

РОЗДІЛ 5 СТРУКТУРНА, ПРИКЛАДНА ТА МАТЕМАТИЧНА ЛІНГВІСТИКА

УДК 81.26

DOI <https://doi.org/10.32782/tps2663-4880/2022.26.2.31>

МЕТОДОЛОГІЧНА БАЗА ДЛЯ ДОСЛІДЖЕНЬ ПИТАНЬ ЛІНГВІСТИЧНОЇ ІНДЕКСАЦІЇ

LINGUISTIC INDEXATION ISSUES: METHODOLOGICAL BASE

Власюк Л.С.,

orcid.org/0000-0003-1020-0076

*викладачка кафедри англійської мови технічного спрямування № 1,
аспірантка кафедри теорії, практики та перекладу англійської мови
Національного технічного університету України
«Київського політехнічного інституту ім. Ігоря Сікорського»*

Демиденко О.П.,

orcid.org/0000-0002-0643-5510

*кандидат педагогічних наук,
доцентка кафедри теорії практики та перекладу англійської мови
Національного технічного університету України
«Київського політехнічного інституту ім. Ігоря Сікорського»*

В статті розглянуто питання лінгвістичної індексації, зокрема її методологічної бази (в тому числі, аналіз лінгвістичної індексації за допомогою методу безпосередніх складників, дистрибутивного аналізу, корпусного аналізу, композиційного методу та фреймового аналізу).

Глобальна інформатизація всіх сфер діяльності людини зумовила необхідність як структуризації інформаційних екосистем текстів, так і вдосконалення наявних пошукових систем. В свою чергу, стрімке зростання вітчизняних і зарубіжних інформаційних ресурсів, а також потреба в підвищенні ефективності їх використання також вимагають осучаснення підходів до методології збирання, перероблення, зберігання й надання інформації користувачам. Оскільки найбільш використовуваним та найбільш ефективним методом інформаційного пошуку є пошук за ключовими словами, важливість вивчення явища лінгвістичної індексації зростає та набирає ще більшої актуальності. Лінгвістичну індексацію визначають як процес перетворення змістових та формальних ознак документів/текстів та інформаційних запитів засобами інформаційно-пошукової мови для забезпечення ефективного інформаційного пошуку. Щоб дослідити функціонування лінгвістичної індексації, необхідно, в першу чергу, дослідити її методологічну базу. Найбільш доцільним є аналіз лінгвістичної індексації в структурно-системній парадигмі. В статті було досліджено явище лінгвістичної індексації за допомогою методу безпосередніх складників для аналізу процесу кластеризації тексту, тобто поділ цілого тексту на окремі групи; дистрибутивного аналізу для вивчення різних типів індексів; корпусного аналізу, який дозволяє створити корпус текстів та полегшує визначення ключових слів; композиційного методу, який уможливорює вивчення п'яти етапів процесу індексації тексту; а також фреймового аналізу, що дозволяє побудувати модель індексації тексту.

Ключові слова: лінгвістична індексація, медіатекст, методи дослідження, структурна парадигма, індекс, кластеризація, дистрибуція.

The article deals with the issue of linguistic indexation, in particular, with its methodological base (including the research of linguistic indexation with the help of immediate constituent analysis, distributive analysis, corpus analysis, compositional method and frame analysis).

Global informatization of all fields of human activity has given the precondition to both necessity of structuring texts' information ecosystem and improving existing search engines. In its turn, rapid increase in the national and foreign informational resources as well as the necessity of enhancing the effectiveness of their use require modernized approaches to collecting, processing, storing and providing users with the information. Since of the most widely used and effective methods of information search is the one by keywords, the significance of linguistic indexation study is growing and its relevance is greater than ever. Linguistic indexation is defined as the process of converting content and formal characteristics of documents/texts and information requests through query language to secure an effective information search. To address the functioning of linguistic indexation it is needed to study its methodological base in the first place. The most appropriate way is to analyze linguistic indexation within the structural and systemic paradigm. The article represents the study of linguistic indexation with the help of immediate constituents analysis to research text's clusterization that is its division into separate groups; distribution analysis to study different types of indexes; corpus analysis which allows to create corpus of texts and facilitates the process of defining keywords; compositional method which enables the research of five stages of linguistic indexation process; and frame analysis which allows to create a model of text's indexation.

Key words: linguistic indexation, media text, research methods, structural paradigm, index, clusterization, distribution.

Постановка проблеми. Важливою рисою лінгвістичних досліджень ХХІ століття є аналітико-синтетична переробка інформації, яка полягає у змістовному аналізі тексту документа, що включає в себе вивчення, відбір, оцінку та узагальнення найбільш важливої інформації та подання цієї інформації користувачу. Невід'ємною складовою аналітико-синтетичної інформації є лінгвістична індексація тексту, в основі якої лежить ідентифікація змісту документа для його подальшого пошуку в інформаційній екосистемі текстів.

В останні десятиліття відбувається масштабна інформатизація всіх сфер діяльності, починаючи від науки і завершуючи культурою. Саме тому постала проблема вдосконалення інформаційно-пошукових систем, що передбачає виконання наступних кроків:

- здійснення інформаційного пошуку;
- формування та структурування баз даних.

Розв'язання цих завдань неможливе без належного індексування текстів. Саме тому дослідження поняття лінгвістичної індексації є вкрай актуальним питанням для сучасної лінгвістики.

Аналіз останніх досліджень і публікацій. Питаннями структурузації та впорядкування інформаційної екосистеми текстів, зокрема дослідженням лінгвістичної індексації, займалися Г.В. Власова, чий науковий доробок присвячений аналітико-синтетичній переробці інформації; Н.М. Кушнарєнко, головні праці якої стосуються питань наукової обробки інформації; І.Г. Лобановська, яка в багатьох своїх працях розкривала проблематику систематизації та організації документів в інформаційних інфраструктурах; також значний доробок у вивчення явища лінгвістичної індексації здійснив О.Л. Сухий, який досліджував алгоритми пошуку в інформаційних системах.

Постановка завдання. Ключове завдання статті полягає у аналізі лінгвістичних методів, які уможливають вивчення функціонування лінгвістичної індексації, що забезпечить максимальне уніфікування й досягнення одноманітності під час систематизування документів. В свою чергу, високоточний, повний та швидкий пошук інформації значно покращить якість наукового опрацювання документів та підвищить його ефективність.

Виклад основного матеріалу. Найбільш популярним видом інформаційного пошуку вважається пошук за ключовими словами, тобто лінгвістична індексація тексту.

Лінгвістична індексація – це процес визначення мовних особливостей, які безпосередньо стосуються контексту, в якому розгортається

висловлювання [7, с. 8]. Іншими словами, це присвоєння тексту певних ключових слів або кодів, які слугують вказівником змісту документа і використовуються для його пошуку.

Лінгвістична індексація забезпечує високу ефективність пошуку необхідних документів і даних за двома найбільш вагомими показниками «точність» та «повнота». Уможливлення швидкого та високоефективного пошуку необхідної інформації в екосистемі текстів є ключовою метою лінгвістичної індексації.

Ключовою парадигмою в межах якої буде проводитися дослідження є **структурна**, яка ґрунтується на системності мови, її рівневої ієрархії, наявності системних відношень на всіх мовних рівнях, дуалізму варіантів та інваріантів мовних одиниць [10, с. 28]. Вивчення явища лінгвістичної індексації медіатексту в структурно-системній парадигмі зумовлено метою дослідження: вивченням структурування інформаційної екосистеми текстів завдяки дослідженню особливостей пошуку за ключовими словами та визначенням ефективної моделі індексації медіатексту.

В ході здійснення дослідження для досягнення поставленої мети будуть використовуватися наступні методи: аналізу та синтезу, індукції та дедукції, зіставний метод, контекстологічний метод, корпусний метод, метод безпосередніх складників, дистрибутивний аналіз, фреймовий аналіз, композиційний. Розглянемо більш детально деякі із обраних методів:

1) *метод безпосередніх складників* являє собою сукупність процедур представлення синтаксичної структури словосполучення або речення у вигляді розкладення цілого на його складники, компоненти яких, в свою чергу, також можуть бути розкладені на компоненти до рівня кінцевих складників (*ultimate constituents*) [10, с. 106]. Результатом кожного кроку членування є виокремлення бінарної конструкції, компонентами якої є ядро (головний елемент) і маргінал (залежний елемент), що узгоджується із загальними правилами синтаксичної будови. Методика безпосередніх складників репрезентує речення шляхом виділення двох домінуючих вершин (ядерної структури): імені й дієслова, – які мають власні групи залежностей. Цей метод буде застосовуватися з метою аналізу явища **кластеризації тексту**, тобто поділу множини текстових документів на групи на основі схожості їх змісту [1, с. 456]. Для цього необхідно розкласти текст на окремі частини, іншими словами, виділити ядро та маргінал, що повністю відображає суть процесу кластеризації тексту: визначення

семантично поєднаних текстів в багатовимірному інформаційному просторі та центрів кластерів, які являють собою тематичні рубрики. До прикладу, розглянемо невеличкий уривок тексту:

The World Health Organization (WHO) has warned that the Omicron coronavirus variant poses a high risk of infection surges around the globe. The variant could lead to severe consequences in some regions, the WHO said on Monday. The variant was detected in South Africa earlier this month with initial evidence suggesting it has a higher re-infection risk. South Africa has been praised for

its prompt reporting of the variant. "Omicron has an unprecedented number of spike mutations that are like poison, some of which are concerning for their potential impact on the trajectory of the pandemic as well as staggering index of potential risk of re-catching the virus", the WHO said. Speaking on Monday, Dr Tedros said scientists around the world are working to discover if the new variant is associated with higher transmission, risk of reinfection and how it reacts to vaccines.

Поділимо на головні та залежні частини перше речення:



Таким чином, шляхом поступового розби-вання вищезазначеного уривку тексту на ядра, від яких відходять маргінали, можна визначити ключові слова (рис. 1).

Наш аналіз тексту за допомогою методу безпосередніх складників показав, що текст має два ядра 'pandemic' та 'consequences', від кожного з яких відходять три залежні елементи, які, власне, і становитимуть ключові слова для обраного уривку тексту.

2) **дистрибутивний аналіз**. Дистрибуція – це сума всіх оточень, у яких трапляється той чи інший елемент мови (фонема, морфема, слово), тобто сума всіх можливих позицій елемента стосовно інших елементів цього ж рівня, його сполучуваність [10, с. 61]. Цей метод буде застосовуватися для аналізу **різних видів індексів**, що є ключовим поняттям, коли мова йде про явище лінгвістичної індексації. Зокрема, дистрибутивний аналіз буде використано для дослідження **N-грамів (N-grams)** [2, с. 57]. До прикладу, розглянемо вище вказаний уривок тексту – дистрибуцію слова «ризик» у ньому:

- 1) **Adj. + risk + of + N** (*high risk of infection*);
- 2) **Adj. + N + risk** (*higher reinfection risk*);
- 3) **Risk + of + N** (*risk of reinfection*);
- 4) **Adj. + risk + of + V + N** (*potential risk of re-catching the virus*).

Доцільним також є використання цього методу разом із корпусним аналізом.

3) для вивчення інших двох видів індексів буде застосовано **корпусний аналіз**:

- **інвертований індекс** (структура даних, в якій для кожного слова колекції текстів у відповідному списку зазначено всі тексти, які вміщують це слово) [6, с. 47];

- **матриця «документ-слово»** (матриця, яка описує частоту вживання слів у колекції текстів) [4, с. 73].

Для цього в створеному корпусі медіатекстів буде визначено ключові слова, згруповані за частотою їх вживання, а також здійснено пошук текстів, де вживаються визначені ключові слова.

4) **композиційний метод** (увага фокусується на композиції: виділення частин, семантичних центрів, аналіз епізодів; структурний взаємо-

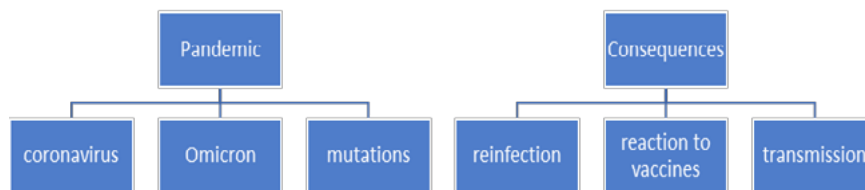


Рис. 1. Виокремлення ключових слів

зв'язок між різними структурними елементами тексту: елемент, система, відношення, ієрархія, опозиція, модель тощо). Метод композиційного аналізу буде застосовано для дослідження **процесу лінгвістичної індексації тексту**, який складається із 5 етапів [8, с. 41]:

- модуль збору текстів та виділення мета інформації з цих текстів;
- модуль збереження індексних структур текстів;
- модуль формування та надання користувачу переліку текстів, згрупованих за їхньою відповідністю запиту користувача;
- модуль формування маркерів, які вміщують мета інформацію;
- модуль лінгвістичного аналізу тексту та формування відповідних маркерів.

5) **фреймовий аналіз** для визначення **ефективної моделі індексації медіатексту**, що передбачає структурування понятійної концептуальної інформації [3, с. 67].

Візьмемо, до прикладу, концепт '*Virus*' з попереднього уривку тексту; структура концепту формується на основі 5 фреймів, виділених С.А. Жаботинською: акціонального, посесивного, предметного, таксономічного й компаративного [5, с. 85]:

1) **Таксономічний фрейм** подає відносини категоризації, які виявляються у двох варіантах: (1) [ХТОСЬ рід є ХТОСЬ вид]; (2) [ХТОСЬ рід є ХТОСЬ роль]: у нашому прикладі фрейм дозволяє встановити вид вірусу: *Omicron virus*, *Delta virus*;

2) **Предметний фрейм** – характеризується за якісними, буттєвими, локативними, темпоральними й кількісними параметрами: *dangerous virus*, *mutated virus*;

3) **Акціональний фрейм** – актанти предметно-референтної ситуації наділені аргументними ролями і поєднані зв'язками, визначеними дією актанта-агенса а й позначеними дієсловом діє/робить, що супроводжується прийменниками: діє за допомогою (інструмента чи помічника); діє на (пацієнс/об'єкт); діє в напрямку (реципієнта); діє заради (мети чи причини); діє задля

(результату чи бенефіціанта): в нашому випадку Найбільш частотними діями референта-агенса *Virus* є наступні: *surges around the globe, leads to severe consequences*;

4) **Посесивний фрейм**: демонструє зв'язок [ХТОСЬ власник має ЩОСЬ предмет власності]. Відношення посесивності специфіковані в трьох субфреймах: (1) власник має власність; (2) ціле має частину; (3) контейнер має вміст. В нашому прикладі реалізується перша пропозиція: *, has an unprecedented number of spike mutations*;

5) **Компаративний фрейм** – виникає на підґрунті таксономічного, сформований зв'язками тотожності, подібності й подоби: (1) тотожність: [ХТОСЬ є ХТОСЬ корелят]; (2) подібність: [ХТОСЬ є як ХТОСЬ корелят]; (3) подоба: [ХТОСЬ є немовби ХТОСЬ корелят]. У вказаному прикладі зв'язок формується на основі подібності: *mutations that are like poison*.

Висновки. Організація та систематизація інформаційної інфраструктури є основною умовою здійснення ефективного пошуку необхідних документів/текстів/матеріалів тощо. Швидкий, повний та високоефективний пошук інформації уможливорюється правильно визначеними та підібраними ключовими словами, тобто лінгвістичною індексацією.

Лінгвістична індексація є одним з видів аналітико-синтетичного опрацювання документів, яке полягає у переведенні змісту документів з природної мови на інформаційно-пошукову, фактично створюючи пошуковий образ документа і запиту.

Вивчення лінгвістичної індексації неможливе без дослідження її методологічної бази. Належний аналіз лінгвістичної індексації забезпечується використанням наступних лінгвістичних методів: дистрибутивний аналіз, метод безпосередніх складників, корпусний аналіз, композиційний метод та фреймовий аналіз, що зумовлено дослідженням лінгвістичної індексації в структурно-системній парадигмі, оскільки її головною метою є структуризація інформаційної екосистеми текстів завдяки пошуку за ключовими словами.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ:

1. Eckert Penelope. Variation and the indexical field. *Journal of Sociolinguistics*. 2008. № 12/4. P. 453–476.
2. Giorgi Alessandra. About the Speaker: Towards a Syntax of Indexicality. New York : Oxford University Press, 2010. 229 p.
3. Власова Г.В. Аналітико-синтетична переробка інформації: навч. посібник. Київ : ДАККІМ, 2006. 290 с.
4. Дарчук Н. П. Комп'ютерне анотування українського тексту: результати і перспективи. Київ : Освіта України, 2013. 544 с.
5. Жаботинська С.А. Концептуальний аналіз мови: фреймові мережі. Мова. *Науково-теоретичний часопис із мовознавства*. 2004. № 9. С. 81–92.
6. Кушнарченко Н.М. Наукова обробка документів: підручник. Київ : Знання, 2006. 334 с.

7. Лобановська І.Г. Індексування документів ключовими словами. Київ : Нілан-ЛТД, 2011. 32 с.
8. Лобановська І.Г. Систематизація документів та організація каталогів та картотек освітянських бібліотек за таблицями Універсальної десятикової класифікації. Київ : Ліра-К, 2019. 105 с.
9. Селіванова О.О. Сучасна лінгвістика: напрями та проблеми. Полтава : Довкілля-К, 2008. 711 с.
10. Сухий О.Л., Міленін В.М., Тарадайнік В.М. Алгоритми пошуку в інформаційних системах. Київ, 2005. 70 с.