Національний технічний університет України
«Київський політехнічний інститут ім. Ігоря Сікорського»
Міністерство освіти і науки України

Національний технічний університет України
«Київський політехнічний інститут ім. Ігоря Сікорського»
Міністерство освіти і науки України

Кваліфікаційна наукова
праця на правах рукопису

Сергеєв Данило Сергійович

УДК 004.93

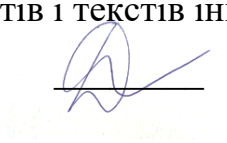
ДИСЕРТАЦІЯ
ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ОБРОБКИ ПРИРОДНОМОВНИХ ТЕКСТІВ НА ОСНОВІ ІНТЕГРАЦІЙНОГО ПІДХОДУ

05.13.06 – інформаційні технології

технічні науки

Подається на здобуття наукового ступеня кандидата технічних наук

Дисертація містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело.



Сергеєв Д.С.

Науковий керівник: Кисленко Юрій Іванович, к.т.н., доцент

Київ – 2019

АНОТАЦІЯ

Сергеєв Д.С. Інформаційна технологія обробки природномовних текстів на основі інтеграційного підходу. – Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 05.13.06 «Інформаційні технології». – Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського». – Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, 2019.

За останні роки дослідження у галузі ОПМ досягли значних практичних результатів, зокрема природномовний голосовий інтерфейс користувача для мобільних пристроїв, суттєвий прогрес у технологіях машинного перекладу, розпізнавання рукописного тексту та голосу тощо. При цьому актуальною залишається задача покращення якості роботи цих систем. Так, дослідження показують, що складніші прикладні технології ОПМ, зокрема машинний переклад та природномовний пошук, показують гірші результати ніж окремі технології нижчих рівнів, і є можливим їх удосконалення на основі існуючого стеку технологій. Ключовим елементом такого удосконалення постають бази знань, які одночасно взаємодіють з прикладними технологіями ОПМ різних рівнів, зокрема – природномовні бази знань.

Основним напрямком сучасних досліджень в галузі ПМБЗ є розробка гібридних ПМБЗ, що поєднують різні моделі ПМБЗ, наприклад мережеву модель з частково рекурсивними елементами. Втім, такі системи часто наслідують не лише переваги, але й недоліки систем, на яких вони засновані. Крім того, розробка гібридної ПМБЗ є складною задачею, і така система часто обмежена умовами прикладної задачі, для вирішення якої вона створюються – а саме, може містити принципові недоліки, які не впливають на вирішення даної задачі, але є суттєвими для ПМБЗ взагалі.

Перевагами існуючих підходів є їх висока ефективність у вирішенні відповідних спеціалізованих задач, як-то розпізнавання символів або статистичного аналізу текстів. Водночас, суттєвими їх недоліком є відсутність системної взаємодії між технологіями різних рівнів, що призводить до недостатньо ефективної роботи комплексних технологій повного циклу обробки природної мови, зокрема систем природномовного пошуку та машинного перекладу.

Наукове завдання дослідження полягає у розробці інформаційної технології обробки природномовних текстів на основі інтеграційного підходу з метою підвищення ефективності роботи технологій обробки природної мови.

Метою дисертаційної роботи є підвищення ефективності обробки природномовної інформації за рахунок інформаційної технології обробки природномовних текстів на основі інтеграційного підходу.

Об'єкт дослідження – процес обробки природномовної інформації.

Предмет дослідження – моделі, методи, алгоритми та інформаційні технології обробки природномовної інформації.

У вступі обґрунтована актуальність теми дисертаційної роботи, сформована мета, ідея і задачі дослідження, наукова новизна й практичне значення отриманих результатів, наведені наукові положення, що виносяться на захист, наукове та практичне значення роботи, дані про публікації, апробацію та впровадження розробок і результатів дослідження.

У *першому* розділі виконано аналіз задач та процесів комп'ютерної обробки природної мови, проаналізовано прикладні аспекти використання технологій обробки природної мови в інформаційних технологіях. Визначено роль баз знань в інформаційних технологіях обробки природної мови як компонента, необхідного для взаємодії різних їх систем, охарактеризовано та проаналізовано існуючи підходи до проектування природномовних баз знань.

Показано, що актуальною є задача розробки інформаційної технології обробки природномовних текстів на основі інтеграційного підходу.

У другому розділі визначено особливості розробки природномовної бази знань для інформаційної технології обробки природномовних текстів на основі інтеграційного підходу. З урахуванням цих особливостей створено формальну модель представлення знань у природномовній базі знань та розроблено моделі її основних елементів, якими є квант знань, або найменший елемент знань, та відношення, яке описує зв'язок між квантами знань. З використанням моделі представлення знань створено метод обробки природномовних текстів для інформаційної технології та процедури використання інформаційної технології у прикладних задачах обробки природної мови.

У третьому розділі розроблено структурну схему інформаційної технології обробки природномовних текстів на основі інтеграційного підходу, виконано аналіз процесів обробки даних в інформаційній технології, зокрема визначено етапи роботи інформаційної технології в режимах аналізу та синтезу та розроблено процедури записування та пошуку природномовних знань у базі знань у складі інформаційної технології. Також наведено приклади використання інформаційної технології обробки природномовних текстів у прикладних задачах природномовного пошуку та машинного перекладу.

У четвертому розділі визначено технічні вимоги до інформаційної системи, яка реалізує інформаційну технологію обробки природномовних знань на основі інтеграційного підходу, зокрема визначено підсистеми та операції і розроблено схему бази даних для інформаційної системи. Проаналізовано обчислювальну складність пошуку природномовних знань у інформаційній системі та проведено її порівняння з аналогами. Виконано експериментальну перевірку використання інформаційної системи для підвищення релевантності результатів природномовного пошуку та виконано аналіз отриманих даних.

У ході виконання роботи одержано нові наукові та практичні результати.

Вперше запропоновано модель представлення природномовних знань на основі інтеграційного підходу до моделювання мовленнєвої діяльності людини, в основі якої лежить формалізоване представлення кванту знань у вигляді фрагменту

довільного природномовного тексту, що дозволяє зберігати структуру тексту у вигляді однотипних структур даних.

Вперше розроблено метод опрацювання природномовних текстів з використанням запропонованої моделі представлення природномовних знань, який виділяє з тексту складові окремих квантів знань, що дозволяє виділити структуру знань довільного природномовного тексту.

Вперше розроблено інформаційну технологію обробки природномовних текстів на основі використання запропонованої моделі представлення знань та методу опрацювання природномовних текстів, що дозволяє виконувати повнотекстовий природномовний пошук за логарифмічний обчислювальний час.

Розроблений в рамках даної роботи метод обробки природномовних текстів може бути використаний для покращення роботи технологій обробки природної мови, зокрема систем природномовного пошуку в мережі Інтернет, систем машинного перекладу, природномовних інтерфейсів користувача.

Створена в рамках даної роботи модель представлення природномовних знань може бути використана для розробки універсальних природномовних баз знань.

Окремі результати розробленої в рамках даного дослідження інформаційної технології було впроваджено в робочий процес ТОВ «Діджитал принт» та ТОВ «Міжнародна текстильна корпорація». Розроблені в рамках даного дослідження моделі і методи впроваджено в матеріалах курсів «Теорія алгоритмів» та «Візуальне програмування» у навчальний процес кафедри технічної кібернетики ФІОТ КПІ ім. Ігоря Сікорського.

В рамках роботи поставлені та вирішені наступні науково-технічні задачі.

На основі аналізу підходів до розробки природномовних баз знань обґрунтовано потребу у розробці універсальної моделі представлення знань природномовного тексту для використання в технологіях обробки природної мови, що поєднує переваги існуючих підходів, та процедури використання такої моделі в інформаційних технологіях обробки природномовних текстів.

На основі інтеграційного підходу до моделювання мовленнєвої діяльності людини розроблено модель представлення знань для використання в технологіях обробки природної мови. Ця модель відрізняється від аналогів тим, що дозволяє представити фрагмент знань довільного природномовного тексту у вигляді універсальної структури. Перевагою розробленої моделі представлення знань є її незалежність від синтаксичної структури тексту та семантичного контексту фрагменту знань.

Розроблено процедури записування та пошуку знань з використанням розробленої моделі представлення знань в технологіях обробки природної мови, які дозволяють встановити зв'язки на структурному рівні між синтаксичною структурою тексту та довільною структурою метаданих.

Розроблено інформаційну технологію обробки природномовних текстів на основі інтеграційного підходу, для якої теоретично показано, що складність пошуку не перевищує так для аналогів, і в середньому є на 5-12% меншою для складних пошукових запитів.

Експериментально показано, що використання природномовної бази знань на основі інтеграційного підходу для природномовного пошуку дозволяє покращити якість роботи систем природномовного пошуку, а саме підвищити середню релевантність результатів на 14%. Впровадження результатів роботи у виробництві призвело до зменшення витрат часу працівників на контроль за складом продукції на 25% та збільшення конверсії природномовного пошуку на 8% відповідно.

Дисертаційна робота виконувалась у відповідності з планами науково-дослідницької роботи кафедри технічної кібернетики Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського» і виконувалась в рамках НДР «Проектування лінгвістичного процесора та бази знань як складових індивідуальної мовної системи для формування цілого кластеру інформаційних природномовних технологій» (номер держреєстрації 0117U004327), в якій автору належить формалізована модель природномовних знань та структура моделі світу на її основі.

Основні положення та результати дисертаційної роботи викладено в 12 друкованих працях, у тому числі: 6 статей у наукових фахових виданнях, з них 1 у міжнародному науковому виданні, що входить до наукометричної бази Scopus, 6 тез доповідей в збірниках матеріалів міжнародних конференцій.

Ключові слова: інформаційна технологія, природна мова, обробка природної мови, інтеграційний підхід, база знань, квант знань, пошук, машинний переклад.

ANNOTATION

Serheiev D.S. Information technology for processing of natural-language texts based on the integrational approach. – Qualifying scientific work published as a manuscript.

In the recent years, the research in the field of natural language processing (NLP) has achieved significant practical results, including natural-language voice user interface for mobile devices, significant progress in machine translation technologies, handwriting and voice recognition, etc. At the same time, the task of improving the performance of these systems remains relevant. Research shows that more complex applied NLP technologies, in particular machine translation and natural-language search, are less efficient than individual lower-level technologies, and it is possible to improve them based on existing technological solutions. The key element of such improvement appears to be knowledge bases, which simultaneously interact with applied NLP technologies of different levels, in particular – natural language knowledge bases (NLKB).

The ongoing research in the field of NLKB is mostly focused on the development of hybrid NLKB that combine different models of NLKB, such as network model of knowledge with partially recursive elements. However, such systems often inherit not only the advantages but also the drawbacks of the systems they are based on. In addition, development of hybrid NLKB is a complex task, and systems of this type are often limited by specific conditions of planned application – namely, they may contain fundamental defects that do not affect the solution of a given problem, but are critical for the NKLB model in general.

The scientific goal of this study is to develop information technology for processing natural language texts based on the integrational approach in order to increase the efficiency of natural language processing technologies.

The object of the research is the process of natural language processing.

The subject of the research is models, methods, algorithms and information technologies for natural language processing.

The introduction establishes relevance of the topic of the dissertation; the purpose, idea and tasks of the research; scientific novelty and practical significance of the obtained results; main scientific results of the study; scientific and practical value of the work; as well as data on publications, testing and introduction of developments and research results.

The first section focuses on analyzing the tasks and processes of computer processing of a natural language. Applied aspects of use of technologies of natural language processing in information technologies are analyzed. The role of knowledge bases in information technologies of natural language processing is determined as a necessary component for the interaction of different systems. Existing approaches to developing natural-language knowledge bases are characterized and analyzed.

The second section deals with specifics of developing natural-language knowledge bases for the information technology in question. A formal model of knowledge representation for a natural-language knowledge base is created, including models of its main elements, namely the quantum of knowledge, or the smallest element of knowledge, and the relation objects that describe connections between quanta of knowledge. A method for processing natural language texts based on this model of knowledge is developed, including the procedure for using information technology in applied natural language processing problems.

The third section describes the structural scheme of information technology for processing natural language texts based on the integrational approach. Data processing operations in information technology are analyzed. Special attention is given to stages of work of the information technology in analysis and synthesis modes. Procedures for recording and searching natural language knowledge are developed, using the knowledge base included in the information technology. Examples are given of use of the information technology for processing natural language texts in practical problems, namely natural language search and machine translation.

The fourth section defines technical requirements for an information system that implements information technology for processing natural language knowledge based on

the integrational approach. Subsystems and operations of such system are defined, and database scheme is developed. Computational complexity of natural language knowledge search in the information system is analyzed and compared with the existing alternatives. Experimental testing of the information system is conducted and the acquired data are analyzed, demonstrating increased relevance of search results of natural language search.

During the course of the work, new scientific and practical results are obtained.

For the first time, a model of representation of natural language knowledge has been created on the basis of the integrational approach to modeling of speech activity of a person, based on the formalized representation of a quantum of knowledge in the form of a fragment of arbitrary natural language text, which allows to store structure of the text in the form of uniform data structures.

For the first time, a method for processing natural language texts using the proposed model of presentation of natural language knowledge has been developed, allowing to acquire components of individual quanta of knowledge from the text and to allocate structure of knowledge of any natural language text.

For the first time, an information technology for processing natural language texts based on the proposed model of knowledge representation and method of processing natural language texts has been developed, allowing to perform full-text natural language search in logarithmic computational time.

The method of natural language text processing developed in this work can be used to improve natural language processing technologies, in particular, systems of natural language search in the Internet, machine translation systems and natural language user interface solutions.

The model of presentation of natural-language knowledge created within the framework of this work can be used to develop universal natural-language knowledge bases.

The information technology developed within this study was partly incorporated into the workflow of LLC "Digital Print" and LLC "International Textile Corporation". The models and methods developed in the framework of this research have been

incorporated in the materials of the courses "Theory of algorithms" and "Visual programming" in the educational process of the Department of Technical Cybernetics, FIOT, Igor Sikorsky Kyiv Polytechnic Institute.

Within the framework of the work, the following scientific and technical problems are set and solved.

Based on the analysis of approaches to development of natural-language knowledge bases, the need to develop a universal model of presentation of knowledge of natural language text for use in natural language processing technologies is established, requiring combination of advantages of existing approaches, and the procedures for using such a model in information technology for processing natural language texts.

On the basis of the integrational approach to the modeling of speech activity of a person, a model of presentation of knowledge for use in natural language processing technologies is developed. The model differs from the existing alternatives as it allows to represent a fragment of knowledge of an arbitrary natural-language text in the form of a universal structure. The advantage of this knowledge representation model is its independence from both syntactic structure of the text and semantic context of the knowledge fragment.

Writing and searching methods are created based on the knowledge representation model, allowing to establish links at the structural level between syntactic structure of the text and arbitrary structure of the metadata in natural language processing technologies.

An information technology for processing natural language texts based on the integrational approach is developed, for which it is theoretically proven that the search complexity does not exceed that of the existing alternatives, and is on average 5-12% lower for complex search queries.

The use of the natural-language knowledge base on the basis of the integrational approach for the natural-language search has been experimentally proven to improve the performance of natural-language search systems, namely to raise average relevance of search results by 14%. Implementation of results of work in production has led to a

reduction in employee time costs on warehouse management by 25% and an increase in the conversion of natural-language search by 8%, respectively.

The dissertation work has been carried out in accordance with the plans of research of the Department of Technical Cybernetics of National Technical University of Ukraine " Igor Sikorsky Kyiv Polytechnic Institute" and within the framework of the research project "Designing of a linguistic processor and a knowledge base as individual components of individual speech system to form a cluster of natural language information technology", registration number 0117U004327, within which the author has created formalized model of knowledge of natural language text and the structure of world model based thereof.

The main provisions and results of the dissertation work are described in 12 publications, including: 6 articles in specialized scientific journals, 1 of them in an international scientific journal included in the scientific base of Scopus, 6 theses in materials of international conferences.

Keywords: information technology, natural language, natural language processing, integrational approach, knowledge base, quantum of knowledge, search, machine translation.

СПИСОК ПУБЛІКАЦІЙ

У виданнях іноземних держав:

1. Kyslenko Y, Sergeiev D. Cognitive architecture of speech activity and modelling thereof. Biologically Inspired Cognitive Architectures. 2015;12.
(автору належить: модель бази знань, функціональна модель інтерфейсу).

У наукових фахових виданнях України:

2. Кисленко ЮІ, Сергеєв ДС. Порівняння способів збереження слів в ІТ. Адаптивні системи автоматичного управління. 2016;(28(1)):33–41.
(автору належить: формалізація проблеми, формалізація використаного підходу, збір та опрацювання даних)
3. Кисленко ЮІ, Сергеєв ДС. Структурний підхід до пошуку природно-мовної інформації. Радіоелектроніка та інформатика. 2015;(3):45–9.
(автору належить: методика аналізу та опрацювання даних)
4. Сергеєв ДС. A model of relation object for the natural language knowledge base [Модель об’єкту відношення для природно-мовної бази знань]. Адаптивні системи автоматичного управління. 2017;(30(1)):106–13.
5. Сергеєв ДС, Хіміч, А.В. Визначення категорії «знання» та її використання в інформаційних природно-мовних технологіях. Адаптивні системи автоматичного управління. 2016;(29(2)):140–6.
(автору належить: формалізація та обгрунтування використаної моделі)
6. Сергеєв ДС. Комп’ютерне моделювання когнітивного аспекту обробки природної мови на основі природно-мовної бази знань. Штучний інтелект. 2016;(4):42–8.

Матеріали конференцій:

7. Сергеєв ДС. Особливості моделювання бази природно-мовних знань. В: Електроніка та інформаційні технології ЕлІТ-2015. Львів: Львів. нац. ун-т ім. І. Франка, ф-т електроніки; 2015. с. 17–20.
8. Сергеєв ДС. Методика оцінки якості роботи природно-мовних пошукових систем. In: SAIT-2016 [Internet]. Київ: НТУУ «КПІ», ІПСА; 2016. p. 413–6. Режим доступу: http://sait.kpi.ua/media/filer_public/73/32/7332a68e-e93b-4c57-a3c8-66f11ee074cd/sait2016ebook.pdf
9. Сергеєв ДС. Оптимізація використання природно-мовних баз знань шляхом тематичної декомпозиції. В: ЕлІТ-2016. Львів: Львів. нац. ун-т ім. І. Франка, ф-т електроніки; 2016. с. 25–8.
10. Сергеєв ДС. Виділення концептів у природно-мовному тексті як спосіб наповнення бази знань. В: SAIT-2017. Київ: НТУУ «КПІ», ІПСА; 2017. с. 321–3.
11. Сергеєв ДС. Природно-мовна база знань як основа моделювання окремих аспектів мовленнєвої діяльності людини. В: Системи та засоби штучного інтелекту AIIS'2017. Київ: КНУ ім. Т.Шевченка, ф-т комп. наук та кібернетики; 2017. с. 171–8.
12. Сергеєв ДС. Integration model for knowledge representation for semantic WEB [Інтеграційна модель представлення знань для семантичного WEB]. В: ICSFTI2018. Київ: КПІ ім.Ігоря Сікорського, ф-т інформатики та обчислювальної техніки; 2018. с. 284–7.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

БД – база даних

БЗ – база знань

БССС – базова семантико-синтаксична структура

ЗОС – знання-орієнтовані системи

ІМС – індивідуальна мовна система

ІТ – інформаційна технологія

ІП – інтеграційний підхід

КЗ – квант знань

ЛП – лінгвістичний процесор

МПЗ – модель представлення знань

ОПМ – обробка природної мови

ПМБЗ – природномовна база знань

ПМІ – природномовна інформація

API – application programming interface

ERM – entity-relationship model

ВСТУП

Обґрунтування вибору теми дослідження.

Обробка природної мови (ОПМ) – загальний напрямок інформатики, штучного інтелекту та математичної лінгвістики, який вивчає проблеми комп'ютерного аналізу та синтезу природної мови. Прикладні задачі ОПМ включають в себе багато підгалузей, зокрема виділення інформації, природномовний пошук, машинний переклад, ідентифікація плагіату та багато інших.

За останні роки дослідження у галузі ОПМ досягли значних практичних результатів, зокрема природномовний голосовий інтерфейс користувача для мобільних пристроїв, суттєвий прогрес у технологіях машинного перекладу [1], розпізнавання рукописного тексту та голосу тощо. При цьому актуальною залишається задача покращення якості роботи цих систем. Так, дослідження показують, що складніші прикладні технології ОПМ, зокрема машинний переклад та природномовний пошук, показують гірші результати ніж окремі технології нижчих рівнів, і є можливим їх удосконалення на основі існуючого стеку технологій. Ключовим елементом такого удосконалення постають бази знань, які одночасно взаємодіють з прикладними технологіями ОПМ різних рівнів, зокрема – природномовні бази знань.

Основи досліджень в цій галузі заклали такі вчені як E.D. Liddy, J.F. Sowa, R. Harris, K. Brown, R. Tadeusiewicz, N. Chomsky, D. Herrmann, N. Chapman, J. Lyons, О.В. Бармак, М.З. Згуровський, О.А. Кришталь, Д.В. Ланде, В.А. Лефевр, В.М. Томашевський, В.А. Широков.

Природномовні бази знань (ПМБЗ) – це клас баз знань, об’єктом роботи яких є природномовна інформація (ПМІ) – тобто, знання в яких зберігається безпосередньо у вигляді ПМІ, на відміну від інших класів баз знань, що використовують спеціалізовані формальні моделі представлення знань у базах знань. Основним напрямком сучасних досліджень в галузі ПМБЗ є розробка

гібридних ПМБЗ, що поєднують різні моделі ПМБЗ, наприклад мережеву модель з частково рекурсивними елементами. Втім, такі системи часто наслідують не лише переваги, але й недоліки систем, на яких вони засновані. Крім того, розробка гібридної ПМБЗ є складною задачею, і така система часто обмежена умовами прикладної задачі, для вирішення якої вона створюються – а саме, може містити принципові недоліки, які не впливають на вирішення даної задачі, але є суттєвими для ПМБЗ взагалі.

Значний внесок в розвиток цих ідей зробили M. Weigt, C. Baker, B. Cronin, R. Brachman, G. Antoniou, C. Fillmore, J.M. Hellerstein, M. Blanton, A. Pable, V.E. Wolfengagen, А. Левицький, В.В. Бочкаров, А.В. Анісімов, П.І. Федорчук, О.В. Іванов та інші. Розроблений на основі робіт цих вчених інтеграційний підхід дозволяє вирішити деякі з відомих проблем комп'ютерної лінгвістики, а розроблені на його засадах природномовні бази знань надають нові можливості для їх використання в технологіях обробки природної мови.

Перевагами існуючих підходів є їх висока ефективність у вирішенні відповідних спеціалізованих задач, як-то розпізнавання символів або статистичного аналізу текстів. Водночас, суттєвими їх недоліком є відсутність системної взаємодії між технологіями різних рівнів, що призводить до недостатньо ефективної роботи комплексних технологій повного циклу обробки природної мови, зокрема систем природномовного пошуку та машинного перекладу.

Наукове завдання дослідження полягає у розробці інформаційної технології обробки природномовних текстів на основі інтеграційного підходу з метою підвищення ефективності роботи технологій обробки природної мови.

Зв'язок роботи з науковими програмами, планами, темами. Дисертаційна робота виконувалась у відповідності з планами науково-дослідницької роботи кафедри технічної кібернетики Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського» в рамках НДР «Проектування лінгвістичного процесора та бази знань як складових індивідуальної мовної системи для формування цілого кластеру інформаційних

природномовних технологій» (номер держреєстрації 0117U004327), в якій автору належить формалізована модель природномовних знань та структура моделі світу на її основі.

Мета і задачі дослідження. Метою дисертаційної роботи є підвищення ефективності обробки природномовної інформації за рахунок використання гнучкої моделі представлення природномовних знань шляхом розробки інформаційної технології обробки природномовних текстів на основі інтеграційного підходу.

Об'єкт дослідження – процес обробки природномовної інформації.

Предмет дослідження – моделі, методи, алгоритми та інформаційні технології обробки природномовної інформації.

Відповідно до мети були поставлені і виконані наступні завдання:

1) проаналізовано існуючі технології обробки природної мови та виконано аналіз особливостей використання природномовних баз знань у цих технологіях;
2) розроблено модель представлення природномовних знань на основі інтеграційного підходу, що враховує особливості природної мови як об'єкту роботи бази знань;
3) розроблено метод виділення знань з природномовного тексту та пошуку знань з урахуванням розробленої моделі представлення знань;
4) розроблено інформаційну технологію обробки природномовних текстів на основі моделі представлення знань та методу їх обробки;
5) виконано експериментальну перевірку розробленої інформаційної технології.

Методи дослідження. У даній роботі були використані такі методи: методи розробки баз даних та баз знань; метод ієрархічної декомпозиції; методи системного аналізу; аксіоматичний метод у частині вихідних теоретичних положень; метод аналітичної оцінки обчислювальної складності алгоритму; методи моделювання та експерименту для перевірки створеної інформаційної технології.

Наукова новизна одержаних результатів.

1. Вперше запропоновано модель представлення природномовних знань на основі інтеграційного підходу до моделювання мовленнєвої діяльності людини, в основі якої лежить формалізоване представлення кванту знань у вигляді фрагменту довільного природномовного тексту, що дозволяє зберігати структуру тексту у вигляді однотипних структур даних.
2. Вперше розроблено метод опрацювання природномовних текстів з використанням запропонованої моделі представлення природномовних знань, який виділяє з тексту складові окремих квантів знань, що дозволяє виділити структуру знань довільного природномовного тексту.
3. Вперше розроблено інформаційну технологію обробки природномовних текстів на основі використання запропонованої моделі представлення знань та методу опрацювання природномовних текстів, що дозволяє виконувати повнотекстовий природномовний пошук за логарифмічний обчислювальний час.

Практичне значення одержаних результатів.

1. Розроблений в рамках даної роботи метод обробки природномовних текстів може бути використаний для покращення роботи технологій обробки природної мови, зокрема систем природномовного пошуку в мережі Інтернет, систем машинного перекладу, природномовних інтерфейсів користувача.
2. Створена в рамках даної роботи модель представлення природномовних знань може бути використана для розробки універсальних природномовних баз знань.
3. Окремі результати розробленої в рамках даного дослідження інформаційної технології було впроваджено в робочий процес ТОВ «Діджитал принт» та ТОВ «Міжнародна текстильна корпорація».
4. Розроблені в рамках даного дослідження моделі і методи впроваджено в матеріалах курсів «Теорія алгоритмів» та «Візуальне програмування» у

навчальний процес кафедри технічної кібернетики ФІОТ КПІ ім. Ігоря Сікорського.

Особистий внесок здобувача. Усі результати, що складають основний зміст дисертаційної роботи, отримані автором самостійно.

У спільних роботах автору належить:

Модель бази знань, функціональна модель інтерфейсу у *[Cognitive architecture of speech activity and modelling thereof, BICA, 2015]*; методика аналізу та опрацювання даних у *[Структурний підхід до пошуку природно-мовної інформації, Радіоелектроніка та інформатика. 2015]*; формалізація проблеми, формалізація використаного підходу, збір та опрацювання даних у *[Порівняння способів збереження слів в ІТ, АСАУ, 2016]*; формалізація та обґрунтування використаної моделі у *[Визначення категорії «знання» та її використання в інформаційних природно-мовних технологіях, АСАУ, 2016]*.

Апробація результатів дисертації. Основні результати роботи доповідалися та обговорювалися на конференціях.

Міжнародній конференції «Електроніка та інформаційні технології / ЕлІТ-2015» з темою доповіді «Особливості моделювання бази природно-мовних знань» (Львів-Чинадієво, 27-30 серп. 2015 р.);

Міжнародній конференції «System Analysis and Information Technology / SAIT-2016» з темою доповіді «Методика оцінки якості природно-мовних пошукових систем»;

Міжнародній конференції «Електроніка та інформаційні технології / ЕлІТ-2016» з темою доповіді «Оптимізація використання природно-мовних баз знань шляхом тематичної декомпозиції» (Львів-Чинадієво, 27-30 серп. 2016 р.);

Міжнародній конференції «Штучний інтелект та інтелектуальні системи AIIS’2016» з темою доповіді «Комп’ютерне моделювання когнітивного аспекту обробки природної мови на основі природно-мовної бази знань» (Київ, 29.11-2.12.2016 р.)

Міжнародній конференції «System Analysis and Information Technology / SAIT-2017» з темою доповіді «Виділення концептів у природно-мовному тексті як спосіб наповнення бази знань» (Київ, 22.05-25.05.2017 р.).

Публікації. За результатами досліджень опубліковано 12 наукових праць, у тому числі 6 статей у наукових фахових виданнях (з них 1 стаття у виданнях іноземних держав, 4 у наукових фахових виданнях України, які входять до міжнародних наукометричних баз), 6 тез доповідей в збірниках матеріалів конференцій.

Структура та обсяг дисертації. Дисертація складається зі вступу, чотирьох розділів, списку використаних джерел (132 найменування) та 6 додатків. Загальний обсяг роботи складає 165 сторінок. Основна частина дисертації займає 118 сторінок, містить 28 рисунків та 12 таблиць.

1 АНАЛІЗ ПРОЦЕСІВ ТА ТЕХНОЛОГІЙ ОБРОБКИ ПРИРОДНОМОВНИХ ТЕКСТІВ

У цьому розділі виконано аналіз задач та процесів комп’ютерної обробки природної мови, проаналізовано прикладні аспекти використання технологій обробки природної мови в інформаційних технологіях. Визначено роль баз знань в інформаційних технологіях обробки природної мови як компонента, необхідного для взаємодії різних їх систем, охарактеризовано та проаналізовано існуючи підходи до проектування природномовних баз знань. Показано, що актуальною є задача розробки інформаційної технології обробки природномовних текстів на основі інтеграційного підходу.

1.1 Аналіз проблематики комп’ютерної обробки природної мови

З часів появи перших формальних систем математики та логіки людство прагнуло автоматизувати вирішення усіх відомих йому проблем, звести складний процес мислення до виконання типових кроків для досягнення конкретного результату. Вершиною цього підходу стала концепція комп’ютера як інструмента, обчислювальна потужність та обсяг пам’яті якого дозволяють надзвичайно ефективно виконувати обчислення, складність яких є практично недосяжною для виконання людиною.

Кількість користувачів комп’ютерів збільшилась з 1% у 1995 році до понад 48% у 2016 [2], що призвело до зміни ролі комп’ютера від вузькоспеціального засобу для вирішення обмеженої кількості задач до універсального допоміжного інструменту, що став де-факто стандартом як у сфері наукових досліджень, так і в прикладному виробництві та навіть у повсякденному житті. Одним з напрямків розробки, що поєднує інформаційні технології та незалежну від них галузь досліджень, є обробка природної мови.

Обробка природної мови (ОПМ) [*Natural Language Processing / NLP*] – загальний напрямок інформатики, штучного інтелекту та математичної лінгвістики, який вивчає проблеми комп'ютерного аналізу та синтезу природної мови [3]. Прикладні задачі ОПМ включають в себе багато підгалузей, зокрема виділення інформації, природномовний пошук, машинний переклад, ідентифікація плагіату та багато інших [4].

За останні роки дослідження у галузі ОПМ досягли значних практичних результатів, зокрема природномовний голосовий інтерфейс користувача для мобільних пристроїв [5], суттєвий прогрес у технологіях статистичного машинного перекладу [6], розпізнавання рукописного тексту [7] та голосу [8] тощо. При цьому актуальною залишається задача покращення якості роботи цих систем. Так, дослідження показують, що складніші прикладні технології ОПМ, зокрема машинний переклад [1] та природномовний пошук [9], показують гірші результати ніж окремі технології нижчих рівнів, і є можливим їх удосконалення на основі існуючого стеку технологій.

Ключовими проблемами ОПМ є задачі синтезу [10] природномовного тексту на основі нетекстової інформації та його аналізу, тобто розуміння [11]. Ці задачі є протилежними за порядком обробки, але при їх вирішенні використовуються одні й ті ж самі інструменти на різних рівнях обробки даних.

До технологій ОПМ входять усі технології, об’єктом роботи яких є ПМ: обробка тексту та голосу, синтаксичний та семантичний аналіз, прагматичний аналіз та обробка знань [12]. Разом ці технології формують повний цикл обробки природної мови, етапи якого показані на рис.1.1. Усі ці технології тісно пов’язані і використовуються при синтезі та аналізі ПМ даних, змінюється при цьому лише порядок їх використання.

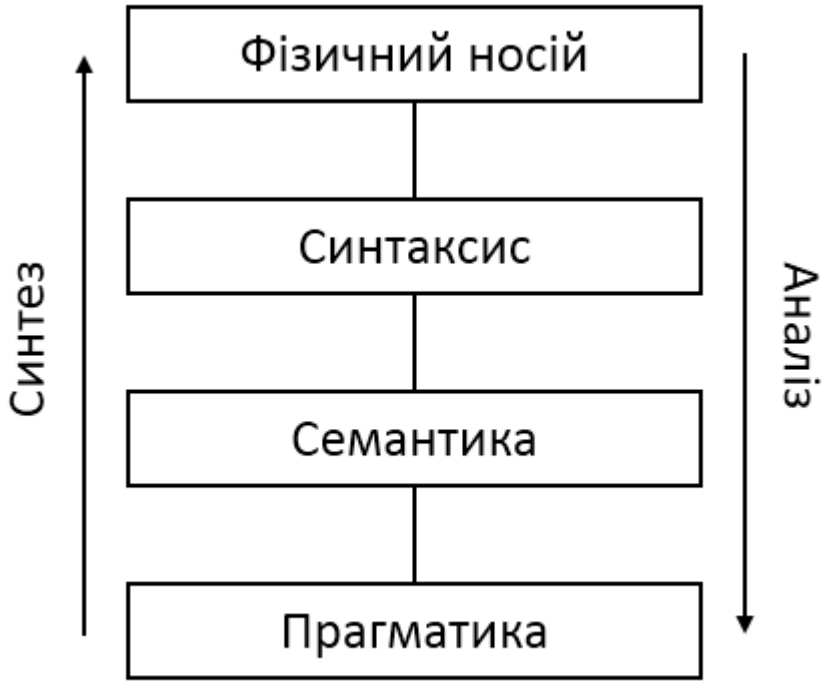


Рисунок 1.1 – Етапи обробки природної мови

Вирішення будь-якої практичної задачі ОПМ вимагає вирішення підзадач на одному або кількох з цих рівнів. Проаналізуємо практичні задачі, для вирішення яких використовуються технології ОПМ, та оцінимо загальний стан цієї галузі.

1.1.1 Аналіз технологій синтаксичного аналізу

Задача найнижчого рівня ОПМ – ідентифікація тексту серед не-текстових даних, тобто обробка тексту на рівні фізичного носія. Вона включає в себе задачі оптичного розпізнавання символів *(optical character recognition, OCR)* [13] та розпізнавання мовлення *(speech recognition, SR)* [14]. Задачі розпізнавання мовлення, символів та слів є самодостатніми і відносяться до допоміжних задач ОПМ, оскільки їх вирішення зводиться до порівняння вхідних даних (звуків, зображень тощо) з обмеженим масивом еталонів (словники, бібліотеки звуків і т.д.) і не потребує більш глибокого аналізу ПМ як об’єкту дослідження.

Оптичне розпізнавання тексту – це механічне або електронне переведення зображень рукописного, машинописного або друкованого тексту в послідовність кодів, що використовуються для представлення тексту в комп'ютері. Задачею *OCR* є ідентифікація текстової інформації на зображенні (фотокопії документу, фото довільної сцени, кадрі з відео тощо), розпізнавання окремих символів в знайденому тексті та розпізнавання слів, що їх формують ці символи.

Критеріями якості роботи систем *OCR* є точність ідентифікації та розпізнавання окремого символу та окремого слова. Задачу *OCR* у її класичному розумінні можемо вважати вирішеною – точність розпізнавання відсканованих документів, що містять друкований текст або текст, набраний у машинний спосіб, сягає 99% для символів та 97% для слів [7]. Такі показники свідчать про вирішення усіх принципових проблем та готовність технології *OCR* для практичного використання.

Сучасний напрям досліджень *OCR* включає розпізнавання тексту з довільної сцени, розширення набору підтримуваних мов, розпізнавання курсивного тексту та розпізнавання рукописного тексту. Станом на 2013 рік точність розпізнавання рукописних текстів становила 70%, а вже у 2014 році – до 84% [15,16]. Зауважимо, що проблеми обробки складних текстових матеріалів вимагають залучення в більшій мірі загальних алгоритмів та методів обробки зображень, ніж безпосередньо розпізнавання символів та слів.

Розпізнавання мовлення (*SR*) - це процес перетворення мовленнєвого сигналу в текстовий потік. На відміну від *OCR*, де одиницями розпізнавання є символ та слово, у *SR* в якості таких одиниць виступають окреме слово та їх сукупність. Критеріями оцінки, відповідно, є точність розпізнавання окремого слова та точність розпізнавання слів у звуковому потоці.

Точність розпізнавання мовлення станом на сьогоднішній день становить 94.2% для слів та фраз зі спеціально сформованого корпусу та 88.7% для розпізнавання довільної прямої мови. Хоча ці показники об'єктивно нижчі за

відповідні показники *OCR*, вони вже досягли точності розпізнавання мовлення людиною – 94.1% та 89.0% відповідно [17].

Використання слів для покращення якості розпізнавання їх елементів є характерним не лише для систем *SR,* але й для систем *OCR* останнього покоління. Так, *Tesseract*, одна з найпотужніших сучасних систем *OCR*, використовує словники для підвищення точності розпізнавання одного символу на основі його оточення – слова [18]. Зауважимо, що для задач цього класу словник виступає первинним джерелом даних, процес створення та наповнення якого не має впливу на роботу системи.

Наступний рівень практичних задач, які вирішує ОПМ – це аналіз тексту на рівні зв’язків між його окремими елементами. До цих задач належать зокрема аналіз запитів користувача для розпізнавання складних голосових команд, в тому числі голосовий інтерфейс користувача [5], системи управління комп’ютером для людей з обмеженими можливостями (системи текст – голос [19] та голос – текст [20]), системи для виділення та аналізу складних концептів у тексті, в тому числі опрацювання емоційного навантаження тексту [21], ідентифікація спеціалізованої термінології [22], визначення абревіатур [23] та власних імен тощо.

Задачі ОПМ, які відповідають цим задачам, належать до лексичного та синтаксичного аналізу.

Синтаксичний аналіз (парсинг, *parsing*) – це процес аналізу вхідної послідовності символів з метою розбору граматичної структури згідно із заданою формальною граматикою [24]. Виділяють дві основні задачі синтаксичного аналізу – аналіз окремих слів і їх граматичних особливостей та аналіз синтаксичної структури фрагменту тексту [25].

Задача синтаксичного аналізу окремих слів відома як стемінг *(stemming)* – визначення морфологічного кореня слова [26] та, більш широко, як лематизація *(lemmatization)* – визначення частини мови, до якої належить дане слово, та його граматичної форми [27]. Цю задачу можемо вважати вирішеною достатньо добре для практичного використання. Точність алгоритмів та прикладних систем

лематизації сягає 98% і більше правильно розпізнаних слів [28,29]. Зазначимо, що лематизація має не лише наукове, але й велике практичне значення – лематизація є однією з основних операцій обробки природномовних пошукових запитів [30,31], використовується в інтерфейсах сучасних мобільних пристроїв, таких як мобільні клавіатури *T9* та *MultiTap* [32], є важливою операцією для попередньої обробки текстів у базах знань [33] тощо.

Втім, висока точність розпізнавання слів у сучасних системах є можливою лише за умови використання стохастичних методів для вибору одного з кількох коректних варіантів інтерпретації тексту та використанню додаткових алгоритмів уточнення результатів його обробки [34], оскільки окреме слово без урахування його оточення в тексті часто не містить достатньої інформації для однозначної його ідентифікації. Більш того, найскладніші проблеми лематизації, такі як розпізнавання омонімів, неологізмів, жаргонних слів, абревіатур тощо, не можуть бути вирішені на матеріалі окремого слова принципово.

Наступним етапом обробки після ідентифікації та розпізнавання окремих символів та слів є синтаксичний аналіз тексту, метою якого є отримання певної інформації про структуру фрагменту тексту на рівні зв’язків між словами. Найчастіше результат синтаксичного аналізу являє собою об’єкт з чітко визначеною структуру – граф, дерево, таблицю тощо. Ця задача добре висвітлена як у наукових дослідженнях, так і у відомих прикладних програмних продуктах.

Так, різні аспекти синтаксичного аналізу реалізують системи перевірки орфографії (в тому числі у продуктах *Google*, *Microsoft*, *OpenOffice* [35], *Grammarly* [36]), спеціалізовані системи синтаксичного аналізу тексту (*Stanford Parser* [37–39]), (*LALR Parsers* [40,41]) та окремих його компонентів [42] і ціла низка наукових та експериментальних продуктів [43,44].

На рис.1.2. наведено приклад синтаксичного розбору фрагменту тексту *«The white cat sat under the chair»* системою *Stanford Parser,* де для кожного фрагменту речення ідентифіковано його роль у структурі тексту *(S, NP, VP, PP),* для кожного слова – його частина мови (*det, adjective, noun* і т.д.) та його місце у цій структурі.

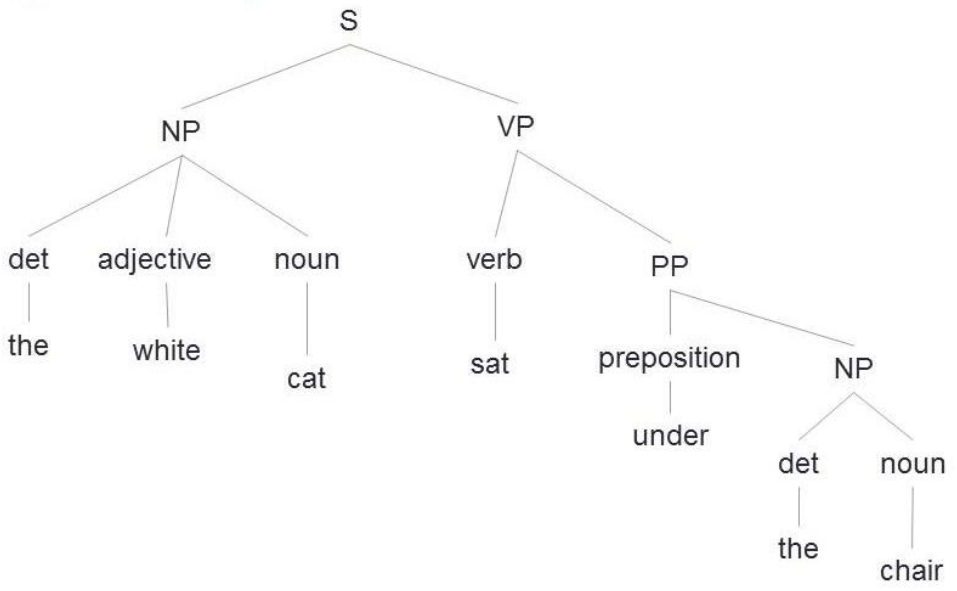


Рисунок 1.2 – Приклад синтаксичного розбору фрагменту тексту, *Stanford Parser* [45]

Оскільки структура природномовного тексту представляє собою складний об’єкт дослідження, оцінка якості роботи синтаксичних аналізаторів також є складною задачею, адже, за умови відсутності чітких правил, які могли б бути використаними для автоматичної обробки тексту, еталонний зразок формується в ручному режимі. Точність роботи систем синтаксичного аналізу тексту загалом нижча, ніж точність розпізнавання окремих символів чи слів – 85-90% [46].

Отже, ефективність вирішення задач лексичного та синтаксичного аналізу в ОПМ та відповідних їх практичних задач є високою. Втім, лексичний та синтаксичний аналіз дозволяють в кращому випадку визначити лише усі можливі варіанти розпізнавання даного слова або фрагменту тексту. При цьому визначити структуру, правильну з точки зору оператора, можливо лише враховуючи семантичне навантаження тексту [47].

1.1.2 Аналіз технологій семантичного аналізу

Задачі найвищого рівня ОПМ – ті задачі, які вимагають розуміння тексту не лише на рівні визначення можливих синтаксичних зв’язки між окремими його

елементами, але й дослідження смислового навантаження тексту та виділення з нього знань. Сюди входить широкий спектр практичних задач, кожна з яких може бути також представлена як безпосередня складова ПМ інтерфейсу користувача. Це, в першу чергу, природномовний пошук, тобто автоматичний пошук, при якому запит користувача представлений у вигляді ПМ повідомлення [48]; автоматичний машинний переклад зі збереженням смислового навантаження вихідного ПМ тексту та його відтворення адекватними засобами іншої мови [6]; моделювання спілкування, в тому числі чат-боти [49,50], електронні помічники [8], технології «розумний дім» [51], експертні системи [52].

Задачі ОПМ, що використовуються для вирішення цих задач, належать до семантичного аналізу, в який входять семантичний аналіз – аналіз зв’язків між елементами тексту та прагматичний аналіз – аналіз впливу контексту на їх значення. Зазначимо, що усі задачі цього рівня є основними задачами ОПМ.

Семантичний аналіз – це процес співставлення елементів тексту – від слів та речень до абзаців, текстів та мови в цілому – та їх смислового навантаження [4]. На відміну від задач синтаксичного аналізу, де необхідна інформація міститься у структурі самого тексту, семантичний аналіз вимагає загального розуміння тексту на рівні його сприйняття людиною.

На практиці семантичний аналіз зводиться до визначення зв’язків між окремими елементами тексту – визначення належності прикметників, ідентифікації границь речення та абзацу, визначення семантичних зв’язків між абзацами тощо. Джерелом семантичної інформації може бути як більш складний аналіз того ж самого тексту, так і використання зовнішніх джерел семантичних даних.

Системи на основі N-грам [53] – один з найпростіших варіантів систем семантичного аналізу. N-грама — це послідовність скінченної кількості (N) елементів тексту (як правило, слів), яка часто повторюється у тексті. В рамках семантичного аналізу вони найчастіше використовуються як допоміжний засіб у комплексі з іншими методами, наприклад у комплексі з ймовірнісними методами

або попередньо сформованими семантичними БД для визначення наступного слова у послідовності на основі кількох наявних.

Ефективність використання такого підходу для семантичного аналізу довільного тексту є досить низькою – точність розпізнавання становить близько 85-90% [54]. Оскільки підходи з використанням *N*-грам використовують стохастичні методи для власне аналізу тексту, внутрішні зв'язки довільної *N*-грами є однорідними, а структура її формується автоматично, можемо зробити висновок про те, що використання *N*-грам у семантичному аналізі виправдане лише для вирішення окремих задач, але не як самостійного інструменту семантичного аналізу. Зокрема, *N*-грами дозволяють уникнути розкриття значення власних назв [55] та обробки певні усталених мовних засобів, як-то ідіом та абревіатур [56].

Мережі суміжності *(co-occurrence networks)* [57], як і *N*-грами, використовуються для виділення зв'язків між елементами тексту. На відміну від *N*-грам, де уся послідовність слів розглядається як одна структура, мережі суміжності формуються на основі сукупності зв'язків між парами елементів, що, відкриває набагато більші можливості для їх використання. Підкреслимо дві суттєві відмінності мереж суміжності від *N*-грам: по-перше, у мережах суміжності для визначення характеру зв'язку між елементами можуть використовуватися не лише стохастичні, але й детерміновані алгоритми; по-друге, на відміну від *N*-грам, у мережах суміжності з'являється чіткий розподіл між елементами мережі, що відповідають елементам тексту, та зв'язками між ними.

Мережі суміжності є найбільш ефективними при ідентифікації та розпізнаванні зв'язків між уже визначеними концептами — наприклад, людьми, організаціями, термінами тощо [58]. Водночас, спроби використовувати їх для роботи з поширеними термінами та концептами та для вирішення складних задач, таких як визначення семантики даного концепту або аналіз структури тексту, наразі є малоефективними, з максимальною точністю у 72,5% [59].

Як бачимо, *N*-грами та мережі суміжності добре працюють з невеликими фрагментами тексту, які містять у собі усі необхідні зв'язки, але потребують

зовнішніх джерел семантичної інформації для ідентифікації зовнішніх зв’язків. Проаналізуємо основні підходи до накопичення такої інформації.

Першим підходом є виділення концептів, або концептуалізація [60] – об’єднання смислових та ситуативних синонімів у «концепт» і визначення їх смислових зв’язків у тексті [61]. Ефективність автоматичного виділення концептів найбільша в рамках окремого тексту або обмеженої бази текстів [62], коли задача зводиться до класичної задачі кластеризації з відомою кількістю об’єктів.

Очевидно, що в такому випадку ефективність концептуалізації відповідає ефективності кластеризації, тобто основним критерієм її ефективності є можливість чітко виділити окремі значення даного концепту та достовірно відрізнити їх між собою. Відповідно, на матеріалі динамічних та великих баз текстів можливо отримати лише неточні, узагальнені значення концепту, а для його уточнення необхідно обробляти концепти та структуру тексту, що їх оточує, як окремі сутності.

Побудова семантичної структури тексту фактично є реалізацією тієї ж самої задачі, яку вирішує визначення концептів, у зворотному напрямі. Якщо визначення концептів має на меті пошук окремих смислових одиниць та формування структури тексту на основі концептів та зв’язків між ними, то побудова семантичної структури являє собою створення певного фіксованого каркасу тексту, який наповнюється знайденими семантичними одиницями згідно набору відомих правил. Цей підхід показує гарні результати при виділенні зв’язків в рамках невеликого фрагменту тексту, такого як речення [63], але на більших обсягах тексту виникають вже інші проблеми на рівні обробки знань.

Прагматика - вивчення мови, яке фокусує увагу на користувачах та контексті використання мови, а не довідкових матеріалах, аксіоматичних твердженнях та граматиці [64]. Якщо семантика визначає смислові відношення між елементами тексту, прагматика звертається до знань, що зберігаються поза цим текстом, для вибору найбільш ймовірного з можливих варіантів його розуміння, отриманих за

допомогою синтаксичного та семантичного аналізу, та визначення його значення в контексті знань читача.

Прикладні задачі, в яких використовується семантичний аналіз, потребують зовнішнього відносно задачі джерела семантичної інформації. Так, для виконання якісного машинного перекладу необхідно враховувати не лише безпосередньо переклад слів, але й смислові еквіваленти слів та виразів у різних мовах та враховувати зміну їх значення в залежності від контексту; аналогічно, в задачі природномовного пошуку важливу роль займають усталені вирази та тематичні галузі використання слів, що включені у запит.

Відповідно, для використання семантичної інформації в задачах ОПМ її необхідно представити у вигляді, придатному для автоматичної обробки. Таким чином, вирішення прикладних задач ОПМ потребує використання джерел семантичної інформації, які задовольняють таким вимогам, а саме: є глобальними, тобто зберігають більше інформації про світ, ніж міститься безпосередньо в запиті; та є структурованими, тобто семантична інформація у цих джерелах представлена у вигляді, придатному для автоматичної обробки комп’ютером.

1.2 Аналіз ролі баз знань в технологіях обробки природної мови

1.2.1 Визначення бази знань та природномовної бази знань

Знання-орієнтовані системи (ЗОС) – це комп’ютерні системи, які зберігають і створюють знання та використовують його для вирішення складних задач [65]. До складу ЗОС, як правило, входить база знань, що містить знання про світ, та підсистема логіки ЗОС, яка забезпечує відповідність роботи системи заданим правилам. В задачах ОПМ роль ЗОС зводиться до надання інтерфейсу використання БЗ [66], яка є одним з інструментів ОПМ рівня семантики.

Оскільки БЗ як складова ЗОС є глобальним джерелом знань відносно конкретного ПМ тексту, який є об’єктом роботи ОПМ, якість роботи БЗ прямо залежить від її попереднього наповнення.

База знань (БЗ) – це особливого роду база даних, розроблена для управління знаннями, а саме збором, зберіганням, пошуком і видачою знань [67].

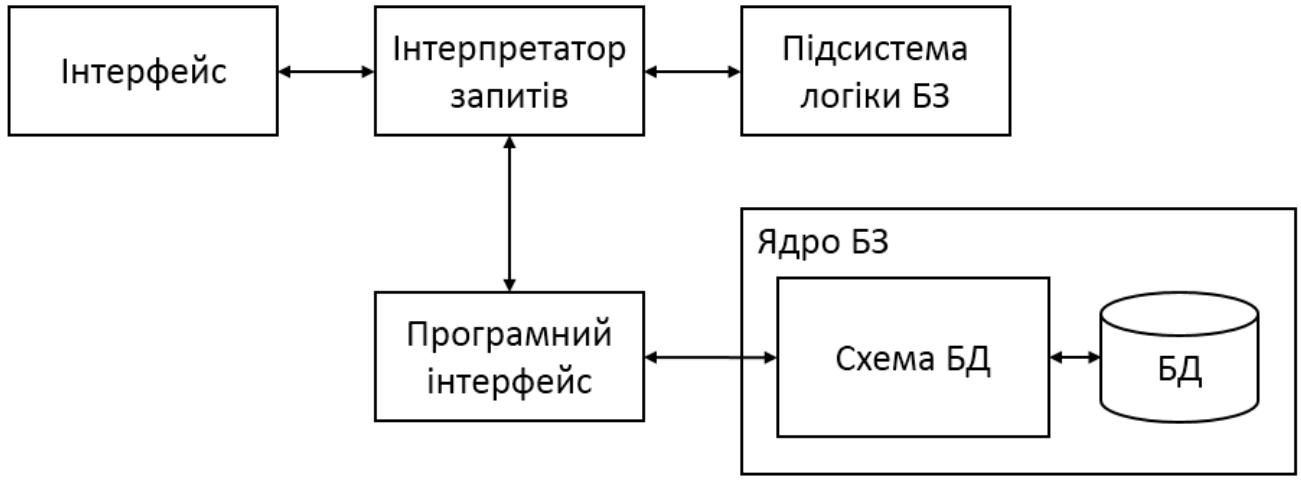


Рисунок 1.3 – Структурна схема бази знань

Структурними компонентами бази знань (БЗ, рис.1.3) в загальному випадку є:

1) Інтерфейс: підсистема, що забезпечує обмін інформацією між БЗ та користувачем;
2) Інтерпретатор запитів: підсистема, яка координує обробку запитів у БЗ;
3) Ядро БЗ та його програмний інтерфейс: підсистема, яка обробляє та виконує запити на рівні БД;
4) Підсистема логіки, яка модифікує запити та результати їх виконання згідно певної системи правил [68].

Проаналізуємо детальніше функції кожної з цих підсистем.

Основною складовою БЗ є *ядро БЗ* – підсистема, що відповідає за виконання основних операцій над наповненням БЗ згідно з визначеним форматом знань БЗ. Ядро БЗ координує взаємодію між БЗ та базою даних (БД), на основі якої цю БЗ побудовано. База даних як частина ядра БЗ – це сукупність даних, організованих

відповідно до концепції, яка описує характеристику цих даних і взаємозв'язки між їх елементами [69].

Модель представлення знань БЗ (МПЗ БЗ) – модель, що описує структуру БД БЗ, а саме об’єкти даних БД, зв’язки між ними та можливі операції над ними. Таким чином, до складу ядра БЗ входить БД, яка реалізує МПЗ БЗ, та програмний інтерфейс ядра БЗ – допоміжна підсистема, який перетворює запити до БЗ, що відповідають МПЗ БЗ, у низькорівневі запити до БД у складі цієї БЗ.

Другою складовою БЗ є *підсистема логіки БЗ* – підсистема, що забезпечує консистентність БЗ, тобто керує обмеженнями на дані в БЗ та зв’язками між ними, зберігає правила обробки цих даних та виведення нових знань на основі існуючих з урахуванням їх відповідності певним формальним вимогам [70].

Інтерпретатор запитів – підсистема, який виступає в якості посередника між потребами користувача та можливостями БЗ. До функцій інтерфейсу запитів належить формування плану виконання запиту користувача у вигляді послідовності операцій БЗ, контроль над процесом виконання цих операцій та формування з них відповіді на запит користувача.

Інтерфейс користувача – підсистема, що забезпечує обробку запитів користувача, передачу їх до БЗ та отримання результатів їх виконання від БЗ. Інтерфейс користувача є точкою входу для зовнішніх запитів до БЗ, причому це може бути як *UI* (*user interface*, інтерфейс користувача), так і *API* (*application programming interface,* інтерфейс прикладного програмування) – тобто, в залежності від реалізації, цей модуль забезпечує взаємодію БЗ з людиною-оператором або з іншими системами.

Визначимо природномовні бази знань (ПМБЗ) як клас БЗ, об’єктом роботи яких є природномовна інформація (ПМІ) – тобто, знання в яких зберігається безпосередньо у вигляді ПМІ, на відміну від інших класів БЗ, що використовують спеціалізовані формальні моделі представлення знань у БЗ. Розширенням ПМБЗ відносно БЗ є підсистема обробки природної мови (підсистема ОПМ), до функцій якої входять усі операції над ПМ даними у БЗ.

Фактично ПМБЗ (рис.1.4.) це будь-яка БЗ, яка здатна зберігати ПМІ та обробляти ПМ запити користувача. Тобто, в широкому розумінні до класу ПМБЗ можемо віднести будь-яку систему, яка складається з БЗ та підсистеми ОПМ, між якими встановлено функціональні зв'язки – навіть якщо структурно це різні, незалежні системи. Параметр, за яким ПМБЗ відрізняється від інших подібних систем, це сила таких зв'язків у системі: тобто, якщо для коректної роботи БЗ і усієї системи використання підсистеми ОПМ є необхідним кроком, така система є ПМБЗ [71].

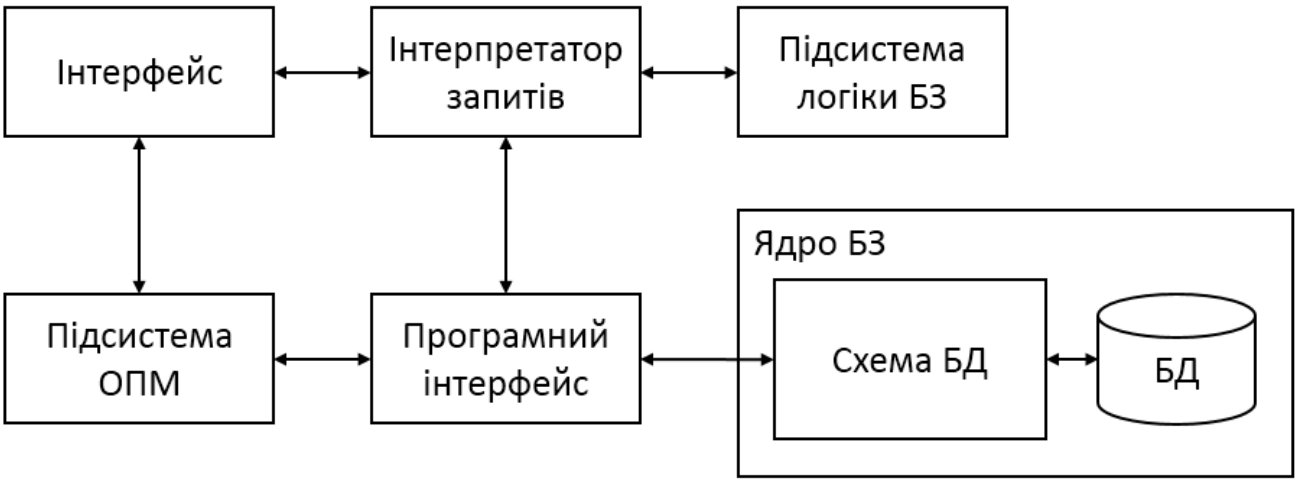


Рисунок 1.4 – Структурна схема природномовної бази знань

Підсистема ОПМ у ПМБЗ використовується ядром ПМБЗ та інтерфейсом ПМБЗ. Ядро ПМБЗ використовує підсистему ОПМ для зберігання та обробки ПМІ, що пов'язана зі знаннями, які зберігаються у БЗ – фрагментів тексту та слів. Інтерфейс БЗ використовує підсистему ОПМ для обробки ПМ запитів користувача ПМБЗ для виділення з них знань згідно МПЗ у БЗ та для перетворення результату виконання запиту у вигляд ПМІ. Система логіки ПМБЗ не має суттєвих відмінностей від системи логіки БЗ. Процес відтворення знань, що зберігаються у тексті це, власне, і є процес наповнення БЗ, який виконує людина-оператор або який відбувається автоматично.

1.2.2 Аналіз існуючих природномовних баз знань

Використання знання як об’єкту роботи БЗ вимагає визначення його чіткої формальної структури; водночас, структура ПМІ є окремим складним об’єктом дослідження. Для взаємодії між цими структурами ПМБЗ повинна зберігати консистентність – тобто, забезпечувати відповідність наповнення ПМБЗ певним формальним вимогам [70], основними з яких є:

1) несуперечливість – здатність зберігати в одній системі твердження, які несумісні на рівні логіки;
2) повнота – здатність зберігати ПМ текст довільної структури та значення;
3) гнучкість – здатність приєднувати нові знання до існуючих.

На основі цих критеріїв оцінимо основні існуючі підходи до проектування ПМБЗ, проаналізуємо особливості їх використання в ОПМ та визначимо їх переваги та недоліки. Для цього класифікуємо ПМБЗ за ступенем зв’язності структурованих знань у БЗ та вхідних текстів:

1) неструктуровані – структура знань на основі граматичної структури тексту;
2) структуровані – формальна структура фрагменту знань, яка заповнюється з тексту;
3) мережеві – структура знань, представлена у вигляді мережі елементів тексту та зв’язків між ними.

Проаналізуємо моделі представлення знань існуючих ПМБЗ згідно визначених вище критеріїв.

Неструктуровані ПМБЗ. У таких ПМБЗ найменшим елементом знань у ПМБЗ є стаття, а зв’язки між цими елементами визначаються взаємними посиланнями, які прив’язані до окремих слів або термінів у тексті статті. Системи цього класу (такі як *Wikipedia* [72], *Wolfram Alpha* [73] та ін.)., дозволяють розбити великі масиви природномовної інформації на менші фрагменти, кожний з яких є більш зручним для пошуку, категоризації і подальшої обробки. Основними проблемами проектування неструктурованих ПМБЗ є розробка системи метаданих,

тобто структури зв’язків між статтями, системи їх категоризації та розподіл ПМ знань між наповненням статей.

Перевагою неструктурованих ПМБЗ є простота роботи для користувача. Недоліком неструктурованих ПМБЗ є їх негнучкість. Оскільки МПЗ таких ПМБЗ не є добре структурованою, задачі аналізу даних перекладається на користувача, що робить неефективним автоматичне наповнення бази знань та додає помилку оператора до негативних факторів роботи системи.

Оцінимо неструктуровані ПМБЗ згідно з описаними вище критеріями.

Повнота: висока. Вміст окремої статті у вигляді ПМ тексту не має технічних обмежень.

Несуперечливість: низька. Логічна обробка наповнення неструктурованих ПМБЗ засобами самої БЗ не є можливим.

Гнучкість: середня. Оскільки усі зв’язки у тексті статті прив’язані до слів та термінів, які можуть змінювати значення в залежності від контексту, модифікація таких посилань може бути автоматизована лише частково.

Структуровані ПМБЗ. В основі МПЗ цього класу лежить деяка штучна структура знань, яка відокремлена від тексту.

Найбільш відомими варіантами ПМБЗ на основі штучної структури є онтології та фреймові системи, в тому числі такі відомі розробки як OWL [74] та OWL2 [75], системи зберігання знань на основі фреймів та системи зберігання природномовних знань на основі об’єктно-орієнтованих баз даних взагалі.

Задачею проектування структурованих ПМБЗ є уніфікація природномовної інформації та підготовка її для подальшої автоматичної обробки комп’ютером [76]. Відповідно, знання у структурованих ПМБЗ зберігається у формі деякого об’єкта, структура якого задана безпосередньо або через набір правил та обмежень, а сама система працює з ПМІ лише після її попередньої обробки зовнішньою системою ОПМ, у якій аналіз тексту відбувається без використання БЗ [77,78], або взагалі обмежений згідно вимог щодо вирішення конкретної практичної задачі [79].

Головна перевага структурованих ПМБЗ, а саме використання чітко формалізованої структури окремого елементу знань, є одночасно і їх головним недоліком. Оскільки природномовний текст в загальному випадку має складну структуру, а повнота знань, що зберігаються у БЗ, прямо залежить від спроможності БЗ відтворити таку структуру, при заповненні ПМБЗ з тексту відбувається втрата або спотворення знань, які виходять за формат представлення знань такої ПМБЗ.

Оцінимо структуровані ПМБЗ згідно з описаними вище критеріями.

Повнота: низька. Деяка частина знань втрачається при заповненні елементу знань.

Несуперечливість: висока. Структурована ПМБЗ добре взаємодіє з системами комп'ютерної логіки, які можуть вирішувати конфлікти наповнення в автоматичному режимі.

Гнучкість: середня. Модифікація окремого елементу знань є базовою операцією, але при цьому необхідно рекурсивно модифікувати усі пов'язані з ним елементи.

Мережеві ПМБЗ. ПМБЗ на основі семантичних мереж є найбільш формалізованим класом ПМБЗ. Системи цього класу (як-то *Google Knowledge Graph* [80], *Palantir* [58] тощо), дозволяють доповнити формальне представлення знань, характерне для ПМБЗ на основі фреймів, системою їх автоматичної обробки.

Очевидною перевагою ПМБЗ на основі семантичних мереж є можливість представлення знань у вигляді графа, що забезпечує високу повноту та гнучкість ПМБЗ. Більш того, відношення або зв'язки між окремими елементами знань так само входять у структуру графу або можуть бути представлені аналогічними йому структурами, що дозволяє делегувати забезпечення несуперечливості знань окремій підсистемі. Таким чином, ПМБЗ на основі семантичних мереж дозволяють розглядати зв'язки «текст-знання» та «знання-метадані» як окремі підсистеми ПМБЗ без об'єднання їх у систему логіки БЗ.

Хоча ПМБЗ на основі семантичних мереж не вирішують проблему зберігання ПМ знань в загальному вигляді, вони дозволяють перенести її вирішення на рівень представлення знань – вищий рівень абстракції, який реалізовано на основі графів і який не вимагає прив’язки до структури ПМБЗ. Фактично, проблема структури та функції елементу знань, притаманна структурованим ПМБЗ, зводиться до проблеми визначення структури графу знань, структури та функцій його елементів та зв’язків між ними.

Оцінимо мережеві ПМБЗ згідно з описаними вище критеріями.

Повнота: середня. Граф як структура даних є достатньо гнучким для відтворення довільного ПМ повідомлення, але його наповнення залежить від роботи системи ОПМ.

Несуперечливість: низька. Логічна обробка наповнення неструктурованих ПМБЗ засобами самої БЗ не є можливим.

Гнучкість: висока. Модифікація певної частини графу в загальному випадку не призводить до змін у інших його частинах.

Таблиця 1.1 – Порівняння класів природномовних баз знань

Клас ПМБЗ	**Повнота**	**Несуперечливість**	**Гнучкість**
структуровані	висока	низька	середня
неструктуровані	низька	висока	середня
мережеві	середня	низька	висока

Порівняємо існуючі підходи до розробки ПМБЗ на основі критеріїв повноти, несуперечливості та гнучкості. З табл. 1.1 бачимо, що в усіх розглянутих підходах немає спільного недоліку – тобто, кожний підхід має певні сильні сторони, але жодний з них не поєднує їх усі.

1.3 Обґрунтування актуальності розробки інформаційної технології обробки природномовних текстів на основі інтеграційного підходу

1.3.1 Формулювання вимог до нового типу природномовної бази знань

Важливим напрямком сучасних досліджень в галузі ПМБЗ є розробка гібридних ПМБЗ, що поєднують різні моделі ПМБЗ, наприклад мережеву модель з частково рекурсивними елементами [81,82]. Втім, такі системи часто наслідують не лише переваги, але й недоліки систем, на яких вони засновані. Крім того, розробка гібридної ПМБЗ є складною задачею, і така система часто обмежена умовами прикладної задачі, для вирішення якої вона створюються – а саме, може містити принципові недоліки, які не впливають на вирішення даної задачі, але є суттєвими для ПМБЗ взагалі.

Актуальною проблемою проектування ПМБЗ є розробка МПЗ, здатної з високою точністю і повнотою представити довільний ПМ текст. Оскільки МПЗ у ПМБЗ на основі семантичної мережі складається з моделі елементу знань та моделі сукупності відношень між ними, задача розробки МПЗ зводиться до визначення структури елементу знань, мережі відношень між ними та механізму прив’язки наповнення БЗ до ПМ текстів.

Представимо більш широко розробку ПМБЗ як моделювання сховища знань, представлених у вигляді ПМІ. Виходячи з цього, можемо представити ПМБЗ як один з компонентів моделювання мовленнєвої діяльності людини взагалі, який має зв’язки з іншими компонентами. Відповідно, для вдосконалення ПМБЗ необхідно враховувати особливості структури ПМ тексту як об’єкту роботи такої БЗ, але при цьому МПЗ не обов’язково є похідною від цієї структури.

У нейропсихології, лінгвістиці та філософії мови, природною мовою або звичайною мовою називають будь-яку мову, яка еволюціонувала у людини природно, через використання і повторення без свідомого планування чи попереднього наміру. Природні мови можуть мати різні форми, такі як мовлення

або спів. Вони відрізняються від конструюються та формальних мов, таких як ті, що використовуються для програмування комп'ютерів або для вивчення логіки.

Для існуючих ПМБЗ характерне представлення ПМІ як незалежного об’єкту, який характеризується граматичною структурою. Найменші структурні елементи цих ПМБЗ так само або є похідними від граматичної структури тексту (як-то *n*-грами [55] та структури на основі предикатів Хомського [83]), або взагалі незалежні від неї (наприклад, концепт як одиниця знань [84]). Ці підходи часто є незручними для практичного використання, і часто елемент знань описується через суто функціональне визначення (наприклад, *«квант знань»* як *«найменша неподільна смислова порція інформації»* [85]). Отже, в основі нового типу ПМБЗ повинна лежати така МПЗ, яка одночасно включає в себе структуру універсальної моделі знань та структуру довільного природномовного тексту.

1.3.2 Інтеграційний підхід до обробки природної мови

Для вирішення поставлених задач звернемося до лінгвістики, а саме до інтеграціонізму, який задовольняє більшості поставлених вимог.

Інтеграціонізм (*integrationism*) – це підхід до теорії комунікацій взагалі і зокрема лінгвістики, який зародився у 1980-і роки паралельно в кількох групах вчених, найбільш широко відомою з яких є колектив дослідників в Оксфордському Університеті на чолі з Roy Harris [86]. На сьогоднішній день інтеграціонізм залишається активним напрямом досліджень, вагомі внески у який зробили такі вчені як H.G. Davis [87], Adrian Pable [88], Christopher Hutton [89] та інші.

Однією з основних ідей інтеграціонізму є перехід від сприйняття тексту як окремої сутності, що строго описується деякими внутрішніми правилами, до тексту як продукту відображення процесів більш високого рівня, як-то процесу мислення та сприйняття людиною дійсності.

В контексті письмового тексту це означає наступне:

1) Текст є способом опису дійсності, а не самостійним об’єктом
2) Текст є похідним від сенсорного сприйняття
3) Формування та сприйняття тексту залежать від власних знань/досвіду

Відповідно, для вирішення задач обробки природної мови потрібно опрацьовувати не лише текст як граматичну конструкцію, але й об’єкти та процеси, які є первинними відносно нього.

В рамках інтеграціонізму на кафедрі технічної кібернетики КПІ ім. Ігоря Сікорського протягом кількох десятиріч, починаючи з 90-х років формувався інтеграційний підхід (далі – ІП) до структурної організації мови, що враховує сучасні досягнення у багатьох помежованих напрямках досліджень мовленнєвої діяльності людини. Основні положення цього підходу викладені у роботах Ю. Кисленка [90–92].

Інтеграційний підхід завоював авторитет в КПІ, на його основі були сформовані відповідні курси «Системна організація мови» та «Інформаційні природномовні технології», був презентований на міжнародному рівні, зокрема у концентрованому вигляді він оприлюднений в роботах *«Back to basics of speech activity»* [93] та *«Cognitive architecture of speech activity and modelling thereof»* [94].

Індивідуальна мовна система. Першою принциповою особливістю ІП є використання концепції індивідуальної мовної системи, яка була запропонована ще у 20-х роках ХХ сторіччя Л. Щербою [95].

Індивідуальна мовна система (ІМС) представляє собою сукупні знання окремої людини про мовлення і складається з двох основних частин: *лінгвістичного процесору* (ЛП ІМС), що містить знання про мовну організацію, які можуть бути представлені на свідомому і підсвідомому рівнях, та *бази знань* (БЗ ІМС), де зберігається уся сукупність накопичених знань мовного рівня про довкілля, в якому ця людина живе.

На відміну від підходів класичної лінгвістики, де текст представляє собою незалежний об’єкт дослідження, а знання є похідним від нього, в рамках ІП об’єктом моделювання є сама мовленнєва діяльність людини, а текст розглядається

як результат трансляції когнітивного потенціалу на мовний рівень. Це дозволяє з нових позицій підійти до усього спектру задач обробки природної мови та, зокрема, проблеми комп'ютерного моделювання окремих аспектів мовленнєвої діяльності людини – спілкування, пізнання, накопичення знань тощо.

Структура ІМС представлена на рис.1.5

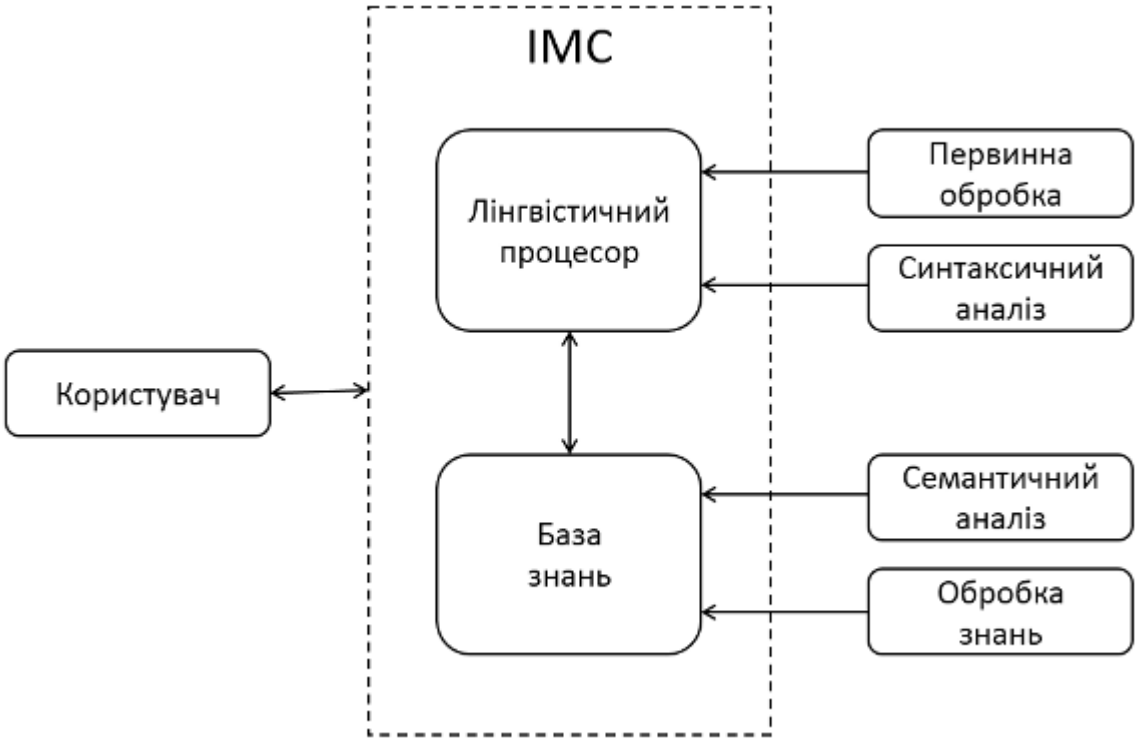


Рисунок 1.5 – Структура та функції індивідуальної мовної системи [93]

Використання ІМС як узагальненої моделі мовленнєвої діяльності людини, де БЗ ІМС зберігає когнітивний потенціал, а ЛП ІМС забезпечує його трансляцію на мовний рівень, дозволяє:

1) встановити через взаємодію ЛП ІМС та БЗ ІМС чітку взаємну залежність між технологіями ОПМ, де ПМІ розглядається як результат мовленнєвої діяльності людини, та технологіями обробки знань, де знання представляють собою результат її розумової діяльності;
2) розділити в окремі підсистеми ПМБЗ роботу з текстом та роботу зі знаннями, не відкидаючи зв'язки між цими підсистемами.

Сенсорний квант знань. Друга принципова концепція ІП – поняття універсального сенсорного кванту знань.

Квант знань сенсорного рівня, або *сенсорний квант знань* (СКЗ) – це двоскладова структура $Subj/Mov$, включно з її атрибутивним оточенням, яка відтворює модель ситуації сенсорного рівня, що існує у триєдності часу, простору та дії [90]. Структура СКЗ є похідною від структури ситуації зорового рівня, що випливає з нейрофізіології обробки сенсорної інформації людиною. Зір людини, що забезпечує до 90% сенсорної інформації, яку вона отримує, працює за дискретною схемою [96], де окрема дискрета охоплює ситуацію.

Ситуація (ситуація зорового рівня) – це фрагмент зорової складової довкілля, що потрапляє на центральну ямку сітківки та опрацьовується за повною програмою [90]. Функціональне навантаження зорового тракту на сьогоднішній день ретельно досліджене. Зокрема, експериментально доведено, що у зоровому тракті людини існують ансамблі нейронів, що виконують чітко визначені функції.

Як показав С. Зекі [97], ці ансамблі нейронів виділяють з окремої ситуації об'єкти $Obj,$ суб'єкти $Subj$, прикмети цих складових $Attr(Obj/Subj)$ та міру цих прикмет $Attr(Attr)$. Аналогічним чином опрацьовується і динаміка рухомих об'єктів – Mov, $Attr(Mov)$ та $Attr(Attr)$. Більш того, як показав Дж. Хокінз [98], обробка усіх сенсорних даних людиною відбувається подібно до оброки візуальних даних, згідно схеми, описаної С. Зекі – а отже, однієї цієї структури достатньо для моделювання повної картини навколишнього світу, сприйнятої людиною.

Таблиця 1.2 – Співставлення класів слів та елементів СКЗ

Елемент СКЗ	Граматичний клас слова
$Subj$	іменник
$Attr(Subj)$	прикметник
$Attr(Attr)$	прислівник
Mov	дієслово
$Attr(Mov)$	прислівник
$Attr(Attr)$	прислівник

Аналогічні структурні елементи спостерігаємо й на рівні тексту. У лінгвістиці прийнято виділяти 4 основні граматичні ролі – класи слів, що відповідають ролям іменника, дієслова, прикметника та прислівника [99]. Усі змістовні частини мови у тексті мають ролі, що відповідають цим основним граматичним ролям. Така сама відповідність є і між елементами СКЗ та граматичних класами, що показано у табл. 1.2.

Таким чином, між СКЗ та деяким фрагментом тексту можливо встановити відповідність, де елементам СКЗ відповідають елементи тексту, що належать до аналогічного граматичного класу слів.

Використання СКЗ як основи для побудови МПЗ у ПМБЗ дозволяє:

1) чітко визначити і обмежити структуру та границі окремого елементу знань у тексті;
2) встановити формальну відповідність на структурному рівні між знаннями сенсорного рівня та знаннями у вигляді ПМІ;

Мовний квант знань. Третя важлива складова ІП – поняття кванту знань мовного рівня.

Квант знань мовного рівня, або *мовний квант знань (МКЗ)* – це структура ПМ тексту, яка є вербалізацією окремого СКЗ. В рамках ІП в якості МКЗ виступає базова семантико-синтаксична структура.

Базова семантико-синтаксична структура (БССС) – це двоскладова схема опису довільної ситуації реального чи віртуального світу, всі складові якої

актуалізовані на атрибутивному рівні [90]. Структура БССС є похідною від наведеної вище структури СКЗ та підтверджується дослідженнями, презентованими у роботі А. Гвоздєва [100], де представлена послідовність етапів опанування дитиною мовного ладу: від слів, що повністю представляють окрему ситуацію, через двоскладові структури вигляду «іменник-дієслово», включення ситуаційних та предикативних складових та формування атрибутивного рівня до повної структури БССС.

В основі БССС полягає елементарна МКЗ, що включає в себе суб’єкт, предикатор та їх атрибутивне оточення. Структура ядра БССС представлена на рис.1.6.

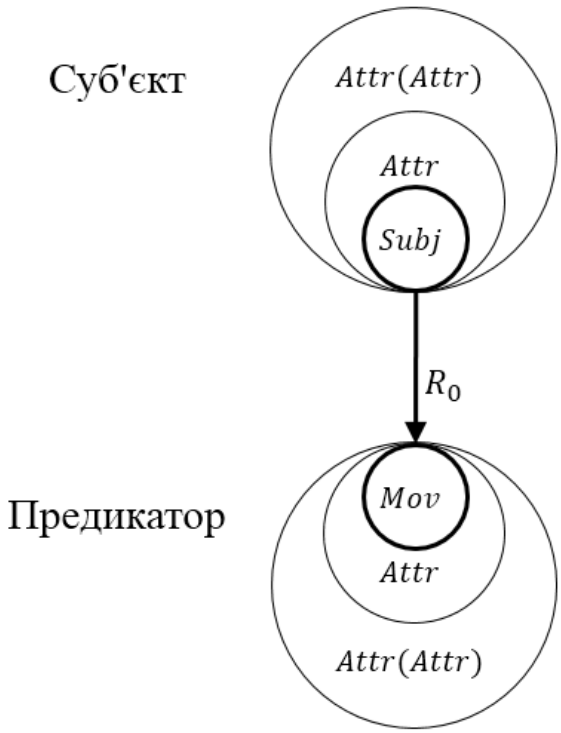


Рисунок 1.6 – Ядро БССС [90]

Ядро БССС описує ізольовану ситуацію, в якій суб’єкт та предикатор пов’язані єдиним основним відношенням R_0 «мати предикатор».

У тексті БССС пов’язані між собою відношеннями через предикати, які належать до наступних типів [101]:

1) ситуаційні відношення визначають місце даної БССС у часі, просторі тощо;

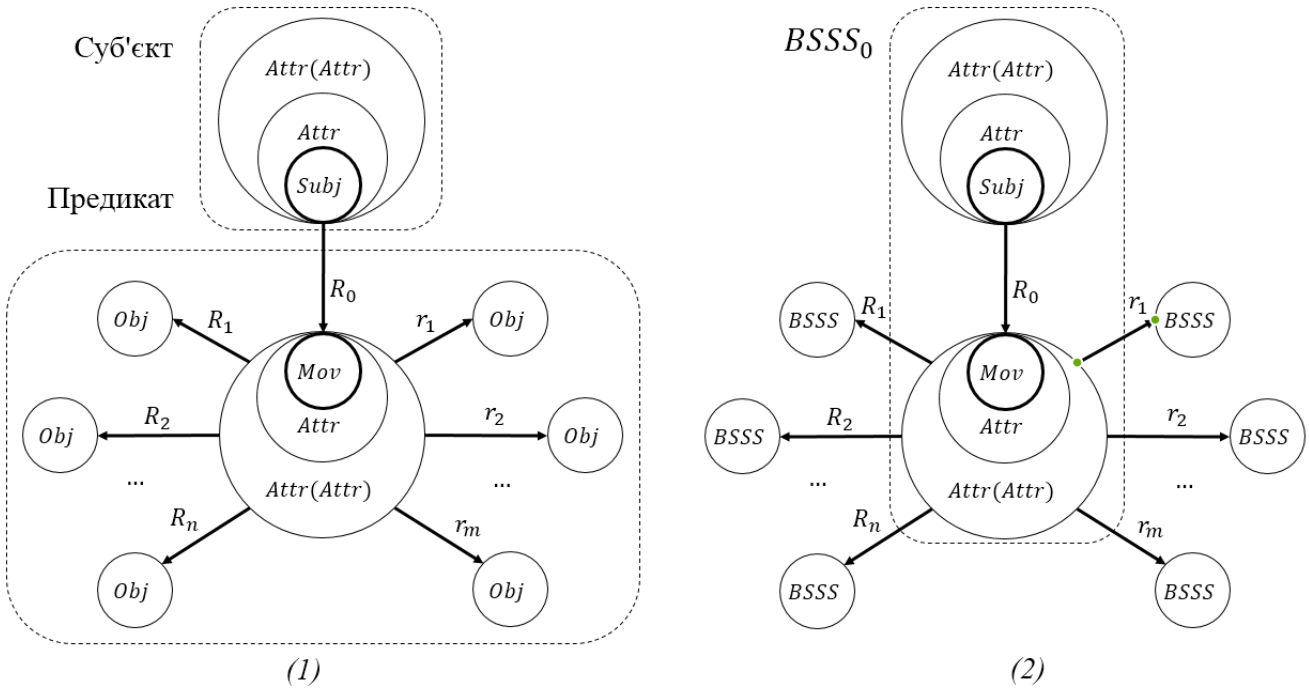
2) предикативні відношення описують взаємодію даної БССС з іншими елементами знань.

Додатково БССС поділяються на два класи в залежності від цілей їх відношень:

1) монопредиктані БССС мають відношення, які зв’язані виключно з елементами типу Obj, що за структурою аналогічні $Subj$ але не мають власних предикатів;
2) поліпредикатні БССС мають відношення, які зв’язують їх з іншими повноцінними БССС.

На рис.1.7 представлені монопредикатна *(1)* та поліпредикатна *(2)* БССС, де R_i – предикативні відношення, r_i – ситуаційні відношення. Зауважимо, що структура монопредикатної БССС описує усі пов’язані елементи Obj, в той час як поліпредикатна БССС (показана як $BSSS_0$) описує тільки зв’язки, оскільки структура мережі відношень кожної з пов’язаних з нею БССС не є фіксованою.

Рисунок 1.7 – Структура монопредикатної та поліпредикатної БССС [94]



Використання БССС як елементу знань ПМБЗ дозволяє:

1) формалізувати структуру довільного ПМ тексту у вигляді БССС;

2) прив’язати структуру ПМ тексту до довільної структури знань ПМБЗ.

Відповідно, інформаційна технологія обробки природномовних текстів, яка втілює ці можливості, дозволяє:

1) встановити через взаємодію ЛП ІМС та БЗ ІМС чітку взаємну залежність між технологіями ОПМ, де ПМІ розглядається як результат мовленнєвої діяльності людини, та технологіями обробки знань, де знання представляють собою результат її розумової діяльності;
2) розділити в окремі підсистеми ПМБЗ роботу з текстом та роботу зі знаннями, не відкидаючи зв’язки між цими підсистемами.
3) чітко визначити і обмежити структуру та границі окремого елементу знань у тексті;
4) встановити формальну відповідність на структурному рівні між знаннями сенсорного рівня та знаннями у вигляді ПМІ;
5) формалізувати структуру довільного ПМ тексту у вигляді БССС;
6) прив’язати структуру ПМ тексту до довільної структури знань ПМБЗ.

Висновки до розділу 1

1. На основі аналізу існуючих технологій обробки природної мови показано, що існуючі технології не виконують в повній мірі поставлені задачі. На основі аналізу взаємодії технологій різних рівнів обробки природної мови визначено роль природномовної бази знань як необхідного компонента технологій обробки природної мови, який забезпечує зв'язок між різними їх рівнями.
2. На основі аналізу підходів до розробки природномовних баз знань обґрунтовано потребу у розробці універсальної моделі представлення знань природномовного тексту для використання в технологіях обробки природної мови, що поєднує переваги існуючих підходів, та процедур використання такої моделі в інформаційних технологіях обробки природномовних текстів.

2 РОЗРОБКА МОДЕЛЕЙ ТА МЕТОДІВ ОБРОБКИ ПРИРОДНОЇ МОВИ НА ОСНОВІ ІНТЕГРАЦІЙНОГО ПІДХОДУ

У цьому розділі визначено особливості розробки природномовної бази знань для інформаційної технології обробки природномовних текстів на основі інтеграційного підходу. З урахуванням цих особливостей створено формальну модель представлення знань у природномовній базі знань та розроблено моделі її основних елементів, якими є квант знань, або найменший елемент знань, та відношення, яке описує зв'язок між квантами знань. З використанням моделі представлення знань створено метод обробки природномовних текстів для інформаційної технології та процедури використання інформаційної технології у прикладних задачах обробки природної мови.

2.1 Аналіз особливостей природномовної бази знань на основі інтеграційного підходу

Проаналізуємо, яким чином використання принципів ІП змінює загальну структуру ПМБЗ. Для цього виділимо в окремі підсистеми модулі ПМБЗ, функціональне навантаження яких відповідає функціям ЛП ІМС та БЗ ІМС, та оцінимо, наскільки така заміна змінює функціонування ПМБЗ взагалі.

Введемо позначення *ПМБЗ ІП* (ПМБЗ на основі ІП) для подальшого використання.

Оскільки ПМБЗ ІП забезпечує обробку ПМ знань, представлених у вигляді сукупності БССС та зв’язків між ними, основні зміни відбуваються на рівні представлення та зберігання знань у ядрі ПМБЗ [102]. З особливостей БССС як структури ПМ знань випливає можливість прив’язки ПМІ та знань у ПМБЗ до спільної структури, а отже – на рівні структури ПМБЗ відбувається заміна обов’язкових модулів у складі БЗ на інтерфейси-адаптери, що працюють безпосередньо з ядром ПМБЗ. Це дозволяє зменшити зв’язність між підсистемами

ПМБЗ та розглядати більшість з них як окремі системи, взаємодія яких з ПМБЗ обмежується функціональними зв’язками.

Таким чином, підсистеми ПМБЗ у випадку ПМБЗ ІП (рис.2.1) виходять за рамки ПМБЗ і на рівні самої ПМБЗ представлені лише інтерфейсами обміну запитами. Зазначимо, що для повноцінного використання ПМБЗ ІП для вирішення задач ОПМ ці системи є практично так само необхідними для її адекватної роботи, але формально на структурному рівні вони є окремими системами, і використання ПМБЗ ІП без них є цілком можливим, особливо при роботі на низькому рівні з даними та запитами, які не потребують додаткової обробки та інтерпретації.

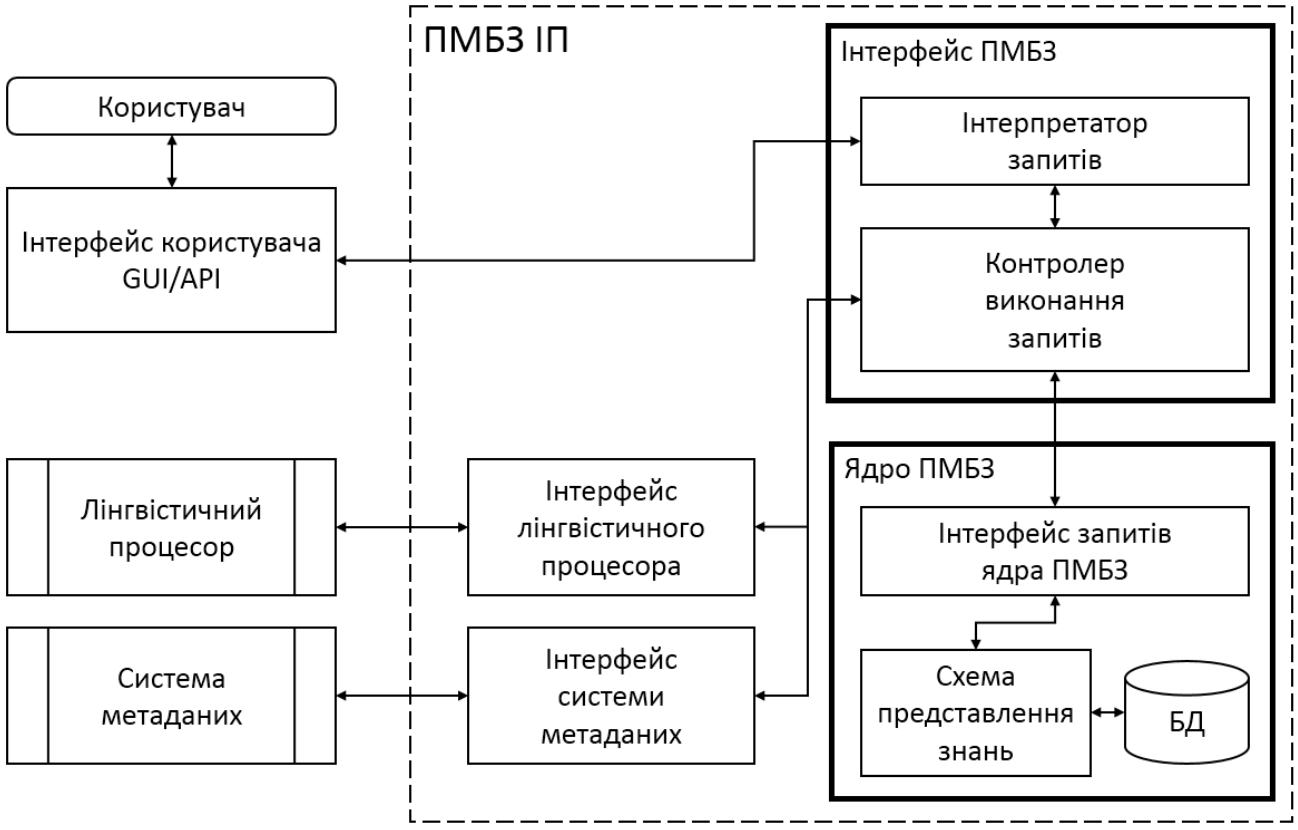


Рисунок 2.1 – Структурна схема природномовної бази знань на основі інтеграційного підходу

Інтерфейс ПМБЗ ІП виконує функції, аналогічні таким у загальній структурі ПМБЗ, а саме керує обміном запитами між ПМБЗ та користувачем та забезпечує декомпозицію складних зовнішніх запитів до формату запитів, які приймає ядро ПМБЗ.

Інтерфейс ЛП повністю забезпечує виконання задач ОПМ, які потребує ПМБЗ ІП, від синтаксичної обробки запитів до зберігання вихідних природномовних текстів, на основі яких заповнено БЗ. Основна функція інтерфейсу ЛП в ПМБЗ ІП – перетворення даних і запитів між форматами ПМІ та БССС.

Інтерфейс системи метаданих забезпечує двосторонній зв'язок між наповненням ПМБЗ ІП та зовнішнім джерелом логічних правил, правил виведення тощо з метою підтримки якості наповнення ПМБЗ, а саме цілісності, несуперечливості та повноти знань, та забезпечення механізму виведення нових знань у БЗ.

Описана вище структурна схема ПМБЗ ІП враховує особливості ІП, а саме забезпечує розділення обробки та зберігання природномовних знань. Будь-які додаткові системи, що розширюють можливості ПМБЗ, ізольовані від ядра БЗ за допомогою інтерфейсів і не впливають на роботу самої ПМБЗ ІП.

Таким чином, для розробки ПМБЗ для ІТ ОПМ необхідно виконати наступні кроки:

1) розробка моделі представлення знань ПМБЗ на основі ІП;
2) розробка методу обробки природної мови з використанням ПМБЗ на основі ІП;
3) розробка технічних вимог для реалізації такої МПЗ у інформаційній технології обробки природномовних текстів;

2.2 Розробка моделі кванту знань

2.2.1 Формалізація кванту знань як об’єкту ПМБЗ ІП

Опишемо найменший елемент ПМБЗ ІП, що є формалізованим представленням ядра БССС, або кванту знань мовного рівня (МКЗ). Для подальшого використання в контексті ПМБЗ ІП визначимо цей елемент як квант знань.

Квант знань (КЗ) – це двоскладова структура, яка включає в себе суб'єкт $Subj$, предикатор Mov та їх атрибутивне оточення, а саме атрибути $Attr(Subj)$ і $Attr(Mov)$ та міри цих атрибутів $Attr(Attr(Subj))$ і $Attr(Attr(Mov))$.

$Subj$ та його атрибутивне оточення визначають *суб'єкт* – текстове представлення предмету або явища реального світу.

Mov та його оточення визначають *предикат* – узагальнену схему опису дії суб'єкту, або динамічну складову ситуації.

Окремий КЗ не включає відношення моно- та поліпредикатного рівня БССС.

Позначимо елементи КЗ як

$$Subj = \{S, Attr(S)\}, \tag{2.1}$$

та

$$Pred = \{P, Attr(P)\}, \tag{2.2}$$

де S та P – суб'єкт та предикатор даного КЗ;

$Attr(S)$ та $Attr(P)$ – атрибутивні оточення суб'єкта та предикатора;

Таким чином, загальна структура КЗ набуває наступного вигляду:

$$Q = \begin{Bmatrix} S, Attr(S) \\ P, Attr(P) \end{Bmatrix}, \tag{2.3}$$

де Q – квант знань.

Оскільки КЗ є відображенням на мовному рівні СКЗ, тобто фрагменту сенсорної інформації реального світу, елементи КЗ – це відображення його елементів на мовному рівні. Виділимо об'єкт, що пов'язує структуру КЗ з його текстовим представленням – лексему.

Лексема – це окреме слово з усією сукупністю властивих йому форм словозміни й значень у різних контекстах [103].

В контексті ПМБЗ ІП це означає, що усі граматичні та смислові форми окремого слова не є самостійними сутностями, а належать до єдиного об'єкту. Використання такої структури даних дозволяє:

1) прив'язати одну й ту ж саму структуру кванту знань до різних варіантів її представлення на рівні тексту;

2) уніфікувати структурне представлення елементів КЗ ПМБЗ як конструкцій ПМІ та відповідних їм елементів СКЗ як елементів сенсорних знань;
3) правильно обробляти ситуації, коли певна форма лексеми на рівні тексту представлена кількома словами.

Визначимо лексему як структуру наступного вигляду:

$$L = \{l_0, [l_1, \dots, l_i, \dots]\}, \quad (2.4)$$

де l_0 – базова форма слова;

l_i – (за наявності) – інші граматичні форми слова.

З точки зору представлення даних, елементи КЗ є елементами структури даних відомого вигляду, в той час як лексема в її нормальному вигляді є елементом ПМІ, тобто елементом типу «рядок». Проаналізуємо, яким чином пов'язані ці елементи.

У найпростішому випадку можливо встановити однозначну відповідність $X \equiv L_X$, тобто даний елемент X, що належить до КЗ (один з S, P, Obj, $Attr$, $Attr(Attr)$) на рівні тексту представлений окремим словом (лексемою) L_X. В такому випадку структура КЗ не виходить за рамки граматичної структури тексту, оскільки кожний елемент тексту (лексема) є також елементом КЗ.

Водночас, зустрічаються і складніші варіанти взаємодії між лексемами та елементами КЗ. Так, в англійській мові граматична конструкція «*have/has been*», хоча й складається з кількох окремих слів, є частинами форми однієї лексеми «*to be*», яка відповідає одному елементу КЗ, а саме предикатору $Pred$. Більш того, в різних мовах складні граматичні конструкції можуть бути представлені і як окремі лексеми, і як складені конструкції, які з точки зору правил граматики є одним словом, що робить визначення їх ролей як елементів КЗ лише на основі синтаксичного розбору неможливим [104].

Таким чином, первинним для визначення ролі даної лексеми у структурі знань ПМБЗ ІП є його роль у КЗ, тобто належність до одного з елементів КЗ та

зв’язки з іншими елементами, незалежно від представлення цієї лексеми у тексті – а отже, кожний з елементів КЗ може складатися з декількох лексем.

Узагальнимо структуру $Subj$ як

$$Subj = \{L_i\},\ i = \overline{1, N_l^S}, \tag{2.5}$$

де L_i – лексеми, що формують текстове представлення $Subj/Obj$;

N_l^S – кількість таких лексем.

Аналогічно визначимо структуру $Pred$

$$Pred = \{L_j\},\ j = \overline{1, N_l^P}, \tag{2.6}$$

де L_j – лексеми, що формують текстове представлення $Pred$;

N_l^P – кількість таких лексем.

Структура об’єкту, що описує атрибутивне оточення $Subj$ або $Pred$ є дещо складнішою, оскільки кожний з його елементів може бути як відсутній (якщо атрибути не вказані), так і містити декілька елементів. Таким чином, об’єкт $Attr(X)$, що описує атрибутивне оточення деякого елементу X у КЗ, складається з множини атрибутів елементу X в рамках даного КЗ, для кожного з яких визначена множина його мір.

Визначимо структуру об’єкту атрибутів як

$$Attr(X) = A_j,\ j = \overline{1, N_{attr}^X}, \tag{2.7}$$

де A_j – об’єкт, що описує окремий атрибут A;

N_{attr}^X – загальна кількість атрибутів X, певного елементу КЗ.

Визначимо загальну структуру атрибуту A_j як

$$A_j = \{a, M(a)\}, \tag{2.8}$$

де a – власне атрибут;

$M(a)$ – множина мір атрибуту a.

Визначимо множину мір атрибута a як

$$M(a) = m_k,\ k = \overline{1, N_m^a}, \tag{2.9}$$

де m – окрема міра атрибуту a;

N_m^a – загальна кількість мір атрибуту a для даного елементу X.

Кожному з атрибутів a та їх мірам m_{Ai} також можемо поставити у відповідність лексеми L, які представляють їх на текстовому рівні, аналогічно $Subj$ та $Pred$.

Представлена вище структура представляє собою формальну модель КЗ, яка пов’язує елементи КЗ, найменших об’єктів МПЗ ПМБЗ ІП та лексем, найменших відповідних об’єктів природномовного тексту.

Використання такої структури дозволяє перейти від аналізу зв’язків між абстрактною структурою знань ПМБЗ та граматичною структурою природномовного тексту до аналізу зв’язків між пов’язаними об’єктами – КЗ та їх елементами у ПМБЗ та лексемами у природномовному тексті.

2.2.2 Процедура виділення квантів знань у тексті

Опишемо загальну процедуру виділення КЗ з вихідного природномовного тексту.

Визначимо елементи, які вказують на наявність КЗ у природномовному тексті, а саме $Subj$, $Pred$ та відповідний зв'язок між ними.

Перевіримо, чи їх наявності у природномовному тексті достатньо для висновку про наявність повного КЗ.

Твердження 1: *фрагмент тексту, що містить 1 суб’єкт та 1 предикатор, можливо представити у вигляді КЗ.*

Доведення:

Структура окремого КЗ є похідною від структури БССС, а саме ядро будь-якої БССС представляє собою КЗ. За визначенням, БССС представляє собою схему опису ситуації сенсорного рівня, а отже, будь-якому текстовому опису ситуації можливо поставити у відповідність хоча б одну еквівалентну БССС. Таким чином, якщо у фрагменті тексту тексті встановлено відношення, яке визначає деякі його елементи як $Subj$ та $Pred$, з такого тексту можливо виділити КЗ, що складається з цих $Subj$ та $Pred$, тобто:

$$\forall (Subj, Pred) \in T \mid \exists\, Q : Q = \{Subj, Pred\}, \qquad (2.10)$$

де T – довільний фрагмент тексту;

$Subj, Pred$ – елементи T, що пов'язані як $Subj$ та $Pred$;

Q – КЗ, що складається з $Subj$ та $Pred$.■

На рівні природномовного тексту елементам $Subj$ та $Pred$ відповідають лексеми, що мають граматичні ролі «іменник» та «дієслово» відповідно і є синтаксично пов'язаними – а отже, для пошуку КЗ у тексті необхідно ідентифікувати відповідні їм елементи тексту [105]. Крім того на цьому ж етапі слід виділити елементи, що відповідають $Attr$ та $Attr(Attr)$) – граматичні класи «прикметник» та «прислівник» відповідно.

Обидві ці задачі є принципово вирішеними в ОПМ, а отже, вважаємо що їх виконує зовнішній відносно ПМБЗ ІП синтаксичний аналізатор, або ж на вхід подається адаптований текст.

Після ідентифікації усіх можливих елементів КЗ потрібно визначити загальну кількість повних КЗ у фрагменті тексту. Для цього необхідно вирішити проблему неповних КЗ, коли у тексті є $Subj$ або $Pred$, для яких відсутня друга частина КЗ.

Фрагмент тексту, що не містить суб'єктів або предикаторів, на рівні тексту відповідає неповному реченню, яке не має підмету або присудку. Таке речення в загальному випадку можемо представити як структуру знань вигляду $K = (Subj)$ або $K = (Pred)$, що є недостатнім для побудови КЗ, який складається з двох частин, $Subj$ та $Pred$.

У випадку, коли в тексті присутній лише $Subj$, можливо розширити його прихованим предикатором «є», що дозволяє в подальшому розширити його до повноцінного КЗ при виникненні відповідного $Pred$, або позначити як елемент Obj монопредикатної БССС. Водночас, КЗ не може містити лише $Pred$ – а отже, фрагменти КЗ без відповідних їм $Pred$ при аналізі ігноруємо.

Іншим особливим випадком є наявність у тексті множини взаємопов'язаних однорідних $Subj$ та $Pred$, зв'язки між якими повторюються.

Твердження 2: *фрагмент тексту, що містить довільну кількість однорідних суб'єктів або предикаторів, можливо представити у вигляді множини КЗ.*

Доведення:

Визначимо фрагмент знань K фрагменту тексту T як таку сукупність елементів ПМБЗ (суб'єкти, предикатори та відношення), яка описує цей фрагмент тексту:

$$T{\sim}K : K \supset (Subj, Pred, R), \tag{2.11}$$

де T – фрагмент тексту;

K – відповідний йому фрагмент знань.

Структура знань фрагменту тексту з одним суб'єктом та одним предикатом речення еквівалентна одному КЗ.

$$K \equiv Q \equiv \{Subj, Pred\}, \tag{2.12}$$

Фрагмент тексту, який містить декілька елементів, ролі яких у тексті відповідають ролям суб'єктів та предикаторів у КЗ, в загальному випадку має наступну структуру:

$$K = (Subj_1, Subj_2, \dots Subj_n, \tag{2.13}$$
$$Pred_1, Pred_2, \dots, Pred_m),$$

де n, m – кількість $Subj$ та $Pred$ у даному фрагменті тексту.

В загальному випадку така структура представляє собою БССС поліпредикатного рівня, де кожна пара $Subj/Pred$ формує окремий КЗ, а усі КЗ у фрагменті тексту пов'язані відношеннями, що ускладнює її аналіз.

Особливим випадком є фрагмент тексту, в якому присутній лише один $Subj$ та кілька однорідних $Pred$ (або, навпаки, кілька однорідних $Subj$ та лише один $Pred$).

Розглянемо фрагмент знань

$$K = (Subj, Pred_1, Pred_2, \dots, Pred_n), \tag{2.14}$$

де n – кількість предикаторів у фрагменті тексту.

Цей фрагмент знань описує фрагмент тексту з одним суб'єктом та множиною пов'язаних з ним однорідних предикаторів, що на рівні тексту відповідає реченню з однорідними присудками, пов'язаними з одним підметом. Так, речення з однорідними присудками: *«діти бігали, кричали, метушилися»* еквівалентно сукупності речень *«діти бігали»*, *«діти кричали»*, *«діти метушилися»*.

На рівні тексту однорідні підмети або присудки можна розбити на множину подібних речень з одним підметом та одним присудком, в яких однорідний член складного речення повторюється, а отже наведений вище фрагмент знань K можемо представити як множину фрагментів знань K_i

$$K = K_1 \cup K_2 \cup \ldots \cup K_n \tag{2.15}$$
$$K_i = (Subj, Pred_i),$$

Враховуючи еквівалентність, встановлену у формулі (2.12), здійснюємо підстановку

$$K_i = (Subj, Pred_i) \rightarrow\ Q_i = (Subj, Pred_i), \tag{2.16}$$

де K_i – фрагмент знань, що входить до K;

Q_i – КЗ, еквівалентний K_i.

Це дозволяє привести K до вигляду

$$K = Q_1 \cup Q_2 \cup \ldots \cup Q_n, \tag{2.17}$$

Так само, речення з однорідними підметами (множина $Subj$) можемо представити як

$$K = (Subj_1, Subj_2, \ldots, Subj_n, Pred), \tag{2.18}$$

Цей фрагмент знань, так само як показано у формулах (2.14) – (2.15), еквівалентний множині квантів знань

$$K_i = (Subj_i, Pred) \rightarrow\ Q_i = (Subj_i, Pred) \tag{2.19}$$
$$K = Q_1 \cup Q_2 \cup \ldots \cup Q_n,$$

Отже, фрагмент тексту, що містить довільну кількість однорідних суб'єктів або предикаторів, можливо представити на рівні монопредикатних КЗ ПМБЗ. ■

Зазначимо, що наявність у фрагменті тексту однорідних суб'єктів та однорідних предикаторів одночасно робить неможливим однозначне його

представлення у вигляді КЗ. Натомість, структура знань такого речення (наведена у формулі (2.13)) при розкладанні спочатку за суб'єктами, а далі за предикаторами, набуває такого вигляду:

$$K = (K_{11} \cup K_{12} \dots \cup K_{1n}) \cup (K_{21} \cup K_{22} \dots \cup K_{2n}) \cup \dots \cup (K_{\mathrm{m}1} \cup K_{\mathrm{m}2} \dots \cup K_{mn}), \quad K_{ij} = (Subj_i, Pred_j), \tag{2.20}$$

де m, n – кількість суб'єктів та предикаторів у фрагменті тексту відповідно.

Отже, структура знань речення з однорідними підметами і присудками є декартовим добутком множин його $Subj$ та $Pred$, подальше спрощення якого виходить за рамки структури представлення знань та вимагає застосування методів семантичного аналізу.

З отриманої таким чином множини $Subj$ та $Pred$, шляхом попарного об'єднання, формуємо сукупність усіх можливих КЗ. Для КЗ, у яких немає відповідного $Pred$, додаємо умовний предикатор «є»; КЗ, у яких немає відповідного $Subj$, ігноруємо. Зазначимо, що процедура формування КЗ з тексту є жадібним – тобто, результатом його виконання є уся множина граматично можливих КЗ, навіть тих, які в даному фрагменті тексту не мають семантичного змісту.

Нижче представлено повну послідовність дій виділення множини КЗ з тексту.

1. Знайти усі лексеми, що відповідають елементам КЗ.
   - 1.1. Знайти лексеми, що мають граматичний клас «іменник» ($Subj$).
   - 1.2. Знайти лексеми, що мають граматичний клас «дієслово» ($Pred$).
   - 1.3. Знайти лексеми, що мають граматичний клас «прикметник» та «прислівник» ($Attr, Attr(Attr)$).
2. Визначити можливі варіанти формування КЗ.
   - 2.1. Для кожного $Subj$:
     - 2.1.1. Для кожного $Pred$, що може бути пов'язаний з цим $Subj$, створити новий КЗ.

2.1.2. Якщо немає жодного $Pred$, створити новий КЗ з даним $Subj$ та $Pred$ «є».

2.2. Для кожного $Attr$:

2.2.1. Додати до цього $Attr$ усі $Attr(Attr)$, що можуть бути з ним пов'язані.

2.2.2. Додати цей $Attr$ до кожного КЗ, до якого він може належати.

3. Повернути усі сформовані КЗ.

2.3 Розробка моделі відношення

2.3.1 Формалізація відношення як об'єкту ПМБЗ ІП

Опишемо наступний рівень МПЗ, а саме відношення, які поєднують окремі КЗ у спільну структуру знань.

Відношення – це логічні зв'язки між об'єктами сенсорного світу. В рамках ІП відношення БССС поділяються на:

- монопредикатні відношення, що пов'язують одне ядро БССС та декілька елементів типу Obj, кожний з яких є еквівалентним частині КЗ $Subj$, що не має пов'язаного $Pred$;
- поліпредикатні відношення, що пов'язують декілька КЗ у поліпредикатну БССС.

Враховуючи встановлену вище відповідність між КЗ, що не містить $Pred$ та КЗ, що містить прихований $Pred$ «є», можемо стверджувати, що монопредикатні відношення є еквівалентними поліпредикатним з прихованим предикатором [106]. Використання монопредикатних відношень є виправданим для аналізу текстів, в той час як для автоматизації ПМБЗ ІП уніфікований вигляд подібних елементів є більш важливим ніж менша складність структури знань.

Визначимо відношення як

$$R = \{Q_1, Q_2, D\}, \tag{2.21}$$

де Q_1, Q_2 – це КЗ, між якими встановлено відношення;

D – додаткова інформація, що описує саме відношення.

Проаналізуємо детальніше, яку саме інформацію зберігає об’єкт відношення.

В першу чергу, відношення містить інформацію щодо його представлення на рівні природномовного тексту. Нехай $g \in D$ – це граматична конструкція або правило, яке ідентифікує дане відношення R у тексті.

Структура відношення з урахуванням цього має наступний вигляд:

$$R = \{Q_1, Q_2, \{g\}\}, \tag{2.22}$$

Зауважимо що деякі відношення є тривіальними і не вимагають додаткових засобів для відтворення на рівні тексту. В такому випадку це окремо зазначено у полі g аналогічно до інших варіантів.

Наступний змістовний елемент об’єкту відношення – це інформація про його належність до предикативних або ситуаційних відношень БССС. Це впливає не лише на семантичне навантаження відношення у БЗ, але й на його граматичну структуру g, а отже – ця інформація є необхідним елементом.

$$R = \{Q_1, Q_2, \{g, t\}\}, \tag{2.23}$$

Крім того, для зберігання відношення у БЗ потрібно враховувати його напрям, оскільки відношення одного й того ж типу може означати різні види зв’язків між КЗ.

На рис.2.2 наведено варіанти напряму відношення між КЗ у ПМБЗ. Це (1) пряме відношення $Q_1 \rightarrow Q_2$, (2) обернене відношення $Q_1 \leftarrow Q_2$, (3) двонапрямлене відношення $Q_1 \leftrightarrow Q_2$ та (4) рекурсивне відношення $Q_1 \leftrightarrow Q_1$ – особливий випадок двонапрямленого відношення, коли КЗ Q_1 має відношення з самим собою.

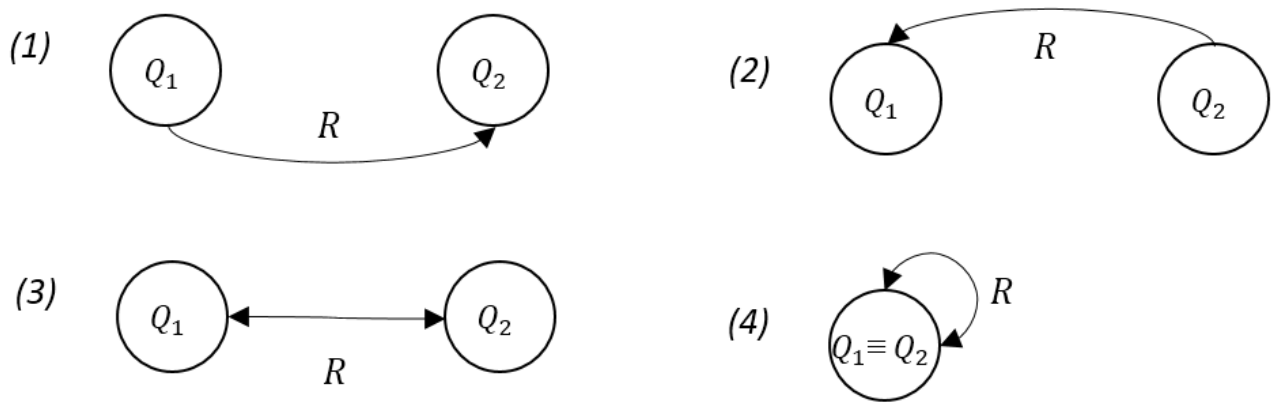


Рисунок 2.2 – Варіанти структури відношення між КЗ

Позначимо інформацію про напрям відношення як $d \in D$. Тоді формула відношення набуває наступного вигляду:

$$R = \{Q_1, Q_2, \{g, t, d\}\}, \quad (2.24)$$

Структура об’єкту відношення, представлена у формулі (2.24), дозволяє в повній мірі описати відношення між квантами знань у БЗ і, таким чином, відтворити модель знань фрагменту тексту, що містить два КЗ, зв’язані відношенням – найпростішу БССС поліпредикатного рівня.

2.3.2 Процедура виділення відношень у тексті

Опишемо загальну процедуру виділення відношень з вихідного природномовного тексту.

На рівні тексту відношення визначаються певними ключовими словами або граматичними правилами, а отже, так само як ідентифікацію лексем для виділення з тексту окремих КЗ, цю задачу виконує зовнішній ЛП на основі синтаксичного аналізу тексту та множини правил, що визначають різні типи відношень. Отримані в результаті цього ідентифікатори відношень необхідно поєднати з отриманою раніше множиною КЗ.

Найпростішим варіантом їх взаємодії є повна структура відношення, що поєднує два КЗ. Фрагмент знань, що описує таку структуру, виглядає наступним чином:

$$K = \{Q_1, R, Q_2\}, \tag{2.25}$$

Такий фрагмент знань може описати довільний фрагмент тексту, що складається з двох КЗ та відношення. Водночас, для природномовних текстів є типовими більш складні структури відношень між КЗ.

Спершу проаналізуємо варіант відношення, що у тексті прив'язане до лише одного КЗ.

Так само як неповний КЗ, що містить лише $Subj$, може бути доповнений прихованим предикатором «є», відношення може бути зв'язане з асоціативним або загальним контекстом даного фрагменту тексту. Це, відповідно:

- відношення до прихованої «поточної ситуації» – наприклад, «дихати глибше» у контексті «глибше, ніж зараз»;
- відношення до попередньо визначеної ситуації – наприклад, «дощ піде пізніше» у контексті «пізніше, ніж прогнозували»;

Таким чином, можемо стверджувати що неповне відношення $\{Q_1, R\}$ відповідає повному відношенню $\{Q_1, R, Q_2\}$, в якому з метою економії мовних засобів пропущене повторення другого КЗ, але відтворити цю залежність можливо тільки за умови контекстуальної неповноти. Якщо другий КЗ відсутній у текстовому оточенні відношення і не може бути визначений за замовчанням з поточної ситуації, таке відношення не несе смислового навантаження і може бути проігнороване.

Інший складний варіант взаємодії між елементами відношення – коли на одному й тому ж КЗ побудовано декілька відношень з іншими КЗ. На відміну від однорідних елементів КЗ, які можуть бути пов'язані лише в один спосіб, для відношень існує два способи взаємодії: горизонтально та вертикально пов'язані відношення.

Горизонтальна взаємодія виникає, коли декілька відношень включають у себе один і той самий КЗ в одній і тій самій ролі:

$$\{Q_0, R_1, Q_1\}, \{Q_0, R_2, Q_2\}, \dots, \{Q_0, R_i, Q_i\}, \dots, \quad (2.26)$$

Перевіримо, чи будь-яку структуру відношень між КЗ у фрагменті тексту можливо описати як множину окремих відношень. Почнемо з горизонтальної взаємодії відношень.

Твердження 3: *довільну кількість відношень, спільним першим елементом яких є даний КЗ, можливо описати як множину окремих відношень.*

Доведення:

Розглянемо структуру знань вигляду, що відповідає фрагменту тексту, в якому описані різні відношення одного КЗ.

Позначимо вихідний КЗ як Q_0, а його відношення - $\{R_i, Q_i\}$.

$$K = \{Q_0, \{R_1, Q_1\}, \{R_2, Q_2\}, \dots, \{R_i, Q_i\}, \dots\}, \quad (2.27)$$

На рівні тексту такі відношення можливо як об’єднати в одному реченні, так і виділити кожний окремо (наприклад, через зв’язки між реченнями, як-то *«цей»*, *«той»*, *«вищезгаданий»* і т.д.). Відповідно, фрагмент знань, в якому відношення мають спільний перший КЗ, можемо шляхом повторення цього КЗ розкласти на окремі фрагменти тексту, кожний з яких містить лише одне відношення:

$$\begin{gathered} K = K_1 \cup K_2 \cup \dots \cup K_i \cup \dots \\ K_i = \{Q_0, R_i, Q_{i+1}\}, \end{gathered} \quad (2.28)$$

що й треба було довести. ■

Тобто, якщо на одному КЗ побудовано декілька відношень, цей КЗ дублюється у кожне з них, в той час як у структурі знань він є їх спільним елементом.

Вертикальна взаємодія виникає, коли відношення поєднані «ланцюжком», де кожний з КЗ поєднує попереднє та наступне відношення:

$$\{Q_0, R_1, Q_1\}, \dots, \{Q_{i-1}, R_i, Q_i\}, \{Q_i, R_{i+1}, Q_{i+1}\}, \dots, \quad (2.29)$$

Твердження 4: *довільна кількість квантів знань, пов'язаних ланцюжком відношень, розкладається на відповідну кількість окремих відношень.*

Доведення:

Розглянемо фрагмент знань, в якому поєднано декілька відношень на різних рівнях – такий, що відповідає фрагменту тексту *«студент С групи Г кафедри К»*.

Структура знань такого тексту виглядає наступним чином:

$$K = \{Q_0, R_1, Q_1, R_2, Q_2, \dots, Q_{i-1}, R_i, Q_i, R_{i+1}, \dots\}, \quad (2.30)$$

Скорочені відношення у такому фрагменті тексту можна розкрити шляхом повтору відповідних КЗ як *«студент С, що належить до групи Г, що належить до кафедри К».*

Таку структуру можемо розкласти так само, як показано вище (*Твердження 3*) для горизонтальних відношень: *«студент С належить до групи Г», «група Г належить до кафедри К»* […].

Цей фрагмент тексту, в свою чергу, має наступну структуру:

$$K = \{Q_0, R_1, Q_1\}, \dots, \{Q_{i-1}, R_i, Q_i\}, \quad (2.31)$$
$$\{Q_i, R_{i+1}, Q_{i+1}\}, \dots,$$

Отже, якщо довільна кількість КЗ попарно з'єднана відношеннями, то така структура знань еквівалентна відповідній кількості окремих відношень, представлених згідно (2.21). ■

Зауважимо, що це справедливо і для рекурсивного поєднання КЗ через відношення: так, навіть текст вигляду «*дім, що збудував Джек…*» може бути скорочений до вигляду, в якому між будь-якими двома його КЗ є лише один зв'язок.

Нижче представлено повну процедуру виділення відношень з тексту.

1. Знайти усі граматичні конструкції, що відповідають відомим відношенням
2. Для кожного відношення знайти відповідні йому КЗ:
   - 2.1. Якщо не знайдено жодного КЗ, ігнорувати відношення
   - 2.2. Якщо знайдено один КЗ, побудувати відношення з посиланням на поточну ситуацію
   - 2.3. Якщо знайдено два КЗ, побудувати повне відношення

2.4. Якщо знайдено більше, ніж два КЗ, побудувати повне відношення з кожним з них

3. Додати кожне відношення у мережу КЗ як об’єкт, що поєднує відповідні КЗ
4. Повернути множину КЗ та відношень

2.4 Розробка моделі представлення знань ПМБЗ ІП

2.4.1 Формалізація моделі представлення знань

Опишемо загальний вигляд фрагменту знань ПМБЗ ІП, що описує структуру знань, відповідну довільному фрагменту природномовного тексту.

Фрагмент знань складається з сукупності КЗ та відношень між ними. Для довільного фрагменту тексту T структура знань приймає наступний вигляд:

$$K_T = \{(Q_1, Q_2, \dots, Q_i, \dots), (R_1, R_2, \dots, R_j, \dots)\}, \quad (2.32)$$

де T – фрагмент тексту;

K – фрагмент знань, що відповідає фрагменту тексту T;

Q_i – КЗ, що належать до K;

R_j – відношення між КЗ у K.

Оскільки КЗ є найменшим змістовним елементом ПМБЗ ІП (як показано у п. 2.2), а відношення є об’єктом заданої структури, що поєднує два КЗ (формула (2.21)), можемо стверджувати, що структура довільного фрагменту знань ПМБЗ ІП представляє собою об’єктний граф [107], тобто граф вигляду $G = (V, E)$, вершинами V якого є множина КЗ, а ребрами E – множина відношень.

Опишемо структуру об’єктного графу, що описує структуру довільного фрагменту знань ПМБЗ ІП:

$$K = (Q, R), \quad (2.33)$$

де Q – множина КЗ у ПМБЗ ІП, які є вершинами графу знань;

R – множина відношень ПМБЗ ІП, які є ребрами графу знань.

Представлення природномовних знань у вигляді графу відповідає загальній структурі природномовного тексту: однією з визначальних характеристик природномовного тексту є його зв’язність, тобто елементи тексту належать до спільної структури (речення, абзацу тощо), яка визначає контекст для їх інтерпретації. Виходячи з цього, дамо формальне визначення фрагменту знань.

Фрагмент знань K_T, який описує структуру знань фрагменту тексту T, що є частиною наповнення ПМБЗ ІП – це зв’язний граф, що є підграфом загального графу знань ПМБЗ:

$$K_T \subseteq K, \tag{2.34}$$

При цьому підграф K_T формально визначає контекст відповідних фрагментів знань, які до нього належать. Тобто, для будь-якого підграфу ПМБЗ ІП може бути встановлена формальна залежність з деяким зовнішнім відносно ПМБЗ об’єктом. Це дозволяє ПМБЗ ІП взаємодіяти з будь-якими іншими технологіями ОПМ, в тому числі системами синтаксичного та семантичного аналізу, не змінюючи структуру представлення знань та попередньо накопичені знання.

Проаналізуємо усі рівні представлення знань, які використовуються ПМБЗ ІП, та встановимо зв’язки, що дозволяють об’єднати їх у єдину систему.

Модель представлення знань ПМБЗ ІП ґрунтується на концепції БССС, причому між відповідними їх елементами можливо встановити однозначну відповідність. Аналогічно, структура БССС тісно пов’язана зі структурою природномовного тексту, і між ними також існує формальний зв'язок. Таким чином, ПМБЗ ІП виступає в ролі посередника між природномовним текстом та формальною структурою знань, що дозволяє використовувати спільну МПЗ для різних природномовних текстів – зокрема, різної мови походження, стилістики тощо [108].

Водночас, МПЗ на основі БССС не містить власних семантичних зв’язків високого порядку – а отже, до мережі знань ПМБЗ ІП у вигляді графу для обробки знань може бути формально підключена довільна зовнішня система обробки метаданих.

Отже, ПМБЗ ІП як загальна структура знань встановлює формальні зв’язки на структурному рівні між, з одного боку, синтаксичною структурою природномовного тексту та, з іншого боку, семантичною структурою його знань. Це дозволяє поєднувати їх для виконання базових операцій над знаннями, таких як записування або пошук у БЗ, що, в свою чергу, відкриває можливості для використання ПМБЗ ІП для усього спектру задач ОПМ.

2.4.2 Дослідження особливостей моделі представлення знань ПМБЗ ІП

Оцінимо, які відмінності запропонована МПЗ має у порівнянні з існуючими системами.

Зв'язок знань з текстом у ПМБЗ ІП реалізовано шляхом виділення з тексту окремих лексем, так само як в таких технологіях аналізу тексту як мережі суміжності [109] та *n*-грами [55]. На відміну від цих технологій, де кожна лексема є незалежною сутністю, а об’єктом роботи виступає уся сукупність зв’язків даної лексеми, у ПМБЗ ІП визначена фіксована структура окремого КЗ, в рамках якого лексема виступає в якості одного з відомих елементів, зв’язки яких описані та формалізовані. Сам КЗ в загальному вигляді складається з суб’єкта та предикатора, які є рівноправними складовими КЗ, включно з їх атрибутивним оточенням – що відрізняє його від елементів знань на основі предиката Н. Хомського [83], де предикат є основою елементу знань, структура якого є похідною від синтаксичної структури тексту.

Згідно з принципами ІП, і структура КЗ, і природномовний текст є похідними від особливостей організації нейромережі людини, а отже КЗ, як вербалізація деякої ситуації образного рівня, є первинною відносно структури тексту. При цьому необхідно відзначити явні паралелі та підкреслити відмінності між ПМБЗ ІП та класичними штучними нейронними мережами [110] – двома технологіями, в основі яких лежить моделювання окремих частин та функцій нервової мережі живого організму. Як штучні нейронні мережі, так і ПМБЗ ІП в першу чергу

спираються на моделювання особливостей нейрофізіології мислення людини; але найменшим об’єктом роботи штучної нейронної мережі є окремий нейрон, роль якого в мережі не залежить від його внутрішньої структури, в той час як у ПМБЗ ІП внутрішня структура кожного окремого КЗ безпосередньо впливає на його роль та можливості.

Відповідно, ПМБЗ ІП представляє собою загальну модель знань людини, розроблену для вирішення проблем ОПМ, а не модель знань природномовного тексту як незалежного об’єкту дослідження. Таким чином, МПЗ ПМБЗ ІП:

- формально визначає КЗ як структуру елементу знань ПМБЗ;
- формально визначає зв’язки між елементами КЗ та елементами природномовного тексту;
- не залежить від синтаксичної структури тексту.

Зв’язки між елементами знань у ПМБЗ ІП представляють відношення – логічні зв’язки між КЗ, що поєднують їх у загальну мережу знань. Схожим чином мережа знань формується у семантичних або фреймових мережах, де відношення встановлюються між концептами [111] та онтологіях, де іменовані відношення поєднують фрейми відомої структури [112]. Найближчим аналогом ПМБЗ ІП є семантичні мережі на основі рекурсивних фреймів [81], що поєднує плюси онтологій та семантичних мереж, а саме складається з елементів знань чіткої структури і при цьому за рахунок рекурсивної організації дозволяє гнучкі зв’язки між ними.

На відміну від цих технологій, знання у ПМБЗ ІП можуть бути на структурному рівні поєднані з відповідними елементами природномовного тексту, в той час як точна структура кожного окремого КЗ допускає різні варіанти взаємодії її складових елементів – але при цьому структура об’єктів відношень ПМБЗ ІП дозволяє використовувати у ній методи роботи з семантичними мережами. Зазначимо, що наповнення ПМБЗ ІП в загальному вигляді може бути представлена у вигляді орієнтованого графу вигляду «тісний світ» (*small world network*) [113], кожна вершина якого має порівняно невелику кількість ребер, але при цьому від

будь-якої з вершин можливо побудувати досить короткий шлях до будь-якої іншої вершини

Семантичні відношення та зв’язки між елементами знань у ПМБЗ ІП реалізовані так само, як і зв’язки з природномовним текстом, а саме шляхом поєднання елементів мережі знань ПМБЗ ІП з довільною зовнішньою системою обробки метаданих. На відміну від аналогів, таких як існуючі семантичні мережі та онтології, ПМБЗ ІП зберігає лише структуру знань, а семантичні зв’язки між ними належать до зовнішньої структури, що може бути прив’язана до структури знань ПМБЗ ІП. Таким чином, МПЗ ПМБЗ ІП:

- не залежить від семантичного оточення певного концепту;
- зберігає повну мережу знань для подальшої обробки;
- дозволяє прив’язати на структурному рівні довільну зовнішню структуру метаданих до власної мережі знань.

Отже, описана вище МПЗ має наступні особливості:

- формально визначає КЗ як структуру елементу знань ПМБЗ;
- формально визначає зв’язки між елементами КЗ та елементами природномовного тексту та водночас дозволяє прив’язати на структурному рівні довільну зовнішню структуру метаданих до власної мережі знань.
- не залежить від синтаксичної структури тексту або семантичного оточення певного концепту;
- зберігає повну мережу знань для подальшої обробки;
- може використовувати існуючі методи роботи з семантичними мережами;

2.5 Розробка методу обробки природномовних текстів з використанням ПМБЗ ІП

Метод обробки природномовних текстів включає у себе два класи задач – задачі синтезу та аналізу. Для їх вирішення використовується ПМБЗ ІП, яка оперує

незалежними об’єктами – природномовним текстом та знаннями. В цьому контексті ці задачі представлені:

– при аналізі – перетворенням «текст-знання»;
– при синтезі – перетворенням «знання-текст».

Основною з цих двох задач є задача аналізу, оскільки при цьому відбувається виділення знання з тексту, який може не містити усієї необхідної для цього інформації. Водночас, задача синтезу зводиться або до відновлення вихідного тексту, що можливо забезпечити його зберіганням у початковому вигляді, або до задачі генерації одного варіанту представлення знань згідно визначених правил.

Етапами обробки природної мови при перетворенні «текст-знання» є:

– синтаксичний аналіз;
– виділення КЗ;
– виділення відношень;
– формування фрагменту знань.

Розглянемо ці етапи детальніше.

2.5.1 Синтаксичний аналіз

Початковий етап обробки природної мови це перетворення тексту в формалізований вигляд.

Мета цього етапу – з вихідного тексту, представленого у вигляді послідовності символів, отримати його синтаксичну структуру.

Крок 1. Синтаксична структура.

Вхідні дані на цьому кроці – текст T_0, сформований згідно граматичних правил деякої природної мови. Таким текстом може бути фрагмент документу, запит користувача тощо. При цьому вхідний текст повинен відповідати правилам граматики відповідної мови, що є необхідною умовою для коректного синтаксичного аналізу, та містити семантичне навантаження, тобто бути змістовним.

Задача, яка вирішується на цьому етапі, це нормалізація текст T_0 до тексту T, тобто виділення змістовних елементів та формування на основі них синтаксичної структури S, що включає в себе слова як елементи тексту, їх граматичні форми, порядок у тексті та синтаксичні зв'язки та граматичні ролі.

$$T_0 \to T, T \to S \tag{2.35}$$
$$S\colon (L, L_i, d)$$

Цю задачу виконує незалежна від ПМБЗ система – синтаксичний аналізатор, спеціалізований для даної мови, як-то *pyMorphy, Stanford Parser* тощо.

Отримана в результаті синтаксична структура S містить так дані:

– лексеми L, в тому числі їх граматичні форми;
– індекси слів w_i, які визначають порядок даного слова у вхідному тексті;
– синтаксичні зв'язки між словами, які визначають належність слів до однієї з груп (*Noun, Verb, Adj, Adv*).

Крок 2. Маркер.

З метою прив'язки фрагменту вхідного тексту T_0 до отриманої з нього синтаксичної структури S, до кожної структури S додається маркер M, який пов'язує слова вхідного тексту w та лексеми синтаксичної структури L.

$$M(T, S)\colon \tag{2.36}$$
$$T(w, w_i) \leftrightarrow M \leftrightarrow S(L, L_i)$$

Це дозволяє формально поєднати елементи вхідного тексту та відповідні їм лексеми, які є елементами мережі знань.

Цю задачу виконує ЛП на основі результатів синтаксичного аналізу. Отримана в результаті структура містить усі необхідні матеріали для формування квантів знань та відношень між ними.

2.5.2 Виділення квантів знань та відношень

Наступний етап це створення та наповнення мережі знань на основі отриманої раніше структури S.

Мета цього етапу – сформувати кванти знань та відношення, які відтворюють структуру знань даного природномовного тексту, та заповнити їх з отриманої раніше синтаксичної структури.

Ці задачі та подальше формування КЗ в загальному випадку виконує ЛП згідно процедур, представлених у підрозділах 2.2, 2.3 з урахуванням особливостей синтаксичного аналізу мови вхідного тексту.

Крок 3. Кванти знань.

На основі лексем, що мають граматичний клас $Noun$ з синтаксичної структури S, ЛП формує кванти знань, в кожному з яких $Noun$ має роль $Subj$.

$$Q = \{\} \tag{2.37}$$
$$Q \leftarrow Noun$$
$$Q = \{Subj(Noun)\}$$

Згідно з твердженням (2) у підрозділі 2.2, отримана в результаті множина квантів знань включає усі кванти знань вхідного тексту, причому кожний з них містить лише один $Subj$.

Далі зі структури S виділяються усі лексеми, що мають граматичний клас $Verb$, які надалі записуються у відповідні їм КЗ. Ця відповідність встановлюється на основі збігу граматичних форм $Noun$ та $Verb$: якщо такий збіг має місце, $Verb$ записується у КЗ як $Pred$.

$$Q = \{Subj(Noun)\} \tag{2.38}$$
$$if(form(Verb) = form(Noun)) : Q \leftarrow Verb$$
$$Q = \{Subj(Noun), Pred(Verb)\}$$

Оскільки в тексті може бути декілька варіантів граматично коректної інтерпретації синтаксичної структури S, що призводить до неоднозначності в ідентифікації зв'язків у КЗ, усі конкуруючі варіанти зберігаються як альтернативні.

$$\forall\, Verb : Q_1(Noun_1, Verb); Q_2(Noun_2, Verb); \tag{2.39}$$
$$K \ni Q_1, Q_2$$

Доповнення КЗ атрибутивним оточенням відбувається аналогічно.

$$Q = \{Subj(Noun), Pred(Verb)\} \tag{2.40}$$

$$if\big(form(Adj) = form(Noun)\big) : Q \leftarrow Adj$$
$$if\big(form(Adv) = form(Adj)\big) : Q \leftarrow Adv$$
$$if\big(form(Adv) = form(Verb)\big) : Q \leftarrow Adv$$
$$if\big(form(Adv) = form(Adv)\big) : Q \leftarrow Adv$$

Таким чином, на основі лексем L синтаксичної структури S, які мають граматичні форми $Noun, Verb, Adj, Adv$, формується повний квант знань.

$$Q = \begin{Bmatrix} Subj(Noun), \\ Attr(Subj)(Adj), \\ Attr\big(Attr(Subj)\big)(Adv), \\ Pred(Verb), \\ Attr(Pred)(Adv) \\ Attr(Attr(Pred))(Adv) \end{Bmatrix} \quad (2.41)$$

При виникненні багатозначності, так само як і у випадку $Subj - Pred$ усі варіанти формування КЗ з тексту зберігаються як конкурентні варіанти.

Аналогічним чином відбувається формування відношень. Основою для цього є отримані вище кванти знань та ідентифікатори відношень – текстові структури або граматичні ознаки, які вказують на наявність відношення у тексті.

Крок 4: Відношення.

Для усіх Q та w, тобто КЗ та слів, які не входять у КЗ (як-то службові частини мови), визначається відповідність їх використання в тексті

Якщо знайдено граматичні ознаки або слова-ідентифікатори – визначити, які КЗ в нього входять, та побудувати відповідне відношення.

$$Q_1, Q_2, [w] \rightarrow R \quad (2.42)$$

Структура об’єкту відношення відповідає структурі, представленій у формулі (2.24). Зазначимо, що як при виділенні КЗ, при виділенні відношень якість обробки залежить від використаного синтаксичного аналізатора, а усі конкуруючі варіанти зберігаються для подальшої обробки.

2.5.3 Формування мережі знань та прив’язка її до тексту

Наступний етап це поєднання отриманих КЗ та відношень між ними у єдину мережу, придатну для зберігання у ПМБЗ. Оскільки ці об’єкти вже визначені на попередніх етапах, на цьому етапі забезпечується їх відповідність вимогам ПМБЗ, зокрема унікальність об’єктів та їх прив’язка до вхідних текстів.

Для створення мережі знань необхідно об’єднати усі однакові сутності – лексеми, кванти знань та фрагменти знань, зберігаючи їх зв’язки між собою та з вхідним текстом.

Крок 5: Формування мережі знань

З існуючих елементів формується єдина мережа знань, яка включає в себе по одному разу кожну унікальну лексему, кожний унікальний КЗ та кожне унікальне відношення.

Наповнення мережі відбувається згідно наступних правил.

Для лексем L.

1. Якщо дана лексема відсутня у ПМБЗ, записати її. Граматичні варіації лексеми при цьому зберігаються у словнику.

Для квантів знань Q, що складаються з лексем.

2. Обробити усі лексеми у складі КЗ згідно правилу (1).
3. Для кожної лексеми у складі КЗ та лексем, пов’язаних з нею: якщо у ПМБЗ відсутня еквівалентна комбінація лексем і їх ролей, записати її. Порядок обробки: $Subj \rightarrow Pred \rightarrow Attr \rightarrow Attr(Attr)$.
4. Якщо у ПМБЗ відсутній еквівалентний КЗ, записати його. Еквівалентним вважати КЗ, який містить ті самі лексеми у тих самих ролях, що й оригінальний КЗ.

Для відношень R, що містять посилання на кванти знань:

5. Обробити усі КЗ згідно правил (3,4).
6. Якщо у ПМБЗ є хоча б один з КЗ, представлених у відношенні – приєднати відношення до цього КЗ.

Результатом виконання цього кроку є мережа знань, структура якої відповідає моделі представлення знань ПМБЗ (2.33).

Крок 6: Формування зв’язків

Отримана мережа знань прив’язується на структурному рівні до визначених раніше структур та інших джерел інформації – синтаксичної структури вхідного тексту, окремих квантів знань, документів, додаткової семантичної інформації тощо.

Це відбувається згідно наступних правил.

1. Для кожного КЗ записати маркер M_Q, що містить унікальний ідентифікатор КЗ, посилання на лексеми, які в нього входять, їх граматичні форми та їх ролі у КЗ.
2. Для кожного відношення записати маркер M_R, що містить унікальний ідентифікатор відношення, посилання на КЗ, які в нього входять, додаткову інформацію про відношення (тип, напрям і т.д.).
3. Для кожного окремого фрагменту тексту записати маркер M_T, що містить унікальний ідентифікатор фрагменту тексту, посилання на КЗ та відношення, що в нього входять, порядок та граматичні форми лексем, посилання на джерело тексту та дані щодо нього, а також допоміжні елементи тексту, що не входять у структуру знань.

Результатом є ієрархічна структура маркерів $M_Q - M_R - M_T$ (квантів знань, відношень та фрагментів тексту відповідно), яка дозволяє прив’язати наповнення мережі знань до вхідних текстів від рівня лексеми до рівня усієї бази знань включно.

Це завершує метод обробки природномовних текстів на основі інтеграційного підходу. Процес формування мережі графічно представлено на ілюстрації (Додаток Г)

2.5.4 Аналіз характеристик отриманої мережі знань

Представлений вище метод обробки природномовних текстів дозволяє заповнити ПМБЗ даними згідно розробленої моделі представлення знань. Результатом обробки є мережа знань, що складається з квантів знань та відношень, та мережа маркерів, в яку входять маркери окремих лексем, маркери квантів знань та маркери фрагментів тексту.

Оцінимо отриману мережу на відповідність критеріям оцінки ПМБЗ.

Повнота. Наповнення ПМБЗ пов’язане з синтаксичною структурою вихідного тексту через лексеми, тобто існує однозначна відповідність між елементами мережі знань та елементами тексту. Таким чином, повнота ПМБЗ визначається можливостями використаного синтаксичного аналізатору і є високою. При виділенні знань з нового тексту може виникати багатозначність, зумовлена характеристиками синтаксичного аналізатору. Порядок такої багатозначності значно менший [114] за порядок варіантів вербалізації одного кванту знань. В такому випадку зберігаються усі можливі варіанти інтерпретації тексту, тобто втрата інформації не відбувається.

Отже, повнота ПМБЗ є високою.

Несуперечливість. Квант знань може частково збігатися з іншим, так само один квант знань може належати до різних фрагментів знань. Це дозволяє зберігати у мережі знань конфліктні елементи, тобто такі, які частково суперечать один одному. При цьому для унікальних елементів зберігаються маркери – тобто, зберігання конфліктних елементів не спричиняє негативного впливу на повноту наповнення.

Отже, несуперечливість ПМБЗ є високою.

Гнучкість. Фактично ПМБЗ включає у себе декілька рівнів пов’язаних об’єктів:

- мережа квантів знань, які включають у себе лексеми;
- мережа фрагментів знань, що складаються з квантів знань;

- вхідні тексти, прив’язані до мережі знань за допомогою маркерів.

Це дозволяє доповнювати та змінювати будь-який з цих рівнів без спотворення інших.

Отже, гнучкість ПМБЗ є високою.

Зазначимо, що властивості ПМБЗ зберігаються при її розширенні – як вниз, до рівня окремих символів та носіїв вхідного тексту, так і вгору, до систем метаданих та прикладних технологій обробки природної мови.

Висновки до розділу 2

1. На засадах інтеграційного підходу до моделювання мовленнєвої діяльності людини розроблено концепцію інформаційної технології обробки природномовних текстів на основі інтеграційного підходу та моделі елементів природномовної бази знань на основі інтеграційного підходу.
2. Розроблено та формалізовано моделі основних елементів знань для інформаційної технології, а саме кванти знань та відношення між ними. Теоретично доведено достатність цих елементів знань для представлення структури знань довільного природномовного тексту. Розроблено процедуру виділення об’єктів знань з тексту та метод обробки природномовних текстів на основі цих процедур.
3. На основі інтеграційного підходу до моделювання мовленнєвої діяльності людини розроблено модель представлення знань для використання в технологіях обробки природної мови. Ця модель відрізняється від аналогів тим, що дозволяє представити фрагмент знань довільного природномовного тексту у вигляді універсальної структури. Перевагою розробленої моделі представлення знань є її незалежність від синтаксичної структури тексту та семантичного контексту фрагменту знань.

4. Розроблено метод обробки природномовних текстів, який дозволяє виділити структуру знань довільного природномовного тексту та сформувати на її основі мережу знань згідно особливостей запропонованої моделі представлення знань.

3 РОЗРОБКА ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ ОБРОБКИ ПРИРОДНОМОВНИХ ТЕКСІВ НА ОСНОВІ ІНТЕГРАЦІЙНОГО ПІДХОДУ

У цьому розділі розроблено структурну схему інформаційної технології обробки природномовних текстів на основі інтеграційного підходу, виконано аналіз процесів обробки даних в інформаційній технології, зокрема визначено етапи роботи інформаційної технології в режимах аналізу та синтезу та розроблено процедури записування та пошуку природномовних знань у базі знань у складі інформаційної технології. Також наведено приклади використання інформаційної технології обробки природномовних текстів у прикладних задачах природномовного пошуку та машинного перекладу.

3.1 Структурна схема ІТ ОПМ

За визначенням, інформаційна технологія це комплекс методичного забезпечення та інструментів, які забезпечують виконання задач певного класу. Основними атрибутами ІТ є вхідні дані, операції, які над ними виконуються, та результати їх обробки (рис. 3.1).

Рисунок 3.1 – Загальна схема роботи інформаційної технології

В залежності від цих атрибутів, інформаційні технології поділяються на різні класи, зокрема за галузями застосування, за характером даних, які обробляються, за процесами обробки тощо.

Інформаційна технологія обробки природної мови це, в загальному випадку, така ІТ, яка приймає на вхід природномовну інформацію у відомому форматі і виконує над нею операції, пов’язані з її аналізом або синтезом. Як окремий випадок інформаційних технологій цього класу, ІТ ОПМ призначена для роботи з прикладними системами обробки природної мови, зокрема системами природномовного пошуку та машинного перекладу.

Відповідно, ІТ ОПМ має такі особливості, які визначають її місце серед інших технологій цього класу:

- вхідними даними є природномовний текст;
- обробка тексту виконується згідно методу, описаному у підрозділі 2.5;
- результатом роботи ІТ ОПМ є природномовні знання, модель яких відповідає моделі, описаній у підрозділі 2.3;
- результати роботи ІТ ОПМ орієнтовані на подальшу обробку іншими системами ОПМ.

Структурно ІТ ОПМ складається з окремих підсистем, представлених на рис. 3.2.

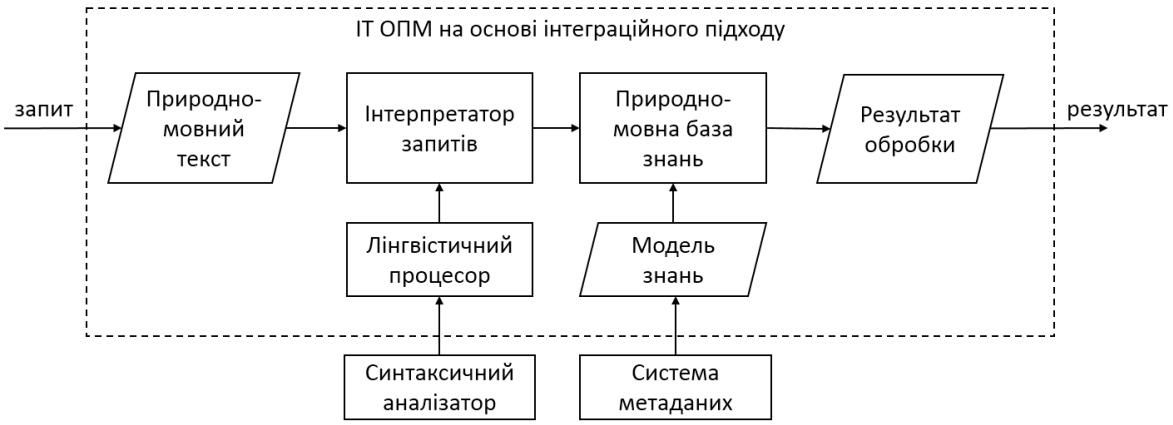


Рисунок 3.2 – Структурна схема інформаційної технології обробки природномовних текстів на основі інтеграційного підходу

Інтерпретатор запитів обробляє вхідні запити у вигляді природномовного тексту, виділяє їх структуру знань та перетворює їх на відповідні елементарні запити до ПМБЗ. В загальному випадку інтерпретатор підтримує базові запити (додавання, редагування, видалення та пошук у ПМБЗ), але за необхідності ця функціональність може бути розширена специфічними запитами, орієнтованими на певну прикладу задачу ОПМ. Ця підсистема є єдиною точкою входу в ІТ ОПМ; тобто, будь-які запити надходять виключно через інтерпретатор.

Лінгвістичний процесор виділяє структуру знань природномовного тексту з використанням методу обробки природної мови, наведеному у підрозділі 2.5. У своїй роботі лінгвістичний процесор використовує сторонній засіб, який не входить у саму ІТ ОПМ – синтаксичний аналізатор, орієнтований на деяку мову, який надає необхідні для роботи лінгвістичного процесора дані.

Природномовна база знань зберігає та надає доступ до природномовних знань, які формують базу знань ІТ ОПМ. Наповнення ПМБЗ організовано згідно моделі, описаній у підрозділі 2.4. У роботі з ПМБЗ за необхідності може використовуватися зовнішня система метаданих, яка виконує додаткову обробку результатів запитів до ПМБЗ – наприклад, відбирає з кожного набору знайдених результатів лише ті, які є найбільш релевантними.

Результатом обробки даних у ІТ ОПМ є структуровані природномовні знання, які надалі передаються у інші системи ОПМ.

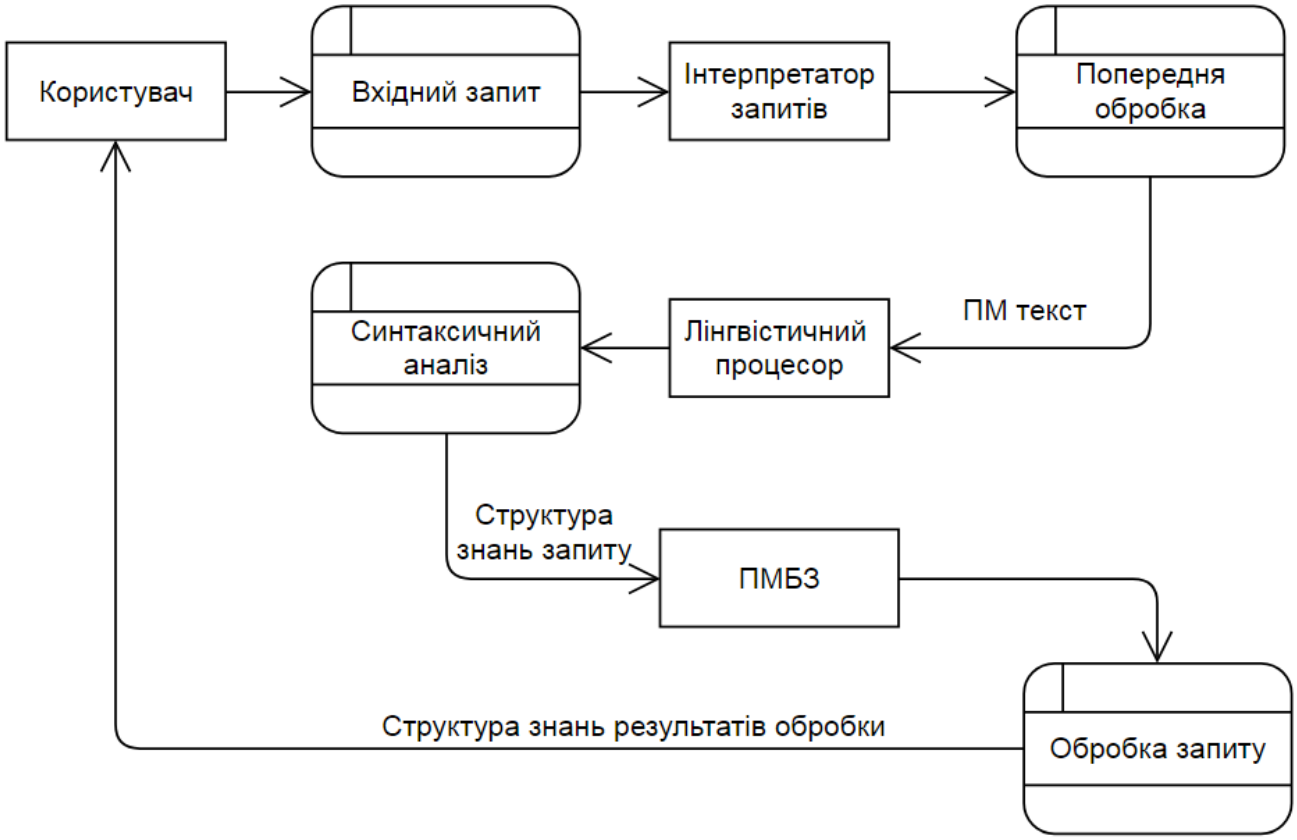
3.2 Аналіз процесів обробки даних в ІТ ОПМ

3.2.1 Аналіз потоків даних в ІТ ОПМ

Проаналізуємо потоки даних, які виникають між розглянутими вище (рис. 3.3) підсистемами ІТ ОПМ. Для цього розглянемо тільки ті кроки обробки, які є спільними для всіх задач ОПМ.

Діаграма потоків даних у форматі UML DFD представлена на рис. 3.3.

Рисунок 3.3 – Діаграма потоків даних у ІТ ОПМ



Вхідний запит надходить від користувача, яким може бути оператор або інша система. Вхідний запит має такі характеристики:

- формат даних запиту – природномовний текст, тобто семантично пов'язаний текст, в якому можливо виділити змістовні зв’язки;
- мова вхідного тексту доступна для обробки лінгвістичним процесором;
- відома задача обробки запиту.

Інтерпретатор запитів виконує попередню обробку отриманого запиту, тобто відокремлює текстові дані від включень даних інших видів, розбиває вхідний запит на менші частини тощо. Результатом обробки є природномовний текст, який має такі характеристики:

- фрагменти даних підготовлені для записування у базу даних;
- текст підготовлений для обробки лінгвістичним процесором.

В залежності від поставленої задачі і використаного лінгвістичного процесора, підготовлений текст може включати додаткові метадані – наприклад,

інформацію про мову кожного фрагменту тексту для вхідного запиту, який може містити тексти різними мовами.

Лінгвістичний процесор виділяє структуру знань вхідного запиту. Результатом обробки є структура знань, яка має таки характеристики:

- формат структури відповідає моделі, описаній у підрозділі 2.3;
- структура знань містить синтаксичну структуру фрагменту тексту, в тому числі посилання на кожне слово, його частину мови та граматичну форму;
- структура знань містить кванти знань та відношення між ними, отримані в результаті роботи ЛП;
- структура знань містить маркери, які пов’язують структуру з вхідним текстом.

Природномовна база знань виконує обробку запиту. Вхідними та вихідними даними для ПМБЗ є структура знань.

Зазначимо, що порядок обробки може змінюватися, але формат даних на відповідних етапах обробки залишається незмінним.

3.2.2 Етапи роботи ІТ ОПМ при аналізі природномовного тексту

В залежності від напряму обробки ПМ текстів, ІТ ОПМ може працювати у режимах аналізу та синтезу тексту. Відповідно, аналіз – це отримання з вхідного ПМ тексту структури знань і її подальша обробка, а синтез – формування з існуючої структури знань ПМ тексту.

Аналіз є більш важливою функцією ІТ ОПМ. По-перше, для багатьох прикладних задач достатньо виділити структуру знань тексту і не обов’язково формувати ПМ текст з результатів обробки вхідного тексту. По-друге, основним способом синтезу тексту є пошук у ПМБЗ фрагментів тексту, які відповідають структурі знань запиту, для чого необхідно попередньо наповнити ПМБЗ у режимі аналізу.

Розглянемо діаграми послідовності обробки даних в цих режимах, визначимо етапи обробки і кроки, які виконуються на кожному етапі.

На рис. 3.4 представлена UML діаграма послідовності обробки даних ІТ ОПМ в режимі аналізу ПМ тексту.

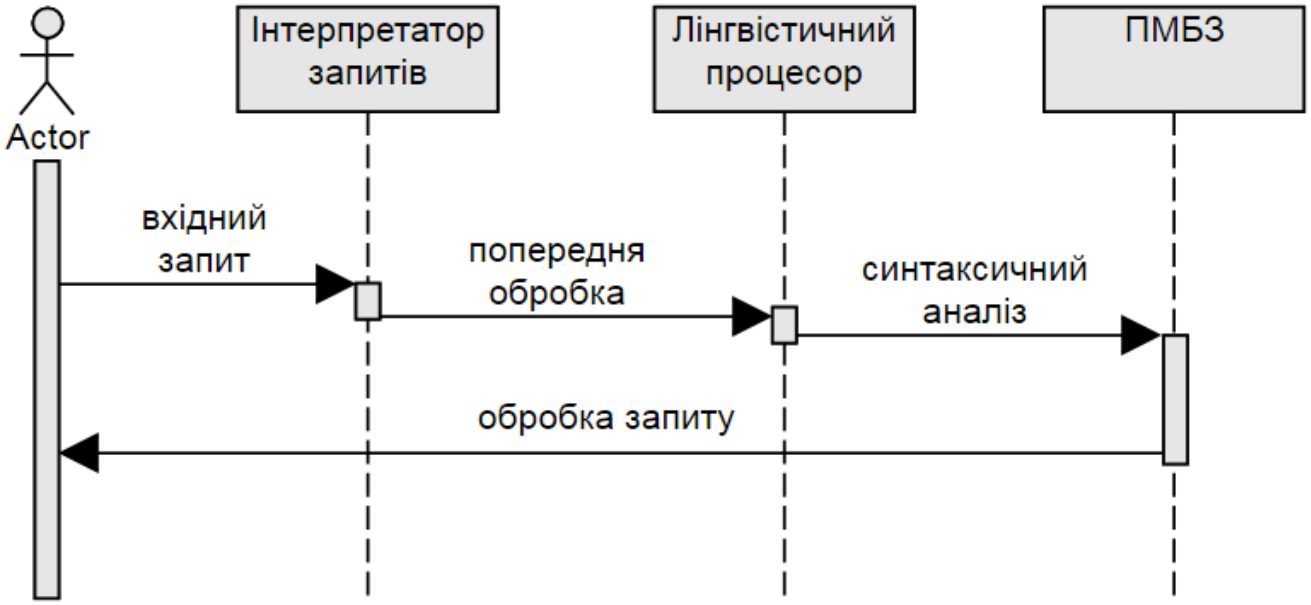


Рисунок 3.4 – Діаграма етапів роботи ІТ ОПМ в режимі аналізу природномовного тексту

При аналізі ПМ тексту виконуються наступні етапи обробки:

- попередня обробка;
- синтаксичний аналіз;
- обробка запиту.

Етап 1. Попередня обробка.

Крок 1. Ідентифікація та виділення текстової інформації.

Крок 2. Видалення нетекстової інформації: формул, ілюстрацій тощо.

Крок 3. Розбивка вхідного тексту на менші фрагменти, як-то речення.

Етап 2. Синтаксичний аналіз. Виконується послідовно для кожного фрагменту.

Крок 1. Виділення слів та їх граматичних форм.

Крок 2. Виділення структури знань тексту згідно методу, представленому у пунктах 2.5.1 – 2.5.3.

Крок 3. Формування маркерів, які прив’язують структуру знань до синтаксичної структури ПМ тексту.

Етап 3. Обробка запиту.

Крок 1. Збереження отриманої структури знань у ПМБЗ.

Крок 2. Збереження фрагменту ПМ тексту у ПМБЗ.

Крок 3. Повернення користувачу посилання на фрагмент знань у ПБМЗ.

3.2.3 Етапи роботи ІТ ОПМ при синтезі природномовного тексту

На рис. 3.5 представлена UML діаграма послідовності обробки даних ІТ ОПМ в режимі синтезу ПМ тексту.

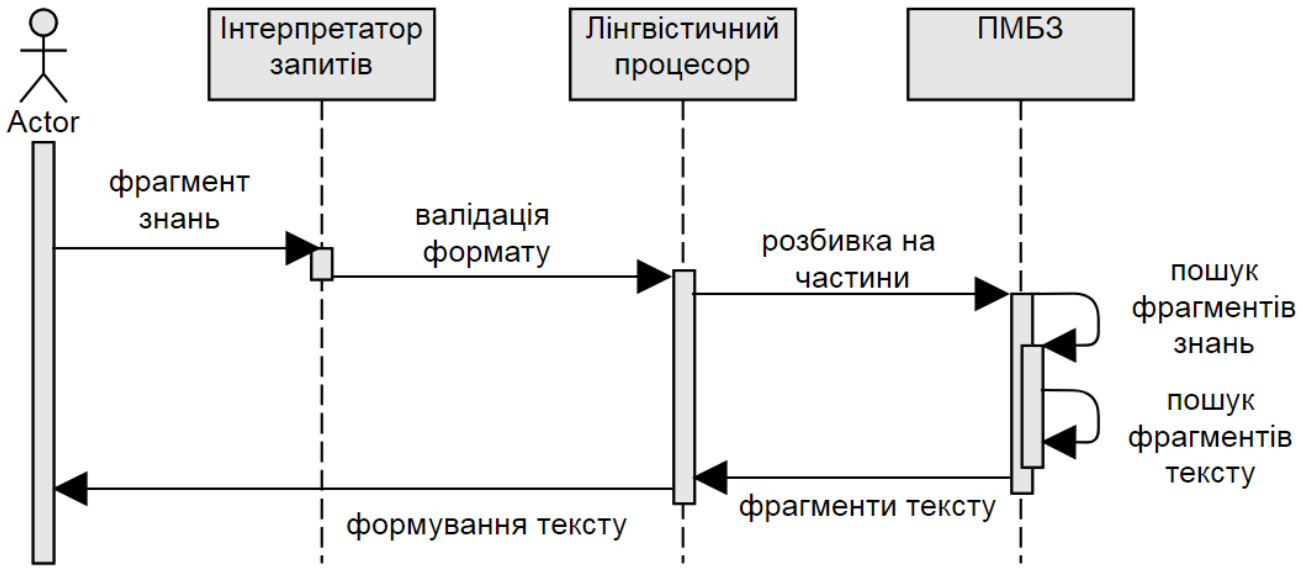


Рисунок 3.5 – Діаграма етапів роботи ІТ ОПМ в режимі синтезу природномовного тексту

При синтезі ПМ тексту виконуються наступні етапи обробки:

- валідація формату;
- розбивка на частини;
- пошук фрагментів знань;
- пошук фрагментів тексту;
- формування тексту.

Етап 1. Валідація формату.

Крок 1. В залежності від джерела даних:

Крок 1а. Якщо дані отримано безпосередньо з ПМБЗ: перейти до наступного кроку.

Крок 1б. Якщо дані отримано зі стороннього джерела: переконатися, що структура даних відповідає моделі знань ПМБЗ.

Етап 2. Розбивка на частини.

Крок 1. Передати структуру знань у ПМБЗ.

Етап 3. Пошук фрагментів знань.

Крок 1. Шукати у ПМБЗ структуру, яка повністю збігається з заданою.

Крок 2. В залежності від результатів виконання Кроку 1:

Крок 2а. Якщо структуру знайдено, перейти до Етапу 4.

Крок 2б. Якщо структуру не знайдено:

Крок 3. Розбити структуру знань на декілька частин.

Крок 4. Для кожної частини послідовно виконати Етап 2, Етап 3.

Етап 4. Пошук фрагментів тексту. Для кожної структури, отриманої під час виконання Етапу 3:

Крок 1. Шукати у ПМБЗ маркер, який відповідає такій структурі.

Крок 2. В залежності від результатів виконання кроку 1:

Крок 2а. Якщо маркер знайдено, зчитати пов'язаний з ним фрагмент ПМ тексту.

Крок 2б. Якщо маркер не знайдено, згенерувати фрагмент ПМ тексту на основі структури знань з використанням лінгвістичного процесора.

Етап 5. Формування тексту. Для кожної структури, отриманої під час виконання Етапу 4:

Крок 1. Додати отриманий фрагмент тексту у масив результатів обробки.

Крок 2. Об’єднати усі фрагменти тексту з використанням засобів лінгвістичного процесора.

3.2.4 Процедура записування знань для ІТ ОПМ

Основними операціями над знаннями в ІТ ОПМ це пошук та записування знань, які є фактично базовими операціями записування/зчитування, адаптованими під формат знань ПМБЗ у складі ІТ.

Задача записування знань у ПМБЗ представляє собою додавання у ПМБЗ знань довільного фрагменту природномовного тексту. Варіантами цієї задачі, у яких не виконуються деякі кроки обробки, є записування знань з підготовленого (адаптованого) тексту або з попередньо структурованого фрагменту знань.

Процедура обробки даних в ІТ ОПМ в режимі записування знань представлена на рис. 3.6.

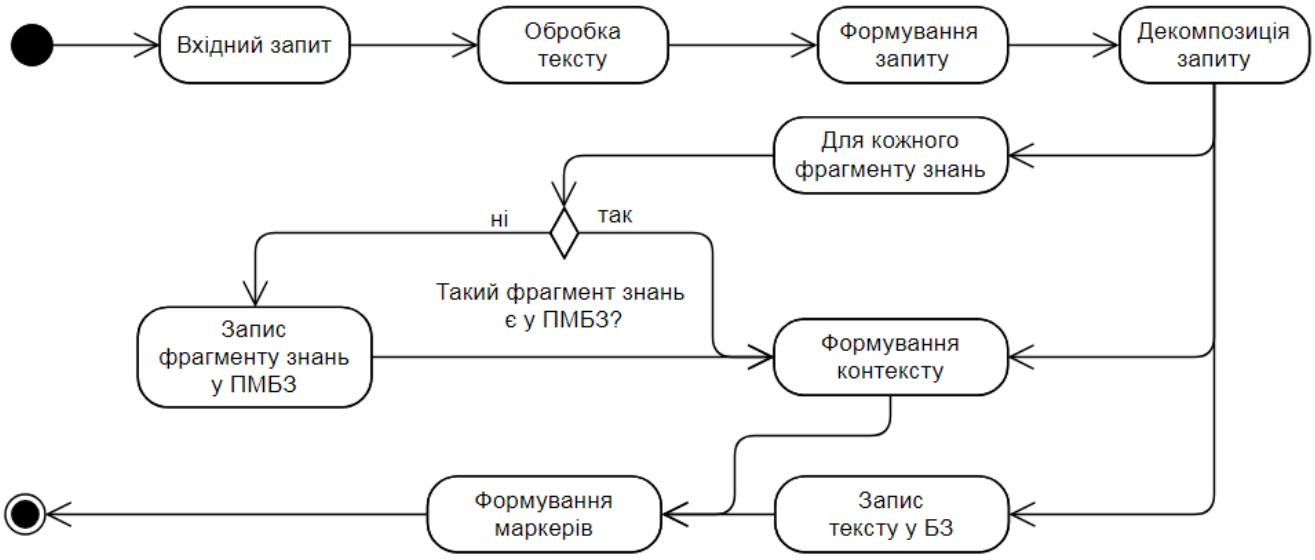


Рисунок 3.6 – Послідовність кроків записування знань в ІТ ОПМ

Процедура записування знань складається з наступних кроків.

1. **Попередня обробка тексту.** Вхідний запит у вигляді ПМ тексту перетворюється у фрагмент ПМ знань згідно процесу, описаному у пункті 3.2.2. Результатом виконання цього кроку є фрагмент знань, який відтворює структуру знань вхідного тексту.

2. **Декомпозиція тексту.** Структура знань, сформована на попередньому кроці, розбивається на окремі КЗ та відношення, при цьому окремо зберігаються зв’язки між ними та вхідний ПМ текст.
3. **Наповнення ПМБЗ.**
   3.1. **Додавання фрагментів знань у ПМБЗ.** Для кожного отриманого фрагменту знань виконується перевірка наявності такого фрагменту у ПМБЗ. Якщо фрагмент не знайдено, він записується як новий елемент.
   3.2. **Додавання тексту у ПМБЗ.** Для кожного фрагменту знань у ПМБЗ записується відповідна частина вхідного ПМ запиту.
4. **Формування контексту.** Після додавання усіх фрагментів вхідного запиту у ПМБЗ формуються зв’язки між ними відповідно до зв’язків, знайдених у вхідному тексті.
5. **Формування маркерів.** Після додавання усіх фрагменті знань до ПМБЗ також додаються зв’язки між знаннями та текстом на рівні лексем, квантів знань, фрагментів тексту та всього запиту відповідно. На цьому кроці за необхідності також додаються метадані, як-то джерело тексту, час та дата коли його було додано тощо.

3.2.5 Процедура пошуку знань для ІТ ОПМ

Задача пошуку знань у ПМБЗ виконується або безпосередньо як пошук фрагменту знань у ПМБЗ, або як пошук фрагменту ПМ тексту з його попередньою обробкою. Особливим випадком є ситуація, коли у ПМБЗ знайдено лише частину фрагменту знань запиту.

Процедура обробки даних в ІТ ОПМ в режимі пошуку знань представлена на рис. 3.7.

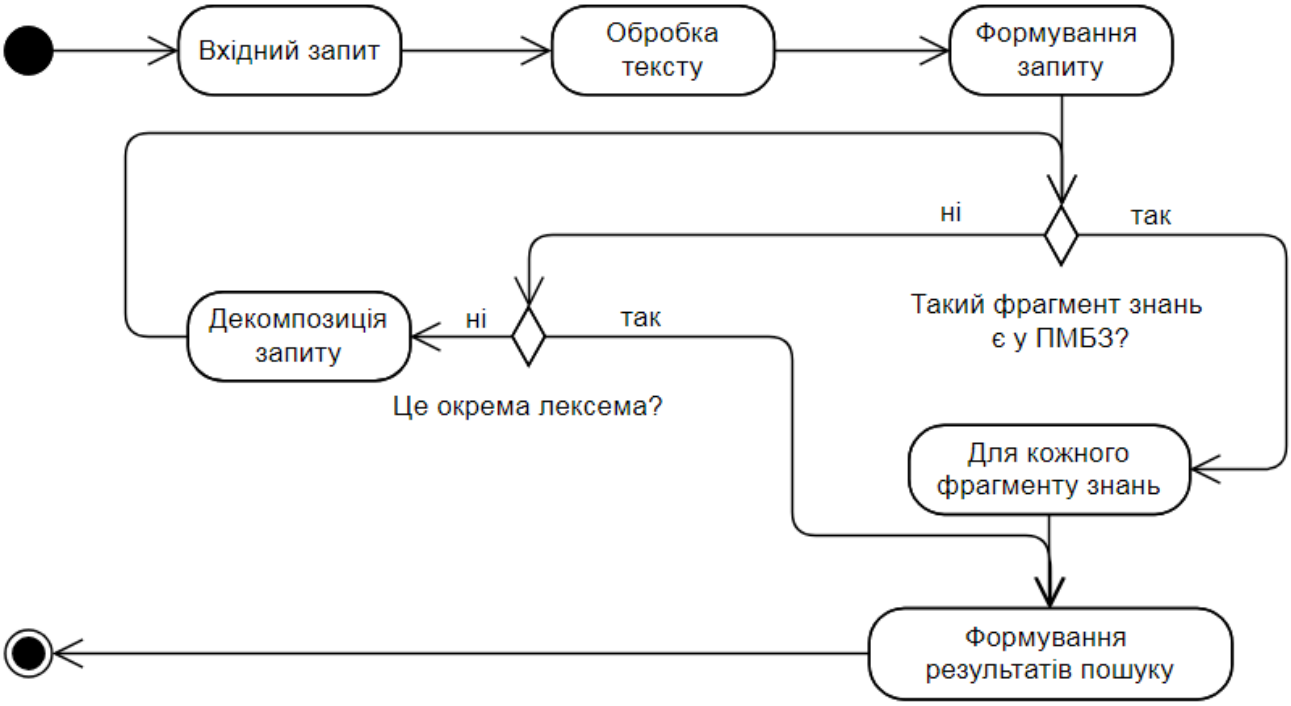


Рисунок 3.7 – Послідовність кроків пошуку знань в ІТ ОПМ

Процедура пошуку знань складається з наступних кроків.

1. **Обробка тексту.** Якщо пошуковий запит представлений у вигляді ПМ тексту, він перетворюється у фрагмент ПМ знань згідно процесу, описаному у пункті 3.2.2. В іншому випадку обробка починається безпосередньо з отриманого в запиті фрагменту знань.
2. **Формування запиту.** Виконується пошук даного фрагменту знань у ПМБЗ.
   2.1. Якщо фрагмент знань знайдено повністю, виконується крок 3.
   2.2. Якщо фрагмент знань не знайдено, виконується його декомпозиція на менші частини, до квантів знань та відношень включно, і для кожної частини виконується крок 2.
   2.3. Якщо частину запиту зведено до окремої лексеми, виконується пошук за цією лексемою як пошук за ключовими словами, без врахування структури знань запиту.
3. **Формування результату.** Отримані результати сортуються у порядку, визначеному програмою, та повертаються до користувача. За необхідності

результати доповнюються додатковими даними, як-то релевантність кожного результату, кількість входжень фрагменту знань у ПМБЗ тощо.

3.3 Приклади використання ІТ ОПМ для вирішення прикладних задач

3.3.1 Процедура використання ІТ ОПМ у задачі природномовного пошуку

Розглянемо приклади процедур використання ІТ ОПМ у задачі природномовного пошуку.

Природномовний пошук – це пошук у базі природномовних документів з використанням природномовного запиту [115]. Вхідними даними для природномовного пошуку є запит користувача у вигляді фрагменту природномовного тексту. Критерієм якості пошуку є релевантність, тобто показник подібності до наповнення документу до запиту користувача [116].

Найбільш поширеним на сьогодні є пошук за ключовими словами, проблемою якого є низька середня релевантність як окремих знайдених документів, так і усього масиву результатів пошуку. Розглянемо процедуру природномовного пошуку з використанням ІТ ОПМ та визначимо її особливості у порівнянні з пошуком за ключовими словами.

Ця процедура складається з наступних кроків.

1. **Обробка запиту**. З вхідного природномовного запиту виділяється фрагмент знань згідно процесу, описаному у пункті 3.2.2.
2. **Пошук.** Над отриманим фрагментом знань виконується процедура пошуку, описана у пункті 3.2.5.
3. **Сортування.** Отримані результати сортуються за релевантністю. Результати однакового рівня релевантності сортуються за зростанням часу їх додавання у ПМБЗ, або згідно встановленим програмою правилам. За замовчанням включаються тільки результати першого рівня.

3.1. Перші результати – фрагменти знань, які повністю збігаються з фрагментом знань запиту.

3.2. Далі – фрагменти знань, які частково збігаються з запитом.

3.3. Далі – фрагменти знань, які містять лексеми, наявні у запиті.

Результатом виконання процедури пошуку є набір документів, відсортованих за релевантністю. Для природномовного пошуку з використанням ІТ ОПМ основними результатами пошуку є ті документи, де лексеми пов’язані такими ж зв’язками, як і у пошуковому запиту. Це є перевагою при порівнянні з пошуком на основі ключових слів, оскільки пошук з використання ІТ ОПМ дозволяє відкинути документи, в яких лексеми запиту розташовані близько у тексті, але не пов’язані змістовно.

У додатку Д наведено приклад процесу пошуку за природномовним запитом з використання ІТ ОПМ.

3.3.2 Процедура використання ІТ ОПМ у задачі машинного перекладу

Розглянемо приклади процедур використання ІТ ОПМ у задачі машинного перекладу.

Машинний переклад – переклад природних мов з використанням комп’ютера [6]. Вхідними даними для машинного перекладу є текст, що відповідає правилам певної природної мови, а задачею – формування іншою мовою тексту, який максимально близький за змістом до вхідного тексту. Критерієм якості машинного перекладу є збереження семантичного та синтаксичного змісту вхідного тексту.

Проблема машинного перекладу це багатозначність, тобто велика кількість технічно правильних варіантів перекладу, з яких лише деякі вірно передають смислове наповнення вхідного тексту. Розглянемо процедуру машинного перекладу з використанням ІТ ОПМ та визначимо її особливості у порівнянні з пошуком за ключовими словами.

Ця процедура складається з наступних кроків.

1. **Обробка запиту**. З вхідного природномовного запиту виділяється фрагмент знань згідно процесу, описаному у пункті 3.2.2.
2. **Пошук.** Над отриманим фрагментом знань виконується процедура пошуку, описана у пункті 3.2.5.
3. **Обробка фрагментів знань**.
   3.1. Для кожного фрагменту знань у запиті відбувається пошук фрагментів знань у ПМБЗ, які пов’язані з фрагментом тексту цільової мови.
   3.2. Якщо знайдено фрагмент тексту з повною відповідністю, він додається у пул можливих перекладів.
   3.3. Якщо знайдено фрагмент тексту з частковою відповідністю, він зберігається для поєднання з іншими фрагментами.
   3.4. Якщо фрагменту тексту з відповідністю запиту не знайдено, виконується дослівний переклад.
4. **Формування результату**.
   4.1. Знайдені на кроці 3.2 результати повертаються як основні результати перекладу.
   4.2. Фрагменти тексту з частковим збігом та результати дослівного перекладу повертаються як додаткові результати для подальшої обробки.

Результатом виконання процедури перекладу є набір фрагментів тексту цільовою мовою, які в найкращому випадку є точним перекладом вхідного запиту, в гіршому випадку – дослівним перекладом.

Оскільки пошуковий запит розглядається як одне ціле, а переклад виконується на якомога більших його частинах, машинний переклад з використанням ІТ ОПМ дозволяє зберегти смислові зв’язки між частинами запиту, а в найкращому випадку навіть надати точний переклад запиту. Крім того, використання ПМБЗ в якості джерела даних для мови-посередника дозволяє успішно перекласти синтаксичні варіації запиту, в яких смислові зв’язки залишаються незмінними.

Зазначимо, що під час написання цієї дисертації компанія *Google* випустила оновлення [117] системи машинного перекладу *Google Translate,* в якому також значну роль відіграє збереження зв’язків між елементами вхідного запиту при перекладі.

Висновки до розділу 3

1. На основі загальної схеми природномовної бази знань розроблено структурну схему інформаційної технології обробки природномовних знань на основі інтеграційного підходу.
2. Розроблено UML діаграми та виконано аналіз етапів роботи інформаційної технології обробки природномовних знань на основі інтеграційного підходу в режимах аналізу та синтезу природномовного тексту.
3. Розроблено UML діаграми та розроблено процедури записування та пошуку знань з використанням розробленої моделі представлення знань в технологіях обробки природної мови, які дозволяють встановити зв’язки на структурному рівні між синтаксичною структурою тексту та довільною структурою метаданих.
4. Розроблено та проаналізовано процедури використання інформаційної технології обробки природномовних знань на основі інтеграційного підходу у задачах природномовного пошуку та машинного перекладу.

4 ЕКСПЕРИМЕНТАЛЬНА ПЕРЕВІРКА ІТ ОПМ

У даному розділі визначено технічні вимоги до інформаційної системи, яка реалізує інформаційну технологію обробки природномовних знань на основі інтеграційного підходу, зокрема визначено підсистеми та операції і розроблено схему бази даних для інформаційної системи. Проаналізовано обчислювальну складність пошуку природномовних знань у інформаційній системі та проведено її порівняння з аналогами. Виконано експериментальну перевірку використання інформаційної системи для підвищення релевантності результатів природномовного пошуку та виконано аналіз отриманих даних.

4.1 Технічні вимоги до інформаційної системи обробки природної мови

4.1.1 Вибір архітектури ІС ОПМ

За визначенням, інформаційна система – це комунікаційна система, що забезпечує збирання, пошук, оброблення та пересилання інформації. В контексті роботи ІТ ОПМ, інформаційна система – це програмна реалізація ІТ ОПМ, спрямована на вирішення певної прикладної задачі.

Інформаційна система обробки природної мови на основі розробленої у даній роботі інформаційної технології (далі – ІС ОПМ) характеризується наступними особливостями:

- ІС ОПМ є універсальною системою, що не має сильних зв’язків з оточенням. Усі процеси обробки даних та результати роботи можуть бути використані у складі різноманітних систем обробки природної мови. Таким чином, ІС ОПМ варто представити у вигляді окремої системи;

- в процесі роботи ІС ОПМ використовує зовнішні синтаксичний аналізатор та систему метаданих, але самі ці системи не входять безпосередньо в ІС ОПМ.

- одночасно ІС ОПМ може обробляти велику кількість різних запитів, більшість з яких обробляється централізовано з використання спільної ПМБЗ у складі ІС.

З огляду на ці особливості в якості архітектури ІС ОПМ логічно обрати веб-архітектуру [118] (рис. 4.1), де основні компоненти ІС (а саме інтерпретатор запитів, ПМБЗ та лінгвістичний процесор) знаходяться на сервері додатків, що пов’язаний з сервером БД, користувачі підключаються до системи з використанням браузера або API, а системи синтаксичного аналізатора та метаданих взаємодіють з ІС ОПМ через Інтернет-підключення.

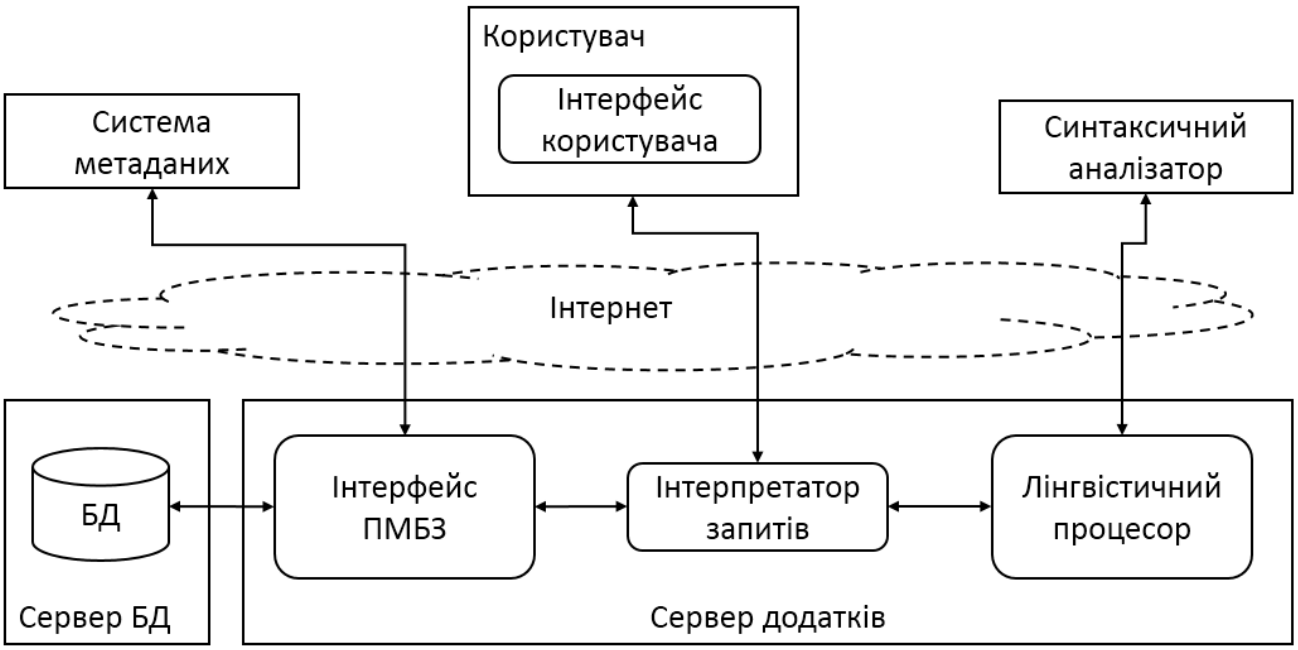


Рисунок 4.1 – Архітектура інформаційної системи ОПМ ІП

Основна частина ІС ОПМ на логічному рівні включає описані нижче компоненти.

1. Сервер БД, на якому розміщено БД системи. Оскільки типи даних ІС ОПМ обмежені природномовним текстом та простими об’єктами, обсяг такої БД є невеликим, але частота запитів (як на читання, так і на записування) висока. Відповідно, для розміщення БД використовується централізований сервер з можливістю горизонтального розширення, на якому встановлено реляційну

СУБД, наприклад *MariaDB* або *PostgreSQL,* або БД будь-якого іншого класу, яка підтримує високу частоту невеликих запитів.

2. Сервер додатків, який є центральним компонентом системи. На сервері додатків розміщено інтерфейс ПМБЗ, інтерпретатор запитів, лінгвістичний процесор та веб-сервер (наприклад *Apache* або *nginx*). За необхідності (зокрема, при виникненні високого навантаження), інтерпретатор запитів та веб-сервер можуть бути перенесені на виділені сервери, а будь-яка з цих підсистем може бути розширена з балансуванням навантаження.
3. Інтерфейс користувача – тонкий клієнт, що представляє собою або браузерний додаток (як-то веб-сайт) або точку входу прикладного програмного інтерфейсу (API). Інтерфейс користувача формує запити згідно формату, який очікує інтерпретатор запитів.
4. Зовнішні система ЛП та система метаданих. Ці системи не входять до складу ІС ОПМ, а натомість є підключеними через адаптери, які дозволяють використовувати існуючі продукти, такі як *Stanford Parser* та *Babelnet.*

4.1.2 Підсистеми та операції ІС ОПМ

В рамках цих компонентів розміщені наступні програмні підсистеми.

На сервері БД:

– база даних, що зберігає знання, структуровані згідно МПЗ, та забезпечує їх низькорівневу обробку.

На серверні додатків:

– ядро ПМБЗ, що забезпечує взаємодію БД та зовнішніх запитів до ПМБЗ.
– контролер запитів, що виконує декомпозицію складних вхідних запитів до елементарних запитів до ПМБЗ та контролює їх виконання.
– інтерфейс запитів, що обробляє вхідні запити користувачів, направляє їх до контролера запитів та забезпечує взаємодію з користувачем.

На машинах-клієнтах:

– інтерфейс користувача та/або інтерфейси прикладних систем.

Зовнішні залежності:

– лінгвістичний процесор, що виконує синтаксичний аналіз текстів.

– система метаданих, що зберігає довільні логічні правила обробки знань та забезпечує їх виконання над вхідними запитами.

Для користувача ІС ОПМ доступні наступні операції.

Робота з текстом:

– декомпозиція (*Text->decompose*) – Обробка фрагменту тексту для виділення його структури знань з використанням ЛП.

– записування (*Text->write*) – Записування фрагменту тексту у ПМБЗ. Використовує *Text->decompose* для декомпозиції тексту у знання та *K->write* для записування отриманого фрагменту знань у ПМБЗ.

– пошук (*Text->search*) – Пошук фрагменту тексту у ПМБЗ. Використовує *Text->decompose* для декомпозиції тексту у знання та *K->search* для пошуку отриманого фрагменту знань у ПМБЗ.

Робота з даними:

– записування (*K->write*) – Записування фрагменту знань у ПМБЗ.

– пошук (*K->search*) – Пошук фрагмента знань у ПМБЗ.

Робота з логікою:

– валідація (*K->validate*) – Перевірка фрагмента знань на відповідність заданим правилам.

Розглянута інформаційна система обробки природномовних текстів є системою обробки інформації, характерними для якої є:

- великий обсяг даних;
- проста структура окремого елементу даних та зв’язків між ними;
- висока інтенсивність обчислень;
- прості процедури обробки даних та операції, які над ними виконуються;
- значна частка типових операцій.

Отже, ІС ОПМ не має особливих вимог до програмного та апаратного забезпечення і може бути розгорнута у типових середовищах (сервери загального призначення, персональні комп'ютери тощо) за умови відповідності ресурсів середовища поставленим задачам.

Нижче наведені рекомендації для розгортання ІС ОПМ.

Сервер: будь-яке комерційне рішення та серверна операційна система.

Клієнт: тонкий клієнт. Для сторонньої системи – інтерфейс запитів типу «міст». Для користувача-людини – браузер або спеціалізований графічний інтерфейс.

База даних: будь-яка БД. Рекомендується *SQL* або *NoSQL* рішення, що дозволяє оптимізувати швидкодію масової обробки простих відношень, як-то *MariaDB, PostgreSQL, Redis* тощо.

Мова програмування: для досліджень традиційна МП, що має готові бібліотеки для вирішення технічних задач, як-то *Java* або *Python*. Для прикладних задач МП з можливістю низькорівневої оптимізації роботи системи, як-то *Java, C#, C++*.

Лінгвістичний процесор: для української мови *pyMorphy2*, для англійської мови *Stanford Parser*.

4.1.3 Розробка схеми бази даних ІС ОПМ

Опишемо об'єкти даних, які використовує ІС ОПМ у своїй роботі, та схему БД, яка їх зберігає. Для опису об'єктів БД та зв'язків між ними використаємо *ERM (Entity-Relationship Model).*

Лексема. Цей об'єкт описує лексему – слово у всій повноті його граматичних форм. Об'єкт лексеми зберігає лише ту інформацію, яка використовується в ІС ОПМ, а саме частину мови, до якої належить слово, та список словоформ, які відповідають його можливим граматичним формам:

$$O_L = \{id, pos, w\}, \tag{4.1}$$

де id – унікальний ідентифікатор лексеми;

pos – ідентифікатор частини мови, до якої належить лексема;

w – множина словоформ лексеми.

В свою чергу, об'єкт словоформи є складовою об'єкту лексеми, включає в себе можливі граматичні форми та словоформи лексеми:

$$O_w = \{w_{nf}, (w_1, w_2, \dots, w_n)\}, \quad (4.2)$$

де w_{nf} – базова (нормальна) форма слова;

w_i – множина граматичних форм слова.

Об'єкт лексеми *word* має наступні поля:

Таблиця 4.1 – Поля об'єкту «лексема»

Поле	Тип
id	int
pos	int
normal_form	string
forms	[string]

Нормальна форма лексеми w_{nf} є обов'язковим елементом об'єкту, інші граматичні форми можуть бути відсутні. Так, в українській мові прислівники мають лише одну форму, яка використовується незалежно від контексту. Крім того, в ролі слова в ІС ОПМ можуть виступати також штучні елементи – абревіатури, власні назви тощо; вони можуть також бути представлені як лексема, що містить лише нормальну форму.

Квант знань. Структура об'єкту кванту знань відповідає його моделі, представленій у підрозділі 2.2. Об'єкт КЗ описує множину лексем, що відповідають його елементам, та функціональні зв'язки між ними.

Об'єкт КЗ має наступний вигляд:

$$O_e = \{id, L, f, r\}, \quad (4.3)$$

де id – унікальний ідентифікатор елементу КЗ;

L – об'єкт лексеми, визначений у словнику;

f – ідентифікатор граматичної форми лексеми;

r – роль даного елементу у КЗ (наприклад Obj, Mov або $Attr$).

Об’єкт КЗ ***structure*** має наступні поля:

Таблиця 4.2 – Поля об’єкту «квант знань»

Поле	Тип
id	int
word	*Word*
form	int
role	int

Відношення. Об’єкт відношення поєднує два об’єкти КЗ та містить додаткову інформацію про саме відношення, яка в загальному випадку представляє собою посилання на певний об’єкт, що описує це відношення.

Об’єкт відношення має наступний вигляд:

$$O_R = \{id, Q_1, Q_2, D\}, \quad (4.4)$$

де id – унікальний ідентифікатор відношення у ПМБЗ;

Q_1, Q_2 – КЗ, які пов’язані цим відношенням;

D – посилання на об’єкт, що містить додаткову інформацію про відношення.

Об’єкт відношення ***relation*** має наступні поля:

Таблиця 4.3 – Поля об’єкту «відношення»

Поле	Тип
id	int
structure_left	*Structure*
structure_right	*Structure*
relation_data	*Object*

Текст. Цей об’єкт описує фрагмент тексту, на основі якого побудовано КЗ у ПМБЗ, та зберігає його для уточнення структури знань у випадку зміни або втрати джерела тексту або суттєвих змін у роботі ІС ОПМ.

Об’єкт тексту має наступний вигляд:

$$O_T = \{id, text, source\}, \quad (4.5)$$

де id – унікальний ідентифікатор фрагменту тексту;

$text$ – тіло фрагменту тексту;

$source$ – посилання на джерело тексту.

Об’єкт тексту *text* має наступні поля:

Таблиця 4.4 – Поля об’єкту «фрагмент тексту»

Поле	Тип
id	int
text	text
source	Object

Маркер. Цей об’єкт зберігає зв'язок між фрагментом тексту та фрагментом знань, що його описує, а саме посилання на КЗ та відношення у ньому.

Об’єкт маркеру *marker* має наступні поля:

Таблиця 4.5 – Поля об’єкту «фрагмент тексту»

Поле	Тип
id	int
text	*Text*
structures	[*Structure*]
relations	[*Relation*]

Повна *ERM* схема ІТ ПМБЗ ІП представлена у додатку В.

4.2 Оцінка обчислювальної складності пошуку в ІС ОПМ

4.2.1 Методика оцінки та визначення обмежень вхідних даних

Оцінимо теоретичну складність пошуку в ІС ОПМ та порівняємо її з аналогічними системами. Для цього виділимо варіанти пошуку в ІС ОПМ, які використовують різні процедури, та визначимо їх обчислювальну складність.

Задача пошуку в ІС ОПМ включає в себе наступні варіанти підзадач, які відрізняються обсягом та типом даних, які передаються у пошуковому запиті:

- повнотекстовий пошук – пошук фрагменту тексту, який містить рядок, що повністю включає даний запит;
- пошук слова або лексеми – пошук фрагменту тексту, який включає в себе дане слово у будь-якій з його словоформ;
- пошук КЗ або його частини – пошук фрагменту тексту, в якому існують усі передані в запиті лексеми, пов’язані відповідними внутрішніми зв’язками КЗ;
- пошук фрагменту знань - пошук фрагменту тексту, в якому існують усі передані в запиті КЗ, пов’язані відповідними відношеннями.

Для порівняння швидкості роботи пошуку у ІС ОПМ та у інших системах створимо модель складності відповідних операцій у пакеті *wxMaxima* на основі отриманих вище формул, які в загальному випадку достатньо точно відтворюють відповідні характеристики реальної системи.

Визначимо відносну складність пошуку як $K = \frac{C_2}{C_1}$, де C_2 – обчислювальна складність виконання задачі у ІС ОПМ, C_1 – відповідний показник її аналогу. Значення $K > 1$ показує відносну перевагу ПМБЗ ІП над аналогом, $K < 1$ – навпаки. Для аналізу цього показника досліджуємо характер функції, що його описує, між мінімальним та максимальним її значенням.

Змінні, на основі якої перевіряємо K – кількість елементів у БЗ та словнику і кількість слів у пошуковому запиті.

Визначимо граничні значення, які можуть приймати ці змінні.

Згідно з оцінками *Google*, у світі існує близько 130 000 000 книг [119], середній обсяг яких сягає 65 000 слів [120]. Таким чином, загальний обсяг БЗ, яка містила б усі ці слова, дорівнює $1{,}3 \cdot 10^8 \cdot 6{,}5 \cdot 10^4 = 8{,}45 \cdot 10^{12}$ слів, або, при округленні до ближчого розряду, 10^{13} окремих слів.

Кількість слів в англійській мові оцінюється як 250 000 окремих слів [121], для української мови це близько 100 000 слів та 1 100 000 словоформ [122]. Отже, розмір словнику ПМБЗ складає, відповідно, 10^5 та 10^6 елементів.

Зазначимо, що при кількості записів на рівні мільйонів включно ($K < 10^6$) існуючі системи є достатньо потужними для виконання усіх перелічених задач, а отже – предметом інтересу є значення цього критерія для БЗ великого обсягу. Для проведення аналізу оцінюємо асимптотичний вимір для найбільшого значення ($K = 10^{13}$) та будуємо графік для визначення характеру функції.

Оцінимо значення K для визначених вище варіантів пошуку.

4.2.2 Оцінка складності повнотекстового пошуку

ІС ОПМ. Задача повнотекстового пошуку в ІС ОПМ складається з пошуку у словнику слів, які представлені символами у запиті, та пошуку у БЗ, що містять ці слова, для їх подальшого аналізу. Такий пошук відповідає операціям бінарного пошуку у масиві слів та масиві КЗ і, відповідно, має клас складності

$$C = \log_2(N_w \cdot n), \tag{4.6}$$

де N_w – кількість слів у словнику;

n – кількість КЗ у БЗ.

Аналог. Аналогом цієї задачі є пошук з повнотекстовим індексом у сучасних БД [123]. За умови використання індексу, який повністю охоплює записи у БД та не зменшує повноти пошуку, обчислювальна складність пошуку у БД залежить від обсягу БД, тобто кількості текстових записів у БД та кількості символів в одному записі:

$$C = \log_2(N_s \cdot n), \tag{4.7}$$

де N_S – кількість символів у одному записі,

n – кількість записів у БД.

Таким чином, повнотекстовий пошук у ПМБЗ та у БД з повнотекстовим індексом належать до одного класу складності.

Порівняння. Для визначеної вище кількості слів у словнику $N_w = 10^6$ це, відповідно,

$$C = \log_2(10^6 \cdot n)$$

, або

$$C_1 = \log_2(10^6) + \log_2(n), \quad (4.8)$$

Для найкращого аналогу, пошуку у БД за повнотекстовим індексом, за умови що середня довжина слова у різних мовах становить від 7 до 12 символів з медіаною близько 10 символів [124], складність пошуку становить

$$C = \log_2(10 \cdot n)$$

, або

$$C_2 = \log_2(10) + \log_2(n), \quad (4.9)$$

Отже, коефіцієнт відносної складності пошуку за символами визначається за формулою

$$K(n) = \frac{C_2}{C_1} = \frac{\log_2(10)+\log_2(n)}{\log_2(10^6)+\log_2(n)}, \quad (4.10)$$

де n – обсяг БЗ,

$K(n)$ – відносна складність пошуку для цього обсягу.

Скоротимо цю формулу до вигляду

$$K(n) = \frac{C_2}{C_1} = \frac{\ln(10 \cdot n)}{\ln(10^6 \cdot n)}, \quad (4.11)$$

Визначимо граничне значення цього коефіцієнту для максимального розрахункового обсягу БЗ:

$$\lim_{n \to max} K(n) = \lim_{n \to 10^{13}} \frac{\ln(10 \cdot n)}{\ln(10^6 \cdot n)} = \frac{14}{19}, \quad (4.12)$$

Отже, для операції пошуку за символами гранична ефективність ПМБЗ становить ~74% від ефективності аналогів.

Оцінимо характер цільової функції на всій області визначення (рис. 3.4)

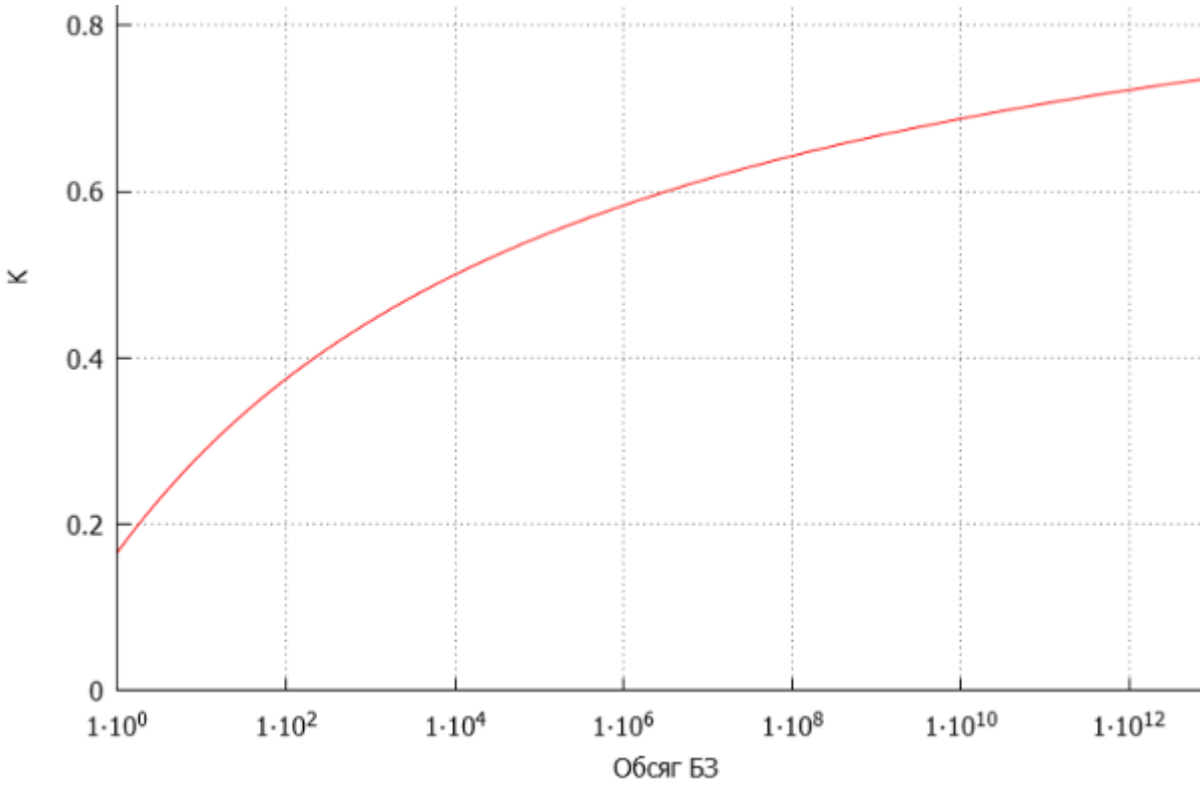


Рисунок 3.4 – Відносна складність пошуку, пошук по символам

По осі абсцис відзначено об’єм ПМБЗ у одиницях слів, по осі ординат – умовна розрахункова складність операції пошуку.

Оцінимо характер цільової функції на уточненій області визначення (рис. 3.5)

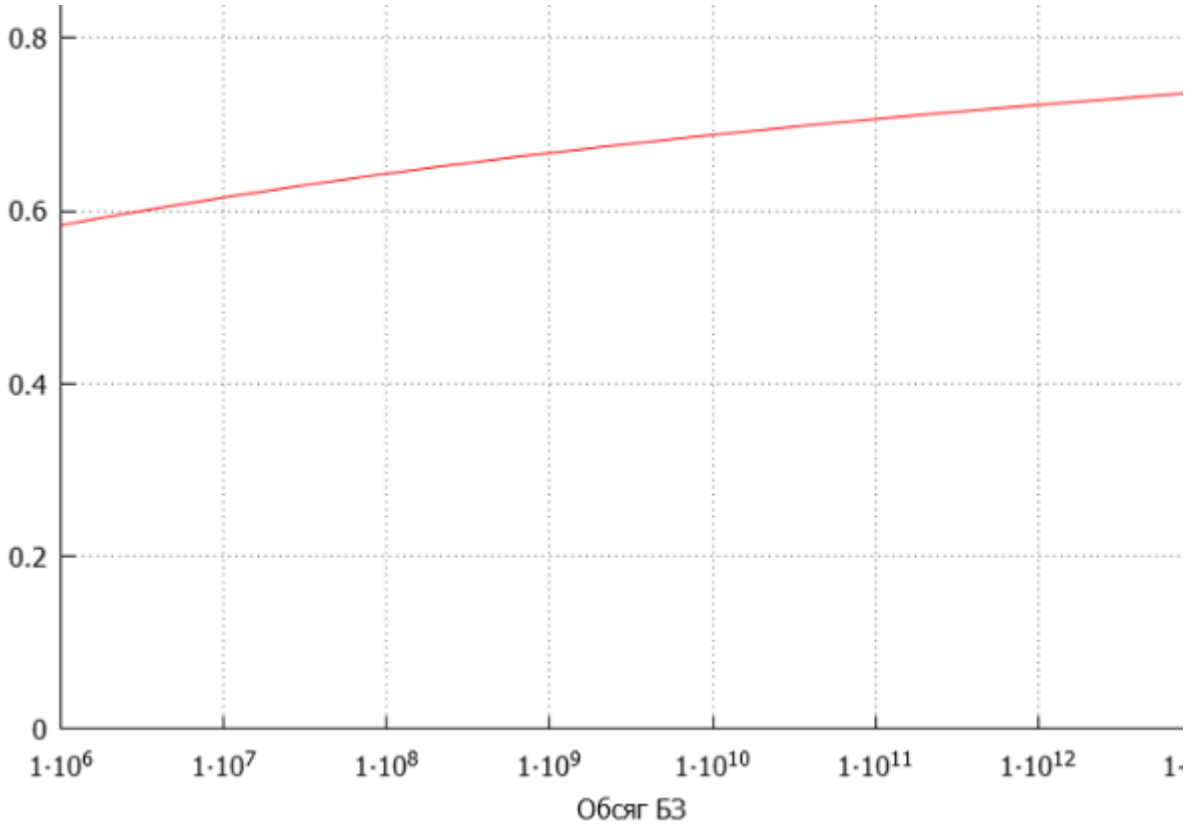


Рисунок 3.5 – Відносна складність пошуку, пошук по символам, уточнена

По осі абсцис відзначено об’єм ПМБЗ в одиницях слів, по осі ординат – умовна розрахункова складність операції пошуку.

Як видно з графіків, швидкість операції пошуку за символами у ІС ОПМ значно менша, ніж у аналогів – на 26% для максимального об’єму БЗ та на 20-40% для уточненого об’єму БЗ. Ефективність ПМБЗ для БЗ малого обсягу до 80% менша за ефективність аналогів.

Отже, ІС ОПМ не є ефективною для задачі пошуку по символам.

4.2.3 Оцінка складності пошуку за окремим словом

ІС ОПМ. Структура знань довільного тексту в ІС ОПМ є еквівалентною графу знань, що складається з КЗ та відношень. Цей граф може бути доповнений окремими масивами його КЗ та відношень.

Якщо слово передано у запиті у вигляді лексеми, яка існує у словнику ІС ОПМ, задача пошуку КЗ, що містять це слово, є задачею пошуку одного елементу у масиві КЗ. Отже, обчислювальна складність цієї задачі становить:

$$C = \log_2(N_Q), \tag{4.13}$$

де N_Q – кількість КЗ у ПМБЗ.

Аналог. Аналогом цієї операції є пошук слова у ПМБЗ, представленій у вигляді семантичної мережі. Оскільки в ІС ОПМ слова представлені лексемами, а у інших ПМБЗ – словоформами, складність операції пошуку за словом для інших ПМБЗ зростає в залежності від середньої кількості словоформ лексеми у мові і становить

$$C = \log_2(N_Q \cdot N_L), \tag{4.14}$$

де N_Q – кількість записів у БЗ,

N_L – середня кількість словоформ у лексемі.

Отже, складність пошуку фрагменту знань за словом залежить, крім безпосередньо кількості записів у ПМБЗ, також від середньої кількості словоформ у лексемі, і є апріорі меншою для ІС ОПМ.

Порівняння. Для української мови, згідно з наведеними вище даними, на кожну лексему припадає в середньому $10^6 \div 10^5 = 10$ словоформ, а отже, пошук будь-якої з них зводиться до пошуку відповідної лексеми.

Таким чином, для семантичних мереж складність цієї операції становить:

$$C_2 = \log_2(10 \cdot n), \tag{4.15}$$

Отже, коефіцієнт відносної складності пошуку за символами визначається за формулою

$$K(n) = \frac{C_2}{C_1} = \frac{\log_2(10 \cdot n)}{\log_2(n)}, \tag{4.16}$$

Визначимо граничне значення цього коефіцієнту для максимального розрахункового обсягу БЗ:

$$\lim_{n \to max} K(n) = \lim_{n \to 10^{13}} \frac{\ln(10 \cdot n)}{\ln(n)} = \frac{14}{13}, \tag{4.17}$$

Отже, для операції пошуку за символами гранична ефективність ПМБЗ становить ~107% від ефективності аналогів.

Оцінимо характер цільової функції на всій області визначення (рис. 3.6)

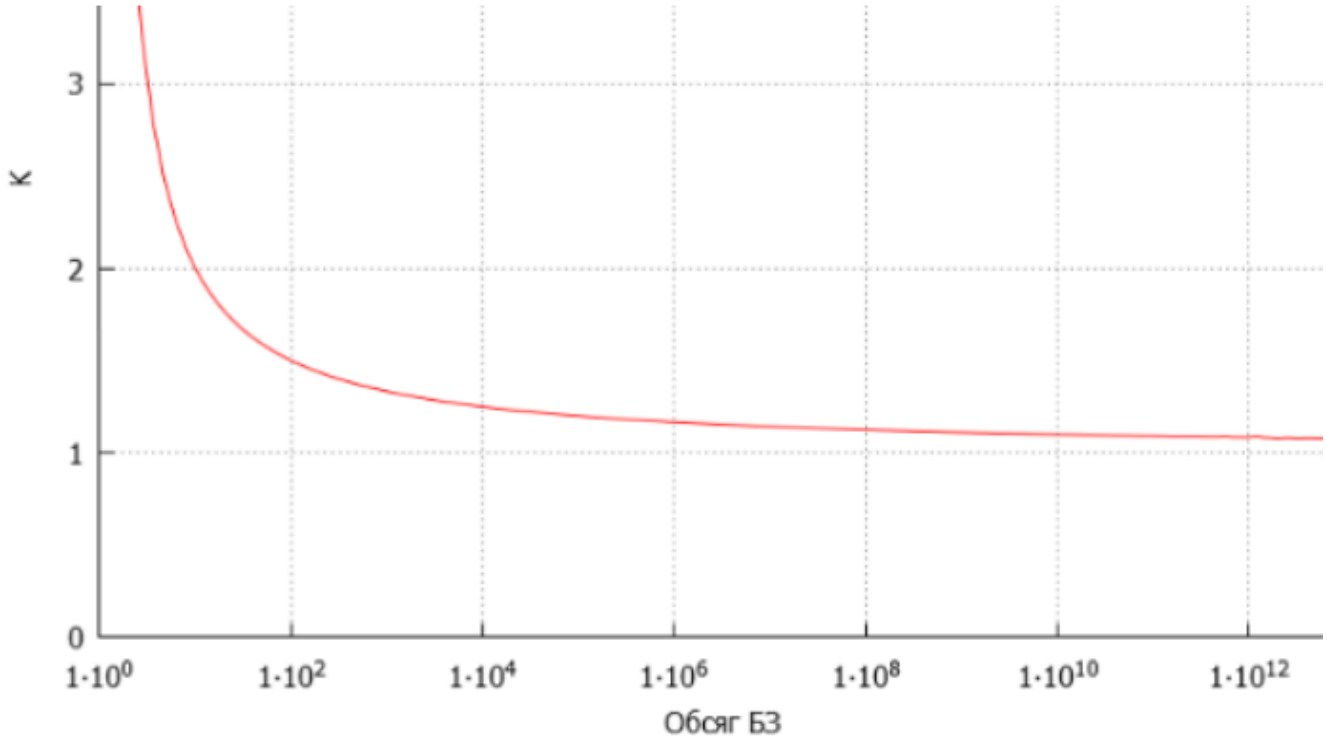


Рисунок 3.6 – Відносна складність пошуку, пошук за словами

По осі абсцис відзначено об’єм ПМБЗ у одиницях слів, по осі ординат – умовна розрахункова складність операції пошуку.

Оцінимо характер цільової функції на уточненій області визначення (рис. 3.7)

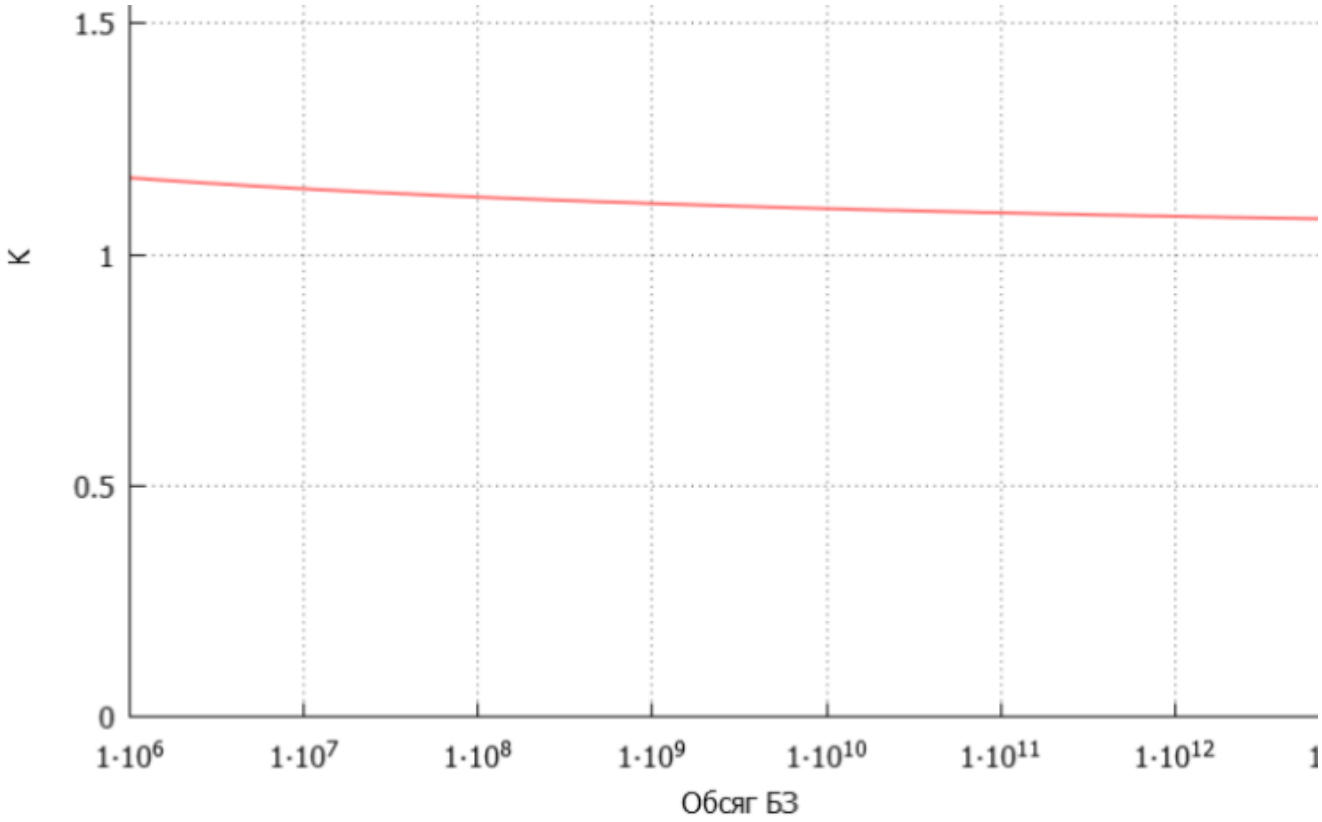


Рисунок 3.7 – Відносна складність пошуку, пошук за словами, уточнена

По осі абсцис відзначено об’єм ПМБЗ в одиницях слів, по осі ординат – умовна розрахункова складність операції пошуку.

Як видно з графіків, швидкість операції пошуку за символами в ІС ОПМ більша, ніж у аналогів – на 7% для максимального об’єму БЗ та на 7-16% для уточненого об’єму БЗ.

4.2.4 Оцінка складності пошуку за простим запитом

ІС ОПМ. Пошук за простим запитом, який не виходить за рамки одного речення природномовного тексту у природномовних пошукових системах або одного КЗ в ІС ОПМ, відбувається за тими ж правилами, що і пошук за словом.

Для ІС ОПМ загальна обчислювальна складність такого пошуку складає

$$C = \log_2\left(N_0^Q\right) + \log_2\left(N_1^Q\right) + \cdots + \log_2\left(N_w^Q\right), \qquad (4.18)$$

де N_0^Q – загальна кількість КЗ у БЗ,

N_i^Q – кількість КЗ у результатах пошуку на $(i-1)$ кроці,

w – кількість слів у запиті.

Тобто, пошук відбувається в ітеративному режимі, і на кожному наступному кроці поле пошуку звужується до результатів пошуку на попередньому кроці.

Представимо його як

$$C = \log_2(N_0^Q \cdot N_1^Q \cdot \ldots \cdot N_w^Q), \quad (4.19)$$

Аналог. Для інших ПМБЗ обчислювальна складність операції пошуку зростає як при наповненні самої бази знань, так і зі збільшенням розміру запиту [125], оскільки кожний елемент у полі пошуку необхідно перевірити на відповідність кожному зі слів у запиті [126].

Відповідно, складність цієї операції складає

$$C = \log_2(N_Q \cdot N_L \cdot w), \quad (4.20)$$

де N_Q – загальна кількість записів у БЗ,

N_L – середня кількість словоформ у лексемі,

w – кількість слів у запиті.

Зауважимо, що використання ітеративного процесу пошуку в ІС ОПМ є можливим завдяки представленню природномовного тексту згідно моделі представлення знань ІТ ОПМ, що дозволяє звертатися за ідентифікатором до КЗ, який однозначно містить або не містить шукане слово, без виконання додаткового запиту.

Порівняння. Задача пошуку по слову у ПМБЗ відповідає повтору операції пошуку даної лексеми у мережі КЗ. Отже, складність такої операції згідно формулі (4.19) становить

$$C = \log_2(N_0^Q \cdot N_1^Q \cdot \ldots \cdot N_w^Q)$$

Складність пошуку КЗ, що містять перше слово даного запиту, відповідає складності пошуку цього слова у БЗ і становить $\log(n)$, де n – кількість елементів у БЗ. Кожна наступна ітерація пошуку виконується на результатах попередньої ітерації. Це, відповідно – кількість КЗ, які містять усі вказані слова.

Для найчастіше вживаного слова, що належить до граматичного класу іменників, «я», абсолютна частота його появи у системі «Частотний словник української мови» [122] становить близько 1,5%, що відповідає зменшенню обсягу вибірки на 2 порядки на кожній ітерації. Таким чином, повна складність операції пошуку

$$C_1 = \sum_{n=0}^{6} 0{,}015^n \cdot \log_2(n) = 1{,}015 \cdot \log_2(n), \quad (4.21)$$

Для пошуку у семантичних мережах це, відповідно,

$$C = \log_2(N_Q \cdot N_L \cdot w)$$

або

$$C_2 = \log_2(6 \cdot 10 \cdot n), \quad (4.22)$$

Отже, коефіцієнт відносної складності пошуку за символами визначається за формулою

$$K(n) = \frac{C_2}{C_1} = \frac{\log_2(60 \cdot n)}{1{,}015 \cdot \log_2(n)}, \quad (4.23)$$

Визначимо граничне значення цього коефіцієнту для максимального розрахункового обсягу БЗ:

$$\lim_{n \to max} K(n) = \lim_{n \to 10^{13}} \frac{\ln(60 \cdot n)}{1{,}015 \cdot \ln(n)} \approx 1{,}11998, \quad (4.24)$$

Отже, для операції пошуку за символами гранична ефективність ІС ОПМ становить ~112% від ефективності аналогів.

Оцінимо характер цільової функції на всій області визначення (рис. 3.8)

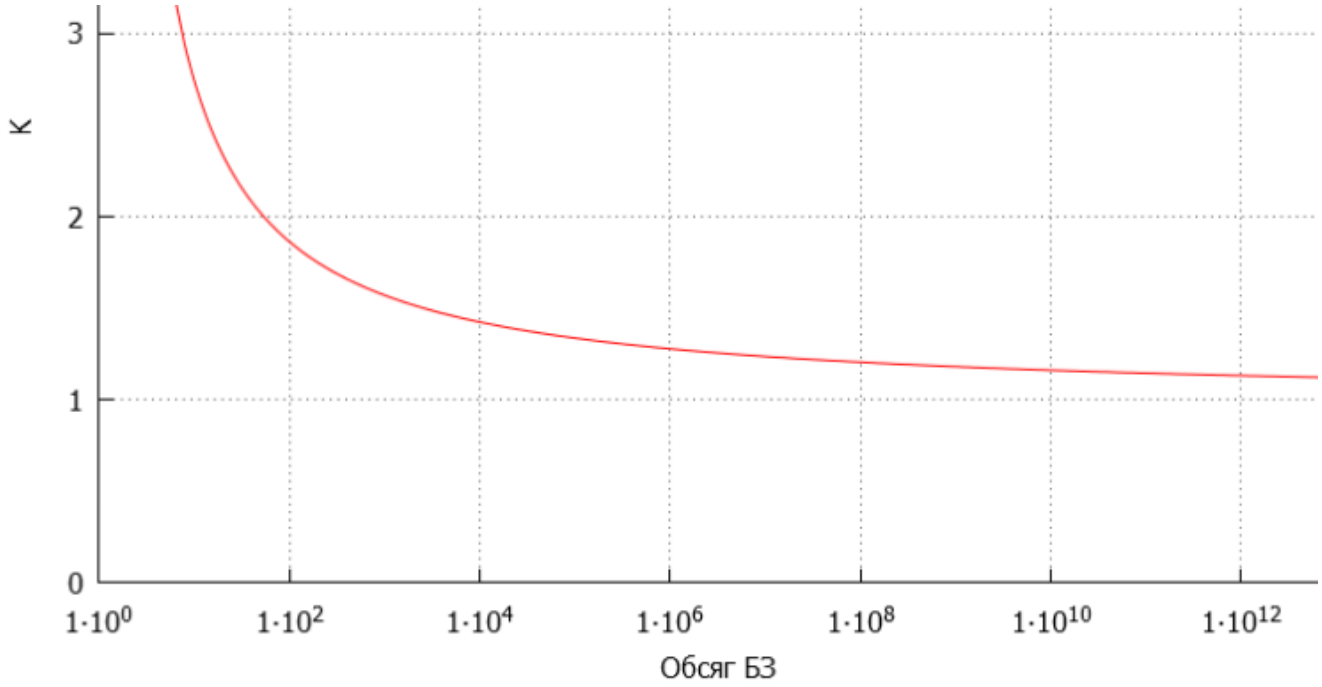


Рисунок 3.8 – Відносна складність пошуку, пошук за простим запитом

По осі абсцис відзначено об’єм ПМБЗ у одиницях слів, по осі ординат – умовна розрахункова складність операції пошуку.

Оцінимо характер цільової функції на уточненій області визначення (рис. 3.9)

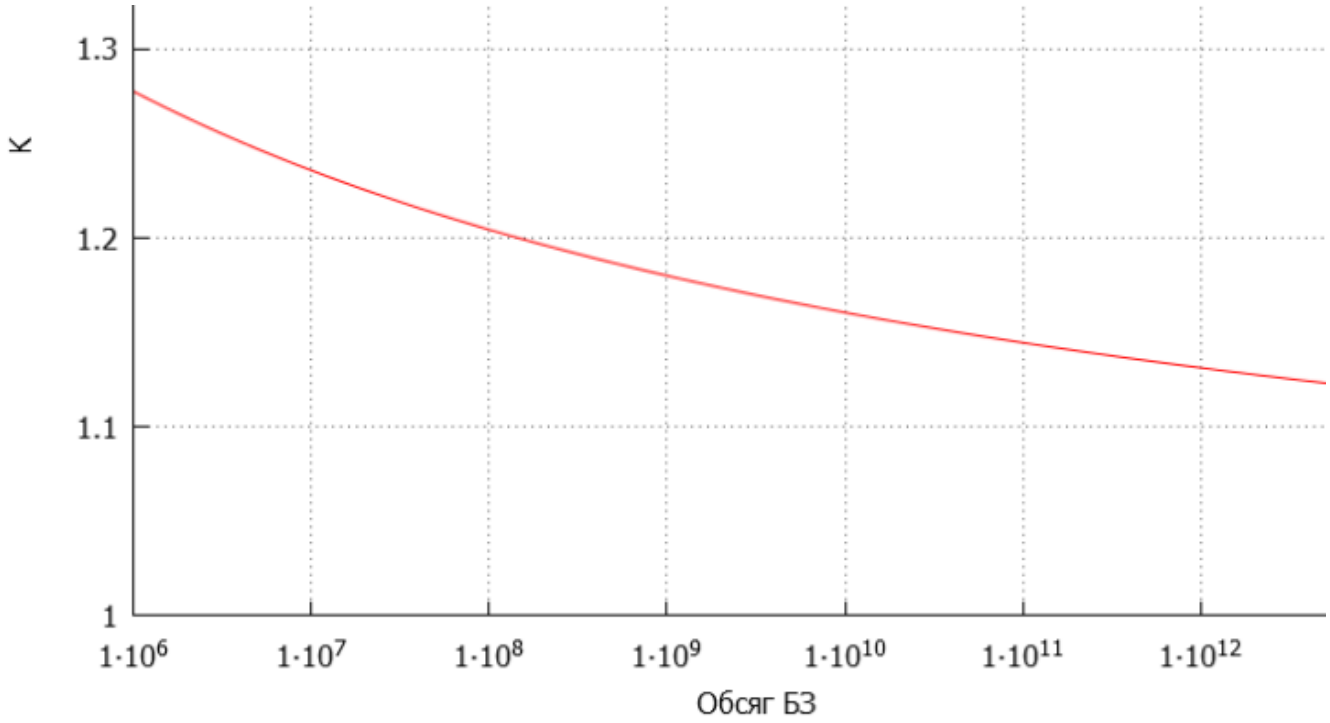


Рисунок 3.9 – Відносна складність пошуку пошук за простим запитом, уточнена

По осі абсцис відзначено об’єм ПМБЗ в одиницях слів, по осі ординат – умовна розрахункова складність операції пошуку.

Як видно з графіків, швидкість операції пошуку за символами в ІС ОПМ більша, ніж у аналогів – на 12% для максимального об’єму БЗ та на 12-27% для уточненого об’єму БЗ.

4.2.5 Оцінка складності пошуку за складним запитом

ПМБЗ ІІІ. В загальному випадку фрагмент знань в ІС ОПМ представляє собою множину КЗ, пов’язаних між собою відношеннями. Виконання такої операції подібне до виконання операції пошуку за КЗ, але крім складових КЗ та зв’язків між ними, необхідно враховувати також відношення, до яких цей КЗ належить.

Отже, обчислювальна складність цієї операції складає

$$C = \sum_{i=1}^{w} \log_2(N_i^Q \cdot N_i^R), \qquad (4.25)$$

де N_i^Q – кількість КЗ у результатах пошуку на $(i-1)$ кроці,

N_i^R – кількість відношень кожного з цих КЗ,

w – кількість слів у запиті.

Аналог. Найкращий аналог цієї операції – пошук у семантичних мережах, який реалізований як пошук у векторному просторі. Обчислювальна складність такого пошуку не змінюється (формула (4.20)).

Порівняння. Задача пошуку за складним запитом в ІС ОПМ відповідає пошуку даного КЗ та його відношень у мережі КЗ.

Отже, складність такої операції згідно формулі (4.25) становить $C = \sum_{i=1}^{w} \log_2(N_i^Q \cdot N_i^R)$. Як показано у [90], в середньому кожна БССС має 7±2 відношення, тобто можемо використати модифіковану формулу (4.21) наступного вигляду:

$$C_1 = \sum_{n=0}^{6} 0{,}015^n \cdot \log_2(7 \cdot n) = 1{,}015 \cdot \log_2(7 \cdot n), \tag{4.26}$$

Для семантичних мереж складність цієї операції становить:

$$C_2 = \log_2(60 \cdot n), \tag{4.27}$$

Отже, коефіцієнт відносної складності пошуку за символами визначається за формулою

$$K(n) = \frac{C_2}{C_1} = \frac{\log_2(60 \cdot n)}{1{,}015 \cdot \log_2(7 \cdot n)}, \tag{4.28}$$

Визначимо граничне значення цього коефіцієнту для максимального розрахункового обсягу БЗ:

$$\lim_{n \to max} K(n) = \lim_{n \to 10^{13}} \frac{\ln(60 \cdot n)}{1{,}015 \cdot \ln(7 \cdot n)} \approx 1{,}05162, \tag{4.29}$$

Отже, для операції пошуку за символами гранична ефективність ІС ОПМ становить ~105% від ефективності аналогів.

Оцінимо характер цільової функції на всій області визначення (рис. 3.10)

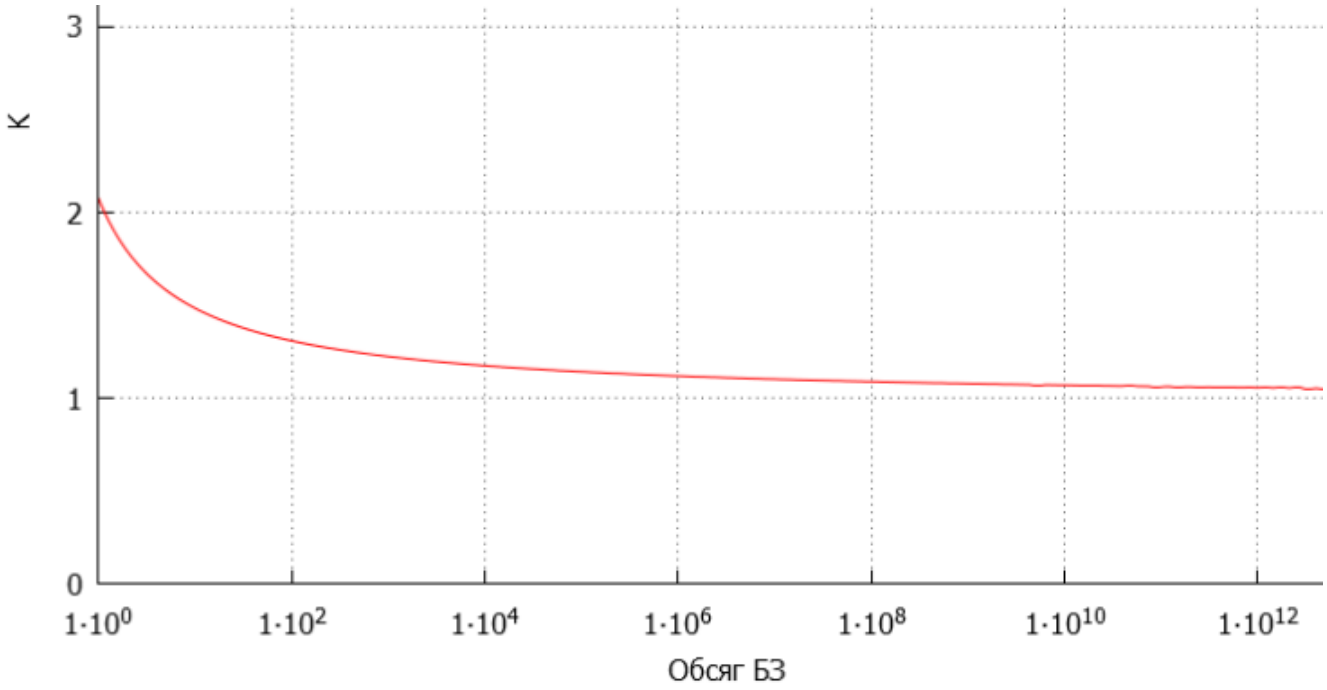


Рисунок 3.10 – Відносна складність пошуку, пошук за складним запитом

По осі абсцис відзначено об’єм ПМБЗ у одиницях слів, по осі ординат – умовна розрахункова складність операції пошуку.

Оцінимо характер цільової функції на уточненій області визначення (рис. 3.11)

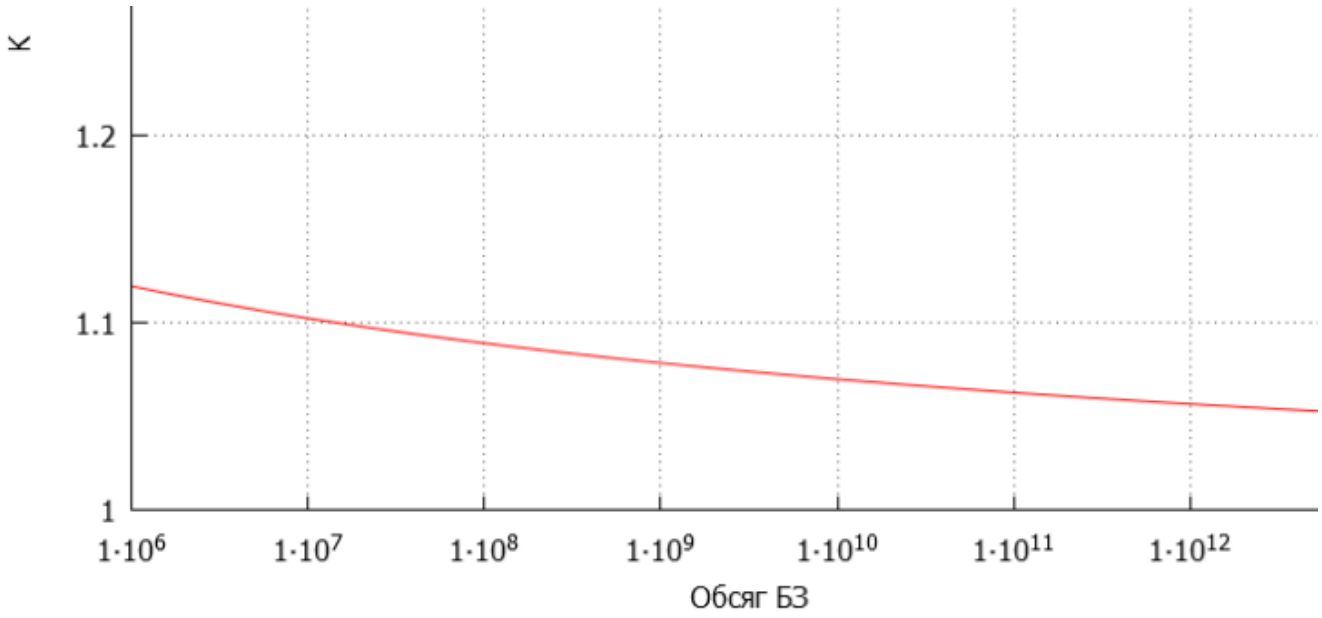


Рисунок 3.11 – Відносна складність пошуку, пошук за складним запитом, уточнена

По осі абсцис відзначено об’єм ПМБЗ в одиницях слів, по осі ординат – умовна розрахункова складність операції пошуку.

Як видно з графіків, швидкість операції пошуку за символами в ІС ОПМ більша, ніж у аналогів – на 5% для максимального об’єму БЗ та на 5-12% для уточненого об’єму БЗ.

Враховуючи, що кількість елементів, які відповідають вмісту пошукового запиту, значно менша за повне оточення для кожного з елементів K_S, а кількість таких з них, які одночасно відповідають ще й маркеру пошукового запиту (тобто складаються з заданих КЗ та відношень, поєднаних заданим чином у заданому порядку), ще звужує поле результатів пошуку, можемо стверджувати, що середня складність цієї операції буде значно менша за отриману складність.

Оскільки кількість КЗ та відношень обмежена, але достатньо велика, для пошуку у БД це відповідає пошуку по одному полю або сутності бази даних з високою кардинальністю (*cardinality*) [127], що є дуже вагомим фактором підвищення його швидкості.

Крім того, необхідно враховувати можливість використання метаданих, пов’язаних з ПМБЗ, у операціях пошуку. Хоча метадані й не є частиною ПМБЗ,

використання їх для пошуку може дозволити значно зменшити обсяг проміжних його результатів на кожному кроці.

Зауважимо також, що середня складність операції пошуку фрагменту знань буде меншою, оскільки ПМБЗ в загальному випадку містить велику кількість унікальних КЗ та відношень, в той час як кожний КЗ або відношення має обмежену кількість власних зв'язків, лише мала частина з яких відповідає даному пошуковому запиту. Таким чином, на кожній ітерації пошуку після першої можемо виділяти підграф результатів цього пошуку K_S, й, ідентифікувавши координати кожного з його елементів у ПМБЗ, виконувати пошук лише у їх безпосередньому оточенні шляхом порівняння усіх їх зв'язків з наявними у пошуковому запиті.

Отже, ІС ОПМ значно гірше за аналоги опрацьовує пошук за символами (гірше на 26%), але на інших видах пошуку показує кращі результати – на 7% краще в найгіршому випадку і до 12-27% краще на прогнозованих обсягах даних.

4.3 Експериментальна оцінка пошуку

4.3.1 Умови експерименту

Виконаємо експериментальну перевірку роботи ІТ ОПМ на прикладі покращення результатів природномовного пошуку шляхом їх обробки за допомогою ІС ОПМ.

Для цього виконуємо наступні кроки.

1. Сформувати пошукові запити.
2. Заповнити ПМБЗ даними, які дозволять використовувати її для обробки результатів пошуку.
3. Виконати пошук та записати результати для їх подальшої обробки з використанням ПМБЗ ІП.
4. Оцінити релевантність результатів без використання ПМБЗ ІП.

5. Обробити результати пошуку за допомогою ПМБЗ ІП та сформувати оброблений масив результатів.
6. Оцінити релевантність результатів після обробки ПМБЗ.
7. Порівняти результати.

Основними критеріями вибору джерела тексту є його семантична оцінка, тобто можливість виділення з тексту максимальної кількості знань без використання додаткових довідкових матеріалів; граматична і стилістична оцінка тексту та мінімальна кількість елементів, які підвищують складність обробки тексту (формул, символів, абревіатур тощо). В якості такого джерела запитів використовуємо набір різноманітних фрагментів тексту з природномовних джерел інформації, зокрема класичних творів української літератури, сайтів новин, юридичних документів, наукових публікацій.

В експерименті аналізуємо тільки перші 100 результатів (10 сторінок) з пошукової видачі природномовної пошукової системи *Google Search*. На основі тестових запусків було встановлено оптимальний розмір пошукового запиту у 2-3 семантично пов’язаних слова, оскільки для більших запитів характерна дуже невелика кількість повністю релевантних результатів пошуку. Пошук виконується з установками за замовчуванням, оскільки оператор точного пошуку «лапки» відкидає велику кількість результатів, де зв’язки між словами відповідають пошуковому запиту, але самі слова в тексті не розміщені у точно такому ж порядку.

Для формування граматичної структури використаємо бібліотеку *pyMorphy2*, принципи роботи якої описані у статті [128]. Головною перевагою цієї бібліотека є можливість опрацювання україномовних текстів, що вигідно відрізняє її від більшості інших аналогічних продуктів, орієнтованих в основному на російську або англійську мову. Бібліотека *pyMorphy2* написана з використанням мови програмування *Python*, яка, крім простоти використання, є популярним інструментом для наукових досліджень. Ця бібліотека використовує алгоритми відмінювання по Залізняку [129] та словник у форматі *OpenCorpora* [130] і дозволяє використовувати усі варіанти словоформ при виникненні неоднозначності.

В якості цільового критерія виберемо релевантність пошуку – кількість результатів, релевантних пошуковому запиту, серед усіх знайдених результатів, у порівнянні з базовими результатами пошуку, тобто пошуком на основі ключових слів.

4.3.2 Приклад обробки результатів природномовного пошуку з використанням ІС ОПМ

Для демонстрації процедури обробки результатів пошуку виберемо наступні пошукові запити:

1) контрольний простий запит для визначення базової релевантності результатів пошуку: «мова».
2) запит вигляду $Subj + Attr$, що описує певний об’єкт, визначений атрибутом: «природна мова».
3) запит вигляду $Subj + Pred$, що описує певний об’єкт та дію, яку він виконує: «мова дозволяє».

Занесемо у ПМБЗ структури знань, що відповідають певним пошуковим запитам.

Для цього:

- виділимо синтаксичну структуру тексту цих запитів;
- сформуємо на основі цієї структури кванти знань;
- запишемо кванти знань у ПМБЗ.

Сформуємо кванти знань цих запитів.

Запит «мова» містить один КЗ, що складається з одного $Subj$:

$$Q_1 = \{\text{мова}\} \tag{4.30}$$

Запишемо його у БЗ з маркером M_1 (рис. 4.1).

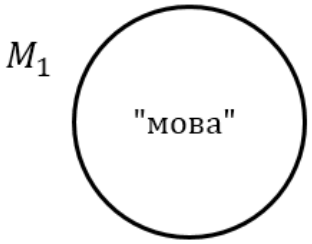


Рисунок 4.1 – Структура знань запиту «мова»

Запит «природна мова» містить один КЗ, що складається з одного $Subj$ та повязанного з ним $Attr$:

$$Q_2 = \{\text{мова}, \text{природна}\} \tag{4.31}$$

Запишемо його у БЗ з маркером M_2 (рис. 4.2).

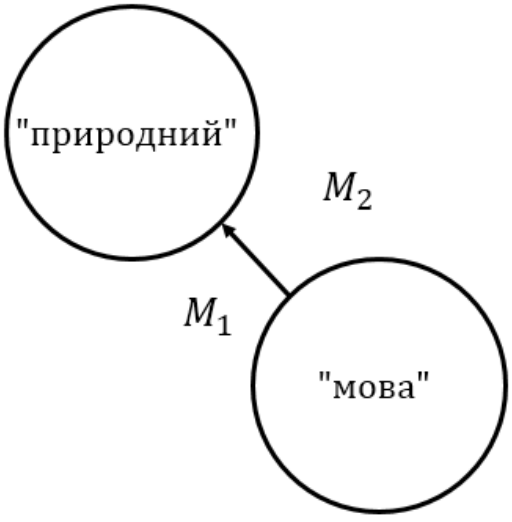


Рисунок 4.2 – Структура знань запиту «природна мова»

Зазначимо, що різні схеми опису одного й того ж КЗ, отримані з різних фрагментів тексту, можуть бути представлені одним КЗ у ПМБЗ, оскільки їх текстове представлення пов’язане з КЗ тільки через відповідні маркери текстів [131].

Запит «мова дозволяє» містить один КЗ, що складається з одного $Subj$ та повязанного з ним $Pred$:

$$Q_3 = \{\text{мова}, \text{дозволяє}\} \quad (4.32)$$

Запишемо його у БЗ з маркером M_3 (рис. 4.3).

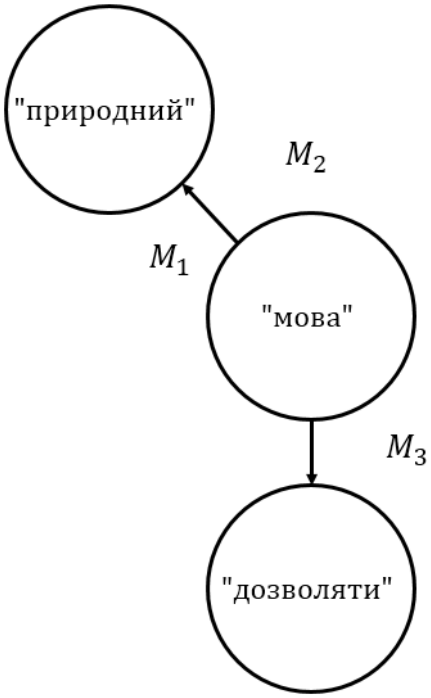


Рисунок 4.3 – Структура знань запиту «мова дозволяє»

Зазначимо, що через лексему «мова» проходить декілька незалежних КЗ, визначені відповідними маркерами. Ці маркери не лише пов’язують структуру знань та вхідний фрагмент тексту, але й зберігають порядок слів у тексті та їх зв’язки в рамках КЗ, що дозволяє точно відтворити відповідні фрагменти тексту на основі даних з БЗ.

Підготуємо матеріали для аналізу та визначимо релевантність результатів пошукової вибірки до їх обробки в ІС ОПМ.

Запит: «мова»

Знайдено результатів: 69,600

Пошук за простим запитом, що містить лише одне слово і не містить зв’язків між ними, показує базову релевантність видачі пошукової системи.

Релевантність пошукової видачі наведено у таблиці:

Таблиця 4.6 – Релевантність результатів пошуку за простим запитом

№ сторінки	Кіль-ть результатів	Релевантних результатів	Релевантність
1	10	10	1
2	10	10	1
3	10	10	1
4	10	10	1
5	10	10	1
6	10	10	1
7	10	10	1
8	10	10	1
9	10	10	1
10	10	10	1
∑	100	100	1

Релевантність видачі становить 1, або 100% – тобто, пошукова система працює добре і не вносить додаткову похибку у оцінку.

Запит: «природна мова»

Знайдено результатів: 17,400

Пошук за запитом, що включає в себе $Subj$ та $Attr$, вимагає аналізу синтаксичних та семантичних зв'язків між елементами тексту.

Релевантність пошукової видачі наведено у таблиці:

Таблиця 4.7 – Релевантність результатів пошуку за $Subj$ та $Attr$

№ сторінки	Кіль-ть результатів	Релевантних результатів	Релевантність
1	10	5	0.5
2	10	4	0.4
3	10	6	0.6
4	10	4	0.4
5	10	3	0.3
6	10	4	0.4
7	10	3	0.3
8	10	5	0.5
9	10	1	0.1
10	10	1	0.1
∑	100	36	0.36

Як видно з таблиці, загальна релевантність становить 36% та зменшується зі зростанням порядкового номеру результату видачі. Отже, пошукова система не враховує в достатній мірі зв'язки між елементами КЗ, навіть в порівняно простому випадку ($Subj$ та $Attr$), коли ці зв'язки є досить однозначними.

Запит: «мова дозволяє»

Результатів всього: 14,400

Пошук за запитом, що включає в себе $Subj$ та $Pred$, є більш специфічним, ніж пошук $Subj$ та $Attr$, оскільки $Subj$ та $Attr$ визначають певний суб'єкт і його повне семантичне оточення, а $Subj$ та $Pred$ – лише одну дію, яку такий суб'єкт виконує.

Релевантність пошукової видачі наведено у таблиці:

Таблиця 4.8 – Релевантність результатів пошуку за $Subj$ та $Pred$

№ сторінки	Кіль-ть результатів	Релевантних результатів	Релевантність
1	10	5	0.5
2	10	2	0.2
3	10	2	0.2
4	10	1	0.1
5	10	1	0.1
6	10	0	0
7	10	0	0
8	10	0	0
9	10	0	0
10	10	0	0
∑	100	11	0.11

Як видно з таблиці, релевантність є найнижчою з розглянутих запитів, причому починаючи з 5 сторінки вона опускається до 0.

Проаналізуємо результати пошуку, сформовані вище, після їх автоматичної обробки з використанням ІТ ОПМ, згідно правилам представленим у [132].

Для кожного запису у результатах було виконано наступні операції:

- розбити запис на фрагменти, обмежені знаками пунктуації;
- якщо у жодному фрагменті не присутні усі лексеми, відкинути результат;
- якщо граматична форма лексем у фрагменті не відповідає одна одній, відкинути результат;
- якщо у хоча б одному фрагменті присутні усі лексеми пошукового запиту у відповідних граматичних формах, підтвердити результат.

Для складних запитів додаємо показник «кількість помилкових результатів» - тобто кількість тих результатів, які були помилково відзначені як релевантні або нерелевантні.

Релевантність обробленої пошукової видачі наведено нижче.

Таблиця 4.9 – Релевантність результатів пошуку за $Subj$ та $Attr$

№ сторінки	Кіль-ть результатів	Релевантних результатів, усього	Релевантних результатів, знайдено	Помилкових результатів	Релевантність
1	10	5	5	0	0.5
2	10	4	4	0	0.4
3	10	6	6	0	0.6
4	10	4	4	0	0.4
5	10	3	3	0	0.3
6	10	4	4	0	0.4
7	10	3	3	0	0.3
8	10	5	5	0	0.5
9	10	1	1	0	0.1
10	10	1	1	0	0.1
∑	100	36	36	0	0.36

Як видно з таблиці, для запиту, що включає в себе $Subj$ та $Attr$, навіть з використанням тривіальної процедури обробки вдалося правильно визначити усі релевантні записи і відкинути 64% початкових записів, які не є релевантними.

Тобто, усі релевантні запити з перших 10 сторінок вихідного пошуку розміщуються на перших 4 сторінках обробленого пошуку.

Таблиця 4.10 – Релевантність результатів пошуку за $Subj$ та $Pred$

№ сторінки	Кіль-ть результатів	Релевантних результатів, до обробки	Релевантних результатів, після обробки	Помилкових результатів	Релевантність
1	10	5	5	0	0.5
2	10	2	2	0	0.2
3	10	2	3	1	0.2
4	10	1	0	1	0.1
5	10	1	0	1	0.1
6	10	0	0	0	0
7	10	0	0	0	0
8	10	0	0	0	0
9	10	0	0	0	0
10	10	0	0	0	0
∑	100	11	8	3	0.08

Для пошукового запиту, що включає в себе $Subj$ та $Pred$, чисельні показники аналогічні попереднім, але при цьому 3 запити було оброблено неправильно:

- в одному випадку пропущено КЗ, елементи якого розділені знаками пунктуації;
- в двох випадках неправильно підтверджено запити, які не є релевантними.

Отже, для більш складної конструкції ($Subj$ та $Pred$) більш актуальним постає розробка та використання більш досконалого лінгвістичного процесору.

В результаті обробки пошукової видачі за запитом, що містить $Subj$ та $Pred$, було відкинуто 92% результатів, 3% з них помилково, тобто релевантні результати з перших 10 сторінок видачі розміщуються на перших 2 сторінках обробленої вибірки.

4.3.3 Аналіз результатів обробки вхідних даних з використанням інформаційної технології

Проаналізуємо результати обробки усіх пошукових запитів з вхідних даних (вхідні дані та результати обробки представлені у Додатку Е).

Для кожного запиту було визначено та записано такі показники:

- текст запиту;
- кількість результатів пошуку; для більшості запитів це число становило 100 (100 перших результатів), але для деяких з них загальна кількість результатів була меншою;
- кількість результатів, які при обробці ІС ОПМ були автоматично визначені релевантними;
- помилкові результати – скільки результатів було помилково віднесено до релевантних або помилково пропущено;
- кількість результатів, які були правильно визначені релевантними.

Діапазон отриманих значень релевантності становить від 3% до 100%, тобто релевантність не обмежена жорсткими рамками, а ефективність додаткової обробки результатів пошуку ІС ОПМ залежить від запиту.

Для більшості запитів кількість помилкових результатів становить 0-1%, максимальне значення 6%. Таким чином, можемо стверджувати що вплив помилок обробки на її результати є відносно невеликим.

Середня релевантність до обробки результатів і після становить 72.5% та 72% відповідно. Це означає що помилкові запити незначно впливають на результати, і

навіть при використанні примітивного лінгвістичного процесора похибка не є суттєвою.

Статистичні оцінки отриманих результатів: медіана дорівнює 86%, нижній квартиль 48.25%, верхній квартиль 97%. З цього можемо зробити такі висновки:

- додаткова обробки пошукових запитів з використанням ІС ОПМ є дуже ефективною для нижніх 25% результатів, де релевантність менша за 50%, і ідентифікація нерелевантних результатів значно зменшує обсяг даних. Це є важливим як для тих задач, які передбачають подальшу обробку цих даних, так і для пошуку за запитами які мають низьку початкову релевантність;
- обробка пошукових запитів з використанням ІС ОПМ не є ефективною для верхніх 25% результатів, релевантність яких сягає понад 97%. Цей показник достатньо хороший для більшості прикладних задач, і для його покращення краще застосовувати більш точні інструменти;
- обробки пошукових запитів з використанням ІС ОПМ дозволяє в середньому покращити ефективність на (100% - 86%) = 14%. Це достатньо хороший показник, який дозволяє стверджувати про ефективність роботи ІС ОПМ та доцільність її використання;
- ІС ОПМ може бути використана для оцінки релевантності результатів взагалі, зокрема при перевірці результатів перекладу.

Висновки до розділу 4

1. Розроблено технічні вимоги до інформаційної системи, яка реалізує інформаційну технологію обробки природномовних текстів на основі інтеграційного підходу, в тому числі обрано веб-архітектуру як архітектуру інформаційної системи, визначено підсистеми інформаційної системи, операції, які в вона виконує та розроблено схему бази даних для неї.

2. Виконано оцінку обчислювальної складності розробленої інформаційної системи та її порівняння зі складністю пошуку у аналогах. Теоретично показано, що складність пошуку в розробленій інформаційній системі не перевищує так для аналогів, і в середньому є на 5-12% меншою для складних пошукових запитів.
3. Експериментально доведено, що використання розробленої інформаційної системи для уточнення результатів природномовного пошуку дозволяє в середньому підвищити релевантність результатів на 14%.

ЗАГАЛЬНІ ВИСНОВКИ

Дисертаційна робота становить собою закінчене наукове дослідження, що вирішує актуальну науково-технічну задачу розробки інформаційної технології обробки природномовних текстів на основі інтеграційного підходу.

В рамках роботи поставлені та вирішені такі завдання.

1. На основі аналізу підходів до розробки природномовних баз знань обґрунтовано потребу у розробленні універсальної моделі представлення знань природномовного тексту для використання в технологіях обробки природної мови, що поєднує переваги існуючих підходів, та процедури використання такої моделі в інформаційних технологіях обробки природномовних текстів.
2. На основі інтеграційного підходу до моделювання мовленнєвої діяльності людини розроблено модель представлення знань для використання в технологіях обробки природної мови. Ця модель відрізняється від аналогів тим, що дозволяє представити фрагмент знань довільного природномовного тексту у вигляді універсальної структури. Перевагою розробленої моделі представлення знань є її незалежність від синтаксичної структури тексту та семантичного контексту фрагменту знань.
3. Розроблено процедури записування та пошуку знань з використанням розробленої моделі представлення знань в технологіях обробки природної мови, які дозволяють встановити зв’язки на структурному рівні між синтаксичною структурою тексту та довільною структурою метаданих.
4. Розроблено інформаційну технологію обробки природномовних текстів на основі інтеграційного підходу, для якої теоретично показано, що складність пошуку не перевищує так для аналогів, і в середньому є на 5-12% меншою для складних пошукових запитів.
5. Експериментально показано, що використання природномовної бази знань на основі інтеграційного підходу для природномовного пошуку дозволяє покращити якість роботи систем природномовного пошуку, а саме підвищити

середню релевантність результатів на 14%. Впровадження результатів роботи у виробництві призвело до зменшення витрат часу працівників на контроль за складом продукції на 25% та збільшення конверсії природномовного пошуку на 8% відповідно.

ЛІТЕРАТУРА

1. Koehn P, Knowles R. Six Challenges for Neural Machine Translation. ACL Workshop on Neural Machine Translation [Internet]. 2017;28–9. Available from: http://arxiv.org/abs/1706.03872
2. Internet Live Stats. Internet Users by Country (2016) [Internet]. Internet Live Stats. 2016. Available from: http://www.internetlivestats.com/internet-users-by-country/
3. Liddy ED, Hovy E, Lin J, Prager J, Radev D, Vanderwende L, et al. Natural Language Processing. Encyclopedia of Library and Information Science. 2003;2126–2136.
4. Manning CD, Schütze H. Foundations of Natural Language Processing [Internet]. MIT Press. 1999. 678 p. Available from: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Foundations+of+Natural+Language+Processing#3
5. Aron J. How innovative is Apple's new voice assistant, Siri? New Scientist [Internet]. 2011;212(2836):24. Available from: http://www.sciencedirect.com/science/article/pii/S026240791162647X
6. Brown K, Isabelle P, Foster G. Machine Translation: Overview. In: Encyclopedia of Language & Linguistics. 2006. p. 404–22.
7. Borovikov E, Lane W. A survey of modern optical character recognition techniques (DRAFT). arXiv:14124183v1 [csCV]. 2004;1(301):1–37.
8. Rawat S, Gupta P, Kumar P. Digital life assistant using automated speech recognition. In: Proceedings of the International Conference on Innovative Applications of Computational Intelligence on Power, Energy and Controls with Their Impact on Humanity, CIPECH 2014. 2014. p. 43–7.
9. Mehrotra S, Kohli S. Application of clustering for improving search result of a website. In: Advances in Intelligent Systems and Computing. 2016. p. 349–56.
10. Harris MD. Building a Large-scale Commercial NLG System for an EMR.
11. Russell SJ (Stuart J, Norvig P, Canny J. Artificial intelligence : a modern approach.

1081 p.

12. Briscoe T. Introduction to Linguistics for Natural Language Processing. October. 2011;1–37.
13. Peng X, Cao H, Setlur S, Govindaraju V, Natarajan P. Multilingual OCR research and applications. 4th International Workshop on Multilingual OCR - MOCR ’13 [Internet]. 2013;(ii):1–8. Available from: http://dl.acm.org/citation.cfm?doid=2505377.2509977
14. Anusuya M, Katti S. Speech recognition by machine: A review. International Journal of Computer Science and Information Security. 2009;6(3):181–205.
15. Bicknese DA. Measuring the Accuracy of the OCR in the Making of America [Internet]. Michigan; 1998 [cited 2017 Mar 14]. Available from: https://quod.lib.umich.edu/m/moagrp/moaocr.html
16. Hannun A, Case C, Casper J, Catanzaro B, Diamos G, Elsen E, et al. Deep Speech: Scaling up end-to-end speech recognition. Arxiv. 2014;1–12.
17. Xiong W, Droppo J, Huang X, Seide F, Seltzer M, Stolcke A, et al. Achieving Human Parity in Conversational Speech Recognition. arXiv [Internet]. 2016; Available from: http://arxiv.org/abs/1610.05256
18. Smith R, Antonova D, Lee D. Adapting the Tesseract Open Source OCR Engine for Multilingual OCR. MOCR: Proceedings of the International Worksop on Multilingual OCR. 2009;(Cc):1–8.
19. Liu KC, Wu CH, Tseng SY, Tsai Y Te. Voice helper: A mobile assistive system for visually impaired persons. In: Proceedings - 15th IEEE International Conference on Computer and Information Technology, CIT 2015. 2015. p. 1400–5.
20. Swetha N, Anuradha K. TEXT-TO-SPEECH CONVERSION. International Journal of Advanced Trends in Computer Science and Engineering. 2013;2(6):269–78.
21. Shaheen S, El-Hajj W, Hajj H, Elbassuoni S. Emotion recognition from text based on automatically generated rules. In: IEEE International Conference on Data Mining Workshops, ICDMW. 2015. p. 383–92.

22. Frantzi K, Ananiadou S, Mima H. Automatic recognition of multi-word terms: The C-value/NC-value method. International Journal on Digital Libraries. 2000;3(2):115–30.
23. Wu Y, Denny JC, Rosenbloom ST, Miller RA, Giuse DA, Xu H. A comparative study of current Clinical Natural Language Processing systems on handling abbreviations in discharge summaries. AMIA . Annual Symposium proceedings / AMIA Symposium AMIA Symposium [Internet]. 2012;2012:997–1003. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23304375%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3540461
24. Chapman NP. LR parsing : theory and practice. Cambridge University Press; 1987. 228 p.
25. Abney SP. Parsing By Chunks. Principle-Based Parsing. 1991;257–78.
26. Frakes WB (William B, Baeza-Yates R (Ricardo). Information retrieval : data structures & algorithms. Information retrieval. Prentice Hall; 1992. 504 p.
27. Gesmundo A, Samardžić T. Lemmatisation as a tagging task. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2. Association for Computational Linguistics; 2012. p. 368–72.
28. Perera P, Witte R. A self-learning context-aware lemmatizer for German. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT '05 [Internet]. Morristown, NJ, USA: Association for Computational Linguistics; 2005 [cited 2016 Nov 9]. p. 636–43. Available from: http://portal.acm.org/citation.cfm?doid=1220575.1220655
29. Rodrigues R, Oliveira HG, Gomes P. LemPORT: a High-Accuracy Cross-Platform Lemmatizer for Portuguese. Maria João Varanda Pereira José Paulo Leal. 2014;267.
30. Apparatus and methods for an information retrieval system that employs natural language processing of search results to improve overall precision. 1997;
31. Jackson P, Moulinier I. Natural language processing for online applications : text retrieval, extraction, and categorization. John Benjamins Pub; 2002. 225 p.

32. James CL, Reischel KM. Text input for mobile devices. In: Proceedings of the SIGCHI conference on Human factors in computing systems - CHI ’01 [Internet]. New York, New York, USA: ACM Press; 2001 [cited 2016 Nov 9]. p. 365–71. Available from: http://portal.acm.org/citation.cfm?doid=365024.365300
33. Busch JE, Lin AD, Graydon PJ, Caudill M. Ontology-based parser for natural language processing [Internet]. Google Patents; 2006. Available from: https://www.google.com/patents/US7027974
34. Ingason AK, Helgadóttir S, Loftsson H, Rögnvaldsson E. A Mixed Method Lemmatization Algorithm Using a Hierarchy of Linguistic Identities (HOLI). In: Advances in Natural Language Processing [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2008 [cited 2016 Nov 9]. p. 205–16. Available from: http://link.springer.com/10.1007/978-3-540-85287-2_20
35. Apache Software Foundation. LanguageTool | Apache OpenOffice Extensions [Internet]. 2016 [cited 2016 Nov 1]. Available from: http://extensions.openoffice.org/en/project/languagetool
36. Grammarly Inc. Free Grammar Checker | Grammarly [Internet]. [cited 2016 Nov 1]. Available from: https://www.grammarly.com/
37. Schuster S, Manning CD. Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016) [Internet]. 2016;2371–8. Available from: http://nlp.stanford.edu/~sebschu/pubs/schuster-manning-lrec2016.pdf
38. De Marneffe M-C, MacCartney B, Manning CD. Generating typed dependency parses from phrase structure parses. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006) [Internet]. 2006;449–54. Available from: http://nlp.stanford.edu/pubs/LREC06_dependencies.pdf
39. Klein D, Manning CD. Fast Exact Inference with a Factored Model for Natural Language Parsing. Advances in Neural Information Processing Systems [Internet]. 2003;15:3–10. Available from:

http://machinelearning.wustl.edu/mlpapers/paper_files/CS01.pdf%5Cnhttp://www-nlp.stanford.edu/~manning/papers/lex-parser.pdf

40. Bermudez ME, Logothetis G. Simple computation of LALR(1) lookahead sets. Information Processing Letters. 1989;31(5):233–8.
41. Grune D, Jacobs CJH. Parsing techniques: A practical guide. Parsing Techniques: A Practical Guide. 2008. 1-662 p.
42. Barker K, Szpakowicz S, Barker K, Szpakowicz S. Semi-automatic recognition of noun modifier relationships. In: Proceedings of the 36th annual meeting on Association for Computational Linguistics - [Internet]. Morristown, NJ, USA: Association for Computational Linguistics; 1998 [cited 2016 Nov 23]. p. 96. Available from: http://portal.acm.org/citation.cfm?doid=980845.980862
43. Hasting AS, Kotz SA, Friederici AD. Setting the Stage for Automatic Syntax Processing: The Mismatch Negativity as an Indicator of Syntactic Priming. Journal of Cognitive Neuroscience [Internet]. 2007 Mar [cited 2016 Nov 23];19(3):386–400. Available from: http://www.mitpressjournals.org/doi/abs/10.1162/jocn.2007.19.3.386
44. Campbell JC, Hindle A, Amaral JN. Syntax errors just aren't natural: improving error reporting with language models. In: Proceedings of the 11th Working Conference on Mining Software Repositories - MSR 2014 [Internet]. New York, New York, USA: ACM Press; 2014 [cited 2016 Nov 23]. p. 252–61. Available from: http://dl.acm.org/citation.cfm?doid=2597073.2597102
45. Neubig G. Document Level Models [Internet]. CS11-747 Neural Networks for NLP. 2018. p. 2. Available from: http://www.phontron.com/class/nn4nlp2018/assets/slides/nn4nlp-21-document.pdf
46. Cer D, Marneffe M De, Jurafsky D, Manning CD, de Marneffe M-C. Parsing to Stanford Dependencies: Trade-offs between Speed and Accuracy. Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). 2010;0:1628–32.
47. Попова І. Досвід загальної типології синтаксичної омонімії. Лінгвістичні

студії [Internet]. 2008 [cited 2017 Mar 7];(18):86–92. Available from: http://www.nbuv.gov.ua/old_jrn/Soc_Gum/Ls/2009_18/18.17_popova_typologiya_omonimii.pdf

48. Tablan V, Damljanovic D, Bontcheva K. A natural language query interface to structured information. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2008. p. 361–75.
49. Weizenbaum J. ELIZA — A Computer Program For the Study of Natural Language Communication Between Man And Machine. Communications of the ACM [Internet]. 1966;9(1):36–45. Available from: http://joi.jlc.jst.go.jp/JST.Journalarchive/jje1965/2.3_1?from=CrossRef
50. Shawar BA, Atwell E. ALICE chatbot: Trials and outputs. Computacion y Sistemas. 2015;19(4):625–32.
51. Luria M, Hoffman G, Zuckerman O. Comparing Social Robot, Screen and Voice Interfaces for Smart-Home Control. Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17 [Internet]. 2017;580–628. Available from: http://dl.acm.org/citation.cfm?doid=3025453.3025786
52. Cimiano P, Haase P, Heizmann J, Mantel M, Studer R. Towards portable natural language interfaces to knowledge bases - The case of the ORAKEL system. Data and Knowledge Engineering. 2008;65(2):325–54.
53. Kroeger PR. Analyzing Grammar An Introduction. Vol. 53, Journal of Chemical Information and Modeling. 2005. 1689-1699 p.
54. Yamamoto H, Isogai S, Sagisaka Y. Multi-Class Composite N-gram language model for spoken language processing using multiple word clusters. In: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics - ACL '01 [Internet]. Morristown, NJ, USA: Association for Computational Linguistics; 2001 [cited 2016 Nov 9]. p. 531–8. Available from: http://portal.acm.org/citation.cfm?doid=1073012.1073080
55. Smadja FA, A. F. From N-grams to collocations. In: Proceedings of the 29th annual

meeting on Association for Computational Linguistics - [Internet]. Morristown, NJ, USA: Association for Computational Linguistics; 1991 [cited 2016 Nov 9]. p. 279–84. Available from: http://portal.acm.org/citation.cfm?doid=981344.981380

56. Kopotev M, Pivovarova L, Kochetkova N, Yangarber R. Automatic Detection of Stable Grammatical Features in N-Grams. 2013;73–81.
57. Dunning T. Statistical Identification of Language. Vol. 94, Computing. 1994. p. 94–273.
58. Khurana H, Basney J, Bakht M, Freemon M, Welch V, Butler R. Palantir. In: Proceedings of the 8th Symposium on Identity and Trust on the Internet - IDtrust '09 [Internet]. New York, New York, USA: ACM Press; 2009 [cited 2016 Nov 9]. p. 38. Available from: http://portal.acm.org/citation.cfm?doid=1527017.1527023
59. Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. 2016. p. 1520–8.
60. Yucong D, Cruz C. Formalizing semantic of natural language through conceptualization from existence. International Journal of Innovation, Management and Technology. 2011;2(1).
61. Camacho-Collados J, Iacobacci I, Navigli R, Pilehvar MT. Semantic Representations of Word Senses and Concepts. 2016 Aug 2 [cited 2016 Nov 9]; Available from: http://arxiv.org/abs/1608.00841
62. Langacker RW. Cognitive Grammar. In: The Oxford Handbook of Cognitive Linguistics. 2012.
63. Bawakid A, Oussalah M. A Semantic Summarization System: University of Birmingham at TAC 2008.
64. Honderich T. The Oxford Companion to Philosophy [Internet]. Vol. 29, History & Philosophy of Logic. 1995. 291-292 p. Available from: http://www.informaworld.com/openurl?genre=article&doi=10.1080/01445340701300429&magic=crossref
65. Akerkar R, Sajja P. Knowledge-Based Systems. Knowledge-Based Systems

[Internet]. 2010;23(5):1–114. Available from: http://proquest.safaribooksonline.com/9780763776473

66. Hirschberg J, Manning CD. Advances in natural language processing. Science [Internet]. 2015;349(6245):261–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26185244
67. Wijnhoven F. Operational knowledge management: identification of knowledge objects, operation methods, and goals and means for the support function. Journal of the Operational Research Society. 2003;54(2):194–203.
68. Писаревська ТА. Інформаційні системи і технології в управлінні трудовими ресурсами: Навч. посібник [Internet]. 2nd ed. Київ: КНЕУ; 2000. 279 p. Available from: http://ubooks.com.ua/books/00092/inx.php
69. Hamilton J, Hellerstein JM, Stonebraker M, Hamilton J. Architecture of a Database System. J M Hellerstein. 2007;1(2):141–259.
70. Jarke M, Neumann B, Vassiliou Y, Wahlster W. KBMS Requirements of Knowledge-Based Systems. Foundations of Knowledge Base Management: Contributions from Logic, Databases, and Artificial Intelligence [Internet]. 1989;1–10. Available from: http://www.wolfgang-wahlster.de/wordpress/wp-content/uploads/KBMS_Requirements_of_Knowledge-Based_Systems.pdf
71. Сергеєв ДС. Особливості моделювання бази природно-мовних знань. In: Електроніка та інформаційні технології ЕлІТ-2015. Львів: Львів. нац. ун-т ім. І. Франка, ф-т електроніки; 2015. p. 17–20.
72. Krötzsch M, Vrandečić D, Völkel M, Haller H, Studer R. Semantic Wikipedia. Web Semantics. 2007;5(4):251–61.
73. Wolfram|Alpha Frequently Asked Questions [Internet]. [cited 2016 Nov 1]. Available from: https://www.wolframalpha.com/faqs2.html
74. Antoniou G, Van Harmelen F. OWL Web Ontology Language. Handbook on Ontologies in Information Systems [Internet]. 2004;2007(September):157–60. Available from: http://www.w3.org/TR/owl-features/
75. Bock C, Fokoue A, Haase P, Hoekstra R, Horrocks I, Ruttenberg A, et al. OWL 2

Web Ontology Language: Structural Specification and Functional-Style Syntax [Internet]. Syntax. 2008. p. 1–123. Available from: http://www.w3.org/TR/2008/WD-owl2-syntax-20081202/

76. Sowa JF. Principles of Semantic Networks: Explorations in the Representation of Knowledge. Morgan Kaufmann series in representation and reasoning. 1991. xi, 582.
77. Baker CF, Fillmore CJ, Cronin B. The structure of the framenet database. International Journal of Lexicography. 2003;16(3):281–96.
78. Emami A, Jelinek F. A neural syntactic language model. Machine Learning. 2005;60(1–3):195–227.
79. Zhuge H. A knowledge grid model and platform for global knowledge sharing. Expert Systems with Applications. 2002;22(4):313–20.
80. Singhal A. Official Google Blog: Introducing the Knowledge Graph: things, not strings [Internet]. 2012 [cited 2017 May 21]. Available from: https://googleblog.blogspot.co.uk/2012/05/introducing-knowledge-graph-things-not.html
81. Wolfengagen V. Frame Theory and Computations. Computers and artificial intelligence [Internet]. 1984 [cited 2017 Mar 21]; Available from: https://www.academia.edu/418210/Frame_Theory_and_Computations
82. Gailly F, Poels G. Conceptual modeling using domain ontologies: Improving the domain-specific quality of conceptual schemas. … of the 10th Workshop on Domain-Specific Modeling [Internet]. 2010; Available from: http://dl.acm.org/citation.cfm?id=2060367
83. Chomsky N. Syntactic structures. Janua linguarum Series minor ; no4 [Internet]. 2002;115. Available from: http://books.google.com/books?hl=en&lr=&id=mrz3TsgLPzQC&pgis=1
84. Stock WG. Concepts and semantic relations in information science. Journal of the American Society for Information Science and Technology [Internet]. 2010 Oct [cited 2016 Nov 3];61(10):1951–69. Available from:

http://doi.wiley.com/10.1002/asi.21382

85. Федорук ПІ. Графо-автоматна модель адаптивної системи дистанційного навчання та контролю знань. Математичні машини і системи [Internet]. 2006 [cited 2017 May 25];(4):144–54. Available from: http://www.immsp.kiev.ua/publications/articles/2006/2006_4/Fedoruk_04_2006.pdf
86. Harris R. The Origin of Writing. Vol. 28, Euphytica. 1986. x + 166.
87. Davis HG, Taylor TJ. Redefining linguistics. Vol. 9, Redefining Linguistics. 2014. 1-172 p.
88. Pablé A. Language, knowledge and reality: The integrationist on name variation. Language and Communication. 2010;30(2):109–22.
89. Hutton CM. Abstraction and Instance. The Type-Token Relation in Linguistic Theory. Pergamon Press; 1990. 180 p.
90. Кисленко ЮІ. Архітектура мови. Київ: ІЗМН; 1998. 344 p.
91. Кисленко ЮІ. От мысли к знанию (нейрофизиологические основания). Київ: Укр. літопис; 2008. 101 p.
92. Кисленко ЮІ. Рекурсивный синтаксический анализатор. Науковий вісник кафедри ЮНЕСКО Київського державного лінгвістичного університету. 2000;(1):157–64.
93. Kyslenko YI. Back to basics of speech activity. Biologically Inspired Cognitive Architectures. 2014;8:46–68.
94. Kyslenko Y, Sergeiev D. Cognitive architecture of speech activity and modelling thereof. Biologically Inspired Cognitive Architectures. 2015;12.
95. Щерба ЛВ. Языковая система и речевая деятельность [Internet]. Ленинград: Наука; 1974. 427 p. Available from: http://elib.gnpbu.ru/textpage/download/html/?bookhl=&book=scherba_yazykovaya-sistema--deyatelnost_1974
96. Rayner K. Eye movements in reading and information processing: 20 years of research. Psychological Bulletin [Internet]. 1998;124(3):372–422. Available from:

http://prx.library.gatech.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=1998-11174-004&site=ehost-live

97. Zeki S. The visual image in mind and brain. Scientific American. 1992;267:68–76.
98. Hawkins J, Blakeslee S. On Intelligence. Book [Internet]. 2005;1–174. Available from: http://www.amazon.com/On-Intelligence-Jeff-Hawkins/dp/0805078533/ref=sr_1_1?ie=UTF8&qid=1363894040&sr=8-1&keywords=on+intelligence%5Cnhttp://books.google.com/books?hl=en&lr=&id=Qg2dmntfxmQC&oi=fnd&pg=PA1&dq=On+Intelligence&ots=6jytL5MrhZ&sig=s6B7XKfYQI-B3
99. Haspelmath M. Word classes and parts of speech. In: International Encyclopedia of the Social & Behavioral Sciences [Internet]. 2001. p. 16538–45. Available from: http://dx.doi.org/10.1016/B0-08-043076-7/02959-4
100. Гвоздев АН. Формирование у ребенка грамматического строя русского языка. Москва: АПН РСФСР; 1949. 268 p.
101. Сергеєв ДС. Природно-мовна база знань як основа моделювання окремих аспектів мовленнєвої діяльності людини. In: Системи та засоби штучного інтелекту AIIS’2017. Київ: КНУ ім. Т.Шевченка, ф-т комп. наук та кібернетики; 2017. p. 171–8.
102. Сергеєв ДС, Хіміч, А.В. Визначення категорії «знання» та її використання в інформаційних природно-мовних технологіях. Адаптивні системи автоматичного управління. 2016;(29(2)):140–6.
103. Левицький АЕ, Сингаївська АВ, Славова ЛЛ. Вступ до мовознавства. Київ: Центр навчальної літератури; 2006. 104 p.
104. Сергеєв ДС. Integration model for knowledge representation for semantic WEB [Інтеграційна модель представлення знань для семантичного WEB]. In: ICSFTI2018. Київ: НТУУ «КПІ ім.Ігоря Сікорського», ф-т інформатики та обчислювальної техніки; 2018. p. 284–7.
105. Сергеєв ДС. Виділення концептів у природно-мовному тексті як спосіб наповнення бази знань. In: SAIT-2017. Київ: НТУУ «КПІ», ІПСА; 2017. p. 321–

3.

106. Сергеєв ДС. A model of relation object for the natural language knowledge base [Модель об’єкту відношення для природно-мовної бази знань]. Адаптивні системи автоматичного управління. 2017;(30(1)):106–13.
107. Angles R, Gutierrez C. Survey of graph database models. ACM Computing Surveys. 2008;40(1):1–39.
108. Сергеєв ДС. Оптимізація використання природно-мовних баз знань шляхом тематичної декомпозиції. In: ЕлІТ-2016. Львів: Львів. нац. ун-т ім. І. Франка, ф-т електроніки; 2016. p. 25–8.
109. Veling A, Van Der Weerd P. Conceptual grouping in word co-occurrence networks.
110. Tadeusiewicz R. Neural networks: A comprehensive foundation. Vol. 3, Control Engineering Practice. 1995. p. 746–7.
111. Sowa JF. Semantic Networks. Encyclopedia of Artificial Intelligence [Internet]. 1992;24(3):291–9. Available from: http://www.jfsowa.com/pubs/semnet.htm
112. Kaufmann E, Bernstein A. Evaluating the usability of natural language query languages and interfaces to Semantic Web knowledge bases. Web Semantics: Science, Services and Agents on the World Wide Web. 2010;8(4):377–93.
113. Barrat A, Weigt M. On the properties of small-world network models. Eur Phys J B. 2000;94:85–94.
114. Анісімов АВ, Марченко ОО, Никоненко АО. Алгоритмічна модель асоціативно-семантичного контекстного аналізу текстів природною мовою. In: Матеріали шостої міжнародної науково-практичної конференції з програмування УкрПРОГ’2008. 2008. p. 379–84.
115. Clark D. Natural language, relevancy ranking, and common sense. IEEE Intelligent Systems [Internet]. 1999;14(4):17–9. Available from: http://search.proquest.com/docview/57542911?accountid=142596
116. Cooper WS. A definition of relevance for information retrieval. Information Storage and Retrieval. 1971;7(1):19–37.
117. Quoc V, Schuster M. A Neural Network for Machine Translation, at Production

Scale. Google Al Blog, Research Scientists, Google Brain Team. 2016.

118. Universit S. Service Oriented Architecture and Web Services. Journal of Computing Sciences in Colleges. 2007;(March):1–28.
119. Taycher L. Inside Google Books: Books of the world, stand up and be counted! All 129,864,880 of you. [Internet]. 2010 [cited 2017 Apr 11]. Available from: http://booksearch.blogspot.com/2010/08/books-of-world-stand-up-and-be-counted.html
120. Habash G. Average Book Length: Guess How Many Words Are In A Novel [Internet]. 2012 [cited 2017 Apr 11]. Available from: http://www.huffingtonpost.com/2012/03/09/book-length_n_1334636.html
121. How many words are there in the Engli... | Oxford Dictionaries [Internet]. [cited 2017 Apr 11]. Available from: https://en.oxforddictionaries.com/explore/how-many-words-are-there-in-the-english-language
122. Влад В. Українська мова: Частотний словник української мови [Internet]. 2013 [cited 2017 Apr 11]. Available from: http://u-mova.blogspot.com/2013/09/blog-post.html
123. Graefe G, Kuno H. Modern B-tree techniques. In: Proceedings - International Conference on Data Engineering. 2011. p. 1370–3.
124. Parikh R. Distribution of Word Lengths in Various Languages [Internet]. [cited 2017 Jun 25]. Available from: http://www.ravi.io/language-word-lengths
125. Kalantari R, Bryant CH. Comparing the performance of object and object relational database systems on objects of varying complexity. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2012. p. 72–83.
126. Peñas A. Apache Lucene. Language Resources and Evaluation [Internet]. 2009;1–22. Available from: http://lucene.apache.org/
127. Rjaibi W, Lohman GM, Haas PJ. Estimation of column cardinality in a partitioned relational database [Internet]. Google Patents; 2004. Available from: https://www.google.com/patents/US6732110

128. Korobov M. Morphological Analyzer and Generator for Russian and Ukrainian Languages. In 2015 [cited 2016 Nov 1]. p. 320–32. Available from: http://link.springer.com/10.1007/978-3-319-26123-2_31
129. Зализняк АА. Грамматический словарь русского языка. Словоизменение. Москва: Русский язык; 1980. 880 p.
130. Бочаров ВВ, Грановский ДВ. Программное обеспечение для коллективной работы над морфологической разметкой корпуса. In: Труды международной конференции «Корпусная лингвистика – 2011» 27–29 июня 2011 г. Санкт-Петербург: С.-Петербургский гос. университет, Филологический факультет; 2011. p. 348.
131. Сергеєв ДС. Комп’ютерне моделювання когнітивного аспекту обробки природної мови на основі природно-мовної бази знань. Штучний інтелект. 2016;(4):42–8.
132. Сергеєв ДС. Методика оцінки якості роботи природно-мовних пошукових систем. In: SAIT-2016 [Internet]. Київ: НТУУ «КПІ», ІПСА; 2016. p. 413–6. Available from: http://sait.kpi.ua/media/filer_public/73/32/7332a68e-e93b-4c57-a3c8-66f11ee074cd/sait2016ebook.pdf

ДОДАТОК А – СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА ЗА ТЕМОЮ ДИСЕРТАЦІЇ

У виданнях іноземних держав:

1. Kyslenko Y, Sergeiev D. Cognitive architecture of speech activity and modelling thereof. Biologically Inspired Cognitive Architectures. 2015;12.
(автору належить: модель бази знань, функціональна модель інтерфейсу).

У наукових фахових виданнях України:

2. Кисленко ЮІ, Сергеєв ДС. Порівняння способів збереження слів в ІТ. Адаптивні системи автоматичного управління. 2016;(28(1)):33–41.
(автору належить: формалізація проблеми, формалізація використаного підходу, збір та опрацювання даних)
3. Кисленко ЮІ, Сергеєв ДС. Структурний підхід до пошуку природно-мовної інформації. Радіоелектроніка та інформатика. 2015;(3):45–9.
(автору належить: методика аналізу та опрацювання даних)
4. Сергеєв ДС. A model of relation object for the natural language knowledge base [Модель об’єкту відношення для природно-мовної бази знань]. Адаптивні системи автоматичного управління. 2017;(30(1)):106–13.
5. Сергеєв ДС, Хіміч, А.В. Визначення категорії «знання» та її використання в інформаційних природно-мовних технологіях. Адаптивні системи автоматичного управління. 2016;(29(2)):140–6.
(автору належить: формалізація та обгрунтування використаної моделі)
6. Сергеєв ДС. Комп’ютерне моделювання когнітивного аспекту обробки природної мови на основі природно-мовної бази знань. Штучний інтелект. 2016;(4):42–8.

Матеріали конференцій:

7. Сергеєв ДС. Особливості моделювання бази природно-мовних знань. В: Електроніка та інформаційні технології ЕлІТ-2015. Львів: Львів. нац. ун-т ім. І. Франка, ф-т електроніки; 2015. с. 17–20.
8. Сергеєв ДС. Методика оцінки якості роботи природно-мовних пошукових систем. In: SAIT-2016 [Internet]. Київ: НТУУ «КПІ», ІПСА; 2016. p. 413–6. Режим доступу: http://sait.kpi.ua/media/filer_public/73/32/7332a68e-e93b-4c57-a3c8-66f11ee074cd/sait2016ebook.pdf
9. Сергеєв ДС. Оптимізація використання природно-мовних баз знань шляхом

тематичної декомпозиції. В: ЕлІТ-2016. Львів: Львів. нац. ун-т ім. І. Франка, ф-т електроніки; 2016. с. 25–8.
10. Сергеєв ДС. Виділення концептів у природно-мовному тексті як спосіб наповнення бази знань. В: SAIT-2017. Київ: НТУУ «КПІ», ІПСА; 2017. с. 321–3.
11. Сергеєв ДС. Природно-мовна база знань як основа моделювання окремих аспектів мовленнєвої діяльності людини. В: Системи та засоби штучного інтелекту AIIS'2017. Київ: КНУ ім. Т.Шевченка, ф-т комп. наук та кібернетики; 2017. с. 171–8.
12. Сергеєв ДС. Integration model for knowledge representation for semantic WEB [Інтеграційна модель представлення знань для семантичного WEB]. В: ICSFTI2018. Київ: КПІ ім.Ігоря Сікорського, ф-т інформатики та обчислювальної техніки; 2018. с. 284–7.

ДОДАТОК Б – АКТИ ПРО ВПРОВАДЖЕННЯ РЕЗУЛЬТАТІВ НАУКОВИХ ДОСЛІДЖЕНЬ

Товариство з обмеженою відповідальністю
“Діджитал Принт”

3680, м. Київ, бульвар Івана Лепсе 8, корпус 58
ЄДРПОУ 38746620, ІНП 387466226580,р/р 26007001364751
МФО 300528 в АТ "ОТП Банк"
тел. (044) 393-43-14, www. ppeurope.com, e-mail: info@ppeurope.com

«03» серпня 2016 року

Акт
про впровадження результатів дисертаційної роботи
Сергеєва Данила Сергійовича

Акт виданий аспіранту Національного технічного університету України «Київський політехнічний інститут» Сергеєву Данилу Сергійовичу в тому, що результати його дисертаційної роботи на здобуття наукового ступеня кандидата технічних наук впроваджені та використовуються в ТОВ "Діджитал Принт".

Склад впровадження:

- База знань, яка забезпечує класифікацію дизайнів нанесення у каталозі веб-порталу компанії;
- Модифікація системи пошуку, що дозволяє використовувати базу знань для пошуку дизайнів.

Цей акт не є основою для фінансових розрахунків.

директор _____ *Бєлєнький О.А.*

ТОВ «МІЖНАРОДНА ТЕКСТИЛЬНА КОРПОРАЦІЯ»

Товариство з обмеженою відповідальністю
"МІЖНАРОДНА ТЕКСТИЛЬНА
КОРПОРАЦІЯ",
03680, м. Київ. Бул Івана Лепсе, 8
Корпус 58
Код ЄДРПОУ 40268890

Р/р 26008455018867
в АТ "ОТП Банк", м.Київ,
МФО 300528
тел. 044-561-65-25

«17» серпня 2016 року

Акт
про впровадження результатів дисертаційної роботи
Сергеєва Данила Сергійовича

Акт виданий аспіранту Національного технічного університету України «Київський політехнічний інститут» Сергеєву Данилу Сергійовичу в тому, що результати його дисертаційної роботи на здобуття наукового ступеня кандидата технічних наук впроваджені та використовуються в ТОВ "Міжнародна текстильна корпорація".

Склад впровадження:

- База знань, що зберігає інформацію про продукцію компанії, як модуль системи бухгалтерського та складського обліку;
- Інтерфейс користувача для бази знань, що забезпечує можливість використання бази знань оператором.

Цей акт не є основою для фінансових розрахунків.

директор ТОВ «МІЖНАРОДНА ТЕКСТИЛЬНА КОРПОРАЦІЯ»

Сілантьєв Олександр Павлович

ЗАТВЕРДЖЕНО

Декан ФІОТ
НТУУ «КПІ імені Ігоря Сікорського»

д.т.н. проф. С.Ф. Теленик

«01» вересня 2018 р.

АКТ

про впровадження в навчальний процес результатів дисертаційної роботи Сергеєва Данила Сергійовича на тему «Інформаційна технологія обробки природно-мовних текстів на основі інтеграційного підходу»

Цим актом підтверджується впровадження результатів дисертаційних досліджень Сергеєва Данила Сергійовича у навчальному процесі кафедри технічної кібернетики НТУУ «КПІ імені Ігоря Сікорського», а саме використання розроблених методів та алгоритмів у викладанні курсів «Теорія алгоритмів» та «Візуальне програмування».

Цей акт не є підставою для фінансових розрахунків.

Завідувач кафедрою
технічної кібернетики

д.т.н., доц. Пархомей І.Р.

ДОДАТОК В – СХЕМА ERM ДЛЯ ІТ ОПМ

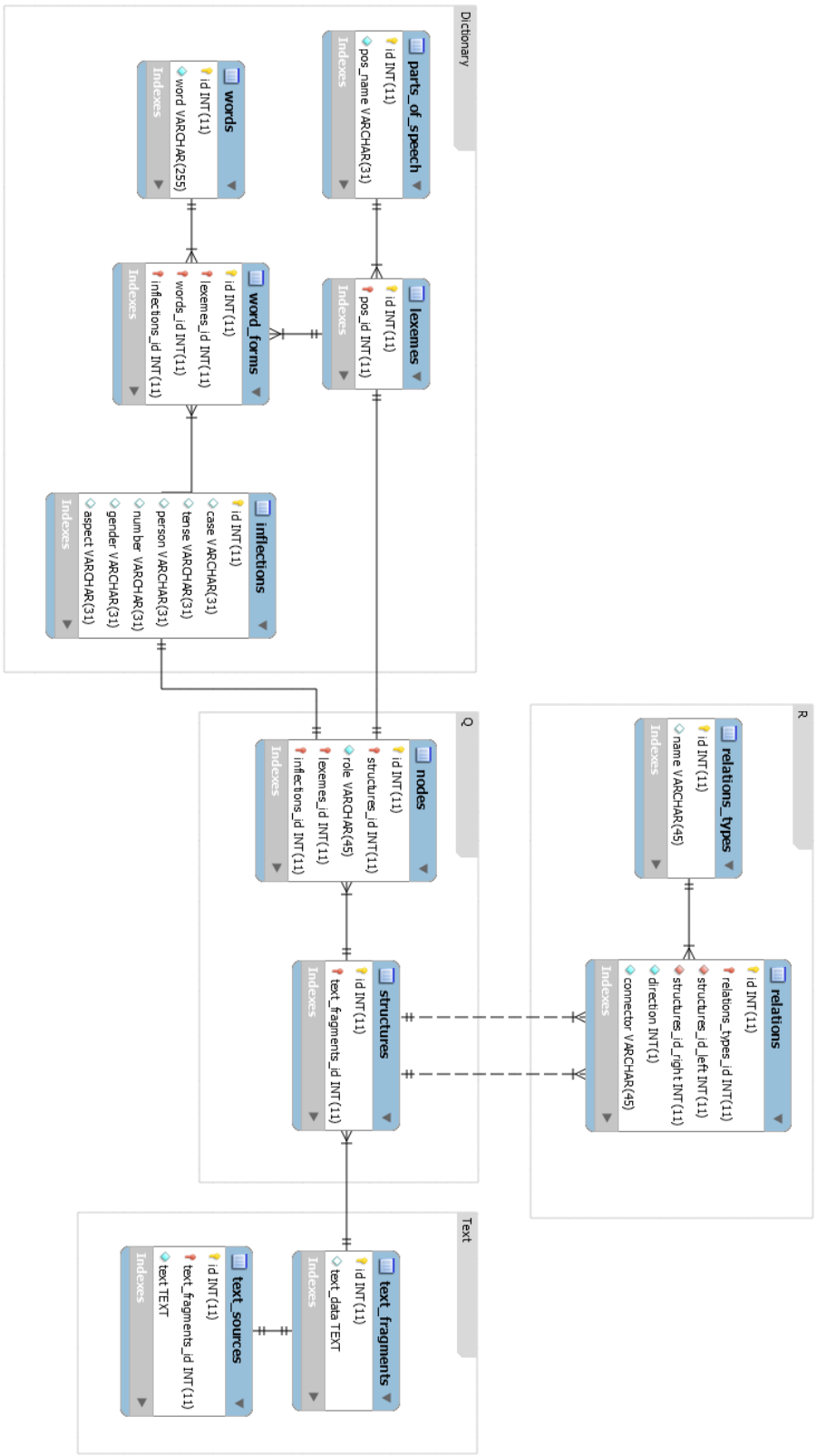


Рисунок Г.1 – Схема ERM для природномовної бази знань на основі інтеграційного підходу

ДОДАТОК Г – ПРОЦЕС ФОРМУВАННЯ НАПОВНЕННЯ ПМБЗ ІП

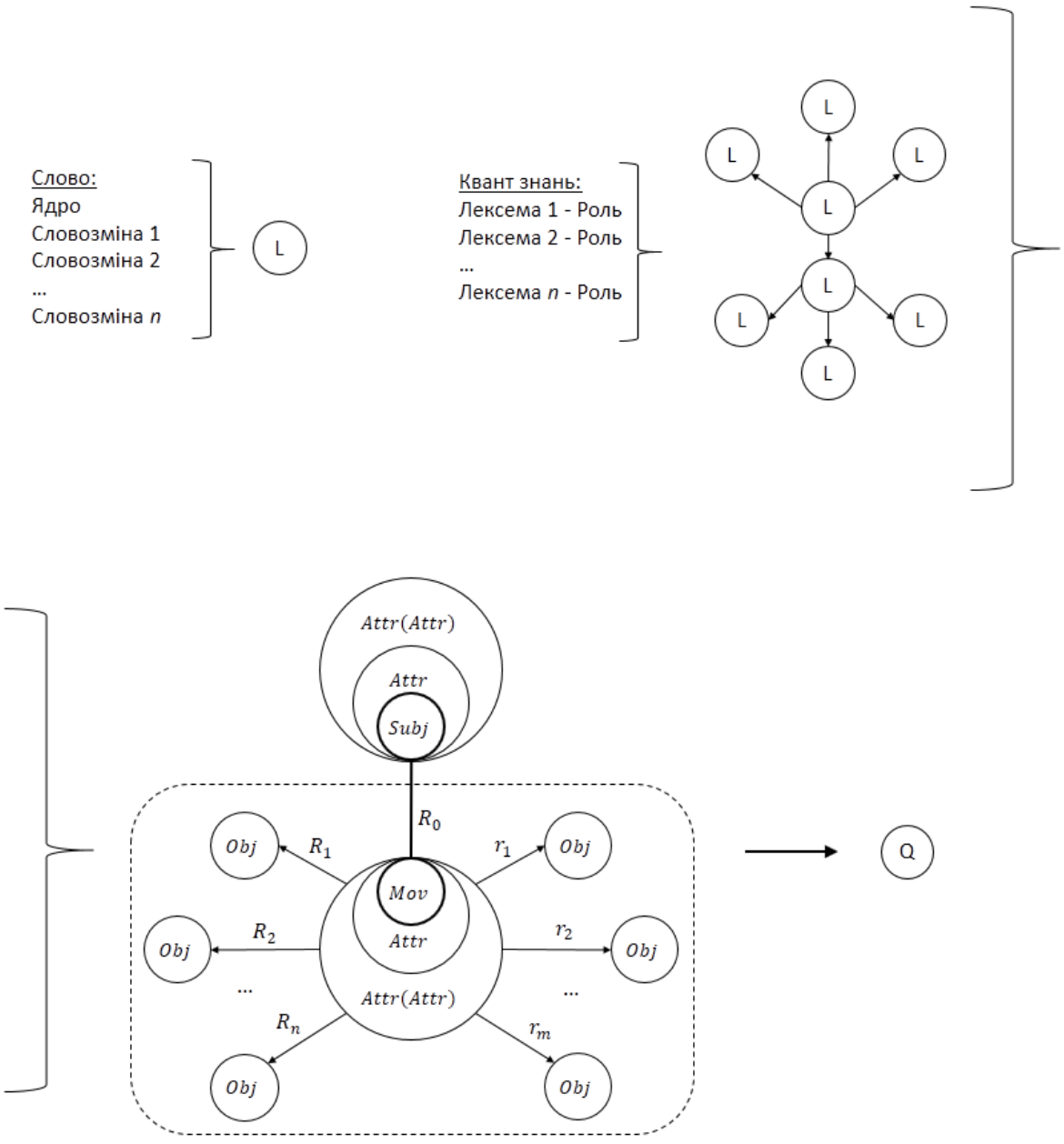


Рисунок Д.1 – Процес формування наповнення ПМБЗ ІП від лексеми до кванту знань

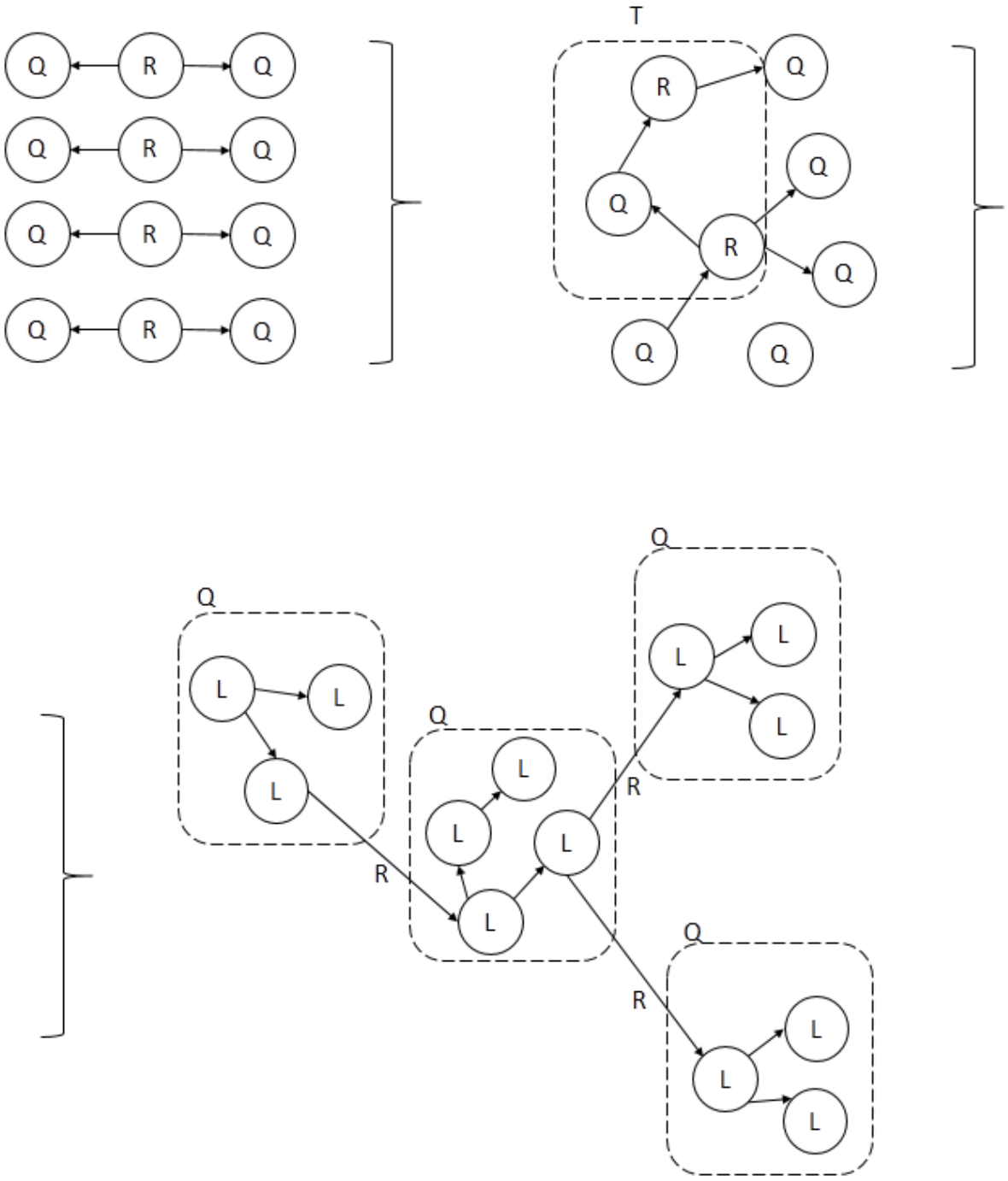


Рисунок Д.2 – Процес формування наповнення ПМБЗ ІП від квантів знань та відношень до мережі знань

ДОДАТОК Д – ПРИКЛАД ВИКОНАННЯ ПРИРОДНОМОВНОГО ПОШУКУ В ІТ ОПМ

Наповнення:
«природні та штучні мови можуть приймати різні форми»
«літературна мова є формою природної мови»
«зміст та форма є важливими характеристиками художнього твору»

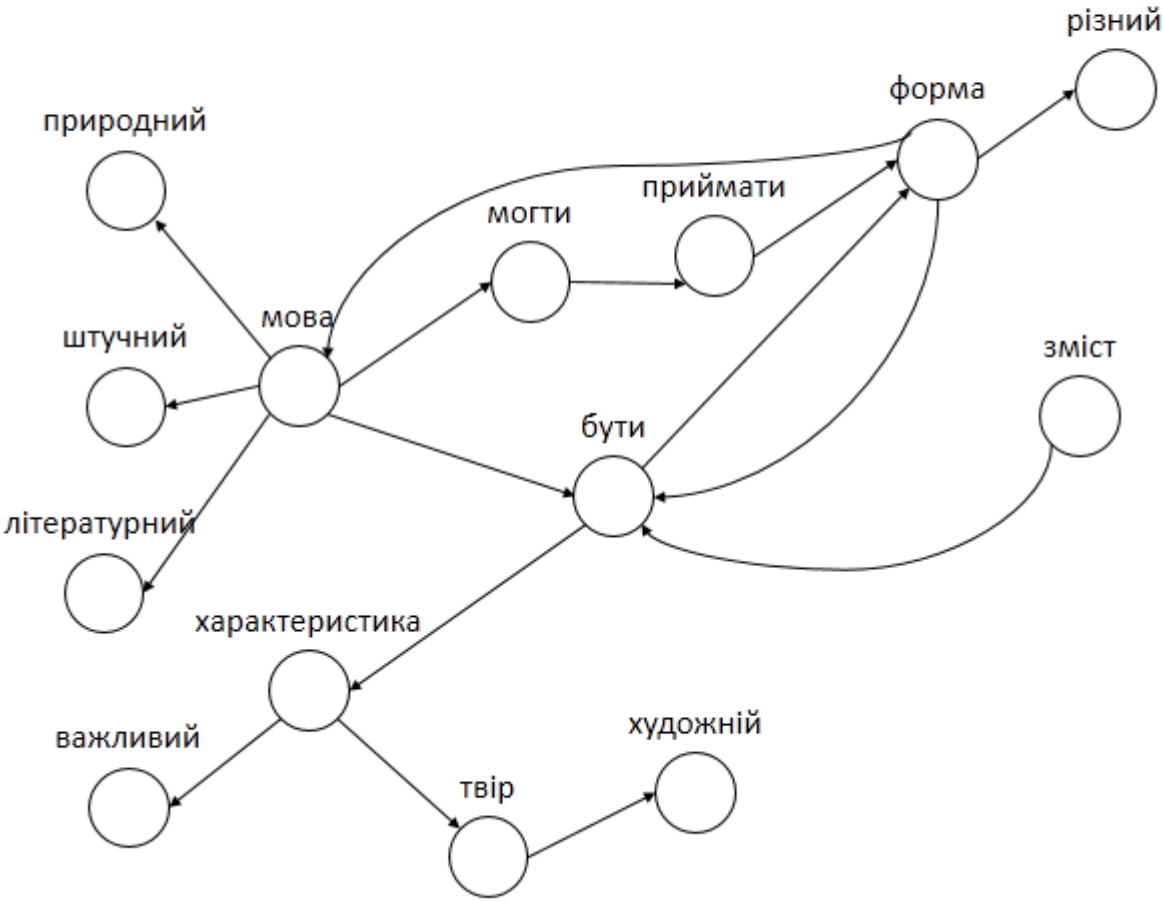


Рисунок Е.1 – Наповнення ПМБЗ ІП вхідними даними для виконання природномовного пошуку

Запит: «мова»
Результати:
«природні та штучні **мови** можуть приймати різні форми»
«літературна **мова** є формою природної мови»

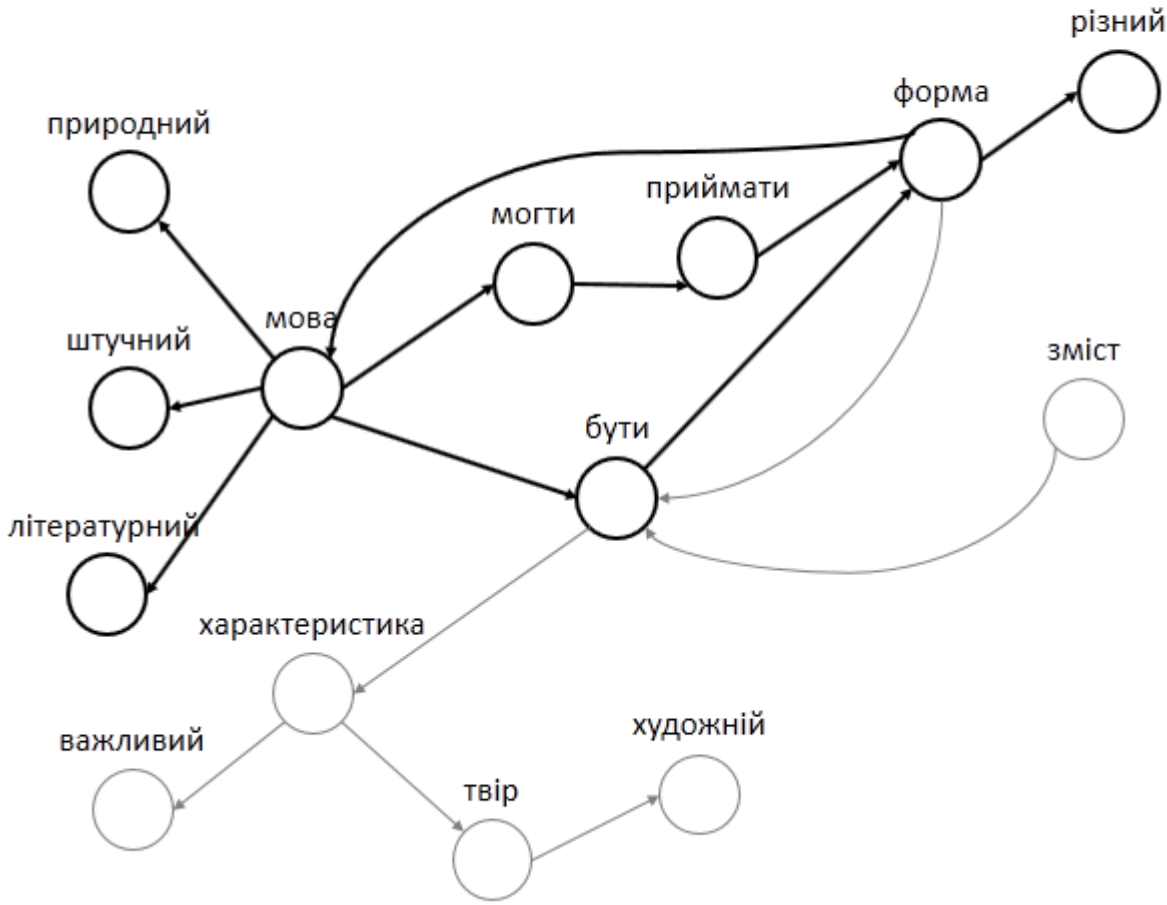


Рисунок Е.2 – Пошук у ПМБЗ ІП за запитом «мова»

Запит: «природна мова»
Результати:
«**природні** та штучні **мови** можуть приймати різні форми»

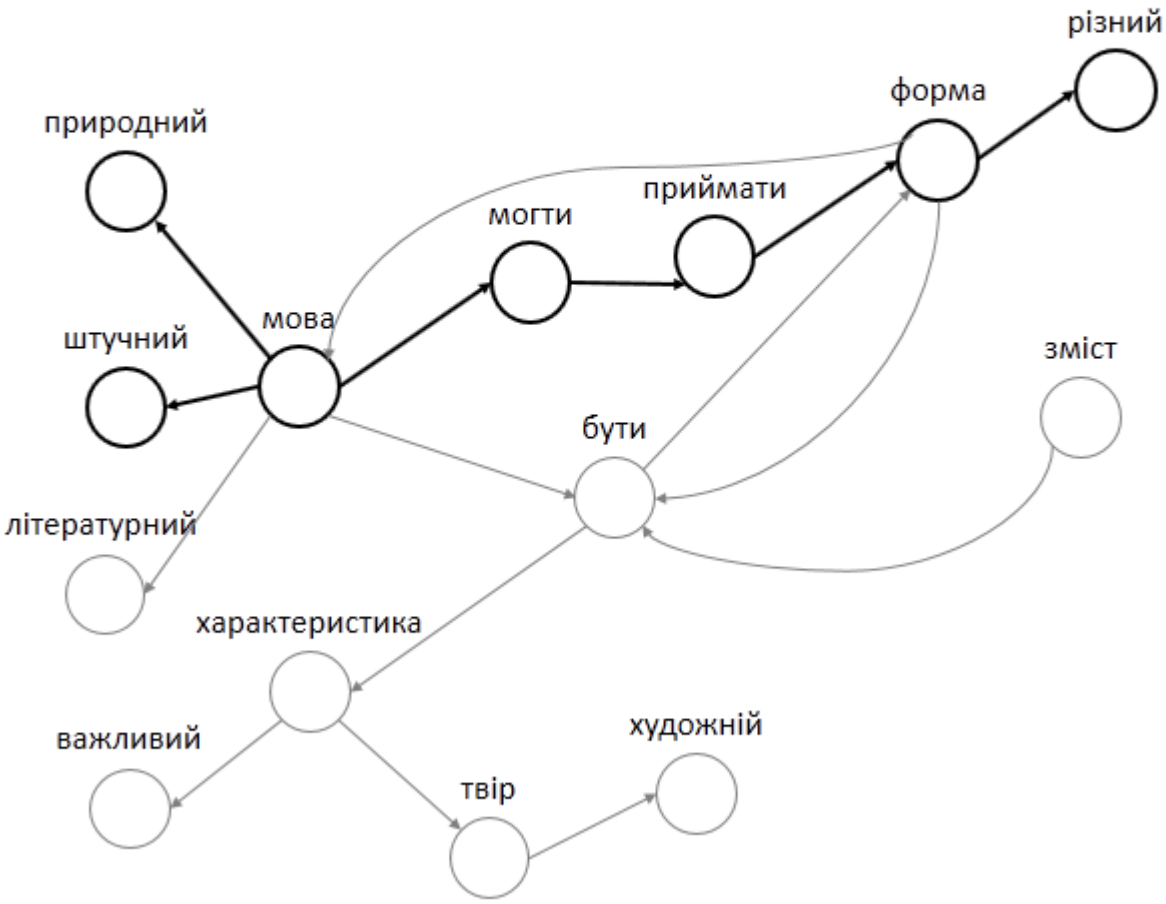


Рисунок Е.3 – Пошук у ПМБЗ ІП за запитом «природна мова»

ДОДАТОК Е – РЕЗУЛЬТАТИ ПРИРОДНОМОВНОГО ПОШУКУ З ВИКОРИСТАННЯМ ІТ ОПМ

№	Запит	Усього	Знайдено	Помилкових	Релевантних
1	база знань	100	92	0	0.92
2	зберігати знання	100	11	2	0.09
3	українська мова	100	98	0	0.98
4	природномовний текст	61	50	0	0.82
5	синтаксичний аналіз	100	75	1	0.74
6	семантичний аналіз	100	46	1	0.45
7	пошуковий запит	100	86	0	0.86
8	аналізувати дані	100	17	2	0.15
9	обробляти дані	100	7	1	0.06
10	база даних	100	94	0	0.94
11	технічні вимоги	100	95	1	0.94
12	експериментальна перевірка	100	95	0	0.95
13	інформаційна технологія	100	85	0	0.85
14	обчислювальна складність	100	48	0	0.48
15	розробка моделі	100	100	0	1.00
16	машинний переклад	100	97	0	0.97
17	ідентифікація плагіату	100	13	4	0.09
18	мене звати	100	98	0	0.98
19	добрий день	100	97	0	0.97
20	права рука	100	98	1	0.97
21	минулий тиждень	100	98	0	0.98
22	виникає проблема	65	35	4	0.48
23	важлива зустріч	100	75	1	0.74
24	старий світ	100	68	2	0.66
25	літак вилітає	100	74	2	0.72
26	пожежна машина	100	86	0	0.86
27	робочий день	100	94	1	0.93
28	накласти санкції	100	72	1	0.71
29	оголосити підозру	100	22	1	0.21
30	військові навчання	70	5	0	0.07
31	сімейний конфлікт	100	99	0	0.99
32	газова турбіна	100	97	1	0.96

33	подвійне громадянство	100	99	0	0.99
34	населений пункт	100	97	0	0.97
35	курс падає	100	30	1	0.29
36	курс зростає	100	32	0	0.32
37	зміцнити позиції	100	93	0	0.93
38	європейська комісія	100	99	0	0.99
39	глибоко стурбований	100	91	0	0.91
40	здобути перемогу	100	100	0	1.00
41	демократичні вибори	100	71	2	0.69
42	попередні результати	100	98	0	0.98
43	міжнародний спостерігач	100	67	0	0.67
44	реальний час	100	29	0	0.29
45	причина катастрофи	100	48	2	0.46
46	соціальні медіа	100	99	0	0.99
47	споживати контент	100	66	3	0.63
48	користувач сервісу	100	29	0	0.29
49	віртуальна реальність	100	100	0	1.00
50	електронна пошта	100	86	0	0.86
51	електронний маркетинг	100	35	0	0.35
52	право(-а) власності	100	100	0	1.00
53	рекламні доходи	100	27	1	0.26
54	програмне забезпечення	100	100	0	1.00
55	основний принцип	100	96	0	0.96
56	абстрактні характеристики	100	3	0	0.03
57	предметна область	100	67	0	0.67
58	повторне використання	100	98	1	0.97
59	базовий клас	100	37	2	0.35
60	мати доступ	100	69	1	0.68
61	статичний метод	100	64	0	0.64
62	динамічний метод	100	23	0	0.23
63	створювати екземпляр	100	50	1	0.49
64	встановити значення	100	49	6	0.43
65	надати доступ	100	91	1	0.90
66	обмежити доступ	100	97	0	0.97
67	дикий кабан	100	100	0	1.00
68	очеретяні хащі	100	87	1	0.86
69	осипаються квіти	100	32	0	0.32

70	чайна церемонія	100	96	0	0.96
71	морський птах	100	26	1	0.25
72	море шумить	100	32	1	0.31
73	варити борщ	100	72	1	0.71
74	зоряне майбутнє	100	21	0	0.21
75	великий кущ	100	42	0	0.42
76	сиве волосся	100	97	0	0.97
77	чисте небо	100	88	0	0.88
78	гостре слово	100	42	1	0.41
79	подвійне життя	100	97	0	0.97
80	славний воїн	100	51	0	0.51
81	темна ніч	100	99	0	0.99
82	сталась подія	100	67	1	0.66
83	сонце гріє	100	100	1	0.99
84	нові чоботи	100	24	0	0.24
85	новий світ	100	99	0	0.99
86	мокрі штані	100	65	0	0.65
87	велика калюжа	100	85	0	0.85
88	тихий плескіт	100	92	0	0.92
89	український народ	100	99	0	0.99
90	права людини	100	100	0	1.00
91	прагнути розвивати	100	65	3	0.62
92	правова держава	100	100	0	1.00
93	діяльність держави	100	78	1	0.77
94	міжнародний договір	100	97	0	0.97
95	пряма дія	100	100	0	1.00
96	державна мова	100	100	1	0.99
97	вільне використання	100	92	1	0.91
98	конституційні права	100	97	0	0.97
99	нагальна необхідність	100	72	0	0.72
100	особисте життя	100	98	0	0.98
101	судовий захист	100	96	0	0.96
102	законні підстави	100	89	0	0.89
103	професійна спілка	100	80	0	0.80
104	рівні права	100	95	0	0.95