

Лахтаріна Наталія Василівна,

ORCID <http://orcid.org/0000-0001-5617-0373>,

молодша наукова співробітниця,

відділ наукового формування національних реферативних ресурсів,

Інститут інформаційних технологій,

Національна бібліотека України імені В. І. Вернадського,

Київ, Україна

e-mail: lakhtarina@nbuv.gov.ua

Лінгвістичні аспекти оптимізації РБД «Україніка наукова» як структурованого текстового корпусу

Закцентовано на пріоритетності РБД «Україніка наукова» у контексті сучасних інноваційних процесів наукової комунікації. Визначено лінгвістичні чинники якісного формування записів реферативної бази даних. Розглянуто напрями лінгвістичної оптимізації РБД «Україніка наукова» з урахуванням перспектив даного інформаційного ресурсу.

Ключові слова: наукова комунікація, реферативна база даних «Україніка наукова», текстовий корпус.

Стрімке поширення штучного інтелекту, розвиток цифрових технологій, автоматизоване опрацювання текстів спричиняють дослідження інформаційних ресурсів не лише як інструментів накопичення й пошуку даних, а й як структурованих текстових масивів для подальшого оброблення природної мови (Natural Language Processing, NLP).

У контексті цих інновацій реферативна база даних «Україніка наукова» становить значний інтерес, оскільки в інформаційному середовищі вирізняється чіткою структурованістю, достовірним науковим контентом, стандартизованою синтаксичною організацією. Реферативна база даних «Україніка наукова» є важливим національним інформаційним ресурсом, що акумулює бібліографічну

та реферативну інформацію про результати наукових досліджень, опублікованих в Україні. Значний обсяг структурованих текстових даних, представлений бібліографічними описами, рефератами, ключовими словами, предметними рубриками та класифікаційними індексами, дає змогу розглядати РБД як структурований текстовий корпус, придатний для лінгвістичного аналізу й удосконалення інформаційно-пошукових процесів. Практика редакційної підготовки записів свідчить, що на якість функціонування бази даних істотно впливають лінгвістичні чинники: варіантність наукової термінології, синонімія, неуніфікованість ключових слів, використання різних граматичних форм одного поняття, неоднорідність авторської термінології. Серед напрямів розвитку РБД є її лінгвістична оптимізація, що передбачає уніфікацію термінології, нормалізацію ключових слів, усунення лексичної та граматичної варіативності, стандартизацію синтаксичної організованості реферативних текстів. Одним із напрямів лінгвістичної оптимізації РБД є уніфікація термінології, що забезпечує однакове використання наукових термінів у реферативних текстах, метаданих, ключових словах та предметних рубриках, сприяючи підвищенню релевантності інформаційного пошуку.

Нормалізація ключових слів може включати: усунення орфографічної та граматичної варіативності, уніфікацію написання іншомовних термінів та аббревіатур, розмежування омонімів (наприклад, *гранат* (плід), *гранат* (мінерал)), усунення дублетів та надлишкових ключових слів, узгодження ключових слів із предметними рубриками. Це сприятиме однорідності лінгвістичного представлення документів та підвищить ефективність інформаційного пошуку. Особливого значення набуває і забезпечення орфографічної та граматичної нормативності ключових слів та термінів у текстах відповідно до чинної редакції Українського правопису (наприклад, унормування написання *вебсайт*, замість *веб-сайт* тощо). Серед лексичної варіативності поширені словосполучення: *цифрова бібліотека* – *електронна бібліотека*, *електронне урядування* – *е-урядування* – *цифрове урядування* тощо – теж зумовлюють необхідність їх уні-

фікації під час редакційного опрацювання відповідно до сучасних лексикологічних норм.

Лінгвістична структурованість РБД Україніка наукова виявляється також на синтаксичному рівні. Реферативні тексти характеризуються використанням типізованих безособових конструкцій. Зокрема, часто використовуються моделі: *проаналізовано результати дослідження; висвітлено проблеми; визначено чинники; запропоновано метод* тощо. Дані моделі забезпечують об'єктивність викладу, концентрують увагу на результаті, а не на виконавцеві дії, роблять реферати більш однорідними за структурою. Подальший розвиток синтаксичної організації реферативних текстів доцільно здійснювати шляхом стандартизації типових синтаксичних моделей, забезпечення єдності мовного оформлення та дотримання редакційних вимог до структури викладу наукової інформації. Для удосконалення структурно-композиційної організації рефератів важливими є такі чинники: чітке виділення об'єкта, мети, основних результатів і висновків; логічна послідовність викладу; усунення надлишкової інформації; використання типових мовних конструкцій, характерних для наукового стилю. Розгляд РБД «Україніка наукова» як структурованого текстового корпусу створює передумови для застосування методів корпусної лінгвістики та технологій автоматизованого опрацювання природної мови.

Аналіз частотності термінів, їхньої сполучуваності, варіантності й семантичних зв'язків дає можливість вдосконалювати редакційну обробку документів, розвивати лінгвістичне забезпечення бази даних та підвищувати ефективність інформаційного пошуку. Перспективним напрямом є використання результатів лінгвістичного аналізу для розвитку контрольованого словника РБД, автоматизованого формування ключових слів і забезпечення термінологічної єдності наукового інформаційного простору. Лінгвістичне забезпечення працює як міст між живою мовою та технічною системою. Тому стандартизація реферативних текстів, класифікаторів, ключових слів допоможуть людям і комп'ютерам розуміти одне одного. Це сприятиме подальшому розвитку РБД

«Україніка наукова» як структурованого текстового корпусу та складника цифрової дослідницької інфраструктури.

Nataliia Lakhtarina,

ORCID: <http://orcid.org/0000-0001-5617-0373>,

Junior Researcher,

V. I. Vernadskyi National Library of Ukraine,

Kyiv, Ukraine

e-mail: lakhtarina@nbuv.gov.ua

Linguistic aspects of optimization of the RDB "Ukrainika Naukova" as a structured text corpus

Attention is drawn to the priority of the RDB "Ukrainika Naukova" in the context of modern innovative processes of scientific communication. Linguistic factors of qualitative formation of abstract database records are determined. Directions of linguistic optimization of the RDB "Ukrainika Naukova" are considered, taking into account the prospects of this information resource.

Keywords: scientific communication, abstract database "Ukrainika Naukova", text corpus.