університет
“Київський політехнічний інститут”

КОЛИ “ВІКНО” НЕ Є ВІКНОМ, АБО ЩЕ РАЗ ПРО СУЧАСНУ УКРАЇНСЬКУ ІТ-ТЕРМІНОЛОГІЮ

Розглядаються особливості перекладу термінів українською мовою, зокрема в ІТ-сфері, в галузі комп’ютерної термінології. Окреслено проблеми, пов’язані з мовою-посередницею та явищем калькування термінів, вказано причини появи термінів, що є нехарактерними для української літературної мови. Подано деякі рекомендації практичного характеру.

Ключові слова: переклад, термін, ІТ-термінологія, ІТ-переклад, стандартизація термінів.

Проблема стандартизації української ІТ-термінології є складною і неоднозначною з багатьох причин. У статті спробуємо окреслити деякі труднощі перекладу і впровадження ІТ-термінів, а також причини виникнення цих труднощів. Приводом для написанні цієї статті стали особисті спостереження у сфері технічного перекладу, зокрема в галузі ІТ-технологій. Складнощі, які виникають під час перекладу й подальшого узгодження термінів між виконавцями (зазвичай в особі перекладачів комерційних фірм), замовниками (компаніями, що створюють і “просувають” продукт на українському ринку) та звичайними користувачами, які не хочуть змиритися з дивними, на їхню думку, українськими новотворами, часто гальмують процес формування саме українського (а не запозиченого іншомовного) термінологічного глосарію. І навіть чинні стандарти [ДСТУ 3966-2009] мало допомагають, коли у практичному застосуванні раптом відбувається бойкотування терміна користувачами продукту.

Утворення і впровадження терміна в загальнонауковий апарат – процес складний і неоднозначний, який проходить декілька стадій: первісне винайдення терміна науковцем або групою вчених чи творців продукту; співвіднесення його з явищем (де-

нотатом); первісне прийняття його галузевою науковою спільнотою; впровадження терміна на загальносвітовому рівні через представлення його в англомовних виданнях; поширення терміна в різних країнах світу шляхом перекладу його з англомовних джерел; нарешті, необов'язковий етап – поширення терміна в широкому вжитку за межами галузевої спільноти (як правило, через користувачів певного продукту або завдяки термінологічній омонімії – проникнення терміна в інші шари лексики, де він може набути інших значень, втративши своє первісне значення, або ж доповнитися новими семантичними відтінками).

Хочемо наголосити на етапі, коли термін перекладається англійською мовою й через неї поширюється в інших мовах. Англійська мова через свій міжнародний статус і найбільшу популярність порівняно з іншими мовами, стає своєрідною посередницею. Отже, дослівний переклад на English передбачає такий самий переклад на інші мови без втрати значення чи набуття нових. Наведемо кілька прикладів з різних галузей науки. На початку XX століття російський математик Марков винайшов явище, яке було названо "цепи Маркова" (рос.). У зарубіжних виданнях термін було перекладено як *Markov chain* (англ.), відповідно далі його перекладали вже з англійської *Markov-Ketten* (нім.), *Markov-keten* (голланд.), *cadena de Markov* (іспан.) тощо.

Отже, англійська мова перебрала на себе роль посередниці, якою довгий час була, скажімо, латина (нині вона залишається інтернаціональною мовою в галузі медицини та фармакології). Неангломовний термін, з'являючись у певній науковій галузі, набуває міжнародного визнання через дослівний переклад або транслітерацію англійською мовою.

Дещо інакше цей процес відбувається у сфері ІТ-технологій, що почали бурхливо розвиватися лише у другій половині XX ст., відповідно, й термінологія даної сфери є новітньою, а процес її уніфікації та застосування ускладнюється багатьма факторами. По-перше, творцем комп'ютерної технології є не одна людина, а переважно колектив авторів. По-друге, ці технології здебільшого розробляються в англомовних країнах, отже, перві-

сне творення термінів відбувається англійською мовою. Потретє, користувачами цих термінів є не лише фахівці певної галузі, а й широкі кола нефахівців в усіх країнах світу. Це веде до того, що шлях популяризації терміна ускладнюється за принципом "скільки людей – стільки думок", а процес його стандартизації стає залежним від так званої "популярності" того чи іншого варіанту назви. Так, компанія Microsoft Ukraine на сайті YouTube запропонувала список термінів для обговорення та голосування. Цікаво, що в подальшій практиці компанія вимагає від перекладачів застосування тих термінів, за які проголосувало найбільше відвідувачів, незалежно від семантичної відповідності терміна його суб'єкту.

Така ситуація часто призводить до нерозуміння природи виникнення деяких термінів, які на широкий загал сприймаються як нехарактерні для української мови, проте відтворюють зміст первісного терміна й за граматичними законами мають право на існування. Наведемо лише один із численних прикладів, який викликає багато суперечок. Термін *application*, що російською мовою був перекладений як 'приложение', в українській мові перекладається як 'застосунок' або 'застосо́вання'. Цим терміном позначається середовище, в якому працюють комп'ютерні програми. Львівський мовознавець М. Гінзбург переконливо доводить право на існування саме цих словоформ, ґрунтуючись на семантичній відповідності слова дії (як недоконаного виду), події (як доконаного виду) або ж наслідку. Application – це програма для комп'ютера, що працює під управлінням операційної системи, отже, вона поєднує в собі суб'єкт і дію, які передані російським віддієслівним іменником 'приложение', а в українській мові процес (як дія), доконаний процес (подія) та його наслідок передаються трьома різними словами: двома дієсловами та похідним віддієслівним іменником, наведеним вище [Гінзбург 2004; Гінзбург 2008].

Проблеми виникають також через те, що в галузі ІТ-перекладів майже немає фахівців з технічною та лінгвістичною освітою. Оскільки "чистий" філолог не може якісно здійснити тер-

мінологічний переклад у певній сфері, компанії надають перевагу людям, що мають відповідну фахову освіту, проте не є професійними перекладачами.

І, нарешті, найголовніше: світ комп'ютерних технологій є світом віртуальним, отже, й більшість термінів цієї сфери має асоціативний характер, коли звичайне слово набуває іншого, "комп'ютерного", проте цілком зрозумілого на інтуїтивному рівні значення. Можливо, найпростішим і найпоширенішим прикладом є термін *вікно*, яке первісно позначає отвір у споруді для проникнення світла та вентиляції, проте в сучасних молодих людей це слово дедалі частіше викликає і комп'ютерні асоціації. І хоча з лінгвістичного погляду це слово увійшло до сфери ІТ як жаргонізм, сьогодні воно внормований термін, що застосовується в підручниках та спеціальній літературі. Отже, бачимо приклад не лише трансформації поняття, а й перехід слова з одного лексичного класу до іншого. Такий прийом термінотворення застосовується, вочевидь, для полегшення користування ІТ-витворами: *процесор* – похідне від *процес*; *материнська плата* – перший елемент у словосполученні вжитий задля асоціативно-емотивного підйому – це те, без чого не може існувати комп'ютерна система, його основа; *меню* – вочевидь, перелік; *архітектура комп'ютера* – компонування, будова його складових; *ядро* – центральна частина операційної системи; *миша* – пристрій для перетворення механічних рухів користувача на рухи курсору на екрані і под.

Спростивши все розмаїття підходів, що застосовуються фахівцями, можна виділити два основні типи перекладу: *трансформаційний та денотативний* [Мирам : 44]. У другому випадку відбувається сприйняття повідомлення, формування образу й подальша інтерпретація його за допомогою мовних засобів – тобто передбачається більш вільне поводження із терміном, оскільки з'являється момент асоціативно-суб'єктивний, відповідність між словом та поняттям вибудовується за правилом непрямого, асоціативного зв'язку. Наведемо ще один приклад – з металургійної галузі, цікавий зворотною відповідністю слова та його значення. У традиційному розумінні слово *збагачення* має семантику приєднання, накопичення, поповнення, а в металургійній промисловос-

ті *збагаченням руди* називають методи розділення металів та мінералів, тобто йдеться якраз про вилучення чогось із загалу.

Трансформаційний метод перекладу базується на "перетворенні об'єктів і структур однієї мови на об'єкти і структури іншої мови за відповідними правилами" [Мирам : 44] на всіх мовних рівнях. Тобто йдеться про певну дослівність із добором відповідного контекстного варіанта. Цей підхід є більш прийнятним для перекладу технічних текстів загалом, але в ІТ-сфері постає нагальна проблема: у ІТ-перекладах, як правило, відсутній контекст. Здійснюється переклад рядків, а не тексту – така особливість цієї галузі. У такому випадку перекладач не розуміє, як саме йому перекласти багатозначне слово або ж яку граматичну форму обрати, наприклад, *select* може бути перекладене як 'виберіть', 'вибір' або ж 'вибрати' – залежно від контексту. Для перевірки правильності зробленого перекладачем вибору існує етап перевірки – контекстний перегляд, що є останнім у процесі перекладу. Рев'ювер бачить усі рядки перекладу, але вже вставлені в інтерфейс, тобто в такому вигляді, як це побачить користувач на екрані, – тому може зробити висновок про хиби або навіть про довжину рядків, що часто мають бути коротшими (оскільки в англійському оригіналі слова зазвичай коротші). Так, англійське *OK* можна перекласти і як 'гаразд', і як 'так', і як 'підтвердити'. Задля уникнення неоднозначності його зазвичай так і залишають неперекладеним.

І тут постає ще одна невідповідність між теорією та практикою перекладу: чи варто перекладати або заміняти розширеними поняттями слова, які звучать у різних мовах однаково і є для однієї з них словами іншомовного походження? Один із головних постулатів теорії перекладу стверджує, що потрібно намагатися максимально замінювати іншомовні слова відповідниками мови перекладу. Проте важливим у перекладі термінів є власне призначення тексту: науковий обіг чи комерційний проект. У другому випадку, окрім традиційних глосаріїв та словникових установок, є ще вимоги замовника, а також особливості власне видавців/перекладачів продукції. Таким чином, перед перекла-

дачами постає проблема: чи можна той самий іншомовний термін, який в одному тексті має різні відтінки значень, перекладати не "дослівно", а підміняти тим чи іншим варіантом його власного визначення. Наприклад, *menu* перекладати як 'група кнопок' або 'група прапорців у діалоговому вікні', або просто залишити іншомовним відповідником 'меню', який буде охоплювати всі варіанти значень у тексті без уточнень.

Що ж до української ІТ-термінології, то вона часто калькує терміни, і не лише з англійської мови, а й з російської, яка в такому випадку стає мовою-посередницею. Яскравим прикладом є термін *check box*, дослівний переклад якого 'коробка/шухлядка контролю'. У російській мові переклали цей термін як *флажок*, зорієнтований на денотат – межу, позначену прапорцями, за яку не можна заходити. Український відповідник цього поняття позначає термін *прапорець*, білоруський – *сцяжок*. Порівняймо з іншими мовами: у польській мові – це *pole wyboru*, у чеській – *zaškrtávací políčko*, у литовській – *žymės langelį*, в італійській – *casella di controllo*, а в есперанто – *kontrolu skatolo*, що є найбільш показовим, адже штучні мови, як відомо, намагаються уникнути асоціативності. Отже, бачимо калькування українського терміна (як і багатьох інших) з російської мови, що в нашій ситуації двомовності може сприйматися неоднозначно. Проте факт залишається фактом: подібні скальковані з російської переклади вже увійшли в галузевий обіг і міцно вкоренилися.

Використання в перекладі мов-посередниць, зокрема російської, як посередниці в перекладі з англійської, може мати вельми негативні наслідки. Повернімося до терміна *application*, правильним українським перекладом якого, як було зазначено вище, є *застосунок* або *застосування*. Цікаво, що раніше був поширений некоректний переклад цього терміна – *додаток*. Це типова помилка подвійного перекладу: термін *приложение* можна тлумачити англійською і як 'application', і як 'appendix'. Якби в українських перекладачів не було звички перекладати через мову-посередницю, цієї прикрої помилки не виникло б. Таким чином, термін *приложение* є неоднозначним:

Ми окреслили основні проблеми ІТ-перекладу, які мають певні особливості порівняно з перекладом художнім. Слід очікувати, що з розвитком технологій (машинного та автоматизованого перекладу) ці відмінності будуть увиразнюватися. Це означає, що з накопиченням практичного досвіду, має бути створена і спеціальна теорія ІТ-перекладу. Сподіваємося, що висловлені міркування спонукатимуть до досліджень у зазначеному напрямку.

1. ДСТУ 3966-2009. – [Чинний від 2010-07-01]. – К., 2010. – (Національний стандарт України). 2. *Гінзбург М. Д.* Application, як це українською? / М. Д. Гінзбург // Інформаційно-керуючі системи на залізничному транспорті. – 2004. – № 3. – С. 81–84 [Електронний ресурс]. – Режим доступу: http://msu.kharkov.ua/tc/cons/applic.html#q16. 3. *Гінзбург М. Д.* Синтаксичні конструкції у фахових текстах. Практичні висновки з рекомендацій мовознавців / М. Д. Гінзбург // Вісник: Проблеми української термінології. – Львів, 2008. – № 620. – С. 26–32 [Електронний ресурс]. – Режим доступу: http://msu.kharkov.ua/tc/zasady/sintax.html. 4. *Мирам Г. Э.* Профессия: переводчик / Г. Э. Мирам. – К., 1999.

Рассматриваются особенности перевода терминов на украинский язык, акцентируется на различиях перевода в сфере компьютерной терминологии и других видов перевода. Также обозначены проблемы, связанные с языком-посредником, указаны причины появления терминов, не характерных для украинского литературного языка, и даны некоторые рекомендации практического характера.

Ключевые слова: перевод, термин, ІТ-терминология, ІТ-перевод, стандартизация терминов.

Some features of terminology translation into Ukrainian are considered. The key point of the article is the difference between IT-terminology translation and other types of translation. Also problems related to use of intermediate translation language are outlined and origination of unnatural, not agreed with language standard, terms in Ukrainian is discussed. Some practical advices are formulated.

Keywords: translation, term, IT terminology, IT translation, standardization of terms.

Стаття надійшла до редакції 19.09.2012

УДК 81’322

Лариса Кислюк
Інститут української мови НАН України

ПОДАННЯ НОВОТВОРІВ У ЛЕКСИЧНІЙ БАЗІ ДАНИХ

Запропоновано використання основоцентричного і формантоцентричного підходів до подання новотворів у лексичній базі даних. Порівняно активність системних словотвірних моделей у колективній та індивідуальній мовній практиці.

Ключові слова: основоцентричний підхід, формантоцентричний підхід, словотвірна модель, новотвір.

Сучасний український лексикон активно поповнюється великою кількістю нових похідних слів. Дослідники сучасних слов’янських мов називають роль словотворення в сучасній номінації провідною, оскільки на одне непохідне слово припадає в середньому 4–5 дериватів. Перевага новотворів в інноваціях підтримує типологічні риси української мови як мови флективної з перевагою синтетизму в її номінації. Безперечно, існує потреба в лексикографічній фіксації нової похідної лексики. Для української мови можна назвати низку словників, які подають нову лексику. Це серія словників: “Лексико-словотвірні інновації” А. Нелюби, які перевидано окремою книгою “Словотворчість незалежної України 1991–2011” (Харків, 2012), “Нові слова та значення” Л. Василькової та Л. Туровської (Київ, 2008), “Нові й актуалізовані слова та значення: словникові матеріали” колективу авторів відділу лексикології та лексикографії Інституту української мови НАН України під керівництвом О. Тищенко (Київ, 2010), словник-довідник “Нове в українській лексиці” Д. Мазурик (Львів, 2002) та інші. Загальний принцип подання таких одиниць – алфавітний, з тлумаченням і/або контекстом уживання реєстрової одиниці та його паспортизацією.

Проте сьогодні назріла потреба перейти від опису окремих нових одиниць до систематизації мовних явищ, оскільки 20 років незалежності – це часовий проміжок, достатній для форму-

вання нового покоління носіїв мови. У відділі структурно-математичної лінгвістики Інституту української мови НАН України запропоновано ідеографічний або тезаурусний принцип упорядкування нової лексики. Співробітники для визначення стабілізації нової лексики успішно застосовують поняття функціонального потенціалу інновації, яке обґрунтувала Є. Карпіловська. Вона розуміє його як сумарний показник парадигматичних (ієрархічних, гіперо-гіпонімічних), епідигматичних (деривацій-них та асоціативних) і синтагматичних (позиційних і комбінаторних) відношень інновації в системі мови й у тексті, її номінаційної та комунікаційної активності [Карпіловська: 6]. Основною дослідницькою базою стали матеріали комп'ютерного фонду лексико-словотвірних інновацій (КФІ) у сучасній українській мові, створеного співробітниками відділу за текстами україномовних друкованих, електронних та інтернет-видань періоду незалежності, обсягом понад 20 тисяч одиниць із так званим діагностичним контекстом і його паспортизацією. До його складу входять номінації, не зафіксовані у словниках радянської доби (виданих до 1991 р.).

КФІ укладено як дослідницький корпус нової лексики, де мікроконтекст і його паспортизація мають важливе значення: по-перше, джерело має бути показовим і авторитетним, чи це періодичне видання, чи зразок авторської белетристики або художньої літератури. Пошук нових одиниць відбувається вручну через аналіз і розмітку тексту. Досвід вивчення динаміки сучасних мов доводить, що навіть за наявності національного корпусу мови у відкритому доступі, у такій дослідницькій роботі він не допоміг би розв'язанню багатьох поставлених завдань, передусім з'ясуванню тенденцій стабілізації нової лексики. Для української мови, на жаль, взагалі досі не маємо такого показового корпусу сучасних текстів.

Друге важливе питання, яке потребує фахового аналізу дослідника – надання новим лексемам статусу: нова, актуалізована, нова форма, нове значення. Недостатньо протестувати пев-

ну лексему на наявність чи відсутність її, скажімо, у реєстрі академічного тлумачного "Словника української мови" в 11-ти томах (далі – СУМ). Якщо це неосемантизм, це можна з'ясувати лише через аналіз достатнього мікроконтексту, який має виконувати діагностичну роль, та зіставлення його з наявними дефініціями у словниках. Якщо це новотвір, лексему треба перевірити і за приступною лексикографічною спадщиною кінця XIX – першої половини XX ст.: можливо, це актуалізована, а не нова лексема тощо.

Актуальною є проблема не лише вияву новотворів, а й упорядкування нової похідної лексики в лексичній базі даних, якою є структурована й формалізована сукупність інформації про лексеми як об'єкт опису. В основі будь-якої формалізації інформації лежить її структурування, докладний опис складників і відношень між ними. Словотвір – це ієрархічно структурована система категорій, розрядів, типів і моделей, тому нову похідну лексику можна системно ієрархічно упорядкувати.

У пропонованій статті описано структуру фрагменту ідеографічного словника нової лексики як лексичної бази нового типу, бази відомостей про нову українську лексику. Об'єктом опису є нові похідні лексеми, засвідчені в сучасній мовній практиці. Їм відведено в структурі словникової статті окрему зону, де описано епідигматичні (деривaційні) відношення реєстрової одиниці.

Заслуговують на увагу два підходи до вивчення новотворів: основоцентричний, запропонований І. Ковaликом [Ковалик 1979 : 11] і пропагований учнями його дериватологічної школи [Грещук : 6–38], де основа є типологізувальним чинником у дериватології, та традиційно поширений формантоцентричний.

Основоцентричний підхід допомагає систематизувати похідну лексику, зокрема й нову, яка об'єднується у словотвірні гнізда навколо так званих "ключових слів доби" (вислів О. Земської [Земская : 92]). Ключові слова, набуваючи піку активності, можуть утворювати гнізда спільнокореневих слів обсягом від деякі-

лькох десятків до декількох сотень одиниць. Тоді реєстровою одиницею буде базове слово гнізда, наприклад лексема *держава* сьогодні формує гніздо з понад сотнею похідних, із яких декілька десятків – новотвори. Нові похідні одиниці входитимуть до правої частини словникової статті конкордансного типу. Новотвори в зоні епідигматики погруповані за способом творення (афіксальним, усіченням, осново- й словоскладанням – враховуючи, в початковій чи кінцевій позиції розміщена реєстрова одиниця, тобто є вона категоризатором чи ідентифікатором), а також за ступенем похідності у словотвірному гнізді. До кожного новотвору подано діагностичний мікроконтекст обсягом від декількох слів, як правило речення, а за потреби й двох речень з паспортизацією джерела. Мікроконтексти з інноваціями КФІ вибрано вручну з допомогою комп'ютера з масиву українськомовних ЗМІ 2000–2010 років, а також із публіцистичної та частково художньої літератури, що допомагає повніше подати активну мовну практику вживання нових одиниць не тільки в узусі – колективній мовній практиці, а й в авторському, індивідуальному, вживанні.

Отже, основоцентричний підхід подання новотворів у лексичній базі передбачає об'єднання в межах однієї словотвірної статті спектру спільнокореневих похідних, належних до одного словотвірного гнізда з можливістю комплексно показати особливості функціонування кожної нової одиниці. Покажемо це на декількох прикладах зони епідигматики словникової статті *держава*. Серед афіксальних простих похідних, утворених суфіксальнм способом, є новотвір *державонька*, використаний О. Забужко в романі "Музей покинутих секретів" (2010). До нього наведено контекст: *...в нашій бандючній державоньці нічого не робиться просто*...(с. 253), словосполучення *бандючна державонька* показує, що вжито це слово із сарказмом, тобто демінутивний суфікс набуває в авторки негативного значення. Інший суфіксальний дериват *державник,* відсутній в СУМі, але наявний в академічному "Російсько-українському словнику"

(1932) за ред. А. Кримського та С. Єфремова та в "Словарі української мови" (1907–1909) Б. Грінченка зі значенням 'державна людина' (Т. 1, с. 370). Отже, новотвір *державник* у цьому значенні – активізована лексема. Наведені з КФІ контексти демонструють спектр його сучасного вживання в узусі: *Державник понад усе має на оці збудування власної держави*; *Політик думає, де він у січні буде, і платить будь-чим, навіть економічною незалежністю, а державник думає, де Україна буде через рік*…. Тексти дозволяють виявити розвиток нового значення в деривата *державник* – 'студент, який вчиться за державний кошт, на противагу *контрактник'*, за приступним контекстом: *Державникам розбігтися не дають. Перед закінченням ВНЗ студент-державник отримує направлення на роботу.* ("Блог дядька Пола", 13 жовтня 2009 р., http://isaidmeow.blogspot.com/2009/10/blog-post_12.html); *Студент, державник, контрактник*… ("Універсум", студентсько-викладацька газета Житомирського державного університету імені Івана Франка, № 55, 2008, http://universum.zu.edu.ua/old/arhive.php?sid=2&yid=55); що надає йому статусу неосемантизму з ремаркою *розмовне*. Формується антонімічна пара новотворів: *державник – недержавник* 'той, хто працює в недержавній, приватній фірмі' (*Охоронці-недержавники даремно нарікають на суворий контроль з боку Державної служби охорони. Державна служба охорони має безперечні переваги перед* "*недержавниками*" (інтернет-видання "UA-Reporter.com", 5 лютого 2009, http://ua-reporter.com)) та омонімічний ряд: *недержавник* 'той, хто ігнорує або не дбає про державу' – *недержавник* 'той, хто працює в недержавній, приватній фірмі' – *недержавник* (розм.) 'студент-контрактник'. Російська дослідниця О. Єрмакова вбачає в цьому повторне використання тієї ж моделі словотворення. Мікроконтексти допомагають виявляти синонімічні відношення новотворів, а також диференціацію їх за значенням: *бездержавник* (і про народ, і про окрему особу) 'хтось (щось) без держави' і *недержавник* 'той, хто ігнорує або не дбає про державу': *Прези-*

дент України <…> переформатовує населення в народ, бездержавника в державника, вибудовує довгострокову стратегію нашого національного успіху (Л. Григорович 08.02.2008, "Украинские Итоги", березень 2008, http://bilozerska.livejournal.com/58734.html); *І коли, наприклад, бездержавник живе в якійсь державі, то він підлягає, безумовно її законам* (www.plastusa.org/ypy/pdfs/VelykaHra.pdf).

Інший новотвір – дієслово *державити* – виявлено лише у творах художньої літератури, тобто в письменницькому ідіолекті: *За тобою правда – державити треба міцно!* (Р. Іванченко); *…бо ти в цей світ державити прийшла…* (В. Стус); *…переливати епос порожнього в модерн пустого, обпиватися кавою, державити…* (Є. Пашковський).

За алфавітом у межах одного ступеня похідності подано окремо афіксальні (суфіксальні, потім префіксальні, конфіксальні) похідні; *бездержавність, недержавний, бездержав'я, міждержавний*; далі, за їхньої наявності, безафіксні прості похідні, потім складні слова: композити, де реєстрову одиницю подано в початковій позиції (функції категоризатора / ідентифікатора) *державотворення, державоспотворення, держаморальність, держгодівниця, держпіар*, та в кінцевій позиції (функція ідентифікатора / категоризатора) *гіпердержава, квазідержавний*, і юкстапозити *держава-донор, держава-монстр* і *президент-державник, політик-державник* тощо. Серед новотворів засвідчено не лише нейтральну, а й експресивно-оцінну лексику. Для неї запропоновано відповідні ремарки.

Формантоцентричний підхід передбачає подання в одній словниковій статті інновацій, утворених за спільною словотвірною моделлю, тобто загальною формулою однотипних утворень, аналогом їхньої словотвірної структури [Кубрякова : 35], [Клименко : 7]. Тоді об'єднувальним компонентом виступає спільний формант в афіксальних простих похідних або активна основа чи слово в складних словах. Словотвірний формант або активна основа стають самостійною реєстровою одиницею слов-

никової статті. Новотвори подано як члени словотвірного ряду, об'єднані спільною словотвірною моделлю.

Наприклад реєстрова одиниця – активізований формант **-изацій(а) / -ізацій(а)** продуктивний в абстрактних іменниках – назвах дій, занять, процесів, станів; виявляє нові ознаки: він активний у дієйменниках, утворених за аналогією, без мотивування відповідними дієсловами, зокрема від власних назв, абревіатур, словосполук з оцінним значенням, стилістичною невідповідністю поєднуваних основ та суфіксів: *берегинізація, гривенізація, жебракізація.*

Матеріал дозволяє розмежувати стилістично нейтральні утворення, терміни або лексеми, наближені до термінів: *бартеризація, ваучеризація, віртуалізація, глобалізація, екологізація, інституалізація,* і розмовні, стилістично-знижені: *мюзиклізація, дебілізація, кайфізація, тюрмізація, перевертнізація, футболізація.* Мікроконтексти допомагають розвести антоніми *доторканізація : недоторкання;* синоніми*: брендування : брендизація; ожебрачування : жебракізація;* або ж пароніми*: політикування : політизація* тощо.

Реєстровою одиницею може бути активізована основа, наприклад **само-**. Можна спостерігати активізацію вживання **само-** зі значенням 'себе' з іменниками (найбільше з віддієслівними іменниками на **-ння**), *самопроголошення, самобачення, самостояння,* а також із дієсловами, прикметниками, прислівниками, дієприкметниками: *самоагресія, самовибір, самовисуванець, самовисуванство, самозберегтися, самопроголошений*. Використання авторських текстів дає змогу порівняти активність уживання певних словотвірних моделей. Наприклад, за матеріалами "Словника поетичної мови Василя Стуса (рідковживані слова та індивідуально-авторські новотвори)" Л. Оліфіренко (К., 2003), можна простежити, що В. Стус активно використовує моделі складних слів з першою основою **само-**. Таких у нього 72 одиниці, з них 39 віддієслівних іменників на **-ння** *самовивищення, самодонищення,*

самонарікання, самопочезання, самопроминання, самоструму-вання та інші, що дає підстави зробити висновок, що словотвірна модель ‘**само-** + віддієслівний іменник на **-ння**’ є найактивнішою, найзатребуванішою серед новотворів поета, виразником антропоцентричності поезій В. Стуса. Серед новотворів КФІ ця модель також присутня, але набагато менше.

Основа **само-** за матеріалами КФІ сформувала неосемантизм, наближений до терміна, внаслідок скорочення від ‘самовільний’ (‘незаконний’). *Самовільне захоплення → самозахоплення (самозахоплений, самозахоплювач, самозахопник, самозахопництво)* у значенні ‘використання чогось (землі, земельної ділянки) незаконно, тобто до виникнення права власності чи права користування нею’, пор. у СУМ: ‘захоплення самим собою’ (Т. IX, с. 37). Це підтверджують контексти: *У документі конкретизуються такі поняття, як самозахоплення землі та використання її без документів* (Пост, 09.02.2007, № 5, с. 13); *Парламент має намір запровадити кримінальну відповідальність за самозахоплення землі, тобто за використання земельної ділянки до виникнення права власності чи права користування нею... Якщо відбудуться зміни на законодавчому полі,* “*самозахопник*” *може отримати до шести років позбавлення волі* (ГУ, № 214, 2006, с. 6).

Формантоцентричний спосіб подання новотворів дає можливість охопити весь спектр функціонування певної основи чи форманта, його сполучуваність, семантичні й стилістичні особливості, формування нових значень, що засвідчують діагностичні мікроконтексти.

Отже, запропонований системно-структурний підхід подання новотворів у лексичній базі не лише охоплює великі масиви нової похідної лексики, а й структурує її за ознаками форми, семантики та функціонування в тексті, встановлює активність і продуктивність моделей словотворення в колективній та індивідуальній мовній практиці, а також дає змогу прогнозувати зміни в будові та функціонуванні новотворів.

1. *Ґрещук В. В.* Нариси з основоцентричної дериватології / В. В. Ґрещук, Р. О Бачкур, І. Ф. Джочка, Н. М. Пославська. – Івано-Франківськ, 2007. 2. *Земская Е. А*. Активные процессы современного словопроизводства / Е. А. Земская // Русский язык в конце XX столетия (1985–1995). – М., 1996. – С. 90–141. 3. *Карпіловська Є. А*. Тенденції розвитку сучасного українського лексикону: чинники стабілізації інновацій / Є. А. Карпіловська // Українська мова. – 2007. – № 4. – С. 3–15; 2008. – № 1. – С. 24–35. 4. *Клименко Н. Ф.* Система афіксального словотворення сучасної української мови / Н. Ф. Клименко. – К., 1973. 5. *Ковалик І. І.* Дериватологія (словотвір) як самостійна лінгвістична дисципліна та її місце у системі науки про мову / І. І. Ковалик // Словотвір сучасної української мови. – К., 1979. – С. 5–56. 6. *Кубрякова Е. С.* Что такое словообразование / Е. С. Кубрякова. – М., 1965.

Предлагается использование основоцентрического и формантоцентрического подходов при подаче новообразований в лексической базе данных. Сравнивается активность системных словообразовательных моделей в коллективной и индивидуальной языковой практике.

Ключевые слова: основоцентрический подход, формантоцентрический подход, словообразовательная модель, новообразование.

The article proposes to use stem-centered and formant-centered approach for new appeared words in lexical database. Activity of systematical word formative models is compared in usage and idiolect.

Key words: stem-centered approach, formant-centered approach, word formative models, new derivate.

Стаття надійшла до редакції 10.09.2012

УДК 81’322.4

Володимир Лісовський
Військовий інститут
Київського національного університету імені Тараса Шевченка

МОДЕЛЮВАННЯ ПРЕФІКСАЛЬНОГО СЛОВОТВОРЕННЯ В СИСТЕМАХ МАШИННОГО ПЕРЕКЛАДУ (на матеріалі англійських військових текстів)

Розглядаються особливості префіксального словотворення військової англійської лексики, аналізуються найбільш продуктивні словотвірні моделі префіксальних дериватів, визначені у процесі створення лінгвістичного забезпечення системи машинного перекладу.

Ключові слова: система машинного перекладу, префіксальне словотворення, словотвірна модель, словотвірний тип.

Участь Збройних Сил України в міжнародних навчаннях, конференціях, миротворчих місіях, розвиток двосторонньої співпраці з іншими державами та військовими відомствами розвинених країн обумовили зростання потреби в здійсненні перекладів військової документації. Особливу роль у цьому відіграє швидкість здійснення перекладів, тому для оброблення інформації, окрім послуг перекладачів, дедалі частіше використовуються системи машинного перекладу (СМП).

Сьогодні існує велика кількість комерційних СМП (PRAGMA, PROMT, PROLING, SYSTRAN тощо), але всі вони мають один суттєвий недолік – невисоку точність перекладу.

Однією з невирішених проблем автоматичного оброблення текстів є невизначеність, яка притаманна природній мові. Під невизначеністю розуміємо лексичні одиниці в початковій формі, яким не відповідає жоден варіант перекладу в перекладному електронному словнику. Аналіз текстів, зокрема результати перекладу СМП, показали, що відсутність перекладних еквівалентів у машинному словнику пов’язана як із випадками введення нових слів (нових фрагментів знань), так і з продуктивністю афіксального словотворення, на частку якого припадає близько

половини нових слів у СМП. Наприклад, *наш потенциал по сдерживанию* он-лайн-перекладач ПРОМТ – перекладає як *our potential on control*, не розпізнаючи термін *deterrent*, який перекладається як *стримуюча сила, засіб/форма стримування*. Термін утворено від дієслова *deter – стримувати, залякувати* за моделлю **V + ant (ent) > N** зі значенням ‘предметність’. Отже, правильним перекладом буде *our deterrent*. Інший приклад: за допомогою онлайн-перекладача ПРОМТ *underemphasize priorities* перекладається як *приоритеты underemphasize*. *СМП* взагалі не переклала слово *underemphasize*, утворене від дієслова *emphasize* – ‘надавати особливого значення’ за моделлю **under + V > V** зі значенням ‘недостатність або відсутність якості, названої мотивувальною основою’. Правильний переклад – *надавати недостатньо уваги*.

Таким чином, одним із напрямків підвищення якості машинного перекладу є формалізація процесу словотворення в природній мові, що дозволить автоматизувати процес розпізнавання “нових” у системі слів, які утворюються в мові продуктивними афіксальними словотвірними моделями.

Як відомо, одним з основних шляхів поповнення й розширення англійської лексики є афіксальне словотворення. “Афіксальне словотворення – це спосіб словотворення за допомогою афіксів” [УМ : 38].

Процес словотворення в англійській мові вже давно є об’єктом лінгвістичних досліджень, серед яких роботи таких мовознавців: І. Арнольд [Арнольд 1986], О. Бортничук [Бортничук 1988], В. Виноградова [Виноградова 1984], П. Каращук [Каращук 1977], О. Кубрякова [Кубрякова 1965], Н. Шевчук [Шевчук 1983] та ін. Проте у сучасних СМП не враховується словотвірний аспект перекладу.

Матеріалом дослідження префіксальних дериватів в англійській мові слугували тексти військово-технічної та військово-політичної тематики загальною кількістю 500 000 лексичних одиниць, що були виокремленні методом суцільної вибірки з базового підручника Військово-технічний переклад (англійська

мова) [Лісовський 2010]; словника військових термінів Department of Defense Dictionary of Military and Associated Terms: Joint Publication 1-02; польових статутів Збройних Сил США FM – 3-22.34, TOW weapon system, HQ DA Washington, DC, 28 November 2003; FM – 6-2, Tactics, Techniques, and Procedures for field artillery survey, HQ DA Washington, DC, 23 September 1993; Стратегії національної безпеки США National Security Strategy 2010; Чотирічного оборонного огляду Quadrennial Defense Review, Washington, DC, February 2010 та ін., за допомогою безкоштовного програмного засобу аналізу текстів (Antconc3.2.1w) (http://www.antlab.sci.waseda.ac.jp/antconc_index.html).

Префіксація як спосіб словотвору модифікує основу, до якої приєднується префікс. Модифікуючи лексичне значення слова, префікс рідко змінює граматичний характер слова, тому мотивувальне й префіксально мотивоване слово належать до однієї частини мови, наприклад: *mount – dis-mount, orbit – dis-orbit, embark – dis-embark* [Кубрякова : 14].

У сучасній військовій термінології, як зазначив В. Шевчук, використовуються лише 33 префіксальні та 39 суфіксальних морфем, які складають лише 45% від усіх афіксальних засобів англійської мови [Шевчук : 26]. У результаті визначення їх продуктивності та регулярності були відібрані 26 префіксів та напівпрефіксів (префіксоїдів): **all-**; **anti-**; **counter-**; **de-**; **dis-**; **ex-**; **half-**; **in-**; **il-**; **im-**; **ir-**; **inter-**; **micro-**; **mini-**; **mis-**; **multi-**; **non-**; **out-**; **over-**; **pre-**; **re-**; **self-**; **semi-**; **sub-**; **super-**; **un-**.

Як відомо, творення простих та складних слів здійснюється за деривaційними моделями та словотвірними типами. Словотвірний тип визначає схему побудови похідних слів певної частини мови, який характеризується спільністю: частини мови для безпосередньо мотивованих слів; форманта, як у морфемному представленні, так і семантичному значенні [УМ : 622]. Під деривaційною моделлю розуміється закономірність перетворення твірних слів у похідні, тобто поєднання твірної основи слова та словотвірних формантів із відображенням їхніх нових значень та визначенням морфологічних і морфонологічних ознак [УМ : 138]. Спі-

льне значення, що відрізняє всі мотивувальні слова від мотивованих, називається словотвірним значенням певного словотвірного типу [УМ : 622]. Носієм словотвірного значення є словотвірний формант. Словотвірні типи відрізняються один від одного різним ступенем регулярності та продуктивності. Ця властивість широко використовується в мові й значною мірою впливає на формування словників різного призначення, у тому числі й перекладних. Продуктивність визначаємо як здатність основ мотивувальних слів сполучатися зі словотвірним формантом (афіксальною морфемою), що має значний словотвірний потенціал, утворюючи похідні слова за певним словотвірним типом у певний період розвитку мови [РГ 1980 : 133]. На відміну від продуктивності, регулярність – це здатність утворювати похідні слова за багатократно повторюваною формально-семантичною моделлю. Основною одиницею дослідження словотвірних моделей є словотвірна пара, яка складається з мотивувального та мотивованого слова. Наприклад, *calibrate – calibrator; interrogate – interrogator*. Порівняння словотвірних пар виступає у нашому дослідженні інструментом визначення граматичної характеристики похідного слова, граматичного значення твірного слова, словотвірного значення формантів, похідності слова, слова для пошуку в словнику, шляхів перекладу похідного слова, можливих морфонологічних і графічних змін та лексичних обмежень. Під лексичним обмеженням розуміються одиниці, які утворені за нетиповою схемою певної словотвірної моделі та які не можуть бути розпізнані СМП під час морфологічного оброблення, внаслідок чого повинні бути обов'язково відзначенні в перекладному словнику.

Отже, розглянемо процедуру виділення словотвірного префікса і визначення продуктивної словотвірної моделі та словотвірного типу на прикладі англійського словотвірного форманта **anti-**.

Група 1. Словотвірний тип зі словотвірним формантом **anti-**.

Спосіб словотворення: афіксальний, префіксальний. Загальне словотвірне значення: 'спрямований проти того, що позначене основою'. Загальна кількість слів: 30.

У межах цієї групи визначається три підгрупи.

Підгрупа 1.1. Відіменникова модель anti-+N->-N.

У цій підгрупі об'єднані словотвірні пари, твірною основою яких є іменник. Розглянемо чотири словотвірні моделі з формантом **anti-** (табл. 1).

Таблиця 1. **Відіменникові моделі anti-+N->-N**

№	Префікс	Словотвірна модель	Семантичне значення префікса	Семантична ознака твірн. слова	Словник	Приклади	Винятки
1	*anti-*	*anti-+N->-N*	сн	о	зв	anti-globalist – *антиглобаліст*	anti-aliasing anti-cosine
2	*anti-*	*anti-+N->-N*	зппо	р	сп	antidim – *суміш, що запобігає запотіванню скла*	
3	*anti-*	*anti-+N->-N*	пнзбз, звдп	опб, п	сп	antiblackout – *боротьба із засобами РЕП* antimissile – *протиракета*	
4	*anti-*	*anti-+N+ism>-N*	сптнптпо	оппгп	зв	anti-terrorism – *антитероризм*	antiart

Умовні позначки: **сн** – 'суперечливість, невідповідність'; **зппо** – 'засіб проти того, що позначено основою'; **пнзбз** – 'протидія, нейтралізація, знешкодження бойових засобів'; **звдп** – 'захист від будь-яких дій противника'; **сптнптпо** – 'суспільна, політична течія, спрямована проти того, що позначено основою'; **о** – 'особа'; **р** – речовина'; **опб** – 'ознака процесуальності боротьби'; **оппгп** – 'ознака процесуальності політичного, громадського процесу'; **п** – 'предмет'; **зв** – 'загальновживаний словник'; **сп** – 'спеціальний словник'.

Дослідження вибірки слів, утворених за цією моделлю, дозволяє зробити такі висновки:

1. Ця підгрупа має вищу продуктивність в англійських текстах військово-технічного характеру (27 лексичних одиниць), ніж у військово-політичних (3 лексичні одиниці). Найвищу продуктивність продемонструвала третя словотвірна модель.

2. Твірні іменники четвертої словотвірної моделі мають закінчення **-ism**.

3. Аналізована підгрупа також утворює похідні іменники зі словотвірним значенням, ‘протилежний напрямок’ *anti-inversion – зворотне обертання*, проте виокремлювати їх у самостійну словотвірну модель вважаємо недоцільним з причини невисокої регулярності.

4. Похідні слова відіменникової моделі можуть утворювати багатокомпонентні терміни і, як правило, є ключовими словами у фінальній позиції складеного терміна, наприклад *radar-homing antimissile – протиракета з радіолокаційною головкою самонаведення*.

5. Словотвірний формант **anti-** перекладається українською мовою*:*

- за допомогою префікса **анти-**, **проти-** (*antidote – антидот)*;
- описовими зворотами *для боротьби, для захисту* (*antiantimissile – ракета для боротьби з протиракетами*);
- словосполученнями *заходи захисту, заходи боротьби* (*antibug – заходи захисту від засобів прослуховування*);
- прийменником **проти** (*antiterror – дії проти терористів*).

Підгрупа 1.2. Відприкметникова модель anti-+A->-A.

У цій підгрупі об’єднані словотвірні пари, твірною основою яких є прикметник. Розглянемо три словотвірні моделі з формантом **anti-** (табл. 2).

Дослідження вибірки слів цієї моделі дозволяє зробити такі висновки:

1. Модель є високопродуктивною в англійських текстах як військово-технічного, так і військово-політичного характеру. Найвищу продуктивність має третя словотвірна модель.

2. Похідне слово третьої словотвірної моделі також часто бере участь у створенні багатокомпонентних термінів і, як правило, є першим словом складеного терміна (*antipersonnel* and *antimateriel fragmentation warhead – осколкова бойова частина для ураження живої сили та матеріальних об’єктів*).

3. Словотвірний формант **anti-** перекладається українською мовою за допомогою префіксів **анти-**, **проти-** (*anti-submarine – протичовновий*; *anti-bribery – антикорупційний*).

Таблиця 2. **Відприкметникові моделі anti-+A->-A**

№	Префікс	Словотвірна модель	Семантичне значення префікса	Семантична ознака твірн. слова	Словник	Приклади	Винятки
1	*anti-*	*anti-+A->-A*	пппісм	вп	зв	anti-American – *антиамериканський*	
2	*anti-*	*anti-+A->-A*	п	яп	зв	anti-damage – *протиаварійний*	
3	*anti-*	*anti-+A->-A*	пнзбз, звдп	вп	сп	antiaerial – *протиповітряний*	

Умовні позначки: **пппісм** – ‘протилежний за політичними поглядами, ідеологією, способом мислення’; **п** – ‘протидіючий’; **пнзбз** – ‘протидія, нейтралізація, знешкодження бойових засобів’; **звдп** – ‘захист від будь-яких дій противника’; **вп** – ‘відносні прикметники’; **яп** – ‘якісні прикметники’; **зв** – ‘загальновживаний словник’; **сп** – ‘спеціальний словник’.

Підгрупа 1.3. Відіменникова модель anti-+N->-A.

У цій підгрупі об’єднані словотвірні пари, твірною основою яких є іменник. Розглянемо три словотвірні моделі з формантом **anti-** (табл. 3).

Таблиця 3. **Відіменникові моделі anti-+N->-A**

№	Префікс	Словотвірна модель	Семантичне значення префікса	Семантична ознака твірн. слова	Словник	Приклади	Винятки
1	*anti-*	*anti-+N->-A*	пппісм	вп	зв	anti-capitalist – *антикапіталістичний*	
2	*anti-*	*anti-+N->-A*	п	яп	зв	anti-foam – *протипінний*	
3	*anti-*	*anti-+N->-A*	пнзбз, звдп	вп	сп	antiradar – *протирадіолокаційний*	

Умовні позначки: **пппісм** – ‘протилежний за політичними поглядами, ідеологією, способом мислення’; **п** – ‘протидіючий’; **пнзбз** – ‘протидія, нейтралізація, знешкодження бойових засобів’; **звдп** – ‘захист від будь-яких дій противника’; **вп** – ‘відносні прикметники’; **яп** – ‘якісні прикметники’; **зв** – ‘загальновживаний словник’; **сп** – ‘спеціальний словник’.

Дослідження вибірки слів цієї моделі дозволяє зробити такі висновки:

1. Модель є високопродуктивною в англійських текстах як військово-технічного, так і військово-політичного характеру. Найбільш продуктивною є третя словотвірна модель (*antirocket – протиракетний*; *antisatellite – протисупутниковий*; *antiboat – протичовновий*).

2. Словотвірний формант **anti-** перекладається українською мовою за допомогою префіксів **анти-**, **проти-**.

Проаналізувавши словотвірні пари англійської мови, які було об'єднано в словотвірні типи за критерієм спільності словотвірного форманта (префікса) в текстах військово-політичного та військово-технічного характеру, можна зробити такі висновки.

1. Більшість аналізованих префіксів підтвердили свою продуктивність та регулярність у системі афіксального словотворення в англійських текстах військової тематики. Проте такі високопродуктивні префікси в системі афіксального військового термінотворення, як **half-**, **self-**, **semi-**, **sub-**, **super-**, виявилися недостатньо продуктивними при утворенні "нових" слів. У зв'язку з цим виникає необхідність аналізу словотвірних типів у текстах військово-спеціального характеру для всебічного та детального дослідження афіксальних засобів утворення "нових" слів.

2. Найбільш продуктивними та регулярними моделями текстів військово-технічного та військово-політичного характеру є моделі з такими префіксальними словотвірними формантами: **anti-**; **de-**; **dis-**; **in-**; **ex-**; **multi-**; **non-**; **out-**; **over-**; **pre-**; **re-**; **semi-**; **super-**; **un-**.

3. Сформований реєстр продуктивних та регулярних префіксів англійської мови буде використаний для побудови словотвірної моделі в розроблюваній системі машинного перекладу як компонент морфологічного аналізу, що дозволить розпізнавати "нові" слова, які утворюються завдяки продуктивним префіксам, навіть за відсутності їх у перекладному машинному словнику.

1. *Арнольд И. В.* Лексикология современного английского языка: Учеб. для ин-тов и фак. иностр. яз. / И. В. Арнольд. – М., 1986. 2. *Бортничук Е. Н.* Словообразование в современном английском языке / Е. Н. Бортничук, М. В Василенко. – К., 1988. 3. *Виноградова В. Н.* Стилистический аспект русского словообразования / В. Н. Виноградова. – М., 1984. 4. *Каращук П. М.* Словообразование английского языка: учебное пособие / П. М. Каращук. – М., 1977. 5. *Кубрякова Е. С.* Что такое словообразование / Е. С. Кубрякова. – М., 1965. 6. *Лісовський В. М.* Військово-технічний переклад: (англ. мова) : Підручник / За ред. В. В. Балабіна / В. М. Лісовський. – К., 2010. 7. Українська мова : Енциклопедія. – К., 2004. 8. *Шевчук Н. В.* Производные военные термины в английском языке / Н. В. Шевчук. – М., 1983. 9. Joint Publication 1-02, Department of Defense Dictionary of Military and Associated Terms. 10. FM – 3-22.34, TOW weapon system, HQ DA Washington, DC, 28 November 2003. 11. FM – 6-2, Tactics, Techniques, and Procedures for field artillery survey, HQ DA Washington, DC, 23 September 1999. 12. National Security Strategy 2010. 13. Quadrennial Defense Review, Washington, DC, February, 2010. 14. Русская грамматика. – Т. 1 : Фонетика. Фонология. Ударение. Интонация. Словообразование. Морфология / Н. Ю. Шведова (гл. ред.). – М., 1980.

Рассматриваются особенности префиксального словообразования в военной лексике, исследуются наиболее продуктивные префиксы и словообразовательные модели производных слов на предварительном этапе создания компонентов лингвистического обеспечения системы машинного перевода.

Ключевые слова: система машинного перевода, префиксальное словообразование, словообразовательная модель, словообразовательный тип.

The article is devoted to research the peculiarities of military lexis' prefix derivation, the evaluation of the most efficient prefixes and the modeling of derivatives' word-formative formants as the preliminary stage of machine translation system's linguistic support components development.

Key words: machine translation system, prefixes derivation, word-building pattern, word-building type.

Стаття надійшла до редакції 12.09.2012

УДК 81'322

Юлія Маковецька-Гудзь
Національний технічний університет України
"***Київський політехнічний інститут***"

ЕЛЕКТРОННИЙ СЛОВНИК ХУДОЖНІХ ПОРІВНЯНЬ

Проаналізовано стан комп'ютерної лексикографії. Обґрунтовано поняття "електронний словник" та вказано основні переваги електронних словників. Представлено методику розробки "Електронного словника художніх порівнянь".

Ключові слова: комп'ютерна лексикографія, електронний словник, словник порівнянь, лінгвістичні бази даних, гіпертекст.

Слово потрапляє в поетичний текст із певним усталеним значенням, яке, трансформуючись, набуває нових відтінків. Імпульсом до цього є незвична сполучуваність з іншими словами, несподівані контексти. Поети, як ніхто, вміють створювати такі контексти, виражаючи себе, і тим самим збагачують і розвивають мову. Лінгвокреативна творчість поетів може бути зафіксована в каталогах тропів та стилістичних фігур, зокрема – у каталозі (словнику) порівнянь. Тому впорядкування словників порівнянь є важливим завданням.

Потреба в словнику порівнянь української мови особливо відчувається на тлі зарубіжної лексикографії, у якій подібні видання представлені досить широко. Такими є, наприклад, великі словники чеських порівнянь (SČF 1983), болгарських (Кювлієва-Мишайкова 1986), англійських (Wilstach 1924), андалузьких (Rodriguez 1899), іспанських (Castillo 1970; Shirley 1977), португальських (Basto 1924) та ін. Зібрання народних порівнянь видані й у наших найближчих слов'янських сусідів – білорусів (Янкоускі 1973) та росіян (В. Мокієнко 2003; Н. Кожевнікова, З. Петрова 2000).

В українській лексикографії є кілька словників порівнянь, укладених на матеріалі українських фольклорних текстів: І. Гурин "Образне слово. Постійні народні порівняння" (1966); О. Юрченко, А. Івченко "Словник стійких народних порівнянь"

(1993); "Українські прислів'я, приказки та порівняння з літературних пам'яток" (упоряд. М. Пазяк) (2001). Серед цього переліку немає жодного електронного словника порівнянь, тому укладання словника такого типу є необхідним та актуальним.

Єдиної думки про те, що варто вважати електронним словником, не було й, на наш погляд, бути не може. Як і самі словники, зміст, вкладений у це поняття, є авторським і відбиває те бачення питання, яким володіє лексикограф, його досвід, знання тощо. Так, під терміном "електронний словник" мають на увазі: електронну копію друкованого словника, комп'ютерну версію друкованого словника та власне електронний (мультимедійний) словник, укладений на основі фактичного матеріалу і створений за допомогою комп'ютерних програм [Карпіловська].

Електронні словники й енциклопедії розробляються як автономні або мережеві програмні продукти. Не викликає сумніву той факт, що електронні словники надають користувачеві безліч додаткових можливостей порівняно з друкованими аналогами. Перевага такого словника полягає у:

- зберіганні великого обсягу інформації за рахунок гіперпосилань: електронний словник передбачає внесення до своєї структури кількох словників різних типів і жанрів (словника сполучуваності, термінології, граматичних норм та ін.);
- доступності словника за рахунок ефективної системи пошуку (повнотекстовий пошук, одночасний пошук у декількох словниках, швидкість пошуку);
- застосуванні засобів мультимедіа для семантизації лексики: озвучування заголовних слів, уведення ілюстративного матеріалу з фотографіями, анімацією, відеофрагментами;
- використанні словників у локальній і глобальній мережах; при цьому робота зі словниками може проводитися багатьма користувачами одночасно;
- економії часу й матеріальних витрат для створення комп'ютерних словників.

Головними вимогами для комп'ютерного (електронного) словника є повнота лексикографічної інформації та зручність

доступу до неї. Комп'ютерний словник, на відміну від паперового, дозволяє здійснювати різного роду вибірки матеріалу. Не переглядаючи весь корпус словника, користувач може відповідно до запиту вибрати слова за одним або кількома параметрами. Обидва вказані компоненти (повнота інформації та зручність доступу) в електронних словниках є поєднанням досягнень філологічної та комп'ютерних наук. У традиційних (паперових) словниках проблеми повного представлення інформації про слово в певному ракурсі та зручності її пошуку на кінець XX століття були в основному вирішені як у вітчизняному, так і в зарубіжному мовознавстві. Натомість можливості, надані в розпорядження лексикографії новітніми комп'ютерними технологіями, лише починають засвоюватися та впроваджуватися.

Електронний словник є певною базою даних (БД) або сукупністю баз даних та інтерфейсів – місце подання запитів та отримання відповідей на моніторі. Інтерфейс виконує функцію конкретизатора команд (побажань) користувача. Розгалужена система команд, які реалізуються за допомогою кнопок меню, дозволяє викликати на екран із БД саме ту інформацію, яка потрібна (і залишити у прихованому вигляді ту, яка не потрібна) на момент пошуку. Для цього необхідно розподілити всю інформацію статті будь-якого словника на найменші значущі параметри та характеристики і вводити їх окремо до БД. Кожна частина інформації про слово (наприклад, дефініція або граматична інформація) буде зберігатися окремо в певному полі БД.

Проблема браку місця, характерна для паперової лексикографії, певною мірою вирішується в електронній лексикографії, тому електронний словник не обов'язково має наслідувати традиційний стиль у плані скорочень інформації.

Одним з актуальних завдань сучасної лексикографії є систематизація художніх тропів. Адже у зв'язку з посиленим інтересом до вивчення тропів виникає необхідність створення такого словника, у якому в найприйнятніший та найзручніший спосіб можна було б отримати всю потрібну інформацію. Окрім цього, сучасні комп'ютерні технології дають змогу створити

продукт у найбільш зручному для користувача форматі. З огляду на широкі можливості інформаційних технологій та їхнє впровадження в сучасну лексикографію, актуалізується проблема створення комп'ютерних (електронних) словників.

Для електронної лексикографії особливе значення має принцип гіпертексту як основного способу організації віртуального текстового простору. На думку А. Баранова, практична цінність гіпертексту полягає в тому, що "він описує тип інтерактивного середовища з можливістю переходу за посиланнями" [Баранов : 87]. Посилання, якими можуть бути слова, фрази або малюнки, дозволяють користувачу вибрати той чи інший текст чи малюнок і виводити на екран відомості, пов'язані з ним. Нелінійний характер гіпертексту уможливлює подачу інформації у вигляді розгалуженої структури, що дозволяє значною мірою розширити рамки словникової статті.

"Електронний словник художніх порівнянь" (ЕСХП) – перший український словник, створений за правилами гіпертексту та укладений за принципом ідеографічного опису лексичної парадигми компонентів порівняння.

Робота над словником передбачала два етапи: розроблення власне лінгвістичного блоку та програмного.

Лінгвістичний блок. На цьому етапі укладено каталог порівнянь у ЛДБ, який оброблено з урахуванням таких параметрів: суб'єкт порівняння; об'єкт порівняння; ознака, на основі якої порівнюються компоненти; модель порівняння; належність до концептів синоптичної схеми. У результаті відбору та аналізу порівнянь було створено їхні моделі.

Отже, результатами роботи першого етапу є опис, типізація та встановлення параметрів аналізу художніх порівнянь, що будуть використані на наступному етапі проектування словника.

Програмний блок. ЕСХП побудовано на основі бази даних Microsoft Access, у якій зафіксовано весь первинний лексикографічний матеріал (контексти з порівняннями) та отриманий в результаті аналізу на першому етапі новий лексикографічний матеріал, що в подальшому будуть використані для створення словникових статей.

В електронних словниках передбачаються різні види пошуку та різні способи подання інформації. У ЕСХП пошук можна здійснювати за такими параметрами:

- за ідеографічним принципом;
- за моделями порівняння.

Тип пошуку за ідеографічним принципом дає змогу користувачеві вибрати лексеми за деталізованими концептами синоптичної схеми, що уможливлює отримання найбільш повної інформації. Задаючи такий вид пошуку, користувачеві для отримання інформації спочатку потрібно обрати складник порівняння (суб'єкт чи об'єкт). Потім користувачеві запропоновано перелік концептів. При виборі конкретного концепту-посилання користувач отримує картку обраного слова. Крім того, можна здійснювати пошук слова за допомогою спеціального поля. Модуль автоматичного аналізу словника дозволяє користувачеві отримати лінгвістичну інформацію про обране слово за такими параметрами: слово; частота його вживання у функції суб'єкта чи об'єкта; корелят (суб'єкт чи об'єкт); ознака порівняння; модель порівняння; контексти вживання визначеної лексеми; автор.

Обираючи пошук за моделями порівняння, користувач може обрати модель простого чи складного порівняння та вибрати з меню модель порівняння зі списку запропонованих. У такому випадку він отримає таку інформацію: модель; частота її вживання; складники порівняння (суб'єкт та об'єкт); ознака порівняння; контексти вживання обраної моделі; автор.

Запропонована методика укладання електронного словника дозволяє досить легко перетворити текст словника на базу даних, придатну для виконання різноманітних інформаційно-пошукових та дослідницьких процедур.

ЕСХП дозволяє розширити можливості опису та аналізу поетичного тексту. Макроструктура словників:

- Словник є багаторівневою лексикографічною працею, у якій знайшли відбиття мовні факти, втілені в поетичному тексті 70–90-х років XX століття.
- Словник побудований за гіпертекстовою структурою й включає інформацію у вигляді лексичної лінгвістичної бази да-

них реляційного типу, коли значеннєвий зміст одиниці розподіляється за полями, причому головна увага приділяється системам подання інформації, характерної для цілого класу лексем.

- Словник організовує гіпертекстовий простір, що є відбиттям когнітивного простору людини.
- У словнику представлено два види пошуку лексем – ідеографічний та формальний (за формалізованими моделями порівнянь), що допомагає отримати найбільш повне уявлення про художнє порівняння.

Таким чином, електронний словник порівнянь може бути базою для укладання нових комп'ютерних словників (наприклад, загального словника тропів), а також разом зі своєю базою даних виконувати функції автоматичної інформаційно-довідкової дослідницької та навчальної системи. ЕСХП може бути використаний для проведення лінгвістичних, соціолінгвістичних, психологічних та інших досліджень.

1. *Баранов А. Н.* Введение в прикладную лингвистику: Учеб. пособие / А. Н. Баранов. – М., 2003. 2. *Гурин І. І.* Образне слово. Постійні народні порівняння / І. І. Гурин. – К., 1966. 3. *Карпіловська Є. А.* Вступ до комп'ютерної лінгвістики / Є. А. Карпіловська. – Донецьк, 2003. 4. Українські прислів'я, приказки та порівняння з літературних пам'яток / [відп.ред. Й. Ю. Федас, упоряд., вступ. ст. М. М. Пазяк]. – К., 2001. 5. *Юрченко О. С.* Словник стійких народних порівнянь / Юрченко О. С., Івченко А. О. – Х., 1993.

Проанализировано состояние компьютерной лексикографии. Обосновано понятие "электронный словарь", указаны основные преимущества электронных словарей. Представлена методика разработки "Электронного словаря художественных сравнений".

Ключевые слова: компьютерная лексикография, электронный словарь, словарь сравнений, лингвистические базы данных, гипертекст.

The article analyzes the state of computer lexicography. The possibility of creating an electronic dictionary are reviewed. The appellation "electronic dictionary" and are the main advantages of electronic dictionaries are substantiated. The technique of developing "Electronic Dictionary of Authors' Similes" is presented.

Keywords: computer lexicography, electronic dictionary, comparisons, linguistic databases, hypertext.

Стаття надійшла до редакції 14.09.2012

УДК 81’322

Алла Міщенко
Кіровоградський державний педагогічний університет імені Володимира Винниченка

СТВОРЕННЯ ПАРАЛЕЛЬНОГО БАНКУ ДЕРЕВ ДЛЯ НІМЕЦЬКОЇ ТА УКРАЇНСЬКОЇ МОВ

Описано створення паралельного банку дерев для української та німецької мов. Було здійснено ручне тегування й лематизацію токенів для української мови, створено набір тегів для анотування українських речень на рівні синтаксичної структури; проведено вирівнювання й визначення повних або часткових відповідників на рівні як термінальних, так і нетермінальних символів. Для анотування банку дерев українською мовою було застосовано формат TIGER-XML, адаптований до потреб формального опису граматичної структури української мови.

Ключові слова: банки дерев, анотація, тегування частин мови, вирівнювання, переклад.

Інновації в галузі економіки, політики й технологій суттєво впливають на зростання попиту на перекладацькі послуги, а глобалізаційні процеси й формування єдиного полілінгвокультурного суспільства обумовлюють необхідність перекладу контенту багатьма мовами, що супроводжується поступовим, але постійним зростанням обсягів перекладу, збільшенням кількості мов перекладу та витрат на нього. З огляду на це необхідно шукати можливостей оптимізації процесу перекладу. Таким “соломоновим рішенням” стала концепція “пам’яті перекладу”, яка ґрунтується на повторному використанні раніше перекладених сегментів контенту. Згідно з визначенням, пам’ять перекладу – це “архів багатомовних сегментованих, вирівняних (aligned), проаналізованих (parsed) та класифікованих текстів <…>, який дозволяє зберігати попередньо вирівняні сегменти текстів й здійснювати їх пошук відповідно до заданих параметрів” [Eagles : 140].

Концепція пам’яті перекладу вперше була реалізована на програмному рівні ще у 60-ті рр. XX ст. Перший її прототип було створено для Європейської спілки вугілля та сталі і використовувався для перекладу багатомовного контенту галузевих текстів.

Ефективність методу "повторного застосування", з одного боку, та обсяги перекладу, які постійно зростали, з іншого боку, значно прискорили процес створення й розбудови програмного забезпечення для оптимізації процесу перекладу; а можливості збереження контенту в електронній формі обумовили створення лінгвістичних ресурсів як для академічного, так і для комерційного застосування.

Асортимент програмних продуктів (Tools), які вможливлюють повторне використання раніше створеного багатомовного контенту, значно розширився. На сучасному етапі він представлений такими інструментами: системи пам'яті перекладу (Translation Memory Systeme, TMS), системи управління контентом (Content Management Systeme, CMS), системи підтримки технічного автора (Authoring Systeme, AS) та ін., які детально охарактеризовані в науковій літературі [Massion 1995; Reinke 2004; Boiko 2005; Closs 2011; Drewer / Ziegler 2011] та технічній документації на CLAT (Controlled Language Authoring Technology) (2010).

У статті описано експеримент зі створення банку дерев (Treebanks) для української мови (Ukrainian language) та його вирівнювання (Alignment) з банком дерев німецької мови. Робота виконувалася під керівництвом проф. О. Капанадзе і була б неможлива без його консультацій, підтримки та редагування.

Паралельні банки дерев розглядаються сьогодні як важливі ресурси для навчання перекладу, а також для вирішення окремих прикладних завдань у галузі комп'ютерної лінгвістики. У процесі навчання банки дерев слугують для унаочнення й ілюстрації контрастивних феноменів мовних пар. У корпусній лінгвістиці вони використовуються як ефективні ресурси для дослідження й аналізу синтаксичної структури мовних пар. В інших галузях комп'ютерної лінгвістики паралельні банки дерев застосовуються, зокрема, для тренування й оцінювання ефективності парсерів. У перспективі паралельні банки дерев після конвертування у стандартизований формат пам'яті пере-

кладу можуть імпортуватися в комерційні системи пам'яті перекладу на кшталт SDL, Across, DejaVu, MemoQ та застосовуватися в них у процесі перекладу.

Для проведення експерименту було відібрано 40 речень із лексикону валентності для NLP німецькою мовою, укладеного в рамках проекту GREG (German-Russian-English-Georgian) [Kapanadze O, Wanner L., Klatt S 2002 *:* 11]. Ці речення перекладені українською мовою і вручну анотовані з урахуванням особливостей української мови.

У процесі створення паралельного банку дерев, що вирівнюється автоматично, ключове значення відіграють програми синтаксичного аналізу (Parser), які на етапі генерування монолінгвальних банків дерев аналізують синтаксичну структуру речень, оскільки для української мови ще не створено ресурсів для автоматизованого лінгвістичного анотування речень із можливістю їх наступної візуалізації у Synphaty й подальшого застосування для побудови паралельних банків дерев. Таким чином, автоматичне створення паралельного банку дерев для німецької та української мов було неможливим, але привабливість для перекладацької галузі ідеї автоматичного генерування банків дерев із паралельних текстів з їх наступною конвертацією у пам'ять перекладу спонукала нас перевірити можливість опрацювання мов із кириличним шрифтом програмними продуктами Synphaty і TreeAligner, розробленими спеціально для побудови й візуалізації паралельних банків дерев для мов з латинським шрифтом, а також потенційні можливості побудови паралельних банків дерев для української та німецької мов з огляду на їхні морфологічні й структурні дивергенції.

Досягнення запланованої мети передбачало вирішення таких завдань:

- створення й візуалізацію монолінгвального банку дерев для української мови;
- конвертування даних, виведених у форматі tig (Synpathy), у формат xml (TreeAligner), вирівнювання (Alignierung) й ві-

зуалізацію німецько-українського банку дерев у програмі TreeAligner з подальшим редагуванням вручну.

На нашу думку, інструменти, використані для генерування й візуалізації паралельного банку дерев, потребують детальнішого пояснення.

Для вирівнювання паралельних речень використано програму Stockholm TreeAligner. Ця програма виводить дані у форматі xml і, таким чином, дозволяє підготувати паралельні речення до експорту в системи пам'яті перекладу. Але на першому етапі речення для кожної мови лінгвістично анотуються й візуалізуються у формі графів у програмі Synphaty.

Synphaty розроблена в інституті психолінгвістики Макса Планка в голландському місті Неймеген. Основним компонентом програми є візуалізатор синтаксичної структури речення Syntax-Viewer, розроблений в інституті машинної обробки природної мови у Штутгарті (Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart) для дослідницького проекту TIGER. Synphaty підтримується операційними системами Windows, MAC OS та Linux, а ліцензія на її застосування з академічною метою безкоштовна. (Детальніша інформація про Synphaty розміщена на сайті інституту психолінгвістики Макса Планка, http://www.mpi.nl/tools/synpathy.html).

Структура формату tig (Synphaty) складається з двох частин: заголовка (Header) та основної частини (Body). Заголовок містить метадані: назва корпусу, дата його укладання, укладач, експлікація використаних тегів тощо. Основна частина складається з: 1) ідентифікації графа (S ID); 2) початкового символа дерева (S); 3) термінальних вузлів (Terminals, рівень слів); 4) не термінальних вузлів (Nonterminals, рівень фраз: NK, NN та ін.); 5) первинних (SB, HD, OA) та вторинних ребер (NK, NN), які пов'язують термінальні й нетермінальні вузли графа та експлікують їхні синтаксичні функції. Крім того, у графі виводяться характеристики термінальних вузлів (terminal node features: PPER, VVFIN, ADJA, NN) та їхні граматичні значення (Mask. 3. Sg. Nom.; 3. Pl.Pres. Ind.; Akk.Pl.).

Для української мови анотування здійснювалося вручну в програмі Notepad++, а потім результати візуалізувалися у вигляді дерева із застосуванням програми Synphaty.

Як для Synphaty, так і для TreeAligner застосовувався набір тегів, спеціально створений для генерування банку дерев корпусу NEGRA (Stuttgart-Tübinger Tagset, STTS, G. Smith, 2003), який доповнено необхідними для анотування українських речень тегами. Він містить теги для анотування одинадцяти класів слів (Part of Speech, POS): Verbs V, Nouns N, Adverbs ADV, Conjunctions KO, Articles ART, Adpositions AP, Adjectives ADJ, Interjections IT, Pronouns P, Particles PTK, Cardinal Numbers CARD.

На рівні термінальних вузлів цим класам присвоюються відповідні характеристики (features) (табл. 1).

Таблиця 1. **Характеристики дієслова**

Клас	Типи дієслова	Форми дієслова	Характеристики дієслів на рівні термінальних вузлів
V	**A** ‘Auxiliar ’	**FIN** ‘finit ’ **INF** ‘infinit ’ **IMP** ‘imperativ ’ **PP** ‘Partizip Perfekt ’	**VAINF** ‘have ’ **VAIMP** ‘be ’ **VAPP** ‘had ’
	M ‘Modal ’	**FIN** ‘finit ’ **INF** ‘infinit ’ **PP** ‘Partizip Perfekt ’	**VMFIN** ‘könnte ’ **VMINF** ‘können ’ **VMPP** ‘gekonnt ’
	F ‘Full ’	**FIN** ‘finit ’ **INF** ‘infinit ’ **IZU** ‘Infinitiv mit zu ’ **IMP** ‘imperativ ’ **PP** ‘Partizip Perfekt ’	**VFFIN** ‘gebt … ab ’ **VFINF** ‘abgeben ’ **VFIZU** ‘abzugeben ’ **VFIMP** ‘gib … zu ’ **VFPP** ‘abzugegeben ’

Характеристики термінальних вузлів експлікуються граматичними значеннями, яких вони можуть набувати у реченні. У табл. 2 наведено граматичні категорії, характеристики термінальних вузлів, які можуть мати ці граматичні категорії, і ті граматичні значення (value), які можуть присвоюватися характеристикам термінальних вузлів.

Таблиця 2. **Елементи анотації для термінальних вузлів**

Грам. категорія	Характеристики термінальних вузлів	Значення характеристик термін. вузлів
Genus	ADJA, ART, APPRART, NE, NN, PDS, PDAT, PIAT, PIS (teilweise), PPER, PPOSAT, PPOSS, PRELS, PRELAT, PWAT, PWS	Masc, Fem, Neut
Kasus	ADJA, ART, APPRART, NE, NN, PDAT, PDS, PIAT, PIS, PPER, PPOSAT, PPOSS, PRELS, PRELAT, PRF, PWAT, PWS	Nom, Gen, Dat, Acc.
Numerus	ADJ, ART, APPRART, NE, NN, PDAT, PDS, PIAT, PIS, PPER, PPOSAT, PPOSS, PRELS, PRELAT, PRF, PWAT, PWS, V.FIN, V.IMP	Sg, Pl
Grad	ADJA, ADJD	Pos, Comp, Sup
Person	VVFIN, VAFIN, VMFIN, PPER, PRF	1, 2, 3
Tempus	VVFIN, VAFIN, VMFIN	Pres, Past
Modus	VVFIN, VAFIN, VMFIN	Ind, Subj
Nichtfinitheit	VVINF, VAINF, VMINF, VVPP, VAPP, VMPP, VVIMP, VAIMP, VVIZU	Inf, Psp, Imp, Infzu

Графічний візуалізатор унаочнює синтаксичну структуру речення у формі графа й дозволяє редагувати його вручну (“rebuildung” und “re-tagging”) як для термінальних, так і для нетермінальних символів. Остаточне опрацювання, наприклад термінальних символів, здійснюється у вікні управління термінальними символами. Тут можна змінювати назву термінальних вузлів, додавати нові або видаляти існуючі термінальні вузли тощо. Результат редагування зберігається й може експортуватись у формі tig-файлу чи графа за необхідності.

На другому етапі файли формату tig (Synphaty-Output) конвертувались у формат xml (TreeAlingner-Input), у якому зберігались анотовані еквівалентні речення німецькою та українською мовами і вирівнювались вручну у Notepad++, після чого візуалізувалися й редагувались у програмі Stockholm TreeAligner. Приклад вирівнювання паралельних речень у формі графа ілюструє рис 1.

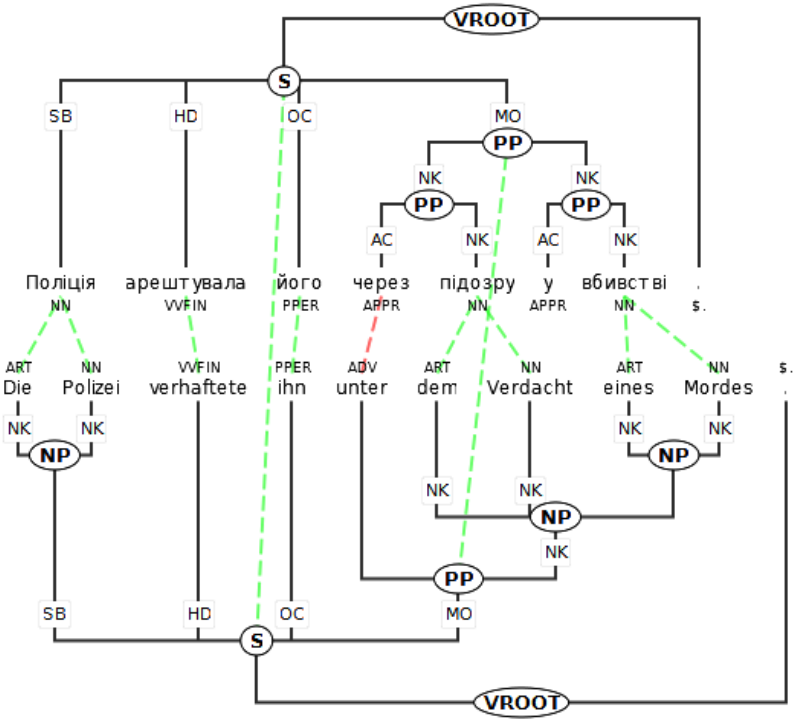


Рис. 1. Візуалізація білінгвального дерева у TreeAligner

Паралельні дерева дозволяють унаочнювати дивергенції у структурі мов. Охарактеризуємо окремі з них для німецької та української мов.

Залежно від типу речення нормативна граматика німецької мови чітко визначає порядок слів у реченні й позиціонування підмета, додатка, присудка та частин складеного присудка. На відміну від регламентованої структури німецького речення, порядок слів у реченнях української мови відносно вільний і визначається, насамперед, функціональною перспективою речення, яка обґрунтовується:

- теорією мовленнєвих актів, яка ґрунтується на гіпотезі, що структура повідомлення – це мовленнєва дія, детермінована ілокуцією автора [Austin 2005 : 17];
- темо-ремною теорією, яка відстоює гіпотезу про те, що лінійна послідовність структурних елементів повідомлення підпорядковується перебігу думок людини у напрямку від відомого (тема) до нового (рема) й визначається такими чинниками: комунікативна ситуація, контекст, ставлення автора до потенційного реципієнта, що є необхідною передумовою для забезпечення процесу комунікації [Lutz 1981 : 12].

Рід іменників у німецькій мові визначається граматичними маркерами і/або артиклями. Артикль як граматична категорія відсутній в українській мові, а його функції (ідентифікації, індивідуалізації та генералізації іменника) передаються лексичними (займенниками) або граматичними засобами (формотворчими морфемами).

Доповнення з прийменником у давальному чи знахідному відмінках, які слугують у німецькій мові для позначення напрямку чи інструмента, подекуди передаються українською мовою прямими або непрямими додатками з прийменниками в орудному чи місцевому відмінках.

Німецькі речення характеризуються трьома типовими ознаками:

- двочленність – обов'язкова присутність двох головних членів речення: підмета та присудка;
- вербальний характер речення: частиною присудка, також складеного іменного присудка, завжди виступає дієслово або дієслово-зв'язка;
- чітко визначене позиціонування присудка у реченні залежно від його типу: розповідне, питальне, спонукальне.

Жодна із зазначених ознак не типова для української мови. Тому присудок в українській мові може випускатись або мати структуру, відмінну від структури німецького присудка:

укр. *Відлига.* – нім. '*Es taut.'* замість *'*Taut.'*

укр. *Це стіл.* – нім. *'Das ist ein Tisch.*' замість *'*Dies Tisch.'*

укр. *Хворіти.* – нім. '*krank' sein.*

Український присудок у формі повнозначного дієслова німецькою мовою може відтворюватися номінальною групою з частково десемантизованим дієсловом, напр.:

укр. *об'єднуються* – нім. '*Verbindungen eingehen'*

укр. *ризикують* – нім. '*Risiken eingehen'*

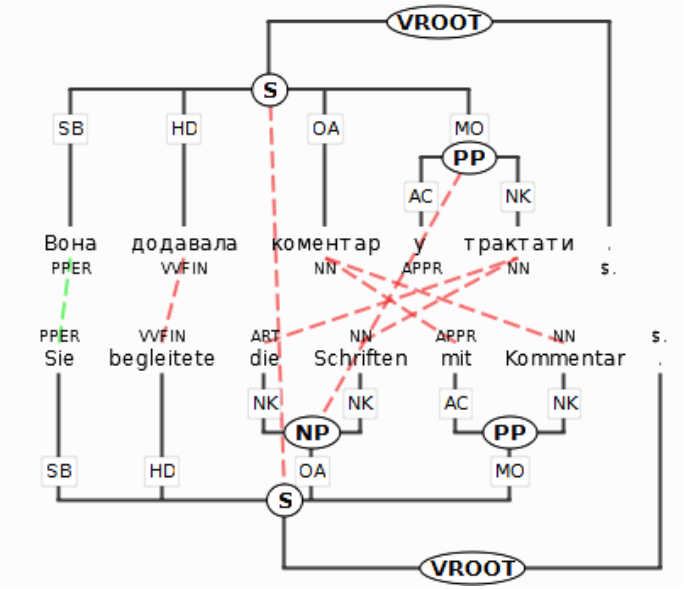


Рис. 2. Дивергенції на рівні заперечення

Присвійні займенники у німецькій мові слугують засобом вираження належності, упорядкування або єднання. Вони ставляться перед іменником і відповідають на питання *чий*, *чия*, *чиє*?

Кожному особовому займенникові у німецькій мові відповідає присвійний займенник, який також узгоджується у роді, числі та відмінку із постопозиційним іменником. Присвійні займенники української мови не мають відповідників серед особових займенників й функціонують як омоніми: укр. *свій* – нім. '*mein'*, '*dein'*, '*sein'* і под.

У заперечних реченнях у німецькій мові використовуються такі лексичні засоби:

- частки *nicht;*
- займенники *kein, keiner, niemand, nichts;*
- прислівники *nirgends, niemals, nie, nimmer, nirgendwo;*
- парні сполучники *weder...noch;*
- префікси: *un-, miss-, a-, il-* u.a.;
- еквівалентні речення *nein, doch.*

В українській мові існують усі відповідники німецьких заперечень за винятком заперечного займенника *kein* та його варіантів, які вживаються перед іменниками. Проте подвійне заперечення, типове для українських речень, виключає нормативна граматика німецької мови: укр. *ніхто не* – нім. *niemand.*

Подекуди конвенціоналізовані мовленнєві зразки настільки відрізняються, що їх неможливо перекладати еквівалентними лексичними засобами, а відтворення семантики речення мовою оригіналу вимагає уживання автентичних засобів мовою перекладу.

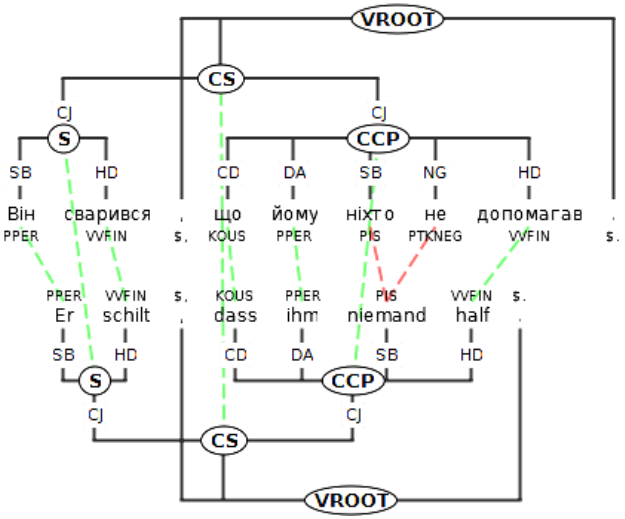


Рис. 3. Дивергенції на рівні прагматики

Проте дивергенції на морфологічному, синтаксичому й прагматичному рівнях німецької та української мов уможливлюють застосування вище зазначених інструментів для генерування паралельного банку дерев для цієї пари мов.

Створення таких лінгвістичних ресурсів має вагоме значення у процесі навчання студентів, для проведення наукових досліджень та прикладного застосуваня. У навчальному процесі вони слугують наочним матеріалом для демонстрації дивергенцій та конвергенцій контрастивної пари мов. У перекладацькій галузі такі ресурси мають значну перевагу порівняно з комерційними системами пам'яті перекладу, які створюються переважно на основі статистичних методів. Інтеграція інтелектуальних лінгвістичних ресурсів такого типу у системи пам'яті перекладу та системи машинного перекладу забезпечує якісний автоматичний переклад контенту.

1. *Austin John L.* How to do things with words / John L. Austin. – Cambridge, 2005. 2. *Boiko B.* Content management bible / B. Boiko. – Indianapolis, 2005. 3. *Drewer P.* Technische Dokumentation : eine Einführung in die übersetzungsgerechte Texterstellung und in das Content-Management / P. Drewer, W. Ziegler. – Würzburg, 2011. 4. CLAT-Client-Manual. – Saarbrücken, 2010. 5. CLAT-Intro. – Saarbrücken, 2010. 6. CLAT-In-For-Word-Manual. – Saarbrücken, 2010. 7. CLAT-UMMT-Manual-EN. – Saarbrücken, 2010. 8. CLAT-UMMT-Manual-DE. – Saarbrücken, 2010. 9. *Closs S.* Single Source Publishing : Modularer Content für EPUB & Co / S. Closs. – Frankfurt a. M., 2011. 10. Eagles. Evaluation of natural language processing systems. Final report (First phase). – Access mode : ftp://issco-ftp.unige.ch/pub/ewg96.pz.gz [Zugriff: 23.12.2002, 22:20 MEZ]. 11. *Kapanadze O.* Towards a semantically motivated organization of a valency lexicon for natural language processing: A GREG Proposal / O. Kapanadze, L. Wanner, S. Klatt // Proceedings of the EURALEX conference, Copenhagen, 2002. 12. *Kapanadze O.* Verbal Valency in Multilingual Lexica / O. Kapanadze // Workshop abstracts of the 7th language resources and evaluation conference-LREC2010. – Valletta, 2010. 13. *Lutz L.* Zum Thema "Thema" : Einführung in die Thema-Rhema-Theorie / L. Lutz. – Hamburg, 1981. 14. *Massion F.* Translation Memory Systeme im Vergleich / F. Massion. – Reutlingen, 2005. 15. *Reinke U.* Translation Memories: Systeme – Konzepte – linguistische Optimierung / U. Reinke // Fachrichtung Angewandter Sprachwissenschaft sowie Übersetzen und Dolmetschen der Universität des Saarlandes. – Sabest : Saarbrücker Beiträge zur Sprache- und Translationswissenschaft. Bd. 2. – Frankf./M.; Berlin [u.a.], 2004. 16. *Samuelsson Y.*

Presentation and representation of parallel tree-banks / Y. Samuelsson, M. Volk // In Proceedings of the Treebank-Workshop at Nodalida, Joensuu, 2005. 17. *Samuelsson Y.* Phrase alignment in parallel treebanks / Y. Samuelsson, M. Volk // In Proceedings of 5th Workshop on Treebanks and Linguistic Theories, – Prague, 2006. 18. *Searle J.* Speech acts: an essay in the philosophy of language / J. Searle. – Cambridge [u.a.], 2005. 19. *Smith G.* (2003), A Brief Introduction to the TIGER Treebank, Version 1. / G. Smith. – Potsdam, 2003. 20. Synphaty: Syntax Editor. – Manual. – Nijmegen: Max Planck Institute for Psycholinguistics, 2006.

Описано создание параллельного банка деревьев для украинского и немецкого языков. Проведено ручное тегирование и лемматизацию токенов для украинского языка; создан набор тегов для аннотирования украинского языка на уровне синтаксической структуры; выровнены и определены полные или частичные совпадения как для терминальных, так и для нетерминальных символов. Для аннотирования банка украинских предложений использован набор тегов в формате TIGER-XML, адаптированный к потребностям формального описания грамматической структуры украинского языка.

Ключевые слова: банки деревьев, аннотация, тегирование частей речи, выравнивание, перевод.

In this paper, we describe outcomes of an experiment on building a parallel Treebank for bridging the Ukrainian language with the German language. The aim of the mentioned experiment was: manually tagging and lemmatization of tokens for Ukrainian corpora; establishing of the compatible tagset for Ukrainian and introduction of the specific syntactic phrasal categories; production of the parallel trees from the bilingual resources; alignment of the German-Ukrainian parallel trees; determining "good" and "fuzzy" matches between the non-terminal and terminal nodes across the syntactic structures of the languages involved. The Ukrainian Treebank was annotated according to an adapted version of the German TIGER guidelines with the necessary changes relevant to the Ukrainian grammar formal description.

Keywords: treebanks, annotation, POS, alignment, translation.

Стаття надійшла до редакції 10.09.2012

УДК 81'322

Андрій Романюк,
Мар'яна Романишин
Національний університет "Львівська Політехніка"

ТОНАЛЬНИЙ СЛОВНИК УКРАЇНСЬКОЇ МОВИ НА ОСНОВІ СЕНТИМЕНТ-АНОТОВАНОГО КОРПУСУ

Докладно розглянуто всі етапи створення сентимент-анотованого корпусу українськомовних відгуків і тонального словника на його основі.

Ключові слова: емоційно-смисловий аналіз, тональний словник, сентимент-анотований корпус, засоби для анотування текстів.

Протягом останніх десяти років у галузі опрацювання природної мови широкої популярності набув емоційно-смисловий аналіз. Про це свідчить численна кількість досліджень у цій сфері. З'явилися такі цікаві проекти, як дослідження емоційного забарвлення відгуків про готелі [Kasper], банки [Deep sentiment analysis], ресторани, коментарів про фільми [Yessenov], продукти, повідомлень у блогах і соцмережах про політичні події тощо.

На сьогодні емоційно-смисловий аналіз застосовується в різних галузях знань: у соціології (наприклад, збирання даних із соціальних мереж про певні вподобання людей), політології (наприклад, збирання даних про політичні погляди), маркетингу (наприклад, збирання даних про найпопулярніші товари), медицині та психології (наприклад, визначення депресивних настроїв) тощо [Давыдов].

Наразі немає доступного емоційно-смислового аналізатора для української мови, але це питання активно опрацьовується. Створення такого аналізатора передбачає насамперед укладання тонального словника та сентимент-анотованого корпусу. У статті описано створення сентимент-анотованого корпусу відгуків українською мовою, а також генерацію тонального словника на основі такого корпусу.

Поняття тонального словника та методи його створення

Тональний словник, у найпростішому його вигляді, становить список слів і словосполучень зі значенням тональності для

кожного слова. Це може бути як числове значення в межах певної шкали (наприклад, від 1 до 10, де 1 – негативне слово, а 10 – позитивне), так і категорія (наприклад, позитивне чи негативне слово). Кожному слову приписується його частиномовна характеристика та початкова форма у випадку, коли словник містить різні словоформи одного слова.

Часто списки слів у словнику обмежують лише самостійними частинами мови. Наприклад, у роботі над повідомленнями російською мовою [Пазельская] використовувався тональний словник із найчастотніших іменників, прикметників, прислівників та дієслів, зібраних зі статей ЗМІ. Для кожного слововживання визначено частину мови, тональність і силу тональності за шкалою від 1 до 3. Прикметники і прислівники поділені на позитивні, негативні та ті, що підсилюють тональність. Іменники поділені на позитивні, негативні, потенційно позитивні та потенційно негативні (залежно від оточення: позитивні в позитивному оточенні, а негативні – в негативному). Дієслова поділені на 8 класів залежно від оточення і ролі в реченні, окремо подані дієслова-зв'язки. Є також слова-інвертори, що змінюють тональність сусідніх слів на протилежне.

Набір слів у тональному словнику переважно стосується однієї тематики, яку опрацьовуватиме аналізатор. Оскільки складання такого списку слів власноруч потребує дуже багато часу, застосовують автоматичні способи видобування слів певної тематики зі сентимент-анотованих корпусів та онтологій, що стосуються певної тематики. Якщо ж тематика словника не передбачена, використовують загальнотематичні семантичні мережі. Наприклад, у праці [Moilanen] послуговуються тональним словником, створеним на основі WordNet 2.1. Прикладом багатомовного тонального словника є SentiWordNet, який часто використовують для здійснення різноманітних завдань у сфері емоційно-смислового аналізу [Denecke]. Достатньо часто послуговуються також General Inquirer Lexicon [Agrin].

Хоч застосування онтологій є досить поширеним способом генерації тональних словників, воно має певні недоліки, яких мож-

на уникнути за умови залучення анотованих корпусів. По-перше, сентимент-анотований корпус є джерелом контекстів для слів, які потраплять до тонального словника, що дає можливість краще визначити емоційне забарвлення кожного слова, а в онтології переважно містять ізольовані слова. По-друге, корпус подає велику кількість нелітературних слів, які важливі для визначення емоційного забарвлення повідомлення, але відсутні в онтологіях.

Оскільки для української мови немає доступного сентимент-анотованого корпусу, було вирішено створити такий корпус власноруч і на його основі згенерувати основу тонального словника.

Створення сентимент-анотованого корпусу

Сентимент-анотований корпус – це корпус текстових повідомлень, у якому кожному повідомленню присвоєно значення емоції, яку воно передає. Створення сентимент-анотованого корпусу – це невід'ємна частина реалізації системи емоційно-смислового аналізу.

По-перше, такий напівавтоматично розмічений корпус дає розуміння того, як за допомогою текстових повідомлень людина висловлює свої емоції щодо певного об'єкта (емоційно забарвлена лексика, ідеограми на кшталт смайликів, пунктуація). По-друге, такий корпус може стати основою для створення тонального словника. По-третє, готовий сентимент-анотований корпус є достатнім засобом для перевірки роботи розробленої системи емоційно-смислового аналізу.

Процес створення сентимент-анотованого корпусу було поділено на такі етапи:

- добір текстових повідомлень для майбутнього корпусу;
- визначення програмних засобів для анотування текстових повідомлень;
- вироблення схеми анотування текстових повідомлень;
- власне анотування текстових повідомлень.

Формування вибірки текстових повідомлень для майбутнього корпусу

Першим кроком для добору повідомлень для майбутнього сентимент-анотованого корпусу є визначення тематики текстів.

Тематика важлива для емоційно-смислового аналізу, оскільки ті самі слова в межах різних тематик можуть мати протилежні тональності. Тематикою повідомлень для нашого дослідження стали відгуки про заклади харчування. На актуальність цієї тематики вказує широке обговорення її на форумах та в соціальних мережах.

Українськомовні відгуки про заклади харчування, які стали основою для корпусу, було взято з популярного форуму http://posydenky.lvivport.com/ та з сайту http://v.lviv.ua/. Вибір припав саме на ці сайти через значну кількість відгуків, які відповідають обраній тематиці, а також тому, що більшість відгуків на цих сайтах написані українською мовою, що дає можливість частково уникнути проблеми фільтрування повідомлень за мовною ознакою та зосередитись власне на створенні корпусу. Структура відгуків на обох сайтах однакова, оскільки кожен відгук містить ідентифікатор автора, час написання відгуку та власне текстове повідомлення.

Визначення програмних засобів для анотування текстів

На сьогодні є багато зручних засобів для анотування текстів. З-поміж найпоширеніших можна назвати: Callisto, WordFreak, GATE, BRAT, DOMEO, CLaRK, Ellogon, UAM та ін [Shapiro]. Коротко про деякі з цих засобів:

Callisto (http://callisto.mitre.org) – це простий засіб анотування, розроблений для підтримки лінгвістичного анотування текстів для будь-якої мови, з підтримкою Unicode. Анотовані тексти зберігаються у форматі ATLAS, який можна імпортувати в xml.

WordFreak (http://wordfreak.sourceforge.net) – це засіб, який підтримує ручне та автоматичне анотування лінгвістичних даних, а також дає можливість застосовувати автоматичне навчання на виправленні анотацій, зроблених людиною. Цей засіб здебільшого використовується для перевірки готового анотованого тексту.

GATE (http://gate.ac.uk/) – це ціла система для опрацювання природної мови, яка, крім інших засобів, містить також і засіб для ручного та автоматичного анотування текстів. GATE дає можливість створювати найрізноманітніші схеми анотування.

DOMEO (http://annotationframework.org/) – це онлайнове середовище, яке допомагає анотувати текст на основі вбудованої онтології. Цей засіб підтримує ручне, напівавтоматичне та автоматичне анотування.

CLaRK (http://www.bultreebank.org/clark/index.html) – це система для розробляння корпусів, основною метою якої є мінімізувати ручну роботу у процесі створення лінгвістичних ресурсів.

Ellogon (http://www.ellogon.org/) – це багатомовне міжплатформне середовище для опрацювання природної мови з відкритим кодом, яке застосовують як окремі науковці, так і компанії, які займаються створенням систем опрацювання природної мови.

Дослідивши описані вище засоби анотування текстів, ми визначили, що для нашого завдання найбільше підходять системи GATE, CLaRK та Ellogon. З-поміж них було вибрано GATE, оскільки ця система проста у використанні, надає зручні засоби редагування анотованого тексту, можливість створення складних схем анотування, можливість зберігати анотовані тексти в форматі xml, працювати з кількома текстовими файлами і кількома схемами анотування одночасно, аналізувати саме українську мову, а також забезпечує підсвічування анотованого тексту різними кольорами, що зручно під час накладання анотацій одна на одну.

Вироблення схеми анотування текстових повідомлень

Схему анотування для сентимент-анотованого корпусу було розроблено за допомогою пакета CREOLE (Collection of REusable Objects for Language Engineering), який має клас AnnotationSchema. Цей пакет дає можливість створювати схеми анотування та діалогові вікна для роботи з ними. Файл конфігурацій creole.xml містить інформацію про ресурси, які використовуються. У нашому випадку – це назви файлів із відповідними мітками [Using GATE Developer].

Розроблена схема анотування українськомовних відгуків про заклади харчування має такі структурні одиниці:

- автор;
- дата;
- відгук;

- цитування попереднього повідомлення;
- речення;
- частина складного речення;
- назва закладу харчування, про який іде мова;
- слово;
- url-адреса.

Автор відгуку позначається міткою nickname. У вхідних даних автор вказаний у першому рядку.

Дату написання відгуку (мітка date) подано після автора.

Власне відгук автора (мітка review) виділено без урахування цитування попереднього повідомлення, якщо таке є. Це потрібно для того, щоб визначати суб'єктивну оцінку автора поточного відгуку, а не автора цитованого повідомлення.

Цитування попереднього повідомлення визначаємо за характерними ознаками (повідомлення починається на "Цитата:") і позначаємо міткою citing.

Окремо виділяється кожне речення відгуку (мітка sentence).

Кожна частина складного речення позначається міткою clause. У випадку простих речень ці дві мітки збігаються. Для кожного такого підречення визначається тональність (атрибут під назвою sentiment): позитивне (positive), негативне (negative) чи нейтральне (neutral). Для цілого складного речення тональність не визначено, оскільки одне складне речення може містити інформацію і з негативним, і з позитивним забарвленням. У відгуку окремо анотуємо також і назву закладу чи закладів харчування, про які йде мова (мітка target).

У кожному емоційно забарвленому підреченні (clause) виділяємо слова чи ідіоматичні словосполучення (позначаються міткою word).

Для кожного виділеного слова було визначено набір атрибутів:

- початкова форма слова (атрибут lemma): значення вписується в автоматизованому режимі;
- частина мови (атрибут part_of_speech) має такі значення: n – іменник, v – дієслово, adj – прикметник, adv – прислівник, pro – займенник, con – сполучник, pre – прийменник, par – частка, exc – вигук, num – числівник, und – інше (напр., смайлики);

- емоційне забарвлення (атрибут sentiment) має такі значення: positive – слова чи словосполучення, що самостійно виражають позитивне значення; negative – слова чи словосполучення, що самостійно виражають негативне значення; neutral – слова чи словосполучення, що самостійно не виражають ні позитивного, ні негативного значення; intensifier – слова-підсилювачі, що не мають самостійного емоційного забарвлення, але підсилюють емоційне забарвлення наступного слова чи цілого підречення (наприклад, такі слова, як *дуже*, *надзвичайно*, *безмежно*, *вкрай*, *досить*); invertor – слова-інвертори, які не мають самостійного емоційного забарвлення, але змінюють емоційне забарвлення наступного слова чи цілого підречення на протилежне (наприклад, такі слова, як *не*, *нема*, *немає*, *неможливо*, *нереально*);
- емоція (атрибут emotion) має такі значення: joy – радість, sadness – сум, anger – злість, fear – страх, disgust – огида, surprise – здивування, none – якщо слово чи словосполучення самостійно не передає емоції.

Атрибути sentiment та emotion матимуть значення neutral і none відповідно для всіх службових частин мови. Лише самостійні частини мови можуть мати інші значення, оскільки лише самостійні частини мови можуть нести певне емоційне забарвлення ізольовано від контексту.

Для того щоб визначити базові емоції для схеми анотування відгуків, було проаналізовано набори базових емоцій за теоріями різних психологів. Шість вищезгаданих емоцій визнані базовими емоціями людини відповідно до концепції відомого психолога Пола Екмана. Базові емоції Екмана у психології – це культурно незалежні емоції, які з'являються в людини впродовж перших шести місяців її життя. Це також такий набір емоцій, які легко передати як мімікою, так і вербально [Ekman]. Звичайно, існують й інші концепції, але базові емоції Екмана вважають у сучасній науковій думці фундаментальними.

Міткою url позначаємо url-адреси, якщо такі є у відгуку.

Інформація про кожну мітку розробленої схеми анотування занесена в окремий xml-файл. На рис. 1 представлено структуру файлу для мітки clause.

```
<?xml version="1.0"?>
<schema>
  <element name="clause" type="string">
    <complexType>
    <attribute name="sentiment" default="neutral"
      <simpleType>
        <restriction base="string">
        <enumeration value="positive"/>
        <enumeration value="negative"/>
        <enumeration value="neutral"/>
        </restriction>
      </simpleType>
    </attribute>
  </complexType>
 </element>
</schema>
```

Рис. 1. Структура файлу clause.xml

На рис. 2 зображено діалогове вікно з назвою і атрибутом мітки. Значення можна автоматично вибрати з випадного списку.

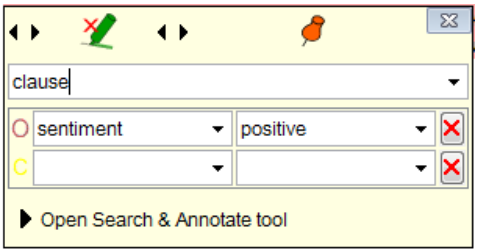


Рис. 2. Діалогове вікно з атрибутом мітки clause в середовищі Gate 7.0

У файл конфігурацій creole.xml також внесено відповідний запис про цю мітку:

```
<AUTOINSTANCE>
    <PARAM NAME ="xmlFileUrl" VALUE ="resources/schema/clause.xml" />
</AUTOINSTANCE>
```

Власне анотування текстових повідомлень

Із метою створення сентимент-анотованого корпусу відгуків українською мовою було залучено студентів кафедри прикладної лінгвістики Національного університету "Львівська політехніка". Для них було розроблено розрахункову роботу в межах курсу "Комп'ютерна лінгвістика". Метою роботи стало ознайомлення студентів із практикою створення корпусів та їх анотування.

Кожен студент отримав методичні вказівки щодо роботи в середовищі GATE, докладний опис розробленої схеми анотування з поясненнями і прикладами, а також власне набір відгуків у текстовому форматі. Після анотування студенти зберігали кожен відгук у форматі xml.

Результатом виконання розрахункової роботи став корпус сентимент-анотованих відгуків обсягом 600 відгуків у зручному для використання форматі. Цей корпус, однак, потребує додаткової перевірки через суттєву кількість помилок.

Приклад анотованого відгуку

Тетянка

24.04.2009, 22:29

а я "Цукерню" люблю і кафешку з тортиками на Дудаєва....

Відгук складається з ідентифікатора автора, дати і часу написання та власне тексту. Після анотування цей відгук було збережено у форматі xml. Структура xml-файлу містить інформацію про кодування, відомості про сам документ; власне відгук із розміткою анотацій і, нарешті, опис міток. Відгук із розміткою анотації має такий вигляд:

<TextWithNodes><Node id="0" />Тетянка
<Node id="8" />
<Node id="9" />
<Node id="11" />24.04.2009, 22:29<Node id="28" />
<Node id="30" />
<Node id="32" />а<Node id="33" /> <Node id="34" />я<Node id="35" /> "<Node id="37" />Цукерню<Node id="44" />" <Node id="46" />люблю<Node id="51" /> <Node id="52" />і<Node id="53" /> <Node id="54" />кафешку <Node id="61" /> <Node id="62" />з<Node id="63" /> <Node id="64" /> тортиками<Node id="73" /> <Node id="74" />на<Node id="76" /> <Node id="77" />Дудаєва<Node id="84" />....<Node id="88" />
<Node id="89" /> <Node id="90" /></TextWithNodes>

Приклад мітки clause у форматі xml:

```
<Annotation Id="7" Type="clause" StartNode="32" EndNode="88">
<Feature>
    <Name className="java.lang.String">sentiment</Name>
    <Value className="java.lang.String">positive</Value>
</Feature>
</Annotation>
```

Приклад мітки word у форматі xml:

```
<Annotation Id="11" Type="word" StartNode="46" EndNode="51">
<Feature>
    <Name className="java.lang.String">part_of_speech</Name>
    <Value className="java.lang.String">v</Value>
</Feature>
<Feature>
    <Name className="java.lang.String">lemma</Name>
    <Value className="java.lang.String">любити</Value>
</Feature>
<Feature>
    <Name className="java.lang.String">sentiment</Name>
    <Value className="java.lang.String">positive</Value>
</Feature>
<Feature>
    <Name className="java.lang.String">emotion</Name>
    <Value className="java.lang.String">joy</Value>
</Feature>
</Annotation>
```

Із наведених вище прикладів можна побачити, що кожна мітка містить початкову і кінцеву позиції фрагменту тексту, анотованого цією міткою, а також назви і значення її атрибутів, якщо такі є. Таке xml-дерево можна легко розпізнати і застосовувати для майбутніх досліджень.

Створення тонального словника

Створений сентимент-анотований корпус відгуків українською мовою дозволяє згенерувати основну частину тонального словника.

Кожне тональне слово в корпусі приведено до нижнього регістру й записано до словника. Таким чином, словник містить

слова і словосполучення, інформацію про частиномовну належність кожного слова, його тональність (позитивну чи негативну) й емоційне забарвлення, якщо таке є. До словника також записується початкова форма слова з такими ж атрибутами. Такі відомості надають можливість застосувати тональний словник для здійснення емоційно-смислового аналізу: у випадку відсутності потрібної словоформи в тональному словнику, слову присвоюватиметься тональність його початкової форми.

Оскільки службові слова ізольовано не несуть емоційного забарвлення, словник міститиме лише самостійні частини мови. Окрім початкових форм слів, словник міститиме й інші словоформи. Це пов'язано з тим, що використовуючи лише початкову форму слова, ми втрачаємо морфологічну інформацію, яка може бути корисною для подальшого емоційно-смислового аналізу. Наприклад, слова *люблю* і *любив* можуть мати різну тональність. Якщо в першому випадку тональність швидше за все буде позитивною, то в другому випадку може бути менш позитивною чи навіть негативною.

Окрім позитивних та негативних слів і словосполучень, тональний словник також міститиме слова-інвертори (наприклад, *не*, *немає*, *неможливо*) і слова-підсилювачі (*дуже*, *надзвичайно*, *безмежно*, *вкрай*).

Таким чином, за допомогою розробленого сентимент-анотованого корпусу відгуків можна згенерувати тональний словник, який міститиме позитивні та негативні слова і словосполучення, слова-інвертори та слова-підсилювачі.

Перспективою презентованого дослідження є подальше поповнення такого словника, зокрема шляхом залучення словників синонімів і антонімів. Надалі цей словник буде використано для реалізації емоційно-смислового аналізатора, метою якого є визначення емоційного забарвлення повідомлень українською мовою.

1. *Давыдов А. А.* Системная социология: Opinion Mining / А. А. Давыдов. – М., 2009. – Режим доступу: http://www.isras.ru/index. php?page_id=1024. 2. *Пазельская А.* Метод определения эмоций в текстах на русском языке / А. Г. Па-

зельская, А. Н. Соловьев // Компьютерная лингвистика и интеллектуальные технологии. Сб. научных статей. Вып. 10 (17). – М., 2011. – С. 510–522. 3. *Agrin N.* Developing a Flexible Sentiment Analysis Technique for Multiple Domains / Nate Agrin. – 2006. – Режим доступу: http://courses.ischool.berkeley.edu/i256/f06/projects/agrin.pdf. 4. Deep sentiment analysis with attensity analyze optimises Lloyds' customer service. – Режим доступу: http://www.attensity.com/wp-content/uploads/2010/09/LloydsSuccessStory.pdf. 5. *Denecke K.* Using SentiWordNet for Multilingual Sentiment Analysis / Kerstin Denecke // ICDE Workshops. – 2008. – P. 507–512. 6. *Ekman P.* Basic Emotions / Paul Ekman // Handbook of Cognition and Emotion. – John Willey & Sons Ltd, 1999. – P. 45–60. 7. *Kasper W.* Sentiment Analysis for Hotel Reviews / Walter Kasper // Proceedings of the Computational Linguistics-Applications Conference. – Poland, 10/2011. – P. 45–52. 8. *Moilanen K.* Multi-entity Sentiment Scoring / Karo Moilane, Stephen Pulman // Proceedings of Recent Advances in Natural Language Processing (RANLP 2009). – Bulgaria (Borovets). – September 14–16 2009. – P. 258–263. 9. *Shapiro S.* Natural Language Tools for Information Extraction for Soft Target Exploitation and Fusion / Stuart C. Shapiro. – NY, 2007. – P. 36–37. – Режим доступу: http://www.cse.buffalo.edu/~shapiro/Papers/shaaxt07.pdf. 10. Using GATE Developer. – Режим доступу: http://gate.ac.uk/sale/tao/splitch3.html#chap:developer. 11. *Yessenov K.* Sentiment Analysis of Movie Review Comments / Kuat Yessenov, Sasa Misailovic. – Massachusetts Institute of Technology, Spring 2009. – Режим доступу: http://people.csail.mit.edu/kuat/courses/6.863/report.pdf.

Представлены все этапы создания сентимент-аннотированного корпуса украиноязычных отзывов и тонального словаря на его основе.

Ключевые слова: сентимент-анализ, тональный словарь, сентимент-аннотированный корпус, средства для аннотирования текстов.

This paper deals with the implementation of sentiment-annotated corpus of Ukrainian reviews and the creation of sentiment dictionary based on it. The paper describes all the stages of sentiment-annotated corpus creation in detail.

Keywords: sentiment analysis, sentiment dictionary, sentiment-annotated corpus, tools for annotating texts.

Стаття надійшла до редакції 1.09.2012

УДК 81’322

Олена Сірук
Київський національний університет імені Тараса Шевченка,
Іван Держанський
Інститут математики й інформатики
Болгарської академії наук

ЛЕКСИЧНІ ПЕРЕКЛАДНІ ЕКВІВАЛЕНТИ В БОЛГАРСЬКИХ І УКРАЇНСЬКИХ ПАРАЛЕЛЬНИХ ТЕКСТАХ

Мова йде про укладання Корпусу болгарських і українських паралельних текстів та його застосування для дослідження лексичної семантики українських і болгарських перекладних еквівалентів.

Ключові слова: корпусна лінгвістика, паралельні тексти, корпус, лексична семантика, перекладні еквіваленти.

Паралельні корпуси в лексичних дослідженнях

На сучасному етапі розвитку лінгвістики для зіставного аналізу лексичної семантики, як і загалом для міжмовних досліджень, усе частіше застосовують корпуси паралельних – оригінальних і перекладних – текстів [Cysouw, Wälchli]. Такі корпуси розробляються для багатьох мов, зокрема слов’янських (наприклад, двомовні корпуси з українськими текстами в складі Національного корпусу російської мови http://ruscorpora.ru, Корпус паралельних російських і болгарських текстів http://rbcorpus.com тощо). Однак українська та болгарська мови досі не ставали об’єктом зіставного корпусного аналізу, а тексти цими мовами не були об’єднані в паралельний корпус. Автори статті мають на меті розповісти про свій досвід створення такого корпусу, а також про деякі аспекти його застосування для лексико-семантичного та статистичного аналізу перекладних еквівалентів.

Корпус болгарських і українських паралельних текстів (КУБ)

Двомовний корпус містить болгарські й українські паралельні тексти, наявні в електронних бібліотеках або відсканковані та розпізнані нами з паперових видань. Цим зумовлюється перевага в корпусі творів художньої літератури, зокрема романів, які домінують у таких джерелах.

Оскільки оригінальних та перекладених паралельних текстів для української та болгарської мов наразі обмаль, а тим більше в комп'ютерному вигляді та мережевому доступі, було вирішено використовувати як матеріал для корпусу також паралельні болгарські й українські художні переклади з інших мов. Тобто поняття паралельності застосовується нами в ширшому значенні, аніж його тлумачать, приміром, розробники паралельних корпусів НКРЯ [Добровольский, Кретов, Шаров], російсько-болгарського, російсько-словацького корпусу [Гарабик, Захаров], лексикографи [Лендау : 314], залучаючи до паралельних корпусів тільки пари текстів "оригінал – переклад". Інші дослідники ототожнюють термін "паралельні тексти" із терміном "бітекст", яким позначають просто дві версії одного тексту, зазвичай різними мовами [Vitas, Krstev, Laporte]. Ми розуміємо під паралельним корпусом як "єдність підмножини оригінальних текстів та підмножини їх перекладів на іншу(і) мову(и)" [Демська : 90], так і те, що дослідники називають взаємопаралельним корпусом. Такі корпуси "містять як оригінали, так і переклади на конститутивні мови корпусу" [Демська : 100]. Тож наш корпус має декілька секторів, кожен з яких охоплює паралельні болгарські й українські тексти, перекладені з однієї мови. Усі сектори приблизно однакові за розміром. На сьогодні українська частина містить приблизно по 700 тис. слововживань із кожного сектору, болгарська – відсотків на 15 більше, з огляду на граматичні особливості мови та стилістичні вподобання багатьох перекладачів (це можна побачити на прикладі паралельних перекладів "Декамерона" Дж. Бокаччо, де італійський оригінал речення та болгарський відповідник Н. Іванова й Д. Петрова вдвічі довші, ніж український М. Лукаша: *Calandrino, essendogli il vino uscito dal capo, si levò la mattina; e come scese giù guardò e non vide il porco suo e vide l'uscio aperto* || *На следната утрин, когато главата му се избистрила от виното, Каландрино станал, слязъл долу, огледал се и видял, че прасето е изчезнало, а вратата – отворена* || *Прочумався рано-вранці Каландріно після випивки, встав, дивиться – кабана нема, а двері одчинені*). Ми працюємо над

забезпеченням корпусу оригіналами всіх перекладів. Застосування такого підходу дозволяє не тільки побудувати корпус корисного розміру, а й надати базу для дослідження ширшого кола проблем порівняльного мовознавства (через появу опосередкованих відповідників).

Розширення набору мов оригіналів почалося із залучення текстів близькоспорідненими мовами. З'ясувалося, що в мережі найбільше перекладених і українською, і болгарською мовами книг польських та російських, тому поки ми обмежилися цими чотирма слов'янськими мовами. З інших європейських мов було віднайдено і залучено до корпусу паралельні переклади з англійської, німецької, французької, іспанської та італійської мов. Видається цікавим розширити корпус перекладами з інших мов Західної Європи, а можливо, і з неєвропейських, хоча можна передбачити, що із доступністю паралельних текстів буде сутужно.

Очевидно, що чим дальша відстань між мовами, тим більшими є розходження між болгарськими й українськими співвідносними перекладами. Але в процесі перекладу з близькоспорідненої мови перекладач під впливом мови оригіналу може свідомо чи несвідомо обрати вираз, близький до наявного в цій мові, але периферійний для мови перекладу. Імовірність цього значно менша для перекладу з неблизькоспорідненої мови, оскільки тотожного виразу в ній, швидше за все, не буде. Виокремити такі лексичні відповідники важче, зате цікавіше, оскільки вони більш "безпосередні", адже на перекладача менше впливає мова оригіналу, а якщо й впливає, характер цього впливу становить інтерес сам по собі.

Варто зазначити, що тексти, знайдені в мережі Інтернет, часто-густо містять багато помилок розпізнавання, і доводиться їх редагувати зі зверненням до паперового джерела, що потребує немало часу. Наразі триває поповнення корпусу відсканованими книгами паперових бібліотек, хоча низька якість друку часто унеможливлює розпізнання тексту, і такий текст виявляється не придатним для корпусних досліджень.

Тексти поділяються на речення за допомогою власних програмних засобів. У випадку деяких текстів з особливо довгими реченнями видається доцільним зарахування крапки з комою, а іноді й двокрапки, до розділових знаків кінця речення. Без цього відповідні одна одній частини тексту можуть виявитися незручно великими, особливо якщо межі речень не збігаються, що трапляється досить часто (до таких текстів можна віднести "Доктор Фаустус" Т. Манна, "Декамерон" Дж. Боккаччо, "Сто років самотності" Г. Маркеса).

Вирівнювання поділених на речення текстів здійснюється за допомогою програми Hunalign [Varga та ін.], http://mokk.bme.hu/resources/hunalign/. Часткова автоматизація цього процесу посприяла виявленню проблеми "непаралельності" текстів оригіналу та перекладу. Численні розбіжності між зіставлюваними болгарськими й українськими текстами можуть зводитися до скорочення оригінального тексту в перекладі (у трьох основних різновидах: пропуски окремих речень, переформулювання речень у бік скорочення абзацу зі збереженням змісту загалом та викидання з тексту значних його частин: у перекладі П. Киневою роману П. Загребельного "Переходимо до любові" пропущено вбудовану п'єсу), перестановки цілих абзаців і навіть розділів (так, розділи 2, 3 та 10 роману А. Гуляшки "Пригода опівночі" в перекладі О. Кеткова відповідають розділам 9, 1 та 8) або повної змістової невідповідності частин тексту. Зараз можна тільки здогадуватися, на якому рівні трансформувався текст у кожному випадку: чи на рівні перекладача, який із власної ініціативи намагався "покращити" оригінал, чи то з волі редактора, який мав певні вимоги щодо обсягів перекладу, чи відповідно до загальної настанови щодо спрощення перекладів задля ознайомлення з ними широкої читацької аудиторії. Іншою причиною таких розбіжностей у текстах оригіналу і перекладу може бути наявність різних оригінальних видань, у тому числі й скорочених. Але це вже тема для окремої розвідки.

Такі "недопаралельні" тексти викликають неоднозначне ставлення. З одного боку, вони є матеріалом для дослідження, на-

приклад, перекладних трансформацій або історії перекладу, а з іншого – ускладнюють корпусне опрацювання текстів, адже смислові зміни часто супроводжуються формальними перегрупуваннями. Там, де пропускаються або зміщуються речення й абзаци, частини текстів доводиться вирівнювати вручну, а це потребує додаткового часу.

До одиниць, які становлять інтерес для теорії перекладу та не піддаються автоматичному опрацюванню, належать співвідносні, але не паралельні уривки тексту. Наприклад, у романі "Ім'я рози" У. Еко латинський текст Кассінського свідоцтва (X ст.) *Sao ko kelle terre per kelle fini ke ki kontene, trenta anni le possette parte sancti Benedicti* в болгарському перекладі Н. Іванова передано дослівно сучасною мовою як *Знам, че тези земи в границите, в които са очертани, в продължение на трийсет години са били притежание на свети Бенедикт*, а в українському перекладі М. Прокопович його замінено на уривок з "Повчання" Володимира Мономаха (кінець XI – початок XII ст.) *Вскую печална єси, душе моя? Вскую смущаєши мя? Уповаю на Бога, яко исповімся єму* (відповідно до авторського задуму, який полягає саме в байдужості до значення сказаного дійовою особою: для порівняння, в англійському перекладі В. Вівера, 1983, доступному на сайті "Google Книги", цієї фрази взагалі немає).

Проблеми виникають і з поетичними вставками у прозових текстах: стрункість віршової форми досягається здебільшого за рахунок точності перекладу, тому було вирішено залишати для опрацювання тільки масиви прози. Скороченню може підлягати і цілеспрямовано перекручена мова окремих героїв (прикладом є мова Сальватора зі згаданого вище роману, який в оригіналі і в українському перекладі "говорив усіма мовами і не говорив жодною", тобто мішанкою слів та конструкцій різних мов; у болгарському перекладі його мова мало відрізняється від літературної).

Хоча й рідко, але трапляються несвідомі розбіжності, спричинені неоднозначністю виразу в оригінальному тексті, який два перекладачі зрозуміли по-різному. Так, польське *Wybiła godzina*

(С. Лем "Фіаско") передано Д. Андрухівим українською мовою як *Час настав*, а Л. Василевою болгарською – як *Удари един часът* 'Пробила перша година'. Спорідненим, частішим (і цікавішим) явищем є неминуча розбіжність у перекладах неоднозначних лексичних одиниць мови оригіналу, як, наприклад, англ. *you* (*ти* чи *ви*) або нім. *Kirsche* (*вишня* чи *черешня*).

Нарешті, у перекладах трапляються неточності та помилки. До неточності можна віднести переклад французького *cochon de lait* 'молочне порося' (Ж. Верн "Таємничий острів") болгарською мовою як *прасенце* (замість *прасе сукалче*): *После с помощта на Наб морякът нагласи ръжена и добре изкорменият кабиай скоро се опече на пламналия буен огън като обикновено прасенце* (переклад Й. Петрова) || *Потім із Набовою допомогою моряк прилаштував рожен, і майстерно випатрана водосвинка, схожа на молочне порося, незабаром смажилася на веселому ясному вогні* (перекладач – В. Омельченко). Помилку перекладу можна виявити, якщо порівняти паралельні речення *Czas biegu sygnałów nie może być dłuższy od czasu reakcji składników komputera* || *Времето за преминаване на сигнала не трябваше да бъде по-голямо от времето, за което реагират съответните съставни части от компютъра* || *Швидкість руху сигналів не може бути більшою від швидкості реакції складових елементів комп'ютера* (С. Лем "Фіаско"). Чим більша швидкість руху, тим менше часу на нього йде; тож в українському перекладі має бути *швидкість руху сигналів не може бути меншою*.

У процесі накопичення текстів було помічено певну кореляцію між мовами і жанрами творів, яка не сприяє збалансованості корпусу. Навіть якщо знайдено багато текстів, перекладених з якоїсь мови, це може означати, що є багато перекладів творів одного автора. А таке не бажано з погляду статистичних досліджень лексики. В окремих випадках у найчастотніші слова потрапляють вигадані авторами власні назви (наприклад, із фантастичних романів І. Єфремова), що зовсім зайве. Коли авторів двоє-троє, з великою імовірністю можна припускати, що вони працювали в одному жанрі. Тож зараз у болгарському секторі

переважає політичний детектив (А. Гуляшки, Б. Райнов), у польському – історичний роман (Б. Прус, Г. Сенкевич), у російському – наукова фантастика (О. Бєляєв, І. Єфремов). Дається взнаки і мала кількість перекладачів (на стиль перекладу впливають їхні індивідуальні вподобання, особливо в лексиці та фразеології). Зрозуміло, що буде нелегко вийти за межі белетристики і домогтися того, щоб у корпусі були наявні паралельні українсько-болгарські публіцистичні, наукові, мистецтвознавчі, мемуарно-біографічні та інші тексти в представницькій кількості, але слід прагнути до зменшення згаданого дисбалансу хоча б у межах художніх жанрів.

Корпусне дослідження лексичних перекладних еквівалентів

Паралельний корпус надає широкі можливості для статистичного вивчення міжмовних лексичних відповідностей з метою уточнення значень слів або кореляцій між словами у певних значеннях та умовами використання цих слів. Послідовність кроків роботи з лексичними еквівалентами в паралельному корпусі може бути такою:

- окреслюємо в одній з двох мов лексико-семантичне поле (ЛСП), лексико-семантичну групу (ЛСГ) чи окремі лексико-семантичні варіанти (ЛСВ), особливості перекладу яких нас цікавлять. Відбір одиниць ЛСП або ЛСГ добре здійснити за допомогою ідеографічного словника, якщо це можливо для досліджуваної мови. Також доцільно уточнити лексичні значення відібраних слів за тлумачним словником;
- робимо пошук ЛСВ у корпусі й отримуємо певну кількість пар речень, що містять якесь із шуканих слів у секторі, відповідному вихідній мові, разом з інформацією про те, у якому тексті вони знайшлися;
- відфільтровуємо омоніми, залишаючи словоформи з потрібними нам лексичними значеннями;
- визначаємо перекладні відповідники словоформ у паралельних реченнях;
- класифікуємо перекладні відповідники й аналізуємо їх за якісними та кількісними параметрами.

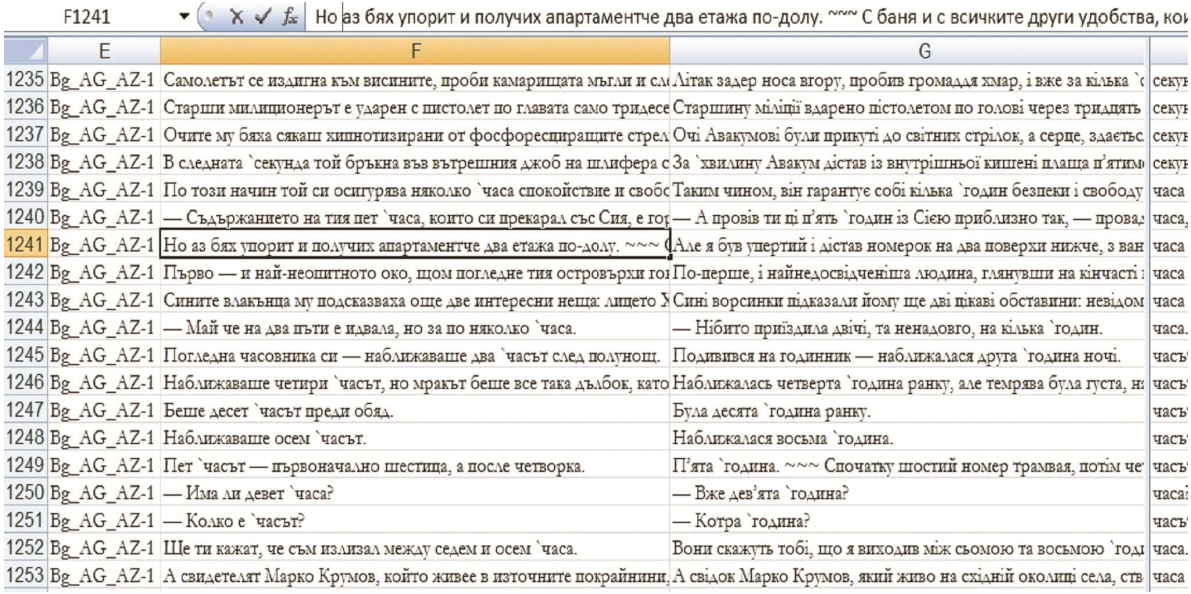


Рис. 1. Вигляд вікна корпусу в процесі дослідження слів часу

Усю послідовність дій можна повторити для іншої мови. Цікавим моментом є аналіз кореляції між частотою пар перекладних еквівалентів та мовою оригіналу, оскільки "навіть коли лексичну одиницю цільової мови можна вжити як перекладний відповідник леми вихідної мови (що не завжди можливо), перекладом цієї лексичної одиниці не завжди є лема вихідної мови оригіналу" [Лендау : 314].

В іншій постановці дослідження ЛСП, ЛСГ чи ЛСВ визначаються одразу в обох мовах, і в корпусі шукаються пари паралельних речень, що містять відібрані одиниці в болгарській, українській чи обох частинах корпусу.

Корпусний аналіз перекладних відповідників можна робити для будь-яких лексико-семантичних одиниць. Ми здійснили його на робочій версії корпусу текстів зі слов'янськими оригіналами (українська частина налічує приблизно 2½ млн. слововживань) для групи болгарських і українських іменників на позначення часу [Держански; Derzhanski, Siruk]. На рис. 1 показано вигляд робочого вікна дослідження, де у стовпчиках вказується порядковий номер пари речень, код джерела, болгарське та українське паралельні речення. Вгорі можна прочитати повний текст виділеного під номером 1241 у стовпчику F болгарського речення. Верхнім штрихом перед початком позначаються слова, які шукає дослідник.

Дослідження лексем на позначення часу дозволило зробити деякі висновки щодо тривалості визначених цими лексемами періодів. Так, лексема *момент* позначає точку (або дуже короткий сегмент) на часовій осі; чим довшим стає сегмент, тим більше імовірності, що він називатиметься не *момент*, а буде визначатися як болгарське *минута* чи українське *хвилина*. Щодо *миг* та *мить*, то ці слова займають середню позицію. Це також пояснює відносно рідке вживання "екстремальних" пар перекладних еквівалентів *минута* : *момент* і *момент* : *хвилина* у порівнянні з парами *минута* : *мить, миг* : *момент, момент* : *мить* і *миг* : *хвилина*. Слово *хвилина* в українській мові вживається на позначення невизначеного короткого проміжку часу активніше, ніж його ос-

новний болгарський відповідник *минута*. Це може мати етимологічне пояснення: *минута* – порівняно нове слово, яке від початку позначає часовий відтинок з точно визначеною і цілком усвідомлюваною тривалістю, тоді як *хвилина* < "*хвиля* 'хвилина; короткий відрізок часу, мить' є давнім запозиченням з німецької через посередництво польської; нім. *Weile* 'певний час' < двн. *hwīl, hwīl(a)* 'час, година' […] припускається спорідненість *хви́ля* (час) і *хви́ля* (на воді)" [Мельничук : 166]. Слово *момент* вживається частіше (тобто глибше засвоєне) в болгарській мові, ніж в українській. Попередні експерименти над близьким за розміром корпусом текстів з англійськими, німецькими, французькими, італійськими й іспанськими оригіналами дають подібні результати.

Цікавим є також дослідження перекладних еквівалентів, які не становлять класичної пари "іменник – іменник". У корпусі іменники на позначення часу та конструкції з ними можуть перекладатися:

- іменником, який не має значення часу: *В подобни мигове не е полезно да се проявява нетърпение* || *В таких випадках не треба гарячкувати* (Б. Райнов "Пан Ніхто");
- прислівником часу: *и в такива мигове пейзажът добива нещо от двусмисления израз на човек* || *й тоді краєвид набував чогось від двозначної гримаси людини* (Б. Райнов "Велика нудьга"); при цьому семантика темпоральності може підкреслюватися вживанням відповідного дієслова зі значенням роду дії: *Найнапред блесна малко, ослепително слънце, което в следните мигове сякаш изпълни всички илюминатори* || *Спочатку спалахнуло невелике сліпуче сонце і відразу засяяло в усіх ілюмінаторах* (П. Вежінов "Коли ти в човні…");
- прислівником, який не має значення часу: *Имал съм и потежки мигове* || *Бувало й важче* (Б. Райнов "Велика нудьга"). Форма вищого ступеня порівняння українського прислівника *важче* мотивована аналогічною формою болгарського прикметника *по-тежки*. Семантика множинності не втрачається, а переходить з іменника у формі множини *мигове* до дієслова із семантикою багатократності *бувало*.

Корпусне знаходження лексичних перекладних відповідників

Ще одне цікаве застосування паралельного корпусу полягає в автоматичному віднайденні лексичних перекладних еквівалентів, кульмінацією якого є автоматизована побудова двомовного словника. Ця процедура ґрунтується на пошуку пар слів, які найчастіше трапляються у співвідносних реченнях паралельного корпусу.

Якість такого словника краща, коли корпус великий, хоча це збільшує вимоги до обчислювальних ресурсів. До факторів, через які якість словника погіршується, належать:

- неточність перекладу, характерна для художнього тексту (особливо для секторів, які містять переклади з третіх мов);
- велика довжина речень, властива деяким авторам, а також створювана автоматично через розходження меж речень у пари вирівнюваних текстів;
- істотна несхожість граматичної будови мов.

З огляду на розвинену словозміну та розгалуженість граматичних парадигм як української, так і болгарської мов конче доцільним буде застосування граматичного словника (лематизатора) на початковому етапі автоматичного визначення лексичних відповідників, для того щоб пошук здійснювався не в просторі пар словоформ, а в істотно вужчому просторі пар лексем.

Розвиток проекту

Роботу над корпусом болгарських та українських паралельних текстів передбачається продовжувати за трьома основними напрямками.

По-перше, розвиватимуться кількісний та якісний склад корпусу. Мова йде як про додавання нових секторів (мов-оригіналів), поповнення корпусу новими творами та їхніми перекладами (у широкому розумінні паралельності текстів), так і про жанрове розширення корпусу, тобто збільшення різноманітності жанрів та зменшення кореляції між мовами і жанрами.

По-друге, доопрацювання вимагає якість текстового матеріалу та його представлення. Це стосується широкого діапазону завдань від пошуку і виправлення технічних помилок до забез-

печення текстів морфологічною та синтаксичною розмітками, виконаних відповідно до прийнятих стандартів.

По-третє, потрібно розробити супровідні допоміжні програмні засоби, а саме лематизатор (на основі граматичних словників) та пошукову систему, і створити онлайнову версію корпусу.

1. *Гарабик Р.* Параллельный русско-словацкий корпус / Р. Гарабик, В. П. Захаров // Труды Международной конференции "Корпусная лингвистика – 2006". – СПб., 2006. – С. 81–87. 2. *Демська О.* Текстовий корпус: ідея іншої форми / О. Демська. – К., 2011. 3. *Держански И.* Думите за време в български и украински език (върху материал от паралелни текстове) (у друці) / И. Держански. 4. *Добровольский Д. О.* Корпус параллельных текстов: архитектура и возможности использования / Д. О. Добровольский, А. А. Кретов, С. А. Шаров // Национальный корпус русского языка: 2003 – 2005. – М., 2005. – С. 263–296. 5. Етимологічний словник української мови : у 7 т. / редкол.: О. С. Мельничук (голов.ред.) [та ін.]. – Т. 6 : У – Я / уклад. Г. П. Півторак [та ін.]. – К., 2012. 6. *Лендау С. І.* Словники: мистецтво та ремесло лексикографії / С. І. Лендау. – К., 2012. 7. Parallel Texts. Using Translational Equivalents in Linguistic Typology. Theme issue in Sprachtypologie and Universalienforschung STUF 60.2. / eds. *M. Cysouw, B. Wälchli.* – 2007. 8. *Derzhanski I.* Brief Time Words in Bulgarian and Ukrainian (Using Evidence from Parallel Texts) / *I. Derzhanski, O. Siruk* // The Eight International Conference "Formal Approaches to South Slavic and Balkan Languages": Book of Abstracts. – Zagreb, 2012. – С. 12. 9. *Varga D.* Parallel Corpora for Medium Density Languages / D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, V. Nagy // Proceedings of the RANLP. – 2005. – P. 590–596. 10. *Vitas D.* Preparation and Exploitation of Bilingual Texts / D. Vitas, C. Krstev, E. Laporte // Lux Coreana. – № 1. – Han-Seine, 2006. – P. 110–132.

Речь идет о создании корпуса болгарских и украинских параллельных текстов и его использовании в исследовании лексической семантики украинских и болгарских переводных эквивалентов.

Ключевые слова: корпусная лингвистика, параллельные тексты, корпус, лексическая семантика, переводные эквиваленты.

The paper introduces the Corpus of Bulgarian and Ukrainian Parallel Texts and discusses its use for the analysis of the lexical semantics of Bulgarian and Ukrainian translation equivalents.

Keywords: corpus linguistics, parallel texts, corpus, lexical semantics, translation equivalents.

Стаття надійшла до редакції 4.09.2012

УДК 81’322

Василь Старко
Східноєвропейський національний університет імені Лесі Українки,
Наталія Чейлитко
Київський національний університет імені Тараса Шевченка

ПАРАМЕТРИЗАЦІЯ КОРПУСУ ЯК СПОСІБ ПІДВИЩЕННЯ ЙОГО РЕПРЕЗЕНТАТИВНОСТІ ТА ЗБАЛАНСОВАНОСТІ

Розглянуто різні методи параметризації корпусу з метою підвищення його репрезентативності та збалансованості. Обґрунтовано, що оптимального ефекту можна досягти завдяки виваженій комбінації різних підходів до параметризації корпусу з огляду на його специфіку.

Ключові слова: корпус, корпусна лінгвістика, параметризація, репрезентативність, збалансованість, вибірка.

Із-поміж вимог, що їх висувають до корпусів, постійно повторюються дві ключові – репрезентативність і збалансованість. Попри те, що вимоги ці були сформульовані ще на етапі становлення корпусної лінгвістики як окремої галузі знань, вони анітрохи не втратили актуальності. За цей час постали нові типи корпусів, вироблено нові підходи до їх укладання, а тому виникла потреба узагальнити, систематизувати напрацювання корпусної лінгвістики.

Основна теза цієї праці полягає в тому, що розв’язання центральних завдань, пов’язаних із репрезентативністю й збалансованістю корпусу, передбачає його докладну та виважену параметризацію. Серед українських мовознавців на потребі параметризувати різні види корпусів наголошували, зокрема, Є. Карпіловська [Карпіловська 2003 : 76] та О. Демська-Кульчицька [Демська-Кульчицька 2005 : 59]. Під параметризацією корпусу маємо на увазі його побудову за визначеним набором параметрів. Параметризація корпусу відбиває (попередньо проведену) параметризацію досліджуваної ділянки мови й має щонайменше такі виміри: 1) широту (охоплення максимальної кількості типів

текстів першого рівня); 2) глибину (виокремлення підкатегорій на нижчих рівнях); 3) пропорційність (заповнення окремих "клітинок" текстами, дібраними в певній пропорції); 4) часовий вимір.

Без репрезентативності будь-які результати дослідження мовного матеріалу, вміщеного в корпусі, чинні лише щодо цього корпусу – їх не можна екстраполювати на відповідну ділянку мови, бо незрозуміло, як корпус із нею співвідноситься. Відтак корпус втрачає свою роль інструмента дослідження мови. Ідеальної репрезентативності можна досягти за умови, що відома чітко визначена й параметризована генеральна сукупність, з якої належить зробити вибірку. Позаяк подати достовірний опис такої сукупності, зокрема визначити реальні пропорції всіх можливих типів текстів, неможливо, творці корпусів мусять задовольнятися наближенням до ідеальної репрезентативності [Leech 2007].

Однією з необхідних складових репрезентативності корпусу є його збалансованість. Збалансованість розуміють переважно як пропорційність різних частин корпусу: частка кожного типу текстів у корпусі має бути пропорційною – однак щодо чого? Справді пропорційний корпус можна побудувати з огляду на три вихідні позиції [Biber 1993]: 1) творці тексту (у цьому разі пропорції визначають відносно кількості створених текстів за певний період часу); 2) реципієнти (пропорції встановлюють відповідно до кількісного співвідношення текстів, сприйнятих носіями мови під час комунікативної діяльності – усної чи письмової); 3) самі тексти (пропорції визначають відносно розподілу текстів за різними жанрами). Однак у жодному серйозному корпусі не зроблено спроби досягти такої пропорційності. Річ у тім, що в кожен момент часу універсум живої мови складається переважно з усного приватного мовлення й повністю пропорційний корпус містив би близько 90 % записів бесід, 3 % листів і нотаток і 7 % решти жанрів [Biber 1993]. На практиці в корпусах усна частина якщо й наявна, то значно поступається писемній. Проте навіть у межах писемного корпусу досягти ідеальної збалансованості фактично неможливо.

Репрезентативність і збалансованість можуть мати множинні виміри й потребують певної точки відліку. Іншими словами, ціл-

ком резонно запитати: Корпус репрезентативний / збалансований щодо чого? Корпус збалансований відносно чого? Один із аспектів, який майже не враховують під час планування корпусів, – розмежування функцій мовця й слухача. Зазвичай два тексти вважають рівноцінними – як результати мовної діяльності їх було створено один раз. Однак із погляду реципієнтів текстів можна говорити про "індекс сприймання", тобто скількох слухачів / читачів текст досягнув. До того ж, різні аспекти цих понять можуть входити в суперечність. Наприклад, корпус, репрезентативний і збалансований щодо типів текстів за обсягом їх продукування, не буде репрезентативним за типами мовних явищ [Biber 1993]. Отже, поняття репрезентативності й збалансованості не лише градуйовані й багатовимірні, а й певною мірою неоднозначні.

Одним із поширених підходів до здійснення параметризації корпусу є **орієнтування на авторитетний корпус**, як-от Браунський корпус (Brown Corpus), цебто побудова нового корпусу за параметрами взірцевого.

Параметризація на основі фахової оцінки. Експертів залучають для оцінювання різних величин, які є базисом вироблених згодом параметрів корпусу. Індивідуальні оцінки бажано усереднювати.

Параметризація із застосуванням формальних методів. В обмежених випадках, коли відома генеральна сукупність текстів, можна скористатися статистичними формулами отримання достовірної вибірки. Коли ж цю сукупність неможливо проаналізувати, послуговуються непрямими методами. Приміром, Д. Байбер [Biber 1989; 1993] запропонував будувати типологію текстів, а отже й параметризувати корпус, на основі статистичних показників, які характеризують дистрибуцію кластерів мовних явищ у різних типах текстів. Соціологічні методи побудови вибірки застосовують, коли вдаються до демографічної моделі (див. нижче).

Параметризація з опертям на списки джерел. За цього підходу використовують національні бібліографічні індекси, списки бестселерів, списки найчастіше запитуваних книжок у бібліотеках, опис фондів конкретної бібліотеки. Можна використати

стохастичні методи й статистичні формули, щоб визначити, уривки зі скількох книжок потрібно ввести до корпусу, щоб отримати репрезентативну вибірку.

Параметризація з позицій отримувачів тексту. Пропорційність можна забезпечити шляхом відтворення кількісного співвідношення різних типів текстів, сприйнятих їх реципієнтами: застосовуючи звичні методи соціологічного опитування, побудувати репрезентативну демографічну вибірку й дослідити, скільки часу реципієнти сприймають усні повідомлення та письмові тексти різних жанрів.

Параметризація на основі внутрішньомовних критеріїв. Єдина відома нам концепція такого типу розроблена в працях Д. Байбера [Biber 1989; 1993]. Автор розробив метод оцінки розподілу мовних явищ за типами текстів на основі статистичних критеріїв із використанням факторного й кластерного аналізу. Суть методу полягає в тому, що типи текстів можна визначати за кластерами типових для них мовних ознак. Статистичні показники також придатні для визначення мінімального обсягу корпусу, репрезентативного щодо однієї або низки мовних ознак.

Параметризація на основі зовнішніх (ситуаційних) критеріїв. На сьогодні це найпоширеніший метод параметризації, застосований у багатьох значних корпусах. За такого підходу важить як широта, так і глибина типології [Leech 2007]. Кожна клітинка утвореної матриці підлягає заповненню відповідними текстами, зазвичай у певній пропорції. Визначення цих пропорцій становить окрему проблему.

Параметризація відповідно до потреб користувачів. Чимало корпусів, зокрема національних, покликані забезпечити найрізноманітніші пошукові запити користувачів. У такому разі важливо забезпечити гнучкість корпусу, зокрема дати користувачам змогу формувати підкорпуси за необхідними параметрами.

Динамічний підхід передбачає методику побудови моніторингових корпусів. Дж. Синклер [Sinclair 1991 : 24–26] обстоював ту ідею, що з лавиноподібним зростанням електронних текстів стане непотрібним конструювання малих корпусів на основі ви-

падкових виборок для інтенсивного дослідження. Натомість доцільним виявиться укладання відкритого корпусу, який постійно поповнюватиметься. Автор назвав корпуси такого типу *моніторинговими* (monitor corpora), позаяк вони схоплюють поточний "стан мови" за кожен інтервал часу. На основі моніторингового корпусу можна формувати менші корпуси за заданими параметрами.

Циклічний підхід. На основі попереднього досвіду й теоретичних положень формується початковий корпус (наприклад, визначається матриця типів текстів і пропорція кожного різновиду), після чого здійснюється емпіричне дослідження вміщеного в ньому матеріалу й визначаються шляхи розвитку корпусу. Д. Байбер запропонував циклічну модель створення корпусу: первісний проект корпусу, збирання текстів, емпіричне дослідження мовної варіативності (ключове поняття для автора), перегляд структури корпусу і наступний цикл [Biber 1993].

Використання ресурсів Інтернету. Нове покоління надвеликих корпусів створюють із використанням ресурсів Інтернету, зокрема за допомогою пошукових сервісів. Про збалансованість чи репрезентативність у цьому випадку не йдеться. Достовірність таких корпусів забезпечена їхнім обсягом. Однак параметризація текстів має бути запроваджена, оскільки вона забезпечує виконання корпусом низки важливих функцій.

Метод стратифікації полягає у формуванні класифікаційної матриці, кожне вічко якої має бути заповненим. У такий спосіб підвищують збалансованість. Цей загальний метод набуває конкретного втілення залежно від об'єкта стратифікації: текстів, суб'єктів усного мовлення чи релевантних ознак досліджуваної ділянки мови.

Метод пропорційної вибірки. Відомий також як метод пропорційного звуження, він має на меті досягти пропорційної репрезентативності, за якої відносна частота досліджуваного явища в корпусі близька до його відносної частоти в досліджуваній ділянці. Однак вищезгаданий метод наражається на кілька перешкод: обсяг ділянки мови найчастіше неможливо встановити (вона, як правило, постійно зростає), існує поріг відображення

(тобто за межами вибірки можуть виявитися менш частотні мовні явища), послідовне застосування пропорційності проблематичне (через значну перевагу усного мовлення).

Метод суцільної вибірки. Цей метод технологічно найпростіший, але призводить до розбалансованості. Повноту охоплення (наприклад, певних мовних явищ) можна оцінити лише постфактум. Проте такий підхід за умови здійснення додаткового редагування корпусу може виявитися дуже корисним, зокрема, під час укладання ілюстративного корпусу.

Метод випадкової вибірки. Проста випадкова вибірка може призвести до того, що периферійні шари лексики буде охоплено недостатньо. Одним із шляхів розв'язання цієї проблеми є застосування *стратифікованої випадкової вибірки*, коли спершу визначають категорії текстів (та їхню частку в загальному корпусі), а потім ці категорії наповнюють випадково вибраними текстами.

Демографічна вибірка полягає в соціальній стратифікації респондентів за віком, статтю, соціальним класом та географічним регіоном. За такими параметрами формують репрезентативну вибірку мовців.

"Опортуністичний" метод полягає в тому, що укладачі корпусу просто користуються з наявних на даний момент технологій і доступних мовних матеріалів і не звертають особливої уваги на принципи побудови корпусів. Це особливо показово щодо корпусів, укладених до настання ери електронних публікацій.

Метод цілеспрямованого доповнення. Мається на увазі використання спеціальних методів здобуття мовного матеріалу від мовців, наприклад, психолінгвістичних методик, які допомагають встановили наявні в репертуарі носіїв мови одиниці, що навряд чи трапляться в корпусі [Фрэнсис 1983].

Доповнення спеціалізованими корпусами. Корпуси загального призначення не завжди достатньо повно охоплюють певні типи текстів, які необхідно скрупульозно досліджувати. Тому укладають спеціалізовані корпуси – вужчі, але водночас репрезентативніші щодо обраної ділянки мови.

Критерій культурної значущості тексту. Творці корпусів застосовують оцінні критерії, щоб розмежувати тексти за важ-

ливістю. У сучасних умовах функціонування української мови (зокрема засилля суржику) такий відбір вкрай потрібен, інакше корпус ризикує перетворитися на смітник. Цей критерій може бути застосовано й до цілих жанрів, наприклад у випадку встановлення бажаного відсотку художньої літератури чи перекладів.

Критерій чистоти вибірки. Проблеми чистоти вибірки австралійського й новозеландського варіантів англійської мови описує Ґ. Кеннеді [Kennedy 1998 : 64–66]. Не менш актуальними вони є і для укладачів українських корпусів, особливо усних, – ідеться про розрізнення явища двомовності й суржику, або ж засміченої, неохайної мови. У випадку писемних творів ідеться, зокрема, про відсіювання неграмотних текстів та текстів із явними ознаками машинного перекладу, переважно з російської мови.

Насамкінець завважмо, що логістичні, фінансові, організаційні, часові та інші обставини накладають обмеження на формування корпусу й змушують його творців до компромісу між бажаним і здійсненним.

Розгляньмо коротко параметризацію **Корпусу сучасної американської англійської мови** (Corpus of Contemporary American English, COCA), що його уклав М. Дейвіс [Davies 2010], – чи не єдиний справді моніторинговий корпус англійської мови. Пропорційність дотримано щодо жанрів, підкатегорій текстів (наприклад, "медицина" в наукових публікаціях) і в часі! Корпус містить у рівній пропорції (по 20 %) п’ять жанрів – спонтанне усне мовлення, художня література, популярні журнали, газети й наукові журнали. Щороку автор корпусу додає приблизно 4 мільйони слововживань до кожного жанру, й наразі на кожен із них припадає по 80 з лишком мільйонів слововживань, що разом становить 450 млн. Подвійна збалансованість (синхронійна між жанрами й піджанрами та діахронійна в межах жанру та піджанру) дає змогу здійснювати статистично надійні порівняльні дослідження. Корпус вдало поєднує параметризацію за зовнішніми критеріями, пропорційність, динамічний підхід, уможливлює порівняння між кількома корпусами, втілює загальну тенденцію до репрезентативності й збалансованості, що дає йому незаперечну перевагу над менш систематично організованими великими корпусами.

Ми окреслили, звісно, лише деякі аспекти складної проблеми. Окремі підходи до параметризації потребують глибшого вивчення, і жоден не є панацеєю. Проте зрозуміло, що створення якомога репрезентативніших і збалансованіших корпусів безпосередньо залежить від продуманої, виваженої їх параметризації, а саме оптимального поєднання різних підходів, методів і критеріїв із врахуванням специфіки конкретного корпусу. Параметризація корпусу залишається частково суб'єктивною, а тому бажано, щоб рішення, покладені в її основу, були прорефлексовані, чітко проартикульовані й спиралися на консенсусну оцінку фахівців.

1. *Демська-Кульчицька О.* Основи національного корпусу української мови: монографія / О. Демська-Кульчицька. – К., 2005. 2. *Карпіловська Є. А.* Вступ до комп'ютерної лінгвістики / Є. А. Карпіловська. – Донецьк, 2003. 3. *Фрэнсис У. Н.* Проблемы формирования и машинного представления большого корпуса текстов / У. Н. Фрэнсис // Новое в зарубежной лингвистике. – М., 1983. – Вып. XIV. Проблемы и методы лексикографии. – С. 334–352. 4. *Biber D.* A typology of English texts / D. Biber // Linguistics. – 1989. – Vol. 27. – P. 3–43. 5. *Biber D.* Representativeness in corpus design / D. Biber // Literary and Linguistic Computing. – 1993. – 8 (4). – P. 243–257. 6. *Davies M.* The Corpus of Contemporary American English as the first reliable monitor corpus of English / M. Davies // Literary and Linguistic Computing. – 2010. – Vol. 25. – № 4. – P. 447–464. 7. *Kennedy G. D.* An introduction to corpus linguistics / G. D. Kennedy. – London, New York, 1998. 8. *Leech G.* New resources, or just better old ones? / G. Leech // Corpus Linguistics and the Web. – Amsterdam, 2007. – P. 134–149. 9. *Sinclair J.* Corpus, Concordance, Collocation / J. Sinclair. – Oxford, 1991.

Рассмотрены разные методы параметризации корпуса как средство повышения его репрезентативности и сбалансированности. Обосновано, что оптимального эффекта можно достичь благодаря взвешенному соотношению различных подходов к параметризации корпуса с учетом его специфики.

Ключевые слова: корпус, корпусная лингвистика, параметризация, репрезентативность, сбалансированность, выборка.

The article considers various methods of corpus parametrization as ways to enhance its representativity and balance. The authors argue that an optimal effect may be achieved through a carefully calculated combination of various approaches to corpus parametrization in view of the special aspects of its nature.

Keywords: corpus, corpus linguistics, parametrization, representativity, balance, sample.

Стаття надійшла до редакції 20.09.2012

УДК 81'322

Марія Шведова,
Дмитро Січінава
Інститут російської мови Російської академії наук

КОРПУСНА ЛІНГВІСТИКА ТА ЛЕКСИКО-ГРАМАТИЧНА ТИПОЛОГІЯ[1]

Наводяться приклади застосування паралельних корпусів Національного корпусу російської мови для розв'язання завдань зіставлення граматики та лексики у перекладних текстах.

Ключові слова: корпусна лінгвістика, паралельний корпус, граматична типологія, лексична типологія, семантика.

Моделі та стимули перекладу

Є багато можливостей практичного застосування Національного корпусу російської мови (НКРМ, http://ruscorpora.ru) в галузі зіставної граматичної та лексичної семантики. Базою для міжмовних (типологічних і контрастивних) досліджень у НКРМ є передусім паралельний корпус – корпус, який складається з оригінальних та перекладних текстів, що вирівняні по реченнях. До складу Національного корпусу російської мови за станом на січень 2013 р. входять 8 двомовних паралельних корпусів: англійсько-російський (та, відповідно, російсько-англійський), німецько-російський, французько-російський, іспансько-російський, італійсько-російський, польсько-російський, українсько-російський та білорусько-російський (для кожної мовної пари подані також тексти, що перекладені у зворотному напрямку) [Добровольский, Кретов, Шаров 2005], [Андреева, Касевич 2005], [Добровольский 2009, Січінава, Шведова, Тищенко-Монастирська 2011], [Sitchinava 2012]. Сукупний обсяг цих корпусів перевищує

[1] Стаття підготована за підтримки програми Президії РАН №36-П "Корпусна лінгвістика" та гранту РФФІ №12-06-33038 "Контрастивні корпусні дослідження російських та французьких дієслівних категорій у поліваріантних паралельних текстах".

37 мільйонів слововживань, і розвиток їх продовжується. На стадії проекту знаходяться ще кілька двомовних корпусів за участю російського (вірменський, латиський, угорський, татарський та інші). Кожна мова паралельного корпусу має морфологічну розмітку, що дозволяє здійснювати точний пошук за словоформою, пошук за лексемою та/або набором грамем, а також за різними комбінаціями цих параметрів.

Крім того, існує багатомовний корпус на базі 25 мов, здебільшого слов'янських, який містить 9 паралельних текстів (граматична розмітка є вже не для кожної мови); його обсяг – 5 мільйонів слововживань. Заплановано створення поліваріантних корпусів, до яких входить кілька різних перекладів тексту на одну мову; нині триває робота над створенням російсько-французького поліваріантного паралельного корпусу.

Тексти вирівнюються по реченнях; основою виділення співвідносних частин є поділ на речення в оригінальному тексті (у перекладі він часто буває іншим). Слід зауважити, що за радянської доби художній переклад із російської мови на мови народів СРСР і навпаки (на сьогодні саме перекладів цього часу найбільше в українсько-російському та білорусько-російському паралельних корпусах) був досить приблизним на різних рівнях, при наповненні та аналізі паралельного корпусу це доводиться постійно враховувати. Причини цього є різними: загальна настанова певних перекладачів, роль автора у створенні авторизованого перекладу (зокрема, він і сам може бути перекладачем), а нерідко й цензурні вимоги. Довільність та різні помилки фіксуються і в перекладах за участю інших мов, хоча й менш систематично. Для зазначення тієї чи іншої неточності в перекладі – пропуску, додавання, заміни тексту зі зміною змісту – у корпусі використовується спеціальний рівень розмітки – атрибут **loose**. Наприклад, тег **omit** використовується на позначення значного пропуску в перекладі (приклад із повісті Л. Толстого "Смерть Ивана Ильича" у французькому перекладі М. Оффмана; не поданий у перекладі фрагмент виділений).

```
<para id="353">
    <se lang="ru">Было лучшее общество, и Иван Ильич танцевал с княги-
    нею Труфоновой, сестрою той, которая известна учреждением обще-
    ства “Унеси ты мое горе”.</se>
    <se lang="fr" loose="omit">Ivan Illitch dansa avec la princesse Troufonova,
    la propre sœur de l'illustre fondatrice de la société “Foin de peines !”.</se>
</para>
```

Використання паралельних корпусів у порівняльних дослідженнях є предметом багатьох праць, назвемо, передусім, спеціальний номер часопису STUF за редакцією М. Сісоува та Б. Вельхлі [Cysouw, Wälchli (eds.) 2007]; див також зазначені вище статті [Андреева, Касевич 2005] і [Добровольский 2009]. При перекладі лексики, як і граматичних конструкцій (чітку межу між ними не завжди можна провести), суттєву роль відіграють **моделі перекладу** та **стимули перекладу**. Зазвичай для конструкції А у мові В існують (реально представлені в текстах) кілька відповідників, які повторюються невипадково, – моделей перекладу (translation patterns; пор. [Hasselgård, Oksefjell (eds.) 1999]). Практично можна стверджувати, що, коли є значний за обсягом корпус перекладних текстів, можна виокремити кілька моделей перекладу конструкції мови А у мові В.

Стимулом перекладу назвемо конструкції в мові оригінальних текстів, що відповідають певним конструкціям у мові перекладу (“що стимулює появу конструкції Х у перекладі на мову В?”).

Відповідники граматичних конструкцій

У праці [Сичинава 2011] представлено результати статистичного дослідження стимулів перекладу в англійських та німецьких текстах для конструкції з часткою *было*, типу *пошёл было, да вернулся*. Було виявлено цікаві стійкі відповідники, причому не тільки і не стільки граматичні, скільки лексичні. Наприклад, у німецькій – частки *eben* та *gleich* ‘рівно’, в англійській – прислівники нетривалої дії типу *for a while* ‘деякий час’. З огляду на ці відповідники можна чимало сказати про семантику російської конструкції (яка не має точного відповідника в англійській та німецькій мовах і переважно позбавлена будь-якого лексичного

стимулу при перекладі). Те ж саме стосується і відповідників іншомовних граматичних категорій у російській мові. Так, перфект прогресива в англійській мові (*I have been doing*) передається російським недоконаним видом (теперішнього або минулого часу) в актуально-тривалому значенні. Разом із цим зазначено такий лексичний засіб для передавання перфекта прогресива, як дієслова типу *начать, решить,* які вказують на актуальність для теперішнього моменту початкової межі ситуації:

Perhaps people *have been celebrating* Bonfire Night early – it's not until next week, folks! [Joanne Kathleen Rowling. Harry Potter and the Sorcerer's Stone (1997)]

Кажется, народ уже начал *праздновать* день Порохового Заговора – рановато, господа, он будет только на следующей неделе! [Дж. К. Роулинг. Гарри Поттер и Волшебный камень / М. Спивак].

Well, it's a new thing the boss *has been trying.* [Stephen King. The Lawnmower Man (1975)]

Это новая методика, которую решил *испробовать* наш босс. [Стивен Кинг. Газонокосильщик / пер. не вказаний]

Відповідники французького часу passé antérieur у російських текстах послуговуються додатковими лексичнми засобами: *лишь, немедленно, сразу:*

Quand Passepartout *eut fini*, il se trouva calme et comme soulagé. [Jules Verne. Le tour du monde en quatre-vingt jours (1872)]

Отведя душу, Паспарту сразу успокоился и почувствовал большое облегчение. [Жюль Верн. Вокруг света за восемьдесят дней / Н. Габинский, Я. Лесюк]

За допомогою паралельного корпусу можна виявити особливості вживання подібних граматичних конструкцій у близькоспоріднених мовах. Так, для російської мови не характерна конструкція "*было* + модальне дієслово" (вирази *могло было, следовало было* замість *могло бы, следовало бы* є помилковими). Натомість сполуки з дієсловом *могти* у давноминулому часі, який етимологічно відповідає російській конструкції, нерідко можна

зустріти в західноукраїнських дореволюційних письменників – І. Франка, О. Мартовича, О. Кобилянської. Наприклад:

І нікого не було тут, кому б *міг був* повірити свій жаль. [Ольга Кобилянська. Земля (1901)]

И никого не было тут, кому он *мог бы* поверить своё горе. [Ольга Кобылянская. Земля / пер. не вказаний].

Із дієсловами повинності ("долженствования") така конструкція в українській мові представлена набагато ширше:

Настал день, когда электронный скальпель *должен был* коснуться разумного мозга… [Б. В. Зубков, Е. С. Муслин. Исповедь после смерти (1966)]

Настав день, коли електронний скальпель *мав був* торкнутися розумного мозку… [Б. В. Зубков, Є. С. Муслін. Німа сповідь / пер. не вказаний]

…в остаточній редакції твою одчайдушну телеграму *мусили були* прийняти без цих двох слів. [Оксана Забужко. Польові дослідження з українського сексу (1996)]

…в окончательной редакции твою отчаянную телеграмму *должны были* принять без этих двух слов [Оксана Забужко. Полевые исследования украинского секса / Е. Мариничева].

У мовленні письменника-білінгва Г. Квітки-Основ'яненка є цікавий приклад граматичної інтерференції. Квітка сам переклав російською мовою свою повість "Конотопська відьма" і відповідно до українського давноминулого часу вжив модальну конструкцію з *было* (в російській літературній мові зустрічалася у XVIII ст.), замінивши лише дистантне положення показника контактним:

Іще *було* уп'ятерить *подобало* за таковоє злодіяніє… [Конотопська відьма (1833)].

Еще *было подобало* упятерить за таковое злодеяние…

Приклади подібної інтерференції між оманливо близькими східнослов'янськими мовами не такі рідкісні, і в неавторському

перекладі також (зокрема, помітний вплив білоруських конструкцій та лексики проглядається в деяких радянських перекладах із білоруської на російську).

Лексичні відповідники

Дослідження моделей та стимулів перекладу в реальному корпусі є, зокрема, інструментом верифікації тверджень про "мовну специфіку" тих чи інших лексем, їх "перекладуваність" тощо. Очевидно, найбільш "специфічній" із семантичного погляду лексиці кожної мови відповідатиме найбільший діапазон моделей і стимулів перекладу – оскільки "головний" відповідник, який підходить до всіх або майже всіх контекстів, в іншій мові відсутній. Так, лексеми, що їх деякі дослідники відносять до "російської мовної картини світу", такі, як *тоска, душа* або *удаль*, мають досить широкий діапазон варіантів перекладу на англійську мову.

Наприклад, кожен із наведених англійських перекладів слова *удаль* у паралельному корпусі поки що трапився по одному разу: *abandonment, assurance, violence, effrontery, jockeying, bravado, daring, fearless confidence*. При цьому характерно, що в перекладах з англійської на російську слово *удаль* та його похідні майже не вживаються. Англійське слово *soul* перекладається практично однозначно як російське *душа* (крім фразеологізмів типу *poor soul*), а російське *душа* може перекладатися по-різному. Так, у контекстах типу *в глубине души* це слово передається англійським *mind.* Цікавим є приклад із *soul* в англійському тексті і *сознание* в російському:

And, all through that silent drive back to Green Street, the *souls* of both of them revolved a single thought: 'Why, oh! why should I have to expose my misfortune to the public like this? [John Galsworthy. In Chancery (1920)]

И все время, пока они молча ехали на Грин-стрит, у обоих *в глубине сознания* неотступно вертелась одна и та же мысль: "Почему, ах, почему мне приходится вот так выставлять напоказ мои невзгоды? [Джон Голсуорси. В петле / М. Богословская]

Перекладні моделі та стимули для слова *тоска* надзвичайно різноманітні (наводимо переклади не тільки за допомогою імен-

ника: *gloom, dismay, grief, agony, excruciating feelings, distracted moods, anguish, ache, despair, woe, loneliness, distress, frustration, desperation, blue, misery, nuisance, sadness, nostalgia, yearning, longing, lust, missing, glumly, sulkily, wistfully*).

You must be as *blue* as I am about our nation's great unvictory. [Kurt Vonnegut. Hocus Pocus (1990)]

Значит, на вас нагнала такую же *тоску*, как и на меня, великая пиррова победа нашего народа. [Курт Воннегут. Фокус-покус / М. Ковалева]

А. Вежбицька вважає найбільш прийнятним, хоч і приблизним, до російського *тоска* англійський відповідник *yearning*. Водночас у корпусі "спектр" відповідників англійського слова *yearning* виявився зміщеним у бік семантики палкого бажання (*желание, жажда, тянуть, томление, томный, жадность, томило желание*).

Так само значна міжмовна варіативність моделей перекладу спостерігається й між близькоспорідненими мовами, наприклад російською та українською. Незважаючи на те, що обсяг паралельного українсько-російського корпусу майже вдвічі менший за англійсько-російський, для російського слова *тоска* в українській знаходимо також значну кількість еквівалентів-іменників (за винятком застарілих): *туга, жаль, жалощі, смуток, нудьга, нудота, сум, журба, скорбота, гризота, зневага, тоскність* (в 11-томному "Словнику української мови" має ремарку "рідковживане"), *скруха*. Крім того, відповідниками до російського *тоска* у предикативній функції виступають українські прислівники на **-о***: тоскно, тужливо, тужно, журно, нудно*.

За допомогою корпусу можна дати хронологічну оцінку таким синонімічним рядам. Деякі українські еквіваленти до російського *тоска*, які нерідко зустрічаються в ранніх літературних текстах, не вживані в сучасних: це "застаріле" (за СУМ) *зануда*, діалектизми *жєль* (не поданий в СУМі, хоч у західноукраїнському мовленні функціонує досі), *жура, туск,* а також слова *печаль, тоска*, які графічно однакові з російськими. Зауважимо, що у творах визначного прозаїка 1920–1930-х років, опонента

русифікації Миколи Хвильового основним засобом вираження цього значення є саме українське слово *тоска*.

Щодо слова *душа*, – яке є і в російській, і в українській, – то у перекладі фразеологізмів міжмовна варіативність досить відчутна: *душой стремился – прагнув, болит душа – уболіває, кто болеет душой – кому болить; горела душа – пекло́; Оленька-душа – Оленька-серце, душа моя – серце; до глубины души – доглибно, от души – щиро, з щирим почуттям, в глубине своей собачьей души – в глибині свого собачого єства.*

Корпус можна використати у процесі дослідження синонімів. Наприклад, російські слова *влажный* і *мокрый* та їхні англійські відповідники розрізняються відповідно до означуваного об'єкта:

(долоні, руки) *damp, wet, moist* (від поту) => *влажный*
(трава, земля, пісок) *damp, humid, wet, soaking* => *влажный, мокрый*
(одяг) *damp, humid, soaking* => *мокрый.*

Семантична несумісність синонімів із означуваним об'єктом зумовлена авторським баченням перекладу:

Trees, forced by the *damp heat*, found too little soil for full growth, fell early and decayed: creepers cradled them, and new saplings searched away up. [William Golding. Lord of the Flies (1954)]

Мокрой жарой деревья толкало в рост, но на хилом слое почвы они не заживались, рано валились и гнили; их обнимали лианы, сквозь них пробивались новые ростки. [Уильям Голдинг. Повелитель мух / Е. Суриц]

У російській мові *мокрая жара* – сполука на межі допустимості; у 230-мільйонному одномовному корпусі НКРМ вона зустрілася лише один раз у контексті *нью-йоркская* "*мокрая жара*", де вона оформлена як елемент "чужої мови".

Таким чином, за допомогою паралельного корпусу можна швидко отримувати велику кількість реальних перекладацьких рішень, що були прийняті носіями мови при створенні перекладу, й аналізувати виявлені відповідники в лексиці та граматиці. Завдання їх інтерпретації (оцінка достовірності відповідників,

інтерференції та помилок, аналіз контексту, виокремлення різних значень тощо), звичайно, залишається справою дослідника. Однак обсяг попередньої підготовчої роботи за допомогою паралельного корпусу може бути значно скорочений.

1. *Андреева Е. Г.* Грамматика и лексика (на материале англо-русского корпуса параллельных текстов) / Е. Г. Андреева, В. Б. Касевич // Национальный корпус русского языка: 2003–2005. – М., 2005. – С. 297—307. 2. *Добровольский Д. О.* Корпус параллельных текстов: архитектура и возможности использования / Д. О. Добровольский, А. А. Кретов, С. А Шаров // Национальный корпус русского языка: 2003–2005. – М., 2005. – С. 263–296. 3. *Добровольский Д. О.* Корпус параллельных текстов в исследовании культурно-специфичной лексики / Д. О. Добровольский // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. – СПб., 2009. – С. 383–401. 4. *Сичинава Д. В.* Комплексное исследование одноязычного и параллельного корпусов в грамматических исследованиях / Д. В. Сичинава // Труды Международной конференции "Корпусная лингвистика–2011". – СПб. – С. 316–332. 5. *Січінава Д. В.* Паралельні українсько-російський та російсько-український корпуси / Д. В. Січінава, О. О. Тищенко-Монастирська, М. О. Шведова // Лексикографічний бюлетень, 20. – К., 2011. – С. 35–38. 6. Parallel Texts. Using Translational Equivalents in Linguistic Typology. Theme issue in Sprachtypologie & Universalienforschung STUF 60.2. / *eds. Cysouw M., Wälchli B.* – 2007. 7. Out of Corpora: Studies in. Honour of Stig Johansson / *H. Hasselgård, S. Oksefjell.* – Amsterdam ; Atlanta, 1999. 8. *Sitchinava D.* Parallel corpora within the Russian National Copus / D. Sitchinava // Prace Filologiczne. – LXIII. – 2012. – S. 271–278.

Приводятся примеры применения параллельных корпусов Национального корпуса русского языка для решения задач сопоставления грамматики и лексики в переводных текстах.

Ключевые слова: корпусная лингвистика, параллельный корпус, грамматическая типология, лексическая типология, семантика.

The paper deals with some examples of using the parallel corpora within the Russian National Corpus for the cross-linguistic analysis of grammar and lexicon in translated texts.

Keywords: corpus linguistics, parallel corpus, grammatical typology, lexical typology, semantics.

Стаття надійшла до редакції 1.09.2012

УДК 81’322

Ольга Шипнівська
Військовий інститут
Київського національного університету імені Тараса Шевченка

ВИЗНАЧЕННЯ ТИПІВ СИНТАКСИЧНОЇ НЕОДНОЗНАЧНОСТІ У ЗНАННЄОРІЄНТОВАНІЙ СИСТЕМІ МАШИННОГО ПЕРЕКЛАДУ

Розглянуто теоретико-практичні засади визначення синтаксично неоднозначних конструкцій в автоматичному синтаксичному аналізі. На матеріалі текстів військової тематики проаналізовано основні типи синтаксичної омонімії.

Ключові слова: автоматичний синтаксичний аналіз, знаннєорієнтована система машинного перекладу, синтаксична неоднозначність, типи синтаксичних омонімів, синтаксична структура.

Синтаксична неоднозначність як один із типів мовної неоднозначності вивчалася багатьма лінгвістами. У традиційному мовознавстві ця проблема розглядалась найчастіше в стилістичному аспекті, теоретичне лінгвістичне обґрунтування неоднозначного тлумачення синтаксичних одиниць висвітлено у працях, присвячених вивченню синтаксичної структури речення [Гладкий 1985 : 110-127; Колесников 1976; Севбо 1981 : 30–65], з автоматичного опрацювання текстів природної мови [Дрейзин 1966; Дрейзин 1966; Иорданская 1967].

У поданому дослідженні, зважаючи на специфіку розроблюваного модуля автоматичного синтаксичного аналізу в знаннєорієнтованій системі машинного перекладу [Замаруева 1999; Толубко 2010], розглядаються теоретико-практичні передумови визначення синтаксично неоднозначних одиниць тексту в автоматичному обробленні та на основі створеної лінгвістичної бази даних таких одиниць проведено аналіз синтаксично неоднозначних одиниць в україномовних текстах військової тематики.

Вивчення синтаксичної омонімії як одного із типів мовної неоднозначності має як теоретичне, так і практичне значення [Колесников 1976; Иорданская 1967]. Особливо актуальним це

питання постає у контексті проблем автоматичного опрацювання текстів [Гладкий 1985 : 110–127; Дрейзин 1966; Дрейзин 1966]. Актуалізація досліджень із автоматичного синтаксичного аналізу ставить завдання формального представлення синтаксичної структури тексту та визначення семантично адекватних синтаксичних структур з-поміж безлічі отриманих у ході автоматичного синтаксичного аналізу правильних синтаксичних структур [Иорданская 1967]. Проте, зважаючи на специфіку явища, принципове вирішення синтаксичної омонімії, на думку дослідників, можливе тільки на рівні семантики. На рівні синтаксису можна говорити лише про зменшення кількості синтаксично неоднозначних одиниць [Иорданская 1967 : 10].

У дослідженні синтаксичними омонімами вважаються омонімічні синтагми речення, відмінність значень яких зумовлена неоднозначністю їхньої синтаксичної структури [Гладкий 1985 : 110]. Неоднозначність синтаксичних структур, на відміну від морфологічних, лексичних омонімів, зумовлена власне синтаксичними засобами: порядком членування речення на синтагми, типом зв'язку слів у реченні [Иорданская 1967; Колесников 1976]. Процедурно ж синтаксичну омонімію визначаємо як можливість приписати фразі більше аніж одну правильну синтаксичну структуру [Иорданская 1967 : 9]. За теоретичним обґрунтуванням Л. Йорданської, правильною синтаксичною структурою називається така синтаксична структура, яку приписує фразі носій мови, базуючись на своєму інтуїтивному уявленні про правильність синтаксичної структури [Иорданская 1966 : 215]. Вона повинна відповідати обраному принципу представлення синтаксичної структури, певному списку встановлених заздалегідь синтагм.

У лінгвістиці запропоновано два підходи зменшення кількості синтаксично неоднозначних структур [Иорданская 1967 : 10–11]. Перший – розроблення правил переваг. В основу цього підходу покладено процедуру порівняння отриманих для даної фрази структур та виявлення тієї, що найточніше відповідає аналізованій фразі. Здійснюється це із залученням процедури синтезу: для кожного із значень, встановлених для окремої фрази, синтезу-

ються всі можливі фрази, які стилістично оцінюються. У наборі фраз для кожного із значень є і вихідна фраза. Із отриманих структур перевага надається тій, за моделлю якої синтезується вихідна фраза з найвищою оцінкою. Другий підхід базується на виборі найбільш достовірного варіанта. При цьому підході алгоритм аналізу повинен мати список неоднозначних ситуацій, для кожної із яких визначено найбільш достовірну інтерпретацію.

У модулі автоматичного синтаксичного аналізу система будує тільки правильні синтаксичні структури, враховуючи можливі неоднозначні інтерпретації [Толубко 2010]. Синтаксична структура представлена за допомогою певного набору бінарних синтаксичних відношень. Список синтагм, за яким визначаються компоненти структури речення, на сьогодні є відкритим. При цьому під синтагмою розуміється сполучення двох елементів речення, для якого визначено головний та залежний елементи і тип синтаксичних відношень між ними [Иорданская 1966 : 216].

Описані у лінгвістиці класифікації типів синтаксичних омонімів демонструють виділення синтаксично неоднозначних одиниць на рівні словосполучень, синтагм, синтаксичних груп, речень [Колесников 1967]. Безперечно, для синтаксичних омонімів, як і для інших лінгвістичних одиниць, характерна знакова двоплановість. В одних випадках план вираження залишається тотожним, а план змісту зумовлює омонімічність, в інших – неоднозначність конструкції зумовлює не тільки план вираження, а й план змісту. У дослідженні за планом вираження визначається декілька типів синтаксичних омонімів: синтаксичні омоніми на рівні словосполучення (в тому числі прийменникових словосполучень), синтаксичні омоніми на рівні синтагми, синтаксичні омоніми на рівні речення (між частинами складного речення), синтаксичні омоніми на рівні синтаксичних конструкцій (як-от, конструкцій із однорідними членами речення, поєднаними сурядним зв'язком). У плані вираження синтаксична омонімія визначається за типом синтаксичних відношень, що існують між головним та залежним компонентами синтаксичної структури (об'єктні, атрибутивні, обставинні) [Мельчук 1974 : 221–236].

Для зменшення кількості встановлених правильних синтаксичних одиниць у нашому дослідженні пропонується аналіз структурно-семантичних типів синтаксичних омонімів, встановлених на основі розробленої бази даних, із метою їх подальшого опрацювання на рівні семантичного аналізу[1].

Класичним прикладом синтаксичної омонімії є іменникові словосполучення із родовим відмінком, які у реченні виступають як неузгоджене означення. Наприклад, у словосполученні *фотографія брата* можна виокремити родовий приналежності (кому належить), родовий автора (хто зробив), родовий тотожності (хто зображений). Значення такого родового відмінка часто залежить від значення головного слова, як, скажімо, у словосполученнях *хатина дядька Івана* та *начальник дядька Івана* – правильне розуміння таких словосполучень не залежить від функцій родового відмінка (*хата* належить дядькові Івану, а начальник не належить).

І. Чередниченко наголошує на синтаксичній омонімії родового відмінка при віддієслівних іменниках: *рішення зборів* (збори вирішили), *рішення справи* (справу вирішили) [Чередниченко 1955 : 32]. Важливо, що дослідник встановив понад двадцять семантичних моделей, в яких родовий відмінок виражає означальну функцію, та близько п'ятнадцяти моделей із родовим відмінком у значенні додатка. Звісно, такі моделі потребують докладного вивчення на значному масиві текстів з урахуванням практичних потреб та можливостей лінгвістичних ресурсів. Наприклад, у межах нашої системи автоматичного синтаксичного аналізу може бути розпізнана модель: головне слово – назва абстрактного поняття, властивості, залежне – особа, якій приписано цю властивість. З урахуванням розробленого нами модуля словотвірного аналізу у словосполученнях *спеціальність студе-*

[1] Докладно про базу даних див. Zamarueva I., Shypnivska O. The instrumental environment for the automatic syntactical analysis of Ukrainian // NLP, Multilinguality / Proceedings of the Sixth International Conference: Corpus Linguistics, Corpus Based Grammar Research. 20–21 October 2011, Bratislava, Smolenice. – 2011. – 163–173.

нта, *професія інженера* для кожної із складових можна буде визначити необхідну семантичну інформацію і в результаті приписати словосполученню атрибутивне значення.

Одним із найпоширеніших типів синтаксичної неоднозначності у проаналізованому матеріалі і є неоднозначність на рівні словосполучень, зокрема іменникових (63,3 %)[1]. Альтернативним для аналізу є здебільшого неоднозначне визначення об'єктних та атрибутивних відношень, причому варто зауважити, що саме така інтерпретація зумовлена їх подальшим опрацюванням у перекладі іншою мовою: *застосуванню сили*, *країнами-членами Альянсу*, *із виконання Договору, керівництво повсякденним життям, військовослужбовці жіночої статі*. Хоча незначну частку серед них становлять саме ті одиниці, для яких у результаті кінцевого аналізу за допомогою семантичного аналізатора можна визначити тільки атрибутивне відношення (5 %): *літаками подвійного призначення, дельтовидне крило досить великої площі, засобів довготривалого зберігання, куртку камуфльованого кольору.* Для решти типів пріоритетним є об'єктне відношення (досить часто із значенням "частина/ціле"): *потенціалу Збройних Сил України, війська берегової охорони, війська протиповітряної оборони, полків армійської авіації, придатності громадян, пенсіонерів Збройних Сил.*

Синтаксичне відношення із об'єктним значенням уже на поверхнево-синтаксичному етапі аналізу тексту можуть отримати словосполучення, в яких головним словом є віддієслівний іменник: *виконання договору, здійснення повітряної розвідки, охорони повітряного простору, вирішення питань, супроводження служби.* Щоправда, не завжди це обмеження може бути визнане пріоритетним, як, скажімо, у словосполученні *звання офіцерського складу,* в якому за розробленою нами системою словотвірного аналізу словоформі *звання* приписується словотвірне значення "віддієслівний іменник зі значенням «звати»".

[1] Загальна кількість встановлених на сьогодні синтаксично неоднозначних конструкцій становить близько 1300 одиниць.

Окрему групу становлять словосполучення, в яких головним словом є віддієслівний іменник, а залежне – іменник на позначення назви істоти: *рішення командира*, *передачі кореспондента*, *втручання противника*, *ураження противника*, *допит полонених*, *показання полонених*. Остаточне рішення щодо визначення синтаксичного відношення як об'єктне чи як атрибутивне здійснюється із урахуванням семантичної інформації. Об'єктні відношення можуть бути встановлені на рівні семантичного аналізу, подібно до випадку зі словосполученнями семантичної моделі "частина/ціле": *комплект предметів*, *вихованці військових ліцеїв.*

Синтаксично неоднозначні прийменникові конструкції ставали об'єктом багатьох лінгвістичних досліджень. Л. Йорданська в російській мові виділяє вісім типів потенційно можливих синтаксично неоднозначних структур речення з прийменником [Иорданская 1967 : 14–15]. Ф. Дрейзин визначає у тексті специфіку розподілу таких конструкцій щодо омонімічного вузла [Дрейзин 1966 : 56–57]. Традиційно прийменникові конструкції розглядаються у контексті розроблення системи автоматичного синтаксичного аналізу. Зокрема, дослідники визначають зони прийменникових зв'язків у тексті [Бугаков 2006; Грязнухина 1985].

Синтаксична неоднозначність структур із прийменником (30 % у базі) визначається типом синтаксичних відношень між прийменником та залежним від нього компонентом при визначенні головного слова, якому підпорядкована прийменникова конструкція. Неоднозначність зумовлюється типом синтаксичних відношень між варіантами додаток/означення за умови, що головним словом конструкції є іменник з дієслівною семантикою: *конфлікт із застосуванням сили, переговорного процесу між сторонами, злочинів проти миру і людства*. Звісно, варіант кінцевого аналізу залежить і від синтаксичного типу речення та від функції сегмента у речення (наприклад, визначена група підмета). Так, у заголовку така конструкція може отримати позначку "атрибутивне": *контроль за ядерними озброєннями.* Визначення неоднозначності прийменникових словосполучень залежить здебільшого від лексичного значення головного слова

прийменникової конструкції та від лексичного значення прийменника. Обмеження, що можна накласти вже на рівні поверхнево-синтаксичного аналізу, зумовлені семантичною характеристикою одного із компонентів словосполучення, яку можна отримати завдяки модулю словотвірного аналізу. Наприклад, якщо головне слово прийменникової конструкції є віддієслівним іменником, для неї приписуємо значення – "об'єктні відношення": *сприяння в наданні, нагляду за дотриманням, зв'язків з громадськістю, зарахування до списків, підготовка до роботи, розгортання у бойові порядки, зіткненні з перешкодою.*

Значну частку аналізованих конструкцій становлять усталені словосполучення із характерним для текстів військової тематики значенням. Такі текстові структури із визначеним типом синтаксичних відношень у межах змодельованого тематичного корпусу можна подавати заздалегідь укладеним списком: *право на носіння, доповідей про хід бойових дій, інформацію про повітряну обстановку, дані про температуру*.

Більшість випадків неоднозначних прийменникових конструкцій отримають правильну синтаксичну ідентифікацію на семантичному рівні лінгвістичного аналізу за семантичною моделлю "частина/ціле": *затворна рама з газовим поршнем, газова трубка зі ствольною накладкою*.

Поширеними серед синтагм із прийменниковими конструкціями є обставинні відношення, зокрема способу дії та місця дії, в яких головним словом є віддієслівний іменник: *пересування на лижах, бій на великих відстанях, досягнення успіху у наступі, забезпечення безпеки в Україні.* У групі присудка в інших прийменникових структурах обставинні відношення способу дії можна визначити у поєднанні з об'єктними відношеннями: *за масштабами виділяють*, *поряд з традиційними джерелами небезпеки виділяють, утворені відповідно до законів України.*

У досліджуваному матеріалі головним словом прийменникового словосполучення вважається перше ліворуч. Проте у лінійній структурі речення часто залишається не з'ясованим питання вибору головного слова прийменникової конструкції з-поміж

декількох, наприклад у реченнях: *(**придатності**) (**громадян**) **України** до професійної військової служби*; *(**резерву**) (**кандидатів**) для проходження служби*; *(**сорочку**) **кольору** (**полину**) з погонами*; *(**придатні**) до проходження військової (**служби**) за станом здоров'я та віком; (**приймання**) військовослужбовців (**командуванням**) з особистих питань*.

Для синтаксично неоднозначних синтагм характерною є здебільшого варіантна визначеність головного слова: *(**застосовується**) для (**ураження**) (**противника**) в ближньому бою; що (**працює**) в ультрафіолетовому (**діапазоні**) з підвищеною стійкістю до теплових пасток; Гармата Т-84 (**оснащена**) ефективною (**стабілізацією**) у двох площинах; (**використовувати**) (**показання**) (**полонених**) в інтересах рішення бойових задач.*

Отримані дані дозволяють зробити висновок, що навіть незначний опрацьований матеріал свідчить про можливість застосування у визначенні синтаксичної неоднозначності ймовірнісного критерію (фільтру). Так, у синтагмах із двома претендентами на роль хазяїна найбільш достовірним є перший ліворуч від омонімічного гнізда.

Особливої уваги в автоматичному синтаксичному аналізі потребують конструкції із однорідними членами речення. Пояснюється це як невизначеністю єдиного підходу формалізації цих зв'язків між членами конструкцій, так і лінгвістичною природою цих конструкції [Дрейзин 1966; Дрейзин 1966; Иорданская 1967; Санников 1989 : 32–79]. Дослідники визначають декілька типів синтаксичної неоднозначності однорідних членів: 1) слово залежить від одного або від декількох однорідних; 2) омонімія за протиставленням однорідність/немає однорідності; 3) омонімія типу "з чим однорідне".

У проаналізованому нами матеріалі незначну частку становлять синтаксично неоднозначні синтагми із сурядним зв'язком (19 випадків). Неоднозначність таких конструкцій полягає у визначенні головного та залежних слів, наприклад:

- ***організація** (взаємодії) та (управління) ЗСУ у мирний та воєнний час* (однорідними можуть бути *організація* та *управлін-*

ня або *взаємодії* та *управління*; неоднозначним є визначення головного слова для прийменникового словосполучення – *ЗСУ, організація та управління, взаємодії та управління*);

- *складається з (**парів**) (розпечених речовин) ядерного боєприпасу, **повітря** і **часток** ґрунту* (визначення однорідності: *складається з парів, повітря, і часток…*чи *складається із з парів … боєприпасу, повітря, і часток …*);
- ***зв'язок** механізованих і танкових **підрозділів** з артилерією* (неоднозначність однорідних членів: що є однорідним *механізованих та танкових* залежать від *підрозділів* чи *механізованих підрозділів та танкових підрозділів з артилерією*; словосполучення *з артилерією* залежить від *танкових підрозділів* чи *від зв'язок*);
- *рухомого стержня і муфти* (чи є слово *рухомого* залежним і від *муфти*). Остаточне визначення такого типу омонімії вимагає більш ґрунтовного дослідження.

Отже, аналіз речень на рівні поверхнево-синтаксичного аналізу демонструє таку властивість тексту, як синтаксично неоднозначні конструкції на рівні словосполучення, синтагми, речення. Простішими щодо однозначного визначення синтаксичних відношень виявились синтаксично неоднозначні словосполучення, для яких вже на етапі поверхнево-синтаксичного аналізу можна застосувати певні фільтри розмежування. Інші типи потребують додаткового опрацювання, зокрема на рівні семантичного аналізу. Одними із пріоритетних завдань є визначення максимально повного списку властивих українській мові типів синтаксичної неоднозначності та розроблення семантичного аналізатора речень.

1. *Бугаков О. В*. Функціонування прийменників в українському тексті: морфологічний та семантико-синтаксичний аспекти / Автореф. дис. канд. філол. наук / О. В. Бугаков. – К., 2006. 2. *Гладкий А. В*. Синтаксические структуры естественного языка в автоматизированных системах общения / А. В. Гладкий. – М., 1985. 3. *Грязнухина Т. А*. Анализ предложных связей в научном тексте / Т. А. Грязнухина. – К., 1985. 4. *Дрейзин Ф. А*. Частота появления основных видов синтаксической омонимии в русских текстах / Ф. А. Дрейзин // НТИ. – № 12. – 1966. – С. 55–59. 5. *Дрейзин Ф. А.* Зависимые слова и группы

слов при однородных существительных с точки зрения автоматического синтаксического анализа (на материале русского языка) / Ф. А. Дрейзин // НТИ. – № 7. – 1966. – С. 43–45. 6. *Замаруєва І. В.* Комп'ютерна модель розуміння природно-мовної текстової інформації / І. В. Замаруєва // Проблемы программирования. – 1999. – № 2. – С. 96–102. 7. *Иорданская Л. Н.* Свойства правильной синтаксической структуры и алгоритм ее обнаружения (на материале русского языка) / Л. Н. Иорданская // Проблемы кибернетики. – Вып. 11. – М., 1964. – С. 215–243. 8. *Иорданская Л. Н.* Синтаксическая омонимия в русском языке (с точки зрения автоматического анализа и синтеза) / Л. Н. Иорданская // НТИ. – № 5. – 1967. – С. 9–17. 9. *Колесников Н. П.* Омонимия в предложении и вопросы ее устранения / Н. П. Колесников. – М., 1976. 10. *Мельчук И. А.* Опыт теории лингвистических моделей "Смысл-Текст". Семантика, синтаксис / И. А. Мельчук. – М., 1974. 11. *Санников В. З.* Русские сочинительные конструкции. Семантика. Прагматика. Синтаксис / В. З. Санников. – М., 1989. 12. *Толубко В. В.* Задачі автоматичної обробки синтаксичної структури в знання-орієнтованій системі машинного перекладу / В. В. Толубко, О. О. Шипнівська, А. В. Ляшенко // Вісник Київського нац. ун-ту ім. Т. Шевченка. Серія: Військово-спеціальні науки. – Вип. 27. – 2010. – С. 136–140. 13. *Севбо И. И.* Графическое представление синтаксических структур и стилистическая диагностика / И. И. Севбо. – К., 1981. 14. *Чередниченко І. Г.* Принципи розрізнення неузгоджених означень і присубстантивних додатків, виражених формою родового відмінка без прийменників / І. Г. Чередниченко // Українська мова. – 1955. – № 2. – С. 33–38.

Рассматриваются теоретико-практические основания определения синтаксически неоднозначных конструкций в автоматическом синтаксическом анализе. На материале текстов военной тематики проанализированы основные типы синтаксической омонимии.

Ключевые слова: автоматический синтаксический анализ, знаниеориентированная система машинного перевода, синтаксическая неоднозначность, типы синтаксических омонимов, синтаксическая структура.

In the paper the theoretical and practical basis of the syntactical ambiguity for the automatic syntactical analysis is described. The description of main types of syntactical homonyms for the automatic syntactical analysis in military texts is presented.

Key words: automatic syntactical analysis, knowledge-based machine translation system, syntactical ambiguity, types of syntactical homonyms, syntactical structure.

Стаття надійшла до редакції 25.09.2012

НАШІ АВТОРИ

Алексієнко Людмила Антонівна – кандидат філологічних наук, доцент кафедри сучасної української мови Київського національного університету імені Тараса Шевченка

Дарчук Наталія Петрівна – кандидат філологічних наук, доцент кафедри сучасної української мови Київського національного університету імені Тараса Шевченка

Держанський Іван Олександрович – доктор, старший науковий співробітник Інституту математики й інформатики Болгарської академії наук

Завадська Вікторія Валеріївна – кандидат філологічних наук, викладач кафедри української мови, літератури та культури факультету лінгвістики Національного технічного університету "Київський політехнічний інститут"

Кислюк Лариса Павлівна – кандидат філологічних наук, старший науковий співробітник відділу структурно-математичної лінгвістики Інституту української мови НАН України

Лісовський Володимир Миколайович – заступник начальника кафедри військового перекладу Військового інституту Київського національного університету імені Тараса Шевченка

Міщенко Алла Леонідівна – кандидат філологічних наук, старший викладач кафедри перекладу та загального мовознавства Кіровоградського державного педагогічного університету імені Володимира Винниченка

Романишин Мар'яна Михайлівна – аспірантка кафедри систем автоматизованого проектування Національного університету "Львівська Політехніка"

Романюк Андрій Богданович – кандидат технічних наук, доцент кафедри систем автоматизованого проектування Національного університету "Львівська Політехніка"

Сірук Олена Борисівна – кандидат філологічних наук, науковий співробітник лабораторії комп'ютерної лінгвістики Інституту філології Київського національного університету імені Тараса Шевченка

Січінава Дмитро Володимирович – кандидат філологічних наук, старший науковий співробітник відділу корпусної лінгвістики та лінгвістичної поетики Інституту російської мови імені В. В. Виноградова РАН

Старко Василь Феодосійович – кандидат філолологічних наук, докторант кафедри прикладної лінгвістики Східноєвропейського національного університету імені Лесі Українки

Чейлитко Наталія Генндіївна – кандидат філологічних наук, асистент кафедри сучасної української мови Київського національного університету імені Тараса Шевченка

Шведова Марія Олексіївна – кандидат філологічних наук

Шипнівська Ольга Олександрівна – кандидат філологічних наук, молодший науковий співробітник, Військовий інститут Київського національного університету імені Тараса Шевченка

ЗМІСТ

Кафедра сучасної української мови Київського національного університету продовжує видання міжвідомчого збірника "Українське мовознавство", що виходить із періодичністю – один випуск збірника на рік.

Статті до збірника можна надіслати *на поштову адресу*:

01601, Київ, бульвар Тараса Шевченка, 14,
Київський національний університет імені Тараса Шевченка,
Інститут філології, кафедра сучасної української мови

або на електронну адресу: *labacompli@gmail.com*

чи подати особисто Ходаківській Ярині Володимирівні (лабораторія комп'ютерної лінгвістики, ауд. 130).

Телефони: 239-33-49 – кафедра сучасної української мови,
239-33-54 – лабораторія комп'ютерної лінгвістики.

Статті для публікації у збірнику мають відповідати вимогам:

1. Файл із текстом статті та відомостями про автора подається в текстовому редакторі Microsoft Word:
– відомості про автора статті: прізвище, ім'я, по батькові, місце роботи, учене звання, науковий ступінь, адреса, контактні телефони;
– резюме і ключові слова – українською, російською та англійською мовами.
– назва – українською, російською та англійською мовами.
– прізвище та ім'я – українською, російською та англійською мовами.

Окремо подається роздрукований примірник тексту, вичитаний та підписаний автором. Статті аспірантів подаються з рекомендацією наукового керівника. **Не вичитані автором тексти (з фактичними або технічними помилками) до друку не приймаються**.

Правила оформлення статті:

1. Обсяг – 0,5 др. арк. (20 000 знаків), **без переносів**.
2. Формат – А5.

3. Шрифт – Times New Roman, кегль – 11 пунктів, через один інтервал.

4. Поля: верхнє – 2,0; нижнє – 2,2; ліве – 2,0; праве – 2,0 см. Відступ абзацу – 0,5 см.

5. Ім'я (повністю) і прізвище, місце роботи (назва навчального закладу чи установи без скорочень) друкуються праворуч вгорі (10 кегль) – ***жирним курсивом***.

6. Назва статті друкується **ВЕЛИКИМИ** літерами **жирним** шрифтом (10 кегль) через інтервал після прізвища автора.

7. Анотація та ключові слова українською мовою друкуються *курсивом* через один інтервал після назви статті (10 кегль).

8. Текст статті подається українською мовою через інтервал після анотації українською мовою.

9. Після тексту статті подаються: література, **оформлена відповідно до вимог**, і анотації та ключові слова російською та англійською мовою.

10. У тексті **не використовувати стильові маркери.**

11. У тексті перед згадуваними прізвищами має бути лише один ініціал. Між ініціалом і прізвищем ставиться **нерозривний** пробіл (одночасне натискання клавіш ctrl + shift + пробіл), напр.: І.ºВихованець, М.ºЖовтобрюх.

12. Приклади виділяються: речення, слова – *курсивом без лапок* (приклади з художньої літератури також); фонеми, морфеми, символи – **жирним** шрифтом.

13. Цитати беруться у подвійні верхні лапки "…" (напр.: Ш. Баллі писав: "Ми уподібнюємо абстрактні поняття предметам чуттєвого світу, бо для нас це єдиний спосіб пізнати їх і познайомити з ними інших" [Баллі : 221]).

14. Значення слів беруться в одиночні верхні лапки '…' (напр.: *впадина* 'яма в річці'; *просвистина* 'прогалина в лісі').

14. Слід чітко диференціювати тире (–) і дефіс (-). Напр.: "Символізм – це…", 5–10 (для сторінок), 2008–2009 (для років), але: науково-технічний, не сьогодні-завтра.

15. Авторські пропуски тексту позначаються трьома крапками в ламаних дужках <…>.

16. Посилання в тексті подавати тільки у квадратних дужках, напр.: [Кучеренко], [Грищенко 1977: 11], [Грищенко 1979: 4–5; Русанівсь-

кий: 31]. Між номером джерела і номером сторінки (номерами сторінок) мають бути двокрапка і пробіл: [Кучеренко: 55–56] або [Грищенко 1977: 4–5; Русанівський: 31–33].

17. Література подається в алфавітному порядку в кінці статті (слово *література* не пишеться). Номер у списку літератури повинен відповідати лише одному джерелу. Між номером і прізвищем автора, а також між ініціалами має бути **нерозривний пробіл**.

Зразок оформлення літератури

1. *Виноградов В. В*. Русский язык: Грамматическое учение о слове / В. В. Виноградов. – М., 1986. 2. *Вихованець І.* Теоретична морфологія української мови / І. Вихованець, К. Городенська. – К., 2004. 3. *Кравченко М. В.* Словотвірний аналіз дериватів із суфіксом -ин(а) і його похідними формантами / М. В. Кравченко // Українська мова і література в школі. – 1987. – № 9. – С. 24–31. 4. Сучасна українська літературна мова / За ред. А. П. Грищенка. – К., 1997. 5. *Терлак З. М*. Ад'єктивні словосполучення з об'єктними відношеннями у сучасній українській літературній мові: Дис. … канд. филол. наук / З. М. Терлак. – К., 1982. 6. *Чейлитко Н. Г.* Відображення синтаксичної та лексичної цілісності речення через зони зв'язків словоформ: Автореф. …дис. … канд. філол. наук / Н. Г. Чейлитко. – К., 2009.

Вимоги до оформлення статей **обов'язкові**.

Рукописи, не прийняті до друку, не повертаються.

Вимоги розміщені на сайті Інституту філології Київського національного університету імені Тараса Шевченка за електронною адресою http://philolog.univ.kiev.ua/php/kafkaf.php?id=1&sid=6.

Редколегія

Наукове видання

УКРАЇНСЬКЕ МОВОЗНАВСТВО

Міжвідомчий
науковий
збірник

Випуск 43

Засновник та видавець – Київський національний університет імені Тараса Шевченка. Свідоцтво Міністерства інформації України про державну реєстрацію друкованого засобу масової інформації КІ № 3745 від 25.03.99. Кафедра сучасної української мови. Головний редактор А. К. Мойсієнко. Адреса: Київський національний університет імені Тараса Шевченка. Інститут філології: 01601, Київ, б-р Тараса Шевченка, 14, кімн. 95. ☎ (38044) 239 3349, 239 3354

Оригінал-макет виготовлено Видавничо-поліграфічним центром "Київський університет"

**Формат 60x84$^{1/16}$. Ум. друк. арк. 6,97. Наклад 100. Зам. № 213-6467.
Гарнітура Times. Папір офсетний. Друк офсетний. Вид. № Іф5.
Підписано до друку 18.04.13**

**Видавець і виготовлювач
Видавничо-поліграфічний центр "Київський університет"
б-р Т. Шевченка, 14, м. Київ, 01601
☎ (38044) 239 32 22; (38044) 239 31 72; факс (38044) 239 31 28
e-mail: vpc@univ.kiev.ua
WWW: http://vpc.univ.kiev.ua
Свідоцтво суб'єкта видавничої справи ДК № 1103 від 31.10.02**