УДК 811.161.2'33

Василь Старко
Східноєвропейський національний університет
імені Лесі Українки, м. Луцьк

КОМП'ЮТЕРНІ ЛІНГВІСТИЧНІ ПРОЕКТИ ГУРТУ R2U: СТАН ТА ЗАСТОСУВАННЯ

У статті розглянуто низку комп'ютерних лінгвістичних проектів, що їх гурт r2u розвиває для української мови: словникові вебсайти r2u.org.ua й e2u.org.ua, Великий електронний словник української мови (ВЕСУМ), що його покладено в основу повнотекстового пошуку в українській Вікіпедії, засіб перевіряння орфографії, граматики й стилю «Правописник LanguageTool» та Браунський український корпус (БрУК). Подано огляд цих проектів, проаналізовано їхні особливості, описано поточний стан та окреслено потреби користувачів, що їх вони покликані задовольнити. У статті наведено посилання на онлайнові ресурси, де можна безпосередньо скористатися напрацюваннями гурту r2u. Зроблено висновки щодо можливостей застосування результатів цих проектів та перспектив їхнього розвитку.

Ключові слова: *комп'ютерна лінгвістика, автоматичне опрацювання природної мови, комп'ютерна лексикографія, корпусна лінгвістика, корпус, r2u, електронний словник, ВЕСУМ, Вікіпедія, Правописник, LanguageTool, Браунський корпус, БрУК.*

У нинішню епоху стрімкого розвитку інформаційних технологій зростає потреба в засобах автоматичного опрацювання природної мови та в оперативному доступі до значних масивів мовних даних, зокрема в машиночитному форматі [9], [10], [25]. Гостро актуальним є формування комп'ютернолінгвістичної «екосистеми» для української мови — інструментарію й мовних даних, що їх фахівці створюють і удоступнюють іншим дослідникам, розробникам і програмувальникам. Без такої екосистеми чимало проектів доводиться починати «з нуля» й важко уникнути дублювання роботи, наприклад, укладання словника словозміни, який виявляється доконечно потрібним для багатьох завдань. Деякі важливі елементи цієї екосистеми вже

створено, наприклад, комп'ютерний морфемно-словотвірний фонд української мови (роботу над ним розпочато ще наприкінці 1980-х років під керівництвом Н.Ф. Клименко в Інституті мовознавства ім. О.О. Потебні й згодом продовжено в Інституті української мови НАНУ) [10], комп'ютерний фонд інновацій у сучасній українській мові (Інститут української мови НАНУ, керівник Є.А. Карпіловська) [12], доробок лабораторії комп'ютерної лінгвістики Інституту філології Київського національного університету імені Тараса Шевченка (керівник Н.П. Дарчук) [20] та інші, однак багатьох ресурсів ще бракує. На заповнення цих прогалин у комп'ютерній лінгвістиці і комп'ютерній лексикографії й спрямовано працю гурту r2u. У статті йтиметься про такі ресурси й засоби: словникові вебсайти r2u.org.ua та e2u.org.ua, Великий електронний словник української мови (ВЕСУМ), засіб перевіряння орфографії, граматики й стилю «Правописник LanguageTool» і Браунський український корпус (БрУК).

Почнімо з короткої історичної довідки. 10 років тому кількох не знайомих між собою людей об'єднала ідея повернути українству заборонений до вживання, вилучений з обігу та з бібліотек [19: 74], частково знищений, а частково замкнений у радянські спецхрани академічний «Російсько-український словник» за редакцією А. Кримського й С. Єфремова (1924—1933, далі РУС) [29]. Цей багатющий словник став стрижнем сайту r2u.org.ua та рушієм і мірилом дальшої праці гурту. Першим кроком стало сканування РУСа — за сприяння Михайлини Коцюбинської, яка допомогла отримати доступ до паперового видання, це зробив київський книжник Валентин Кульков. Через Віктора Кубайчука та Ольгу Кочергу електронна копія віднайденого словникового скарбу дійшла до зацікавлених фахівців, зокрема долучився директор видавництва «К.І.С.» Юрій Марченко, його колега Олександр Телемко, який зробив електронний текстовий файл РУСа, та програмувальник і комп'ютерний лінгвіст Андрій Рисін. До цього ядра гурту, що географічно розташувався в трикутнику Київ-США-Луцьк, долучалося на різних етапах і в різних проектах чимало осіб, яких тут годі перелічити. Першим результатом співпраці став, 2007 року, вебсайт r2u.org.ua, де викладено електронний варіант РУСа із повнотекстовим пошуком і можливістю завантажити текстовий pdf-файл словника. Онлайнова версія стала можливою завдяки гранту від Наукового товариства імені Шевченка у США з Фонду ім. Івана Романюка. Абревіатура-назва сайту r2u (англ. *Russian to Ukrainian*, тобто з російської на українську) окреслювала його спрямування: з часом додати ще низку лексикографічних раритетів та ретельно опрацьованих високоякісних словників. Згодом постав споріднений вебсайт e2u.org.ua з низкою потужних англійсько-українських та українсько-англійських словників. Паралельно тривала робота над словником словозміни української мови (нині ВЕСУМ), який ліг в основу «Правописника» і дав поштовх створенню БрУКу. Всі ці проекти нині активно розвиваються.

1. Словниковий сайт r2u.org.ua

Найцінніший словник сайту — це, безумовно, РУС. Це останній лексикографічний опис української мови до початку кампанії російщення й «злиття мов» у підрадянській Україні. РУС уклали й зредагували фахівці найвищого класу, однак вже навіть тоді вони зазнали утисків і переслідувань. Внаслідок компартійного тиску 1/5 частину надрукованого словника (*над—нять*) було перероблено, а четвертий том словника знищено в усіх формах, попри те, що редакційна колегія цілком підготувала його до друку [40]. Після цього радянська влада заборонила словник і вилучила його з продажу та бібліотек. Нові настанови укладання словника викладено в кінці другого тому РУСа — їхня суть зводиться до того, щоб «не допуститися в цій його частині шкідливих буржуазних і націоналістичних тенденцій попередніх випусків», «подавати поширені міжнародні слова і терміни в їх інтернаціональній формі, не перекладаючи їх штучно, без потреби на українську мову» [29: 1052—1054]. І навіть у такій частково понищеній і поневченій формі РУС розкриває подиву гідне багатство української мови. У 1950-х роках Ю. Шевельов назвав його «найвищим авторитетом у справі норм української літературної мови» [46: 12]. У наш час Є.А. Карпіловська, О.Д. Кочерга та Є.В. Мейнарович так оцінюють значення РУСа: він «є не лише найґрунтовнішим натепер російсько-українським словником, а й джерелом питомої української лексики, взірцевих словотворчих моделей, мовних конструкцій та усталених висловів, зразків запозичування іншомовних слів та їх адаптування до системи української мови. Він не лише не застарів, а по глибшому вивченні напевне постане як найповніше й найдокладніше сучасне лексикографічне джерело, що його значення для дальшого розвитку української мови важко перебільшити» [11: 115].

Онлайнова версія з гнучким пошуковим інтерфейсом дає змогу використовувати цей словник не лише для російсько-українського перекладу, а, наприклад, для пошуку слів, близьких за значенням чи формою. Приміром, пошук на слово *блищати* видасть такі синоніми: *блискотіти, виблискувати, вилискувати, зоріти, горіти, ясніти, світити, сяяти*; (про рівну й гладку поверхню) *вилискувати, лисніти, лиснитися, лощитися*; (мінливим світлом, блиском) *грати, вигравати, мигтіти, жахтіти, бреніти, леліти*; (коли-не-коли, місцями) *блискати, поблискувати, поблискувати*; (загорятися й гаснути) *блимати, бликати*. Пошук на *тель серед «українських слів без цитат» видає десятки слів із суфіксом *-тель*, зокрема таких, що їх годі знайти в сучасних словниках: *воскреситель, всесотворитель, відновитель, гоїтель, землерушитель, зачатель, зловчитель, звістителъ, сповістителъ, вибавитель, вивіритель, миритель, світоправитель* тощо. Загалом РУС містить значний шар питомої, однак призабутої (зокрема внаслідок втручання сумнозвісних позамовних чинників) лексики, що може стати джерелом збагачування сучасної мови. Перші ґрунтовні дослідження в цьому напрямі засвідчують відродження цієї лексики [13], [23], [45]. Втішає той факт, що завдяки сайту r2u та недавньому перевиданню першого тому в паперовому форматі [30] РУС не лише по-

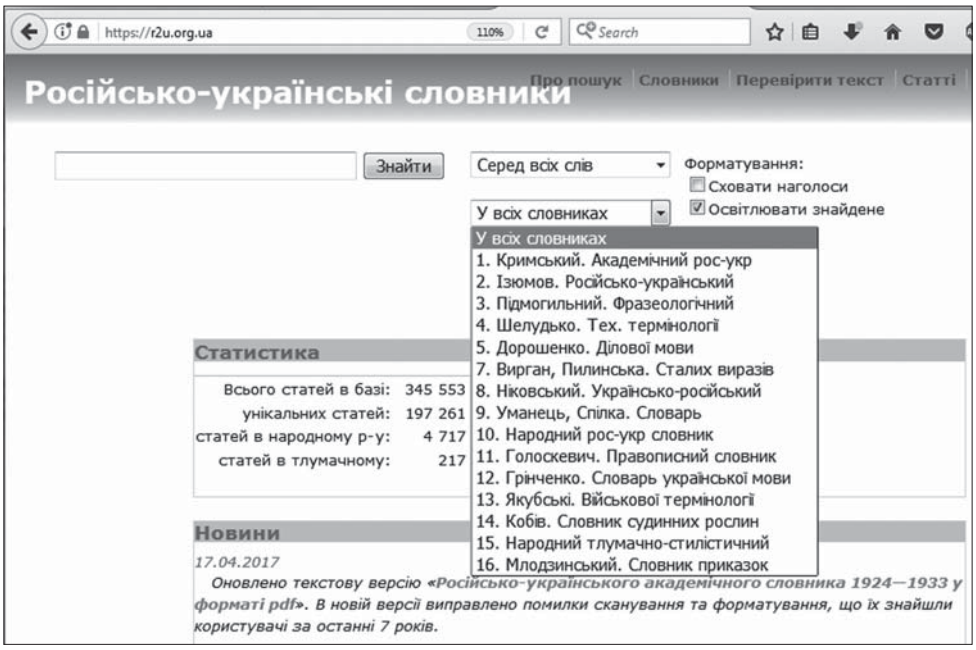

Рис. 1. Перша сторінка сайту r2u.org.ua

вернувся до активного обігу, а й стає об'єктом докладних лінгвістичних досліджень [3], [26], [33], [35], [42], [43]. Дослідники навіть роблять виважені спроби реконструювати втрачений 4-й том словника [44].

Словники сайту r2u стають у пригоді не лише мовознавцям, вчителям і викладачам української мови, студентам, а й перекладачам не з російської мови. Річ у тім, що нерідко трапляється ситуація, коли складно дібрати влучний відповідник до англійського, французького чи іншого чужомовного слова або вислову, а словники між цими мовами й українською не подають достатньої кількості відповідників. Тоді пошук на r2u (за російським чи українським словом, або ж кількаразовий пошук за різними словами) може привести перекладача до шуканої одиниці чи підказати влучний відповідник. Наприклад, автор цих рядків постійно і з великою користю послуговувався ресурсами r2u під час перекладу автобіографії Нельсона Мандели з англійської на українську мову [21].

До бази сайту r2u внесено загалом 16 словників загальним обсягом 345 тис. словникових статей, із яких майже 200 тис. унікальні. Абсолютна більшість словників російсько-українські й належать до періоду українізації [32]. З давніших словників подано двічі вичитаний «Словарь української мови» за редакцією Бориса Грінченка та «Словарь росийсько-український» М. Уманця (М.Ф. Комарова) й А. Спілки. В базі сайту є й сучасний словник Юрія Кобіва, що містить тисячі народних назв рослин [14], а також «Українсько-російський словник» А. Ніковського 1927 року та «Правописний словник» Г. Голоскевича 1929 року. Своєю лексикографічною колекцією сайт завдячує, зокрема, тим, хто переводив словники з оригіналу в електронну форму, — насамперед Вікторові Кубайчуку.

Добірка на сайті таких цінних словників, як академічний РУС, «Російсько-український словник сталих виразів» І. Виргана та М. Пилинської [7], «Російсько-український фразеологічний словник» В. Підмогильного й Є. Плужника [31] та «Практичний російсько-український словник приказок» Г. Млодзинського (за ред. М. Йогансена) [27], слугує прекрасним ресурсом з української фразеології.

Переважну більшість словників користувачі можуть звантажити собі у формі текстових pdf-файлів зі сторінки «Словники» [32], однак досвід показує, що набагато зручніше й швидше користуватися пошуком на сайті, адже він виводить результати пошуку послідовно з кожного словника в базі. До того ж, у розділі «Словники для звантаження» [36] викладено електронні копії (зображення) кільканадцяти словників, що ними можна користуватися офлайн. Це словники української мови П. Білецького-Носенка й Д.І. Яворницького, «Стилістичний словник» І. Огієнка, «Словник чужомовних слів, виразів і приповідок» О. Скалозуба [34], низка термінологічних словників. Однак навіть найбільша колекція словників не може задовольнити всіх потреб користувачів. Тут на допомогу приходить форум сайту, де можна отримати мовні консультації, порадитися щодо вибору відповідника, обговорити складні питання слововжитку, запропонувати свої варіанти перекладу тощо. На основі дописів зареєстрованих користувачів форуму поповнюється сучасний «народний» російсько-український словник. Народним його названо через те, що забраклі статті й відповідники пропонують самі користувачі сайту. Після обговорення на форумі (інколи досить докладно і з посиланнями на текстові джерела) статтю в словник додає редактор. У такий спосіб мовці долучаються до лексикографічного опису гостроактуальної лексики, якої часто бракує в паперових словниках.

У розділі «Про пошук» описано, на що слід звертати увагу, формуючи пошукові запити на сайті. Високу гнучкість пошуку забезпечують символи заміни, що їх можна використовувати в пошукових запитах: ? замінює будь-яку одну літеру (наприклад, на запит *клас?* буде знайдено *класу, класі* тощо), а знак зірочки * замінює нуль або більше літер (*клас* — клас, класу, класі, класом, класами, класний, класти, класифікація* тощо).

На сайті втілено елементи інтерактивності: це не лише форум, а й можливість відсилати адміністраторам звіти про помічені помилки. Такий зворотний зв'язок дає змогу періодично очищати словники від помилок. Завдяки прискіпливому добору високоякісних словників сайт r2u належить до найпопулярніших словникових інтернет-ресурсів України, опрацьовуючи близько 1,5 млн запитів на рік із понад 70 тис. різних IP-адрес, і ці показники постійно зростають.

2. Словниковий вебсайт e2u.org.ua

Як вказує назва (e2u — *English to Ukrainian*), цей ресурс покликаний задовольнити потреби переважно в англійсько-українських словниках. На відміну від r2u тут викладено сучасні словники.

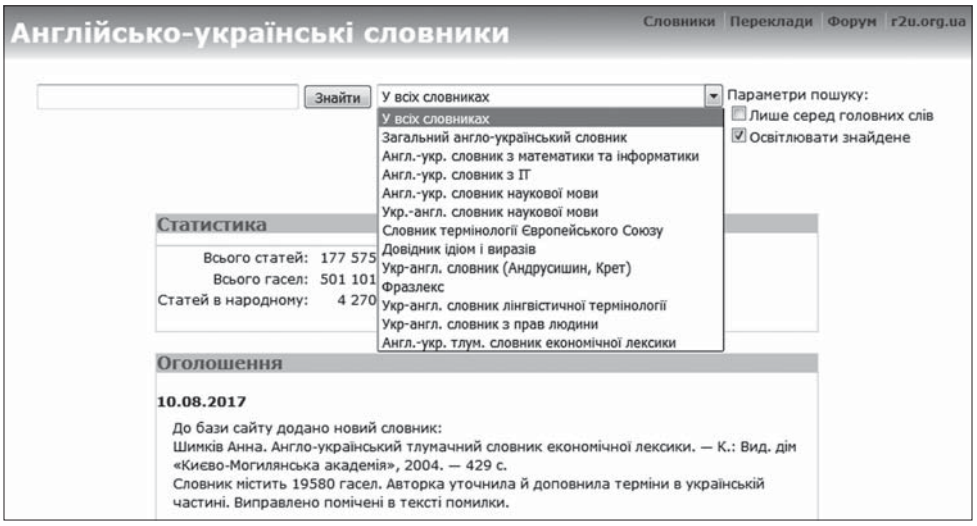

Рис. 2. Перша сторінка сайту e2u.org.ua

Словникова колекція сайту складається з фундаментальних термінологічних словників, що містять великий відсоток загальномовної лексики, а також фразеологічних та загальних словників. Поданий на сайті «Англійсько-українсько-англійський словник наукової мови (фізика та споріднені науки)» О. Кочерги та Є. Мейнаровича [17], що містить понад 280 тис. гасел у двох частинах, — це один з найбільших термінологічних словників України. До бази сайту внесено й ґрунтовні сучасні термінологічні словники в галузі математики та інформатики [22], економіки [47], лінгвістики [15]. Викладено також великий діаспорний українсько-англійський словник К. Андрусишина й Я. Крета на 133 тис. гасел. На основі запитів користувачів триває укладання «народного» англійсько-українського словника (понад 4 тисячі ретельно опрацьованих статей): помічаючи прогалини в наявних на сайті словниках, користувачі пропонують і обговорюють нові статті на форумі, а редактор (Андрій Рисін) укладає їх, спираючись, зокрема, на сучасні тлумачні словники англійської мови й англійсько-українські словники. Окрім того, сайт автоматично веде статистику запитів із нульовим результатом (у словниках нічого не знайдено). Найчастотніші з них вказують на те, яких статей найбільше бракує користувачам, і редактор згодом додає їх до народного словника. Поповнюється новими статтями англійсько-український фразеологічний словник (Фразлекс). У найближчих планах — додання нових словників, зокрема фундаментального енциклопедичного словника з хімії [24]. Наразі словникова база сайту нараховує сумарно понад 500 тис. гасел.

3. Правописник LanguageTool

На платформі languagetool.org створено засоби перевіряння орфографії, граматики й стилю для 28 мов. Гурт r2u розвиває український модуль під назвою «Правописник», що нині містить замалим 500 правил. Основну

частину роботи в цьому напрямку виконує Андрій Рисін. Модуль спирається на чинний правопис, у його основі лежить словник на 300 тис. лем, що дає змогу перевіряти тексти різної тематики й рівня складності, а також список понад 3 тис. однослівних покручів із варіантами виправлення. Перш ніж додати правило до модуля, гурт тестує його на корпусі текстів обсягом понад 100 млн словоформ і вишліфовує на основі отриманих результатів. В основу правил перевіряння покладено принцип точного розпізнавання: правило спрацьовує лише тоді, коли можна з певністю твердити про наявність помилки в тексті. З погляду користувача це вигідно відрізняє «Правописник» від подібних засобів, що застосовують принцип широкого охоплення: правило спрацьовує в усіх місцях потенційних помилок, наприклад, щоразу на слові *даний*, незалежно від того, чи правильно його насправді вжито в реченні. «Правописник» перевіряє різні типи помилок: орфографічні, пунктуаційні, граматичні, стилістичні й логічні. Введений текст він автоматично розбиває на речення, речення — на лексеми (числа, пунктуаційні знаки), до кожного слова встановлює його лему (проводить лематизацію) й граматичні ознаки (наприклад, іменнику надає теги частини мови, роду, числа й відмінка). Проведений у такий спосіб морфологічний аналіз дає змогу гнучко застосовувати розроблені правила, охоплюючи ними всі словоформи потрібного слова. Коли спрацьовує одне з таких правил, засіб виводить на екран повідомлення про помилку, короткий опис та пропозиції виправлення.

Наведімо приклади до кожного типу правил:

мовні покручі: *рани заживають* (замість *загоюються*), *присвоїти* (замість *надати*) *звання*, *грецький* (замість *волоський*) *горіх*, *користуватися попитом* (замість *мати попит*), *переводити* (замість *переказувати*) *гроші*;

граматичні помилки: *згідно чого* (*згідно з чим*), *навчати чому* (*чого*), *завідувач чим* (*чого*);

логічні помилки: *30 лютого*;

орфографічні помилки: *смт.* (зайва крапка в скороченні), *гривна* (*гривня*);

пунктуаційні помилки: *почервоніти, як рак* (зайва кома), *Ви мабуть знаєте* (вставне слово не виділено комами);

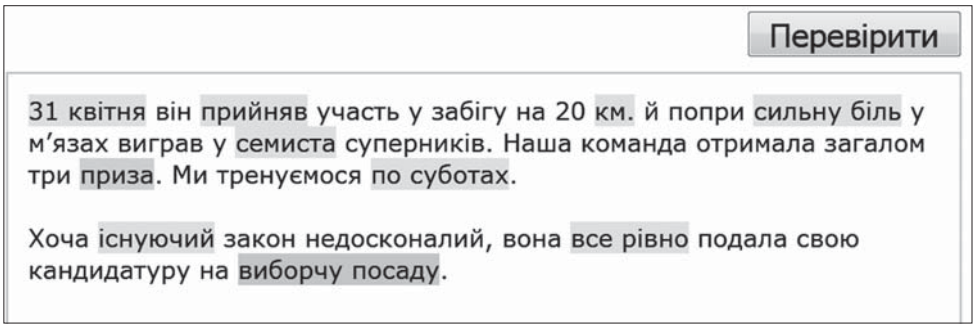

Рис. 3. Приклади різних типів помилок на сторінці «Правописника» http://r2u.org.ua/check

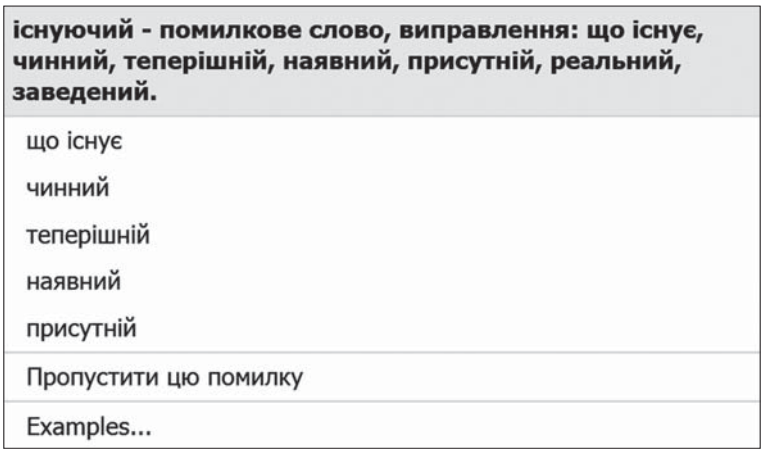

Рис. 4. Приклад пояснення помилки

стилістичні помилки: *більш світліший* (*світліший*), *о одинадцятій* (порушення правил милозвучності), *виборна кампанія*, *виборча посада* (плутання паронімів), *головний пріоритет* (плеоназм).

Скористатися «Правописником» можна в кілька способів: на сайтах languagetool.org/uk/ та http://r2u.org.ua/check, встановивши додаток у браузери «Фаєрфокс» та «Хром» чи в програму «ЛібреОфіс». Існують також додатки для документів Гугла й текстового редактора «Ворд». Принцип роботи засобу такий: вставлений або виділений текст надсилається на сервер, обробляється (але не зберігається на ньому) й повертається користувачеві разом зі звітами про помилки. Виняток — додаток до «ЛібреОфіс», що працює в автономному режимі.

У повідомленнях про деякі помилки подано гіперпосилання на онлайнові ресурси з докладнішим поясненням, наприклад, на книжку Б. Антоненка-Давидовича «Як ми говоримо» [2]. Завдяки цьому «Правописник» виконує й освітню функцію, сприяючи підвищенню мовної культури користувачів. Засіб також сигналізує користувачеві, що слово написано згідно з альтернативним правописом. Таке повідомлення з'являється, приміром, до слів *проєкт*, *радости*, *діягональ* тощо. Виловлює він і мішанину розкладок клавіатури, коли замість українських літер вставлено латинські — така заміна не помітна для людського ока, але збиває алгоритми машинного опрацювання мовних даних, зокрема під час пошуку.

Користувачі засобу мають змогу сформулювати власні правила й відіслати їх на розгляд. Розробники постійно поповнюють набір правил, зокрема додають важкоформалізовні правила перевіряння узгодження між словами в реченні. Завдяки доступності, гнучкості й опертю на великий, постійно поповнюваний словник «Правописник» допомагає редакторам, перекладачам, студентам і всім, хто працює з текстами, не лише позбутися багатьох помилок, а й глибше опанувати багатства української мови.

Отже, «Правописник» — це зручний електронний засіб контролю якості українських текстів, систематизації й практичного застосування мовностилістичних правил, підвищення грамотності й мовної культури. За умови масового й систематичного користування він здатен заощадити час і зусилля багатьох редакторів, коректорів й авторів текстів.

4. Великий електронний словник української мови (ВЕСУМ)

В основі багатьох засобів автоматичного опрацювання природної мови лежить словник словоформ. У мовах аналітичного типу він може мати форму списку повних словоформ, однак для мов із високим ступенем флективності, до яких належить й українська, оптимальна будова словника — це зазвичай список лем із кодами словозміни, на основі яких генеруються всі потрібні словоформи. Саме такий підхід використано у ВЕСУМі. Докладному розгляду цього лексикографічного ресурсу буде присвячено окрему статтю під назвою «Великий електронний словник української мови (ВЕСУМ) як засіб NLP для української мови», яку ми плануємо невдовзі опублікувати, тому тут обмежимося коротким оглядом основних особливостей.

Починаючи від 1990-х років, ВЕСУМ пройшов довгий шлях від словника для перевіряння орфографії у відкритій операційній системі «Лінукс» до сучасного електронного словника лем, словоформ і граматичних ознак (тегів) у машиночитному форматі. З обсягом 316 тис. лем, з яких генеруються понад 4 млн словоформ, ВЕСУМ — найбільший словник такого типу для української мови. Усі виходові дані проекту викладено у вільному доступі онлайн [5]. Словник використано для забезпечення роботи «Правописника», для морфологічного аналізу в Браунському українському корпусі (про це нижче) та в інших проектах із комп'ютерної лінгвістики, зокрема для побудови векторів слів [28]. У червні 2017 р. за допомогою нового українського аналізатора на основі ВЕСУМу переіндексовано базу пошуку української Вікіпедії. Якщо раніше тут застосовували змодифікований російський аналізатор з неуникними прогалинами в пошуку, то тепер у результатах виводиться шукане слово в усіх його словоформах.

У роботі над словником гурт r2u спирався передусім на ґрунтовний «Граматичний словник української літературної мови. Словозміна» колективу авторів під керівництвом В.І. Критської та за редакцією Н.Ф. Клименко [18], залучаючи й інші джерела [1], [6], [37]. Теоретичні підвалини забезпечила академічна «Теоретична морфологія української мови» І. Вихованця і К. Городенської [8]. ВЕСУМ характеризують такі ключові особливості: 1) компактна система кодів відмінювання та тегів слів; 2) охоплення абревіатур і скорочень; 3) подання альтернативних правописних варіантів, рідковживаних слів і форм; 4) понад 47 тис. власних назв, зокрема 22 тис. прізвищ (українських та часто вживаних іноземних), 3 тис. імен та географічні назви, запроваджені внаслідок декомунізації; 5) подання нерекомендованих слів (активних дієприкметників, невдалих кальок тощо)

та варіантів їх заміни; 6) відкритість проекту. Словозміну у ВЕСУМі зреалізовано з використанням таких компонентів: 1) словник лем із кодами парадигм; 2) правила генерування словоформ на основі цих кодів; 3) програмова логіка генерування словоформ; 4) винятки.

Жоден словник не здатен охопити всіх можливих слів, проте електронний формат ВЕСУМу дав змогу впровадити так зване «динамічне тегування», коли засіб розпізнає певні типи слів у реченні за шаблонами замість шукати їх у списку лем. Цей підхід застосовано, зокрема, до таких класів слів: 1) деякі складні прикметники (наприклад, *125-та*, *австронімецький*); 2) прислівники на *по-* (наприклад, *по-чилійськи, по-чилійському*); 3) складні іменники (*лікар-гомеопат, місто-герой*); 4) слова з частотними формантами *арт-*, *інтернет-* тощо (близько 400 формантів). Точність розпізнавання таких типів слів (в усіх відмінкових формах) за допомогою динамічного тегування становить близько 95%. Легко бачити, що простим переліком такі слова в словнику задати важко та й навряд чи доцільно.

ВЕСУМ виконує завдання морфологічного аналізу й синтезу. Синтез передбачає генерування усіх словоформ певної леми, а аналіз полягає в лематизації (зведенні словоформи до леми) й присвоєнні цій словоформі відповідних граматичних тегів. Наприклад, словоформа *розумієте* лематизується до *розуміти* й дістає перелік граматичних тегів verb:imperf:pres:p:2, тобто дієслово, недоконаний вид, теперішній час, множина, друга особа. Відповідно, під час синтезу з леми *розуміти* генеруються словоформи з такими ланцюжками тегів:

розуміти verb:imperf:inf
розумій verb:imperf:impr:s:2
розуміймо verb:imperf:impr:p:1
розумійте verb:imperf:impr:p:2
розумію verb:imperf:pres:s:1
розумієш verb:imperf:pres:s:2

Отже, ВЕСУМ не лише перевіряє орфографію, граматичну правильність і стилістичну витриманість тексту, а й слугує для забезпечення повнотекстового пошуку (у Вікіпедії та на інших платформах) і є ключовим складником проектів у галузі комп'ютерної лінгвістики. Від інших таких словників він відрізняється насамперед форматом (машиночитний, вільно поширюваний), ширшим охопленням лексики (зокрема власних назв) і динамічним характером (постійно поповнюється). На часі є забезпечення зручного доступу до ВЕСУМу як до довідкового джерела: з цією метою на сайті r2u плануємо створити користувацький інтерфейс словника, який виводитиме на екран всю парадигму шуканого слова.

5. Браунський український корпус

Браунський корпус (*англ.* Brown Corpus), що його створили В. Нельсон Френсис та Генрі Кучера в Браунському університеті (США) в 1960-х роках, став взірцем для створення таких корпусів-мільйонників для англійської й інших мов. На сьогодні це корпуси малого обсягу, які, однак,

важливі тим, що на їхній основі можна побудувати статистичну модель мови й натренувати програму-аналізатор, яка далі в автоматичному режимі зможе проаналізувати значно більші обсяги текстів. Із цих міркувань започатковано укладання Браунського корпусу української мови (БрУК) [4].

Наші корпусні дослідження [38], [49] на матеріалі «Корпусу української мови» [16], створеного в лабораторії комп'ютерної лінгвістики КНУ ім. Тараса Шевченка під керівництвом Н.П. Дарчук, засвідчили неоціненне значення згаданого корпусу для розвитку корпусної лінгвістики в Україні й плідність застосування корпусних методів у вивченні української мови. Водночас з'явилося усвідомлення потреби мати хай і невеликий, однак збалансований, репрезентативний і докладно параметризований корпус [39], що був би цілком доступний у машиночитному форматі іншим користувачам. Такий корпус ми будуємо на підвалинах оригінального Браунського корпусу англійської мови з певною адаптацією до українських реалій [41].

В умовах падіння загального рівня текстів в Україні, публікації незредагованих результатів машинного російсько-українського перекладу, чим грішать навіть деякі великі видавництва й потужні ЗМІ, та з огляду на те, що серед мовців є прагнення до розвитку культури мовлення й орієнтації на добірну українську мову, наріжним каменем БрУКу ми поклали вимогу високої якості текстів. Натомість застосування суто дескриптивного підходу без жодного контролю якості й походження текстів може призвести до захаращення корпусу третьосортними текстами. Інші вимоги до фрагментів корпусу загалом відповідають принципам побудови первісного Браунського корпусу [48]: 1) твори мають бути оригінальні (неперекладні), зредаговані, прозові (не більш як 50% діалогічного мовлення у фрагменті); 2) створені й опубліковані за відносно короткий проміжок часу (у нашому випадку — 2010—2017 рр.); 3) до 2 тис. слів з одного твору (у вигляді одного або більше фрагментів). Весь обсяг текстів БрУКу (1 млн слововжитків) складається з 9 категорій у таких пропорціях: преса (25%), художня (25%), наукова (10%), науково-популярна (5%), навчальна (15%), професійно-популярна (7%), релігійна (3%) література, адміністративні документи (3%), інші інформаційні тексти (есеї, мемуари тощо, 7%). У межах кожної категорії забезпечуємо тематичне, жанрове, географічне й авторське розмаїття, щоб досягти збалансованості й репрезентативності корпусу. Корпус має бути пролематизований, проанотований і розомонімізований (знято лексичну й лексико-граматичну омонімію). Ці завдання виконуємо за допомогою описаних вище засобів, зокрема ключову роль відіграє ВЕСУМ, а «Правописник» допомагає контролювати якість текстів. Кожен фрагмент корпусу описано в стандартному переліку метаданих, до яких входять, наприклад, прізвище й ім'я автора, назва твору, місце й рік публікації тощо; в окрему зону виносимо помічені у фрагменті помилки. Наразі обсяг зібраних текстів БрУКу наближається до півмільйона слововживань. Кінцева мета —

створити корпус зі знятою омонімією, що перебуватиме у вільному доступі й стане одним із важливих чинників розвитку систем автоматичного опрацювання української мови.

Підсумовуючи, зазначимо, що проекти гурту r2u некомерційні, мають практичну спрямованість і динамічно розвиваються, сприяючи становленню й розвитку екосистеми української прикладної й комп'ютерної лінгвістики. Гурт r2u відкритий до різнобічної співпраці з зацікавленими фахівцями.

БІБЛІОГРАФІЯ

1. Активні ресурси сучасної української номінації : Ідеографічний словник нової лексики / [Є.А. Карпіловська, Л.П. Кислюк, Н.Ф. Клименко та ін.]; відп. ред. Є.А. Карпіловська. — К. : ТОВ «КММ», 2013. — 416 с.
2. *Антоненко-Давидович Б.* Як ми говоримо / Б. Антоненко-Давидович [Електронний ресурс], режим доступу: http://yak-my-hovorymo.wikidot.com/
3. *Боярова Л.Г.* Українська термінолексика в академічних російсько-українських словниках (20-і рр. XX ст. — початок XXI ст.) / Л.Г. Боярова // Вісник Харківського національного університету ім. В.Н. Каразіна : Серія «Філологія». — 2013. — № 1048. — Вип. 67. — С. 136—140.
4. Браунський український корпус [Електронний ресурс], режим доступу: https://github.com/brown-uk/corpus
5. Великий електронний словник української мови (ВЕСУМ) [Електронний ресурс], режим доступу: https://github.com/brown-uk/dict_uk
6. Великий тлумачний словник сучасної української мови (з дод. і допов.) / Уклад. і гол. ред. В.Т. Бусел. — К.; Ірпінь : ВТФ «Перун», 2005. — 1728 с.
7. *Вирган І.О.* Російсько-український словник сталих виразів / І.О. Вирган, М.М. Пилинська. — Харків : Прапор, 2000. — 864 с.
8. *Вихованець І.* Теоретична морфологія української мови : Академ. граматика укр. мови / І. Вихованець, К. Городенська; За ред. І. Вихованця. — К. : Унів. вид-во «Пульсари», 2004. — 400 с.
9. *Дарчук Н.П.* Комп'ютерна лінгвістика (автоматичне опрацювання тексту): підручник / Н.П. Дарчук. — К. : Вид-поліграф. центр «Київський університет», 2008. — 351 с.
10. *Карпіловська Є.А.* Вступ до прикладної лінгвістики: комп'ютерна лінгвістика: Підручник / Є.А. Карпіловська. — Донецьк : ТОВ «Юго-Восток, Лтд», 2006. — 188 с.
11. *Карпіловська Є.А.* Українська наукова мова в академічному «Російсько-українському словникові» за редакцією А. Кримського та С. Єфремова / Є.А. Карпіловська, О.Д. Кочерга, Є.В. Мейнарович // Вісник Національного університету «Львівська політехніка». Серія : Проблеми української термінології. — 2008. — № 620. — С. 110—15.
12. *Карпіловська Є.* Проблеми комп'ютерного моделювання мовної динаміки / Є. Карпіловська // Лінгвістичні студії: Зб. наук. праць. Вип. 17. — Донецьк: ДонНУ, 2008. — С. 293—297.
13. *Клименко Н.Ф.* Динамічні процеси в сучасному українському лексиконі / Н.Ф. Клименко, Є.А. Карпіловська, Л.П. Кислюк. — К. : Вид. дім Дмитра Бураго, 2008. — 336 с.
14. *Кобів Ю.* Словник українських наукових і народних назв судинних рослин / Ю. Кобів. — К. : Наук. думка, 2004. — 800 с.
15. *Коломієць Л.В.* Українсько-англійський словник лінгвістичної термінології / Л.В. Коломієць, О.Л. Паламарчук, Г.П. Стрельчук, М.В. Шевченко. — К. : Освіта України, 2013. — 455 с.
16. Корпус української мови [Електронний ресурс], режим доступу: http://www.mova.info/corpus.aspx
17. *Кочерга О.* Англійсько-українсько-англійський словник наукової мови (фізика та споріднені науки) / О. Кочерга, Є. Мейнарович. Частина І — англійсько-українська. — Вінниця : Нова Книга, 2010. — XXXIV + 1390 с.; Частина ІІ — українсько-англійська. — Вінниця : Нова Книга, 2010. — XXXIV + 1566 с.

18. *Критська В.І.* Граматичний словник української літературної мови. Словозміна: Близько 140 000 слів / В.І. Критська, Т.І. Недозим, Л.В. Орлова, Т.К. Пуздирєва, Ю.В. Романюк; Відп. ред. Н.Ф. Клименко. — К. : Вид. Дім Дмитра Бураго, 2011. — 760 с.
19. *Кубайчук В.* Хронологія мовних подій в Україні (Зовнішня історія української мови) / В. Кубайчук. — К. : К.І.С., 2004. — 168 с.; режим доступу: http://movahistory.org.ua
20. Лінгвістичний портал mova.info [Електронний ресурс], режим доступу: http://www.mova.info/
21. *Мандела Н.* Довгий шлях до свободи / Нельсон Мандела. Автобіографія. [Пер. з англ. В. Старка] — К. : Наш Формат, 2015. — 568 с.
22. *Мейнарович Є.* Англійсько-український словник з математики та кібернетики: біля 50 000 термінів / Є. Мейнарович, М. Кратко. — К.; Ірпінь : ВТФ «Перун», 2010. — 568 с.
23. Нові й актуалізовані слова та значення: словникові матеріали (2002—2010) / Кер. проекту і відп. ред. О.М. Тищенко. — К. : Вид. дім Дмитра Бураго, 2010. — 280 с.
24. *Опейда Й.* Глосарій термінів з хімії / Й. Опейда, О. Швайка. — Донецьк : Вебер, 2008. — 758 с.
25. *Перебийніс В.І.* Традиційна та комп'ютерна лексикографія / В.І. Перебийніс, В.М. Сорокін. — К. : Видавничий центр КНЛУ, 2009. — 218 с.
26. *Поздрань Ю.В.* «Російсько-український словник» за редакцією А.Ю. Кримського та С.О. Єфремова в історико-лінгвістичному контексті : дис. на здобуття наук. ступеня канд. філол. наук / Ю.В. Поздрань. — К., 2017. — 430 с.
27. Практичний російсько-український словник приказок / Упор. Г. Млодзинський. За ред. М. Йогансена. (Відтворення вид. 1929 р.) — К. : Ін-т енциклопедичних досліджень НАНУ, 2009. — 108 с.
28. Проекти групи lang-uk [Електронний ресурс], режим доступу: http://lang.org.ua/uk/
29. Російсько-український словник. — Т. І : А—Ж / Ред. В.М. Ганцов, Г.К. Голоскевич, М.М. Грінченкова; гол. ред. акад. А.Ю. Кримський. — К. : Червоний шлях, 1924. — XV+290 с.; Т. ІІ : З—Н, вип. 1. З—К / Ред. В. Ганцов, Г. Голоскевич, М. Грінченкова, М. Калинович, А. Ніковський, В. Ярошенко; гол. ред. акад. А. Кримський. — К. : ДВУ, 1929. — 392 с.; Т. ІІ : З—Н, вип. 2 : Л—*намыкивать* / Упоряд.-ред. М. Калинович і В. Ярошенко; гол. ред. акад. А. Кримський. — Харків : УРЕ, 1932. — С. 393—724; Т. ІІ : З—Н, вип. 3 : *Намыл—нять* / Упоряд.-ред. М. Калинович і В. Ярошенко; гол. ред. акад. А. Кримський. — Харків : УРЕ, 1933. — С. 725—1056; Т. ІІІ : О—П, вип. 1 : О—*поле* / Ред. В. Ганцов, Г. Голоскевич, М. Грінченкова, А. Ніковський; гол. ред. акад. С. Єфремов. — К. : ДВУ, 1927. — 336 с.; Т. ІІІ : О—П, вип. 2 : *Поле—пячение* / Ред. В. Ганцов, Г. Голоскевич, М. Грінченкова, А. Ніковський; гол. ред. акад. С. Єфремов. — К. : ДВУ, 1928. — С. 337—654.
30. Російсько-український словник: у 4-х т. — Т. 1. А—Ж / Ред. В. Ганцов, Г. Голоскевич, М. Грінченкова. Гол. ред. акад. А. Кримський. — К. : Вид. дім Дмитра Бураго, 2016. — 12, XIV + 290 с. (Репринт з вид. 1924 р.).
31. Російсько-український фразеологічний словник : Фразеологія ділової мови / Улож. В. Підмогильний, Є. Плужник. (Відтвор. вид. 1927р.). — К. : УКСП «Кобза», 1993. — 248 с.
32. Російсько-українські словники [Електронний ресурс], режим доступу: https://r2u.org.ua/main/dicts
33. *Руссу А.О.* Префіксальне дієслівне термінотворення в «Російсько-українському словнику» (1924—1933 рр.) : дис. на здобуття наук. ступеня канд. філол. наук / А.О. Руссу. — К., 2016. — 219 с.
34. *Скалозуб О.* Словник чужомовних слів, виразів і приповідок, що вживаються в українській мові / О. Скалозуб. — Коломия : Рекорд, 1933. — 476 с.
35. *Скопненко О.* Принципи лексикографічного опрацювання сталих висловів (на матеріалі «Російсько-українського словника» за редакцією А. Кримського й С. Єфремова та білоруських словників 20—30-х рр. ХХ ст.) / О. Скопненко // Лексикографічний бюлетень. — 2008. — Вип. 17. — С. 31—39.
36. Словники для звантаження [Електронний ресурс], режим доступу: http://r2u.org.ua/main/dicts_for_download

37. «Словники України» on-line / Український мовно-інформаційний фонд НАНУ [Електронний ресурс], режим доступу: http://lcorp.ulif.org.ua/dictua/
38. *Старко В.* Корпусні дані в дослідженні українських колоративів / В. Старко // Українська мова. — 2014. — Вип. 1. — С. 51—60.
39. *Старко В.* Параметризація корпусу як спосіб підвищити його репрезентативність та збалансованість / В. Старко, Н. Чейлитко // Українське мовознавство. — Вип. 43. — К., 2013. — С. 87—94.
40. *Старко В.* «Російсько-український словник» (1924—1933) та українське академічне словникарство / В. Старко // Науковий вісник Волинського національного університету імені Лесі Українки, 2008. — № 2. — С. 219—224.
41. *Старко В.* Формування Браунського корпусу української мови / В. Старко // Мовні і концептуальні картини світу. — 2014. — Вип. 48. — С. 415—421.
42. *Тищенко О.М.* Архівна картотека як лексико-ілюстративна база «Російсько-українського словника» за ред. А.Ю. Кримського та С.О. Єфремова. І. Лексична картотека: історія створення та репресій / О.М. Тищенко // Українська мова. — 2016. — № 2. — С. 44—71.
43. *Тищенко О.М.* Архівна картотека як лексико-ілюстративна база «Російсько-українського словника» за ред. А.Ю. Кримського та С.О. Єфремова. ІІ. Мікро- і макроструктура архівної картотеки / О.М. Тищенко // Українська мова. — 2016. — № 3 (59). — С. 57—78.
44. *Тищенко О.М.* Лексикографічний контекст четвертого тому Російсько-українського словника за ред. А.Ю. Кримського та С.О. Єфремова: умовне відтворення реєстру / О.М. Тищенко // Українська мова. — 2015. — № 4. — С. 89—100.
45. *Тищенко О.* Нові й актуалізовані слова та значення (словникові матеріали) у контексті сучасних неословників / О. Тищенко // Українська мова. — 2011. — № 1. — С. 55—68.
46. *Шерех Ю.* Нарис сучасної української літературної мови / Ю. Шерех. — Мюнхен : Молоде життя, 1951. — 404 с.
47. *Шимків А.* Англо-український тлумачний словник економічної лексики / А. Шимків. — К. : Вид. дім «Києво-Могилянська академія», 2004. — 429 с.
48. *Francis W.N.* Brown Corpus Manual [Електронний ресурс] / W.N. Francis, H. Kučera. — Providence, Rhode Island : Brown University, 1979; режим доступу: http://icame.uib.no/brown/bcm.html
49. *Starko V.* Ukrainian Colour Concepts for Blue [Електронний ресурс] / Vasyl' Starko // *Slovo*. Journal of Slavic Languages and Literatures. — 2013. — No. 54. — P. 150—163; режим доступу: http://www.moderna.uu.se/digital Assets/ 591/ c_ 591534-l_1-k_14_starko.pdf

Статтю отримано 02.10.2017

Vasyl Starko
Lesya Ukrainka Eastern European National University, Lutsk

COMPUTATIONAL LINGUISTICS PROJECTS OF THE R2U TEAM: PRESENT STATE AND APPLICATIONS

The article discusses a series of non-commercial, open-source projects developed for Ukrainian by the r2u team. The dictionary website r2u.org.ua provides full-text search for a collection of mostly Russian-Ukrainian and Ukrainian-Russian dictionaries. It brings back into circulation the dictionaries that were banned by the Soviet government as part of its policy to Russianize the Ukrainian language. The pearl of the collection is the academic Russian-Ukrainian Dictionary edited by Ahatanhel Krymsky and Serhii Yefremov (1924-1933). It was the last top-grade general dictionary published before the launch of the Russianization policy and remains an unparalleled source of proper Ukrainian lexis today.

The dictionary website e2u.org.ua offers a series of modern English-Ukrainian and Ukrainian-English dictionaries. The most prominent ones are the terminological dictionaries in physics and related sciences (by Olha Kocherga and Eugen Meinarovich, over 280,000 headwords), mathematics and informatics (by Eugen Meinarovich and Myroslav Kratko, 43,000 terms),

economics (by Anna Shymkiw, 20,000 headwords), and linguistics (by Lada Kolomiiets et al., 9,500 terms). A general English-Ukrainian dictionary is being compiled, one entry at a time, in response to search queries. A large phraseological dictionary is also being added to the website in a piecemeal fashion.

The Large Electronic Dictionary of Ukrainian (VESUM) is a machine-readable POS dictionary. With 316,000 lemmas and over four million generated word forms, it is the biggest of its kind for Ukrainian and has been adopted for full-text search in the Ukrainian-language Wikipedia.

Another project is the Pravopysnyk LanguageTool—an advanced Ukrainian spellchecker which checks also grammar and style (http://languagetool.org/uk/). Finally, the Brown Ukrainian Corpus (BrUK) is a project to build a one-million POS-tagged, lemmatized and disambiguated corpus of modern Ukrainian that can be used, inter alia, to train a Ukrainian POS tagger. All r2u projects are available online, and the corresponding links are provided in the article.

Keywords: *computational linguistics, natural language processing, computer lexicography, corpus linguistics, corpus, r2u, electronic dictionary, VESUM, Wikipedia, Pravopysnyk, LanguageTool, Brown Corpus, BrUK.*

Мовна мозаїка

***ПОЗАЯК* – ЦЕ ЯК?**

В останні десятиріччя помітне активне вживання слова *позаяк*. На нього натрапляємо в сучасній художній літературі, радіо- й телеефірі, пресі, наукових текстах і навіть у розмовному літературному мовленні, напр.: *Аж надто примітивний засіб до підслуховування, очевидно, не дав ніяких результатів,* ***позаяк*** *дідусь вийшов у коридор і став заглядати у шпарину* (М. Гримич); *Кен та Рейчел,* ***позаяк*** *вони пройшли на фестиваль раніше від нашого тріо, сиділи на іншій лаві двома рядами ближче до майданчика* (М. Кідрук); *...восплачте, пророки, згадайте нас на суді,* ***позаяк*** *ближчає літо Господнє!* (Є. Пашковський); *Серед пріоритетів Національної експертної комісії — удосконалення нормативно-правової бази,* ***позаяк*** *у чинному законі багато недоліків* (Віче, 2010, № 13); *Болларди на Андріївському узвозі не вмикають,* ***позаяк*** *місто не приймає їх у комунальну власність* (Телеканал «112 Україна», 18.04.2016); *За ієрархією форм вираження об'єктного значення центральне місце в синтаксичній системі української мови посідає знахідний відмінок, який є спеціалізованим виразником прямого об'єкта,* ***позаяк*** *реалізує всі типи об'єктних синтаксем...* (С. Різник); *...свобода означає вибір між добром і злом, що стоять перед людиною,* ***позаяк*** *без такого вибору немає і бути не може морального життя* (О. Киричук) та ін.

Однак деякі мовці не розуміють, що означає слово *позаяк*, або навіть хибно вважають його новотвором. Насправді *позаяк* — причиновий сполучник. Його синоніми — *тому що, оскільки, через те що, бо*. В українській літературній мові функціонує вже понад сто років. Наприкінці XIX – початку XX ст. *позаяк* досить широко використовували і в художній літературі, і в публіцистиці, і в епістолярному стилі, причому як загальновживане слово. Серед письменників та громадських діячів — І. Франко, В. Стефаник, М. Коцюбинський, В. Самійленко, М. Старицький, О. Пчілка, М. Драгоманов, М. Лисенко та ін. За радянського часу цей сполучник уживали обмежено, переважно в художній літературі як застаріле, жартівливе слово або й з виразною іронією чи взагалі для негативної характеристики персонажів. У перші двоє десятиріч XXI ст. *позаяк* знову повернувся до активного використання в ролі загальновживаного сполучного засобу.

Лариса Колібаба