УДК 81'33

Н. В. Нікуліна

ЗАГАЛЬНИЙ ОГЛЯД КОМП'ЮТЕРНО-МОВОЗНАВЧИХ НАУКОВИХ ПРОЄКТІВ УКРАЇНИ

У пропонованій статті проаналізовано мовознавчі наукові проєкти з використанням комп'ютерних технологій, зокрема онлайновий засіб проведення асоціативних експериментів «Стимулус», зацифрування унікальної архівної лексичної картотеки, цифрові проєкти й засоби для української мови від гурту r2u, Генеральний регіонально анотований корпус української мови, більш відомий під абревіатурою ГРАК.

Описано загальну інформацію про проєкти, проведено аналіз технічних особливостей і деталей сервісів, параметрів та структури інтерфейсів користувача, зібрано інформацію про долучення здобувачів різних рівнів освіти до опрацювання текстових масивів.

Ключові слова: *корпус української мови, лексична картотека, комп'ютерні технології.*

Nikulina N. Overview of Ukrainian Computational Linguistics Projects. *Researchers in the humanities are increasingly integrating computational technologies into their projects, thereby generating new knowledge in linguistic technology, computational and cognitive linguistics, lexicography, and corpus linguistics. This paper provides an analysis of linguistic research projects that use computational technologies, such as the online tool "Stimulus" for conducting associative experiments, the digitisation of a unique archival lexical card index, digital projects and tools for the Ukrainian language developed by the r2u group, and General Regionally Annotated Corpus of Ukrainian, commonly known by its acronym GRAC.*

The article overviews various projects, analysing the technical specifications and service details, including user interface parameters and structures. It discusses the "Stimulus" platform, developed for the interactive administration of experiments within web environments and smart devices. This tool is particularly suited for building a large-scale database of associative surveys, facilitating the study of language consciousness dynamics. The archival lexical database holds significant cultural value. The capability to process these lexical collections in a modern digital format is essential for both specialists and Ukrainian society. A significant achievement of the r2u project team is "Pravopysnyk Language Tool". It is a resource for checking spelling, grammar, and style, which allows users to review texts on diverse subjects and complexity levels. It also includes a list of over 3,000 lexical hybrids with suggested corrections.

The study gathers information on the involvement of students at various educational levels in processing text corpora. The participation of both students

https://doi.org/10.34142/23127546.2024.61.25

and university teachers in corpus-based research and in preparing materials for the corpus enhances teaching techniques of philological disciplines. It improves course content, promotes intellectual development among students, and engages them in experimental work.

Keywords: *Ukrainian language corpus, lexical card index, computational technologies.*

Вступ

Дослідники в царині гуманітарних наук все активніше використовують у своїх наукових проєктах комп'ютерні технології, чим створюють нові знання в галузі лінгвістичної технології, комп'ютерної та когнітивної лінгвістики, лексикографії, корпусної лінгвістики. Усе це свідчить про розвиток цифрової гуманітаристики, що створює нові методології гуманітарних досліджень, будує цифрову дослідницьку інфраструктуру, сприяє глобальній взаємодії вчених.

Переглянувши тематику мовознавчих досліджень з корпусної лінгвістики, можемо констатувати, що тематика науково-дослідницьких вишукувань побіжно й різнопланово була репрезентована в роботах українських учених. Вітчизняні студії з дослідження корпусної лінгвістики доволі широкоформатні за тематикою і репрезентовані дослідженнями таких науковців: Голощук С. Л., Демська-Кульчицька О. М., Карпіловська Є. А., Тищенко О. М., Фокін С. Б., Широков В.А. та ін.

Метою нашої наукової розвідки є прагнення розповісти про новітні корпусні технології, показати теоретичне й практичне значення корпусів української мови для проведення наукових досліджень в галузі лінгвістики, що сприяє інтелектуальному розвитку молоді, долучає до експериментів, а професорсько-викладацькому складу допомагає покращити зміст навчальних курсів і методику викладання.

Серед основних *завдань дослідження* є такі:

1) поінформувати про сучасні комп'ютерно-мовознавчі наукові проєкти, програмні засоби й інформаційні ресурси корпусної лінгвістики;

2) заналізувати технічні особливості й деталі сервісів, параметри й структуру інтерфейсів користувача;

3) ознайомити з практикою долучення студентів до укладання корпусів.

Методи дослідження

Під час дослідження послуговувалися такими загальнонауковими методами: аналіз і синтез, описовий метод, порівняння й зіставлення, моделювання, гіпотетико-дедуктивний, функціональний, що органічно поєднуються з галузевими методиками наукового вивчення й опрацювання теоретичного матеріалу в мовознавстві.

Виклад основного матеріалу

Корпус у лінгвістиці – це «структуроване зібрання машиночитаних текстів найрепрезентативнішого мовного матеріалу тієї або іншої природної мови, є передовсім загальним методом фіксації та дослідження мови, узагальненою моделлю організації та подання фактичного матеріалу, базованим на машинних технологіях» (Демська-Кульчицька, 2003: 40).

Наразі в Україні функціює низка цікавих комп'ютерно-мовознавчих наукових проєктів з корпусної лінгвістики, як от: онлайновий засіб проведення асоціативних експериментів «Стимулус», зацифрування унікальної архівної лексичної картотеки, цифрові проєкти й засоби для української мови від гурту r2u, Генеральний регіонально анотований корпус української мови (більш відомий під абревіатурою ГРАК).

ПРОЄКТ «СТИМУЛУС». Розробницею і координаторкою цього проєкту є *Загородня Ольга* (Національна академія статистики, обліку та аудиту), початок функціювання сервісу датується 2019 роком. Це зручний і легкий у використанні інструмент для проведення лінгвістичних (асоціативних) онлайн-досліджень, який надається авторами безкоштовно. Сервіс створено для інтерактивного проведення експериментів у вебсередовищі з допомогою мобільних пристроїв, зручний для формування масштабної бази асоціативних опитувань для вивчення динаміки показників мовної свідомості.

Деякі технічні особливості й деталі сервісу «Стимулус»:

1. Інтерфейс є полілінгвальним, простим і зрозумілим у використанні.

2. Користувачам не потрібно завантажувати або встановлювати спеціальні додаткові програми – сервіс доступний з будь-якого браузера.

3. «Стимулус» автоматично збирає введені відповіді, опрацювання результатів респондентів покладено на дослідника, а побудову

зведених таблиць і графіків з результатами дослідження сервіс виконує автоматично.

4. Переваги сервісу, порівняно з іншими маркетинговими програмами асоціативних ігор: науковість, багатофункційність, гнучкість у налаштуванні на широкий спектр завдань, візуалізація результатів дослідження (діаграми, графіки, гістограми, конотативні площини), місткість (СТИМУЛУС, 2019).

ПРОЄКТ «АРХІВНА КАРТОТЕКА». Про особливості й етапи зацифрування унікальної архівної лексичної картотеки *Оксана Тищенко* детально описує у своїх наукових працях, наголошуючи, що «..цінність АК для сучасної україністики стала передумовою переведення її матеріалів у цифровий формат, який дає можливість: а) зберегти АК; б) удоступнити її для широкого кола користувачів; в) активізувати науково-дослідчу й застосовну роботу з АК. Підготовчий етап зацифрування АК тривав чотири місяці 2018 р.: сотні волонтерів – від школярів до професорів – долучилися до всеукраїнської акції «Збереження Архівної лексичної картотеки» й опрацювали близько 6 млн. карток...» (Тищенко, 2020: 149).

Електронна система «Архівна картотека» дає користувачеві доступ до всіх типів інформації в картках, а інформацію карток у базі даних цієї системи структуровано на поля за такими параметрами: 1) заголовні одиниці; 2) додаткові одиниці; 3) характеристики одиниць (семантична, граматична, етимологічна); 4) цитата до слова чи стійкої словосполуки; 5) джерело (автентичне джерело, умовне позначення джерела чи автора); 6) тип картки за її вхідною мовою (російсько-українська, українська, українсько-російська; російська; неможливо визначити); 7) тип картки за часом створення; 8) підпис: оператор – автор введення даних картки; дата введення даних картки до електронної системи; 9) зауваги лексикографа і оператора.

Надбання архівної лексичної картотеки має загальнокультурне значення, бо можливість опрацьовувати ці словникові скарби у сучасній цифровій формі важливе як для фахівців, так і для українського суспільства загалом (Тищенко, 2020: 151).

ЦИФРОВІ ПРОЄКТИ ГУРТУ R2U. На особливу увагу заслуговують комп'ютерні лінгвістичні проєкти гурту r2u, діяльність якого очолив *Василь Старко.* Абревіатура-назва гурту, а далі й сайту ***r2u*** *(англ. Russian to Ukrainian),* тобто з російської на українську.

В основу роботи було покладено ідею «повернути українству заборонений до вживання, вилучений з обігу та з бібліотек, частково знищений, а частково замкнений у радянські спецхрани академічний «Російсько-український словник» за редакцією А. Кримського й С. Єфремова (1924–1933, далі РУС). 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 ядра цього гурту долучалося на різних етапах і в різних проєктах чимало осіб, яких тут годі перелічити... Онлайнова версія стала можливою завдяки гранту від Наукового товариства імені Шевченка у США з Фонду ім. Івана Романюка» (Старко, 2019).

Напрацюванням когорти сподвижників ***r2u*** є:

1) словниковий вебсайт r2u.org.ua (http://e2u.org.ua/). До бази цього сайту r2u внесено загалом 16 словників загальним обсягом 345 тис. словникових статей, із яких майже 200 тис. унікальні;

2) словниковий вебсайт e2u.org.ua (http://r2u.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;

3) Великий електронний словник української мови – ВЕСУМ (https://r2u.org.ua/vesum/) 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 проєктів 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);

4) засіб перевіряння орфографії, граматики й стилю «Правописник LanguageTool» (https://languagetool.org/uk/; https://r2u.org.ua/check). На цій платформі створено засоби перевіряння орфографії,

граматики й стилю для 28 мов. Гурт r2u розвиває український модуль під назвою «Правописник», що нині містить 500 правил. Модуль спирається на чинний правопис, в 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;

5) Браунський український корпус – БрУК (https://r2u.org.ua/corpus). Започатковано через усвідомлення потреби мати хоч і невеликий, однак збалансований, репрезентативний і докладно параметризований корпус, що був би цілком доступний у машиночитному форматі іншим користувачам. Такий корпус побудовано на підвалинах оригінального Браунського корпусу англійської мови з певною адаптацією до українських реалій (Старко, 2017).

ГРАК. Вагомі результати у сфері цифрової гуманітаристики, саме у галузі корпусної лінгвістики, репрезентує колектив **Генерального регіонально анотованого корпусу української мови [ГРАК]**, що розроблений спільною командою з лінгвістів і представників ІТ-сектору з України, Норвегії, Німеччини. Основні виконавці – *Марія Шведова, Рупрехт фон Вальденфельс, Василь Старко*; технічною реалізацією опікувалися *Сергій Яригін, Андрій Рисін, Тимофій Ніколаєнко, Арсеній Лукашевський, Кирило Захаров, Дмитро Чаплинський, Міхал Возняк, Михайло Крук*; наукові консультанти проєкту – *Дмитро Січінава, Михайло Назаренко*.

Дуже важливо, що до опрацювання текстових масивів долучаються студенти, магістранти, аспіранти вишів, а також слухачі онлайн-курсу «Корпусна лінгвістика» (Єнський університет за підтримки фонду DAAD), які у такий спосіб тільки розширюють обрії свої дослідницьких навичок.

У підготовці матеріалів Генерального регіонально анотованого корпусу української мови в різні роки брали участь здобувачі освіти 12 закладів вищої освіти України (ГРАК, 2017-2024): Національний університет «Львівська політехніка» (керівники: *Марія Шведова, Олена Левченко, Оксана Таран, Тетяна Шестакевич*); Львівський національний університет імені Івана Франка (керівник: *Соломія Бук*); Київський університет імені Бориса Грінченка (керівник: *Олена Доценко, Тетяна Горохова, Михайло Вінтонів, Ірина Саєвич*); Український католицький університет (керівник: *Василь Старко*); Національний

університет «Києво-Могилянська академія» (керівники: *Інна Ренчка, Людмила Дика, Людмила Підкуймуха, Анастасія Онатій, Наталя Кобченко*); Одеський національний університет імені І. І. Мечникова (керівники: *Марія Шведова, Наталя Кондратенко, Олександра Білінська, Анна Бордовська, Надія Хрустик*); Національний технічний університет «Харківський політехнічний інститут» (керівник: *Ніна Хайрова*); Інститут філології Київського національного університету імені Тараса Шевченка (керівник: *Тетяна Качановська*); Таврійський національний університет імені В. І. Вернадського (керівник: *Наталя Кущ, Людмила Юлдашева),* Херсонський національний технічний університет (керівники: *Наталя Кудрявцева, Євгенія Петренко*); Донецький національний університет імені Василя Стуса (наразі передислокований до Вінниці) (керівник: *Ірина Гарбера*); ННІ «Інститут державного управління» ХНУ імені В. Н. Каразіна (керівник: *Тетяна Ковальова*). До опрацювання матеріалу для ГРАКу долучилися аспірантки Інституту мовознавства АН України *Євгенія Петренко* і *Катерина Шкиль*, а також слухачі онлайн-курсу «Корпусна лінгвістика» (Єнський університет за підтримки фонду DAAD) (керівник: *Наталія Чейлитко*).

Варто зазначити, що «корпус охоплює період з 1816 по 2024 р. і містить понад 130 тисяч текстів різних жанрів, близько 30 тисяч авторів. Пошук слова чи виразу, побудова конкордансу в ньому можлива за такими критеріями: автор тексту, регіон, жанр тексту, місце публікації, роки публікації, мова оригіналу документа та ін. У корпусі можна здійснювати звичайний формальний пошук, пошук з «байдужими символами» (англ. – «wildcard characters»), а також запити CQL («Context Query Language», спеціальна проста мова запитів, розроблена саме для пошуку в розміченому текстовому корпусі)» (Фокін, 2018: 51).

Висновки

Отже, в Україні є багато сподвижників, які працюють в царині корпусної лінгвістики. Зазначимо, що зацифрування корпусів української лексики відіграють вагому роль й у лінгводидактиці: коли студенти опрацьовують корпуси слів української мови, то це ще й сприяє опануванню мови на належному рівні, тому в перспективі маємо на меті долучити студентів ХНАДУ до цієї роботи. Дослідницько-пошукова

співпраця викладацького складу і здобувачів різних рівнів освіти, як, наприклад, з ГРАКом, удосконалює технології гуманітарних досліджень і викладання, сприяє інтелектуальному розвитку молоді, долучає до експериментів, а викладачам допомагає покращити зміст навчальних курсів і методику викладання.

ЛІТЕРАТУРА

1. **ГРАК** (2017–2024) Генеральний регіонально анотований корпус української мови (М. Шведова, Р. Фон Вальденфельс, С. Яригін, А. Рисін, В. Старко, Т. Ніколаєнко та ін.) Київ, Львів, Єна, 2017–2024. uacorpus.org. 2. **Демська-Кульчицька,** О. М. (2003) Базові поняття корпусної лінгвістики. *Українська мова, 1,* 42–47. 3. **Сервіс СТИМУЛУС** (2019). Узято з https://stimulus.tools/uk. 4. **Старко,** В. Ф. (2017). Комп'ютерні лінгвістичні проекти гурту r2u: стан та застосування. *Українська мова, 3,* 86–100. Узято з https://r2u.org.ua/data/other/7_Starko.pdf. 5. **Старко,** В. Ф. (2019). Українська: дух і буква в цифрі. *Українська інтернет-газета «Збруч» від 21.02.2019.* Узято з https://zbruc.eu/node/87161. 6. **Тищенко,** О. М. (2020) Архівна картотека української мови в цифровому форматі: від пам'ятки мови до сучасного лексикографічного інструментарію. *Rocznik Slawistyczny, Wrocław, LXIX,* 185–197. 7. **Фокін,** С. Б. (2018). Корпуси текстів: здобутки України та перспективи врахування закордонного досвіду. *Літературознавство. Мовознавство. Фольклористика. Вісник КНУ. Серія: Іноземна філологія, 1 (28),* 51–54.

REFERENCES

1. **HRAK** (2017–2024) Heneralnyi rehionalno anotovanyi korpus ukrainskoi movy [General regionally annotated corpus of the Ukrainian language] (M. Shvedova, R. Fon Valdenfels, S. Yaryhin, A. Rysin, V. Starko, T. Nikolaienko ta in.) Kyiv, Lviv, Yena, 2017–2024. uacorpus.org [in Ukrainian]. 2. **Demska-Kulchytska**, O. M. (2003) Bazovi poniattia korpusnoi linhvistyky [Basic concepts of corpus linguistics]. *Ukrainska mova – Ukrainian language, 1,* 42–47 [in Ukrainian]. 3. **Servis STYMULUS** (2019) [Service STIMULUS]. Retrieved from https://stimulus.tools/uk [in Ukrainian]. 4. **Starko**, V. F. (2017). Kompiuterni linhvistychni proekty hurtu r2u: stan ta zastosuvannia [Computer-assisted linguistic projects of the r2u group: status and application]. *Ukrainska mova – Ukrainian language, 3,* 86–100. Retrieved from https://r2u.org.ua/data/other/7_Starko.pdf [in Ukrainian]. 5. **Starko**, V. F. (2019). Ukrainska: dukh i bukva v tsyfri [Ukrainian: spirit and letter in numbers]. *Ukrainska internet-hazeta «Zbruch» vid 21.02.2019.* Retrieved from https://zbruc.eu/node/87161 [in Ukrainian]. 6. **Tyshchenko**, O. M. (2020) Arkhivna kartoteka ukrainskoi movy v tsyfrovomu formati: vid pamiatky movy do suchasnoho leksykohrafichnoho instrumentariiu [Archival index of the Ukrainian language in digital format: from language monuments to modern lexicographic tools]. *Rocznik Slawistyczny, LXIX,* 185–197 [in Ukrainian]. 7. **Fokin**, S. B. (2018). Korpusy tekstiv: zdobutky Ukrainy ta perspektyvy vrakhuvannia zakordonnoho dosvidu [Corpora of texts: Ukraine's achievements and prospects for taking into account foreign experience]. *Literaturoznavstvo. Movoznavstvo. Folklorystyka. Visnyk KNU. Seriia: Inozemna filolohiia – Literary studies. Linguistics. Folklore studies. KNU Bulletin. Series: Foreign philology, 1 (28),* 51–54 [in Ukrainian].

Нікуліна Неля Василівна – кандидат філологічних наук, магістр педагогіки вищої школи, доцент, завідувач кафедри українознавства, Харківський національний автомобільно-дорожній університет; вул. Ярослава Мудрого, 25, м. Харків, 61002, Україна.

E-mail: nykulina@ukr.net

http://orcid.org/0000-0001-7832-7407

Nikulina Nelia – Candidate of Philological Sciences (Ph.D in Philology), Master in Pedagogy of Higher School, Associate Professor, Head of the Ukrainian Studies Department, Kharkiv National Automobile and Highway University; 25 Jaroslava Mudroho Str., Kharkiv, 61002, Ukraine.

Стаття надійшла до редакції 30 вересня 2024 року

ДСТУ 8302:2015: Нікуліна Н. В. Загальний огляд комп'ютерно-мовознавчих наукових проєктів України. *Лінгвістичні дослідження*: зб. наук. пр. Харк. нац. пед. ун-ту імені Г. С. Сковороди / гол. ред. Н. В. Піддубна. Харків, 2024. Вип. 61. С. 314–322. DOI: https://doi.org/10.34142/23127546.2024.61.25

APA: Нікуліна, Н. В. (2024). Загальний огляд комп'ютерно-мовознавчих наукових проєктів України. *Лінгвістичні дослідження, 61,* 314–322. DOI: https://doi.org/10.34142/23127546.2024.61.25