МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «ЛЬВІВСЬКА ПОЛІТЕХНІКА»

Кваліфікаційна наукова праця
на правах рукопису

ВИСОЦЬКА ВІКТОРІЯ АНАТОЛІЇВНА

УДК 004.82:004.89:004.91

ДИСЕРТАЦІЯ

АНАЛІЗ ТА СИНТЕЗ КОМП'ЮТЕРНИХ ЛІНГВІСТИЧНИХ СИСТЕМ ОПРАЦЮВАННЯ УКРАЇНОМОВНОГО ТЕКСТОВОГО КОНТЕНТУ

10.02.21 – структурна, прикладна і математична лінгвістика

Подається на здобуття наукового ступеня доктора технічних наук

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

_____ В.А. Висоцька

Науковий консультант –
Литвин Василь Володимирович,
д.т.н., професор

Ідентичність всіх примірників дисертації
ЗАСВІДЧУЮ:
Вчений секретар спеціалізованої вченої ради

Бунь Ростислав Адамович/

Львів – 2023

Анотація

Висоцька В.А. Аналіз та синтез комп'ютерних лінгвістичних систем опрацювання україномовного текстового контенту. – На правах рукопису.

Дисертаційна робота на здобуття наукового ступеня доктора технічних наук за спеціальністю 10.02.21 – структурна, прикладна і математична лінгвістика. – Національний університет «Львівська політехніка», Міністерство освіти і науки України, Львів, 2023.

У дисертації вирішено важливу науково-прикладну проблему аналізу та синтезу комп'ютерних лінгвістичних систем (КЛС) для розв'язання різних задач опрацювання україномовного текстового контенту на основі розроблення нових та удосконаленні відомих моделей, методів та засобів опрацювання природної мови (Natural-Language Processing, NLP). Проведено аналіз сучасного стану та перспективи розвитку ІТ опрацювання природної мови текстового контенту. Визначено поняття КЛС та наведена загальна їх класифікація. Проведений детальний аналіз відомих КЛС, що дало можливість вдосконалити загальну класифікацію відповідних інформаційних систем (ІС). Визначені основні NLP-задачі КЛС, на основі яких наведені приклади та порівняльний аналіз відомих сучасних КЛС. Це дало можливість сформувати загальні напрями дослідження. Описана та проаналізована основна загальна схема процесу лінгвістичного аналізу тексту природньою мовою засобами КЛС. Визначені основні стани та властивості КЛС, їх класифікація та особливості. Проаналізовано відомі класичні підходи та напрями опрацювання природної мови. Наведена загальна класифікація основних NLP-підходів, напрямів та додаткових методів лінгвістичного дослідження для NLP-задач. Проведено аналіз існуючих основних методів та методики опрацювання природної мови засобами машинного навчання (Machine Learning, ML). Вдосконалена їх класифікація та визначені типові проблеми ML-методів для опрацювання україномовних текстів. Зроблений огляд відомих ІТ розроблення КЛС на основі особливостей та технологій інтелектуального аналізу потоку україномовного контенту. Визначені основні вимоги до оцінювання ефективності КЛС на основі ML-

технології та аналізу великих даних (Big Data Analysis, BDA). Визначені основні ML-методи для BDA з множини текстових потоків контенту. Визначені вимоги до кластеризації текстового контенту при неконтрольованому ML. Проаналізовано особливості проектування та розроблення КЛС на основі визначення основних етапів як графемний, морфологічний, лексичний, синтаксичний, семантичний аналіз/синтез україномовного тексту для розв'язку конкретної NLP-задачі. Зроблена та конкретизована постановка проблеми опрацювання україномовного тексту на основі визначення функціональних особливостей інтелектуального аналізу текстового потоку. Загальний аналіз проблеми аналізу україномовного тексту та визначення основних проблем опрацювання україномовного тексту дало можливість сформулювати основні етапи та вимоги до проекту типової КЛС для розв'язку конкретної NLP-задачі. Ідентифікація основних характеристик КЛС та обґрунтування реалізації проекту типової КЛС дало можливість визначити очікувані ефекти від відповідної реалізації проекту. На основі аналізу вхідних/вихідних потоків контенту КЛС визначені та сформульовані функціональні вимоги до проекту типової КЛС, її програмних модулів, мережних, програмних та технічних інструментів програмної реалізації ІС.

Аналіз та синтез КЛС базується на застосуванні лінгвістичного аналізу україномовного текстового контенту, інтелектуальному опрацювання текстового потоку контенту, машинному навчанні системи на достовірних даних та статистичному аналізі для знаходження закономірностей появи лінгвістичних подій. Розроблена інформаційна технологія (ІТ) опрацювання україномовного текстового контенту на відміну від існуючих підтримує принцип модульності типової архітектури КЛС для розв'язку конкретної задачі NLP та аналізу множини параметрів та метрик ефективності функціонування системи відповідно до поведінки цільової аудиторії. Розроблено загальну структуру КЛС для опрацювання текстового контенту українською мовою та концептуальну схему/модель функціонування типової КЛС на основі моделювання взаємодії основних процесів і компонентів системи, що дало змогу вдосконалити ІТ

інтелектуального аналізу текстового потоку на основі опрацювання інформаційних ресурсів. Наведено приклади розроблених КЛС опрацювання україномовного текстового контенту для розв'язку відповідних задач NLP, функціонування яких ґрунтується на розроблених та вдосконалених моделях, методах та алгоритмах.

Удосконалена модель лінгвістичного опрацювання текстового контенту на основі графемного, морфологічного, лексичного, синтаксичного, семантичного, структурного, онтологічного та прагматичного аналізу для вирішення конкретної проблеми NLP. Це дало змогу сформулювати загальні вимоги до процесів опрацювання україномовного контенту. Удосконалення методів опрацювання інформаційних ресурсів, таких як інтеграція, управління та супровід україномовного контенту, дозволило адаптувати процес інтелектуального аналізу текстового потоку до розв'язку різних задач NLP та розробити КЛС, що ефективно функціонують, метрики для розв'язку різних задач NLP. Удосконалені методи NLP на основі регулярних виразів узгодження за шаблоном дозволили адаптувати алгоритми графемного та морфологічного аналізу для опрацювання україномовних текстів. Визначені основні допустимі операції регулярних виразів (Regular Expression, RE) як об'єднання та диз'юнкція символів/ланцюжків/виразів, оператори лічби та прецедентності, а також анкори (Anchor – якір, прив'язка) як спецсимволи ідентифікації присутності/відсутності символів в RE. Удосконалено метод токенізації та нормалізації тексту каскадами простих підстановок регулярних виразів і кінцевих автоматів, що дало змогу адаптувати алгоритм лексичного та синтаксичного аналізів для опрацювання україномовних текстів. Удосконалено метод морфологічного аналізу, заснований на сегментації та нормалізації слів, сегментації речень і модифікованому алгоритмі стемінгу Портера як ефективного засобу ідентифікації афіксів лем для можливості розмічування аналізованого слова, що дозволило підвищити точність пошуку ключових слів на 9%. На відмінну від класичного алгоритму Портера (не має високої точності навіть для англомовних текстів) модифікований є адаптованим для української мови та дає точний

результат в межах 85-93% випадків в залежності від якості, стилю, жанру тексту та відповідно наповнення словників КЛС. Описано алгоритм мінімальної редакційної відстані рядків українських текстів як мінімальна кількість операцій, необхідних для перетворення одного в інший. Вдосконалено метод синтаксичного аналізу україномовного текстового контенту, спрямованого на автоматичне виявлення значущих ключових слів вхідних текстів. Визначено роль і формальні ознаки синтаксичного аналізатора в процесі виявлення ключових слів тематики контенту, проведено декомпозицію процедур запропонованого методу. На відміну від відомих синтаксичних аналізаторів, запропонований метод забезпечує самовдосконалення та самонавчання КЛС-модуля визначення ключових слів за рахунок механізму ідентифікації значущих статистичних параметрів у визначених модератором межах.

Розроблено метод ідентифікації ключових слів в україномовних текстах на основі графемного та морфологічного аналізу основ слів через регулярні вирази та N-грами, що дало змогу підвищити точність пошуку ключових слів на 6-9%, здійснити пошук стійких словосполучень та рубрикацію контенту. Розроблено метод визначення стійких словосполучень на основі ідентифікації ключових слів україномовного тексту та аналізу коефіцієнтів лексичного мовлення автора тексту в еталонних уривках контенту, що дало можливість на основі статистичної лінгвістики покращити точність методу визначення стилю автора тексту на 6-7%. Метод полягає у використанні закону Зіпфа при формуванні стійких словосполучень як ключових з врахуванням наступних правил попереднього лінгвістичного опрацювання тексту: вилучення всіх стопових слів; біграми формувати лише в межах знаків пунктуації; дієслово та займенник вважати знаками пунктуації; дієслова визначати за флексіями; біграми формувати на основі основ без врахування флексій; визначення прикметників за флексіями та вважати, що прикметники є лише зліва у біграмі з україномовних текстів.

Розроблено метод визначення стилю автора тематичного україномовного текстового контенту на основі аналізу ключових слів, стійких словосполучень,

N-грам, лінгвометрії та стилеметрії, що дало змогу визначити стилістичний вклад кожного з авторів та підвищити точність атрибуції науково-технічної публікації на 6-12%. Метод полягає в порівняльному аналізі авторської атрибуції в статистично опрацьованому доробку автора (еталоні) з довільним аналізованим уривком. Метод оцінює ймовірність приналежності публікації до шаблону стилю автора із аналізом відповідних коефіцієнтів лексичного авторського мовлення. Метод працює при умові, що авторський шаблон вже досліджений. Для атрибуції використано аналіз опорних слів, отримані результати подано на основі аналізу коефіцієнтів кореляції.

Розроблено метод обчислення ступеня верифікації автора україномовного тексту із множини можливих на основі порівняльного аналізу стилів потенційних авторів, що дало змогу підвищити точність класифікації за подібністю стилю до [9;34]% із загальної кількості учасників проекту. Для персонального стилю письменника показовими є службові слова, оскільки не пов’язані з змістом і темою текстового контенту та конкретна підмножина таких слів характерна конкретному авторові. Для кожного з уривків проаналізовані та порівняні із еталонним значеннями абсолютні та відносні частоти появи стопових слова. Отже, застосування методу опорних слів дає такі результати: знаходження серед досліджуваних уривків того, що найбільш ймовірно належить до еталону. Інші результати підтверджують дієвість методу опорних слів у авторській атрибуції текстів. Доведено незначущість впливу частки як параметра методу на динаміку зміни результатів експерименту при аналізу україномовних авторських текстів.

Розроблено алгоритм лексичного аналізу україномовних текстів та алгоритм синтаксичного аналізатора текстового контенту. Особливостями алгоритмів є адаптація морфологічного та синтаксичного аналізу лексичних одиниць до особливостей конструкцій україномовних слів/текстів на основі аналізу кожного слова з врахуванням його частини мови та відмінювання. Для цього провадився аналіз флексій слів для класифікації, виділення основи для формування відповідних алфавітно-частотних словників. Наповнення словників

в подальшому враховувалися на наступних кроках визначення авторства тексту як розрахунок лінгвістичних ознак авторського мовлення.

Розроблено методи аналізу та синтезу КЛС на основі створення загальної типової структури системи опрацювання текстового контенту українською мовою через підтримку модульності, моделювання взаємодії основних процесів і компонентів, що дало можливість розширити колекцію розв'язків різних типових задач NLP шляхом реалізації типового програмного забезпечення таких систем. Запропоновано підхід до розроблення програмного забезпечення (ПЗ) контент-моніторингу стопових слів для визначення автора в україномовних текстах стилю на основі Web Mining. Особливість підходу полягає у адаптації лінгвостатистичного аналізу лексем до особливостей морфологічних та синтаксичних конструкцій україномовних слів/текстів. Досліджено результати експериментальної апробації запропонованого методу визначення автора в україномовних наукових текстах технічного профілю в понад 300 одноосібних наукових публікацій зі номерів Вісника НУ «Львівська політехніка» серії «Інформаційні системи та мережі» за період 2001–2021 рр. Виявлено, що для обраної експериментальної бази з понад 300 робіт найкращих результатів за критерієм щільності досягає метод аналізу статті без початкової обов'язкової інформації як анотації та ключові слова різними мовами та списку літератури. Достовірність наукових та практичних результатів підтверджується відповідними матеріалами про впровадження дисертаційних досліджень, а також за рахунок порівняння одержаних практичних результатів на різних вибірках достовірних вхідних даних. КЛС реалізовано на інформаційному ресурсі http://victana.lviv.ua засобами CMS Joomla! (для розроблення е-каркасу сайту), PHP (для реалізації методів опрацювання текстового контенту), HTML (для реалізації розмітки сторінок), CSS (для опису стилів сторінок), MySQL (для зберігання даних та словників). Експериментальне дослідження підтвердило достовірність методу визначення ключових слів – для різних алгоритмів опрацювання первинного тексту середній збіг списків виявлених ключовиків з авторськими змінюється у проміжку 52,6-68,5%. Точність збігу ключових слів із

авторськими коливається в проміжку 43,6-62,9%. Середній збіг змістовних ключових слів порівняно зі всіма знайденими системою коливається в проміжку 38,9-75,8% в залежності від етапів аналізу текстів статей. Точність збігу ключових слів порівняно зі всіма знайденими системою коливається в проміжку 34,3-71,9% в залежності від етапів аналізу текстів статей.

Результати дисертаційної роботи, зокрема: модель КЛС опрацювання текстового контенту українською мовою; інформаційну технологію інтелектуального аналізу текстового потоку на основі опрацювання інформаційних ресурсів; модулі опрацювання україномовного текстового контенту систем підтримки прийняття рішень; методи та моделі лінгвістичного аналізу та опрацювання україномовного текстового контенту; модуль визначення ключових слів в україномовних текстах та метод визначення стійких словосполучень при ідентифікації ключових слів україномовного тексту; модуль визначення автора україномовного тексту та визначення стилю автора тексту. Їх упроваджено під час розроблення інтелектуальної системи автоматичного моніторингу стану елементів конструкцій тривалої експлуатації для ТОВ «Гідравлік-Партнер» (м. Львів) при виконанні госпдоговору 265 „Бізнес-аналіз діяльності ТзОВ «Гідравлік Партнер» на основі онтологічного підходу” та модуля системи розпізнавання цифр у голосі на основі нейронних мереж для ТОВ «А-Солюшнз Девелопмент» (м. Львів) при виконанні госпдоговору 677, а також використані під час виконання науково-дослідних робіт за держбюджетною тематикою у Національному університеті «Львівська політехніка», що підтверджено відповідними актами впровадження.

Результати роботи включені у навчальний процес студентів спеціальностей 124 «Системний аналіз» та 126 «Інформаційні технології та системи», а саме в дисципліни «Комп’ютерна лінгвістика», «Розпізнавання мови», «Методи обчислень та візуалізація даних» за рахунок використання навчальних посібників: Математична лінгвістика. Книга 2. Комбінаторна лінгвістика: навчальний посібник / В. В. Пасічник, Ю. М. Щербина, В. А. Висоцька, Т. В. Шестакевич. Львів: Вид-во Львів. політехніки, 2019. 250 с.; Литвин В. В.

Глибинне навчання: навч. посіб. / В. В. Литвин, Р. М. Пелещак, В. А. Висоцька. – Львів: Видавництво Львівської політехніки, 2021. – 264 с.; Чисельні методи в комп'ютерних науках / В. А. Андруник, В. А. Висоцька, В. В. Пасічник, Л. Б. Чирун, Л. В. Чирун. Львів: Новий Світ – 2000, 2017. Т. 1. 470 с.; Чисельні методи в комп'ютерних науках / В. А. Андруник, В. А. Висоцька, В. В. Пасічник, Л. Б. Чирун, Л. В. Чирун. Львів: Новий Світ – 2000, 2017. Т. 2. 536 с.; Ришковець Ю. В., Висоцька В. А. Алгоритмізація та програмування. Ч. 1: навч. посіб. Львів: Новий Світ – 200, 2018. 337 с.; Ришковець Ю. В., Висоцька В. А. Алгоритмізація та програмування. Ч. 2: навч. посіб. Львів: Новий Світ – 2000, 2018. 316 с.; Висоцька В. А., Литвин В. В., Лозинська О. В. Дискретна математика: практикум: навч. посіб. Львів: Новий Світ – 2000, 2019. 575 с.; Висоцька В. А., Оборська О. В. Python: алгоритмізація та програмування: навчальний посібник. Львів: Новий Світ – 2000, 2020. 516 с.

Ключові слова: NLP, комп'ютерна лінгвістика, текстовий контент, українська мова, графемний аналіз, морфологічний аналіз, лексичний аналіз, синтаксичний аналіз, семантичний аналіз, структурний аналіз, прагматичний аналіз, інформаційна технологія, машинне навчання, опрацювання природної мови, інформаційна система, онтологія, ключові слова, стійкі словосполучення, стиль автора, ідентифікація автора, психологічний аналіз тексту.

ABSTRACT

Vysotska V. Analysis and synthesis of computational linguistic systems for processing Ukrainian textual content. – Manuscript.

Thesis for a Doctoral degree in Technical Science, speciality 10.02.21 – structural, applied and mathematical linguistics. – Lviv Polytechnic National University, Ministry of Education and Science of Ukraine, Lviv, 2023.

The dissertation solves an important scientific and applied problem of analysis and synthesis of computer linguistic systems (CLS) for solving various problems of processing Ukrainian-language text content. It is based on the development and improvement of new and existing models, methods and tools for natural language processing (NLP). The analysis of the current state and prospects for IT development of the textual content natural language processing has been carried out. The concept of CLS is defined and their general classification is given. The detailed analysis of the well-known CLS was carried out, which made it possible to improve the relevant information systems (IS) general classification. The main NLP tasks of CLS are defined, based on which the examples and comparative analysis of the known modern CLS are given. This made it possible to form general directions of the research. The main general scheme of the linguistic analysis process in natural language by means of CLS is described and analysed. The main states and properties of CLS, their classification and features are determined. The well-known classical approaches and trends in natural language processing are analysed. The general classification of the main NLP approaches, directions and additional methods of linguistic research for NLP tasks is presented. The analysis of the existing basic methods and methods of processing natural language by means of the machine learning (ML) was carried out. Their classification has been improved and typical problems of ML methods for processing Ukrainian-language texts have been identified. The overview of the well-known IT development of CLS based on the features and technologies of intellectual analysis of the flow of Ukrainian-language content was made. The main requirements for evaluating the effectiveness of the CLS based on ML technology and big data analysis (BDA) are defined. The basic ML methods for BDA from multiple text content

streams are defined. The requirements for text content clustering in unsupervised ML are defined. The features of the CLS design and development were analysed based on the definition of the main stages such as grapheme, morphological, lexical, syntactic, semantic analysis/synthesis of the Ukrainian-language text for the specific NLP problem solution. The formulation of the problem of processing the Ukrainian-language text based on the definition of the functional features of the text flow intellectual analysis was made and specified. The general analysis of the Ukrainian-language text analysis problem and the definition for the Ukrainian-language text processing made it possible to formulate the main stages and requirements for the typical CLS project for solving the specific NLP problem. The main characteristics identification of the CLS and the justification of the typical CLS project implementation made it possible to determine the expected effects of the project corresponding implementation. Based on the CLS input/output content flows analysis, the functional requirements for the typical CLS project, its software modules, network, software and technical tools for software implementation of IS are defined and formulated.

The analysis and synthesis of CLS is based on the application of linguistic analysis of Ukrainian-language textual content, intelligent processing of textual flow of content, machine learning of the system based on reliable data, and statistical analysis to find patterns in the appearance of linguistic events. Developed information technology (IT) for processing of Ukrainian-language textual content, unlike the existing ones, supports the modularity principle of the typical architecture of the CLS for solving a specific task of the NLP and analysing a set of parameters and metrics of effectiveness of the system in accordance with the behaviour of the target audience. The general structure of the CLS for the processing of text content in the Ukrainian language and the conceptual scheme/model of functioning of a typical CLS based on the modelling of the interaction of the main processes and components of the system were developed, which made possible to improve IT intellectual analysis of the text flow based on the processing of information resources. There are examples of developed CLS for processing Ukrainian-language textual content for solving relevant

tasks of the NLP, functioning of which is based on developed and improved models, methods and algorithms.

An improved model of linguistic processing of textual content based on graphemic, morphological, lexical, syntactic, semantic, structural, ontological and pragmatic analysis to solve a specific problem of NLP is introduced. It has enabled the formulation of general requirements for Ukrainian content processing. Process improvement methodologies for information resources such as integration, management and content support of the Ukrainian language allow to adapt the intellectual analysis of the text stream processing to the solution of various tasks of NLP and develop effective CLS and metrics to solve various NLP problems. NLP methods based on regular pattern-matching expressions are improved and it has allowed the adaptation of grapheme and morphological analysis algorithms to Ukrainian text processing. The main admissible operations of regular expressions (RE) as union and disjunction of symbols/chains/expressions, number and precedence operators, as well as anchors as special symbols for identifying the presence/absence of symbols are defined in RE. A method of tokenisation and normalisation of text by cascades of simple substitutions of regular expressions and finite state machines is upgraded and resulted in the adaptation of the lexical and syntactic analysis algorithm for Ukrainian text processing. The morphological analysis method based on word segmentation and normalisation, sentence segmentation, and a modified Porter stemming algorithm as an effective tool for identifying lemmas affixes to tag the analyzed word is improved. It has resulted in a 9% increase in keyword search accuracy. Unlike the classic Porter algorithm (it does not have the high accuracy even for English-language texts), the modified one is adapted for the Ukrainian language and gives the accurate result in 85-93% of cases, depending on the quality, style, genre of the text and, accordingly, the Ukrainian texts lines is described as the minimum number of operations necessary to transform one into another. The method of syntactic analysis of Ukrainian-language text content aimed at automatic detection of significant keywords of input texts has been improved. The role and formal characteristics of the parser in the process of identifying keywords of the content topic are determined, and

the procedures of the proposed method are decomposed. Unlike well-known parsers, the proposed method provides the self-improvement and self-learning of the CLS module for the definition of keywords due to the mechanism of identification of significant statistical parameters within the limits defined by the moderator.

A method of identifying keywords in Ukrainian texts based on grapheme and morphological analysis of the word base using regular expressions and N-grams is elaborated. It has increased the accuracy of keyword searches by 6-9%, stable word combinations and categorise content search. A method for determining stable word combinations based on the identification of keywords in a Ukrainian text and the lexical coefficients analysis of the text author in the reference text is developed. The accuracy of the method for determining the author's style, based on statistical linguistics, has been improved by 6-7%. The method consists in the use of Zipf law in the formation of stable word combinations as key, taking into account the following rules of preliminary linguistic processing of the text: removal of all stop words; form bigrams only within the limits of punctuation marks; the verb and the pronoun are considered punctuation marks; determine verbs by inflections; form bigrams on the basis of bases without taking into account inflections; determining adjectives by inflections and assuming that adjectives are only on the left in the bigram from Ukrainian-language texts.

A method for determining the author's style of thematic Ukrainian textual content based on the analysis of keywords, stable phrases, N-grams, linguometry and stylometry is developed. It has enabled the recognition of the stylistic contribution of each author and increased the accuracy of scientific and technical publications attribution by 6-12%. The method comprises the comparative analysis of the author attribution in the statistically processed work (standard) with the arbitrary analyzed passage. The method estimates the probability of the publication belonging to the author style template with the analysis of the corresponding coefficients of the author lexical speech. The method works under the condition that the author template has already been researched. The key words analysis was used for attribution, the obtained results are presented based on the analysis of correlation coefficients.

A method is developed to verify the authorship level of a Ukrainian text from the number of possible authors, based on a stylistic comparison analysis of the potential authors. It has improved the classification accuracy of style similarity to [9;34]% of the total number of project participants. The service words are indicative of the writer personal style, since they are not related to the content and topic of the textual content, and the specific subset of such words is characteristic of the specific author. For each of the passages, the absolute and relative frequencies of the stop words were analyzed and compared with the reference values. Therefore, the application of the reference words method gives the following results: finding among the studied passages what most likely belongs to the standard. Other results confirm the effectiveness of the reference words method in the authorial attribution of texts. The insignificance of the influence of the fraction as a parameter of the method on the dynamics of changes in the results of the experiment in Ukrainian-language author texts analysis is proved.

The lexical analysis algorithm for Ukrainian-language texts and the algorithm for the syntactic analyser of text content have been developed. The algorithms peculiarities are the adaptation of the morphological and syntactic analysis of lexical units to the peculiarities of the constructions of Ukrainian words/texts based on each word analysis, taking into account its part of speech and declension. For this purpose, the analysis of the words inflections was carried out for the classification, selection of the basis for the corresponding alphabetic-frequency dictionaries formation. The contents of the dictionaries were subsequently taken into account in the next steps of determining the text authorship as the linguistic features calculation of the author speech.

The analysis and synthesis methods of CLS are developed based on the creation of an organisational structure of the Ukrainian text processing system through the support of modularity, and modelling the main processes and components interaction. It has improved the number of solutions to various typical NLP problems by implementing typical software systems. The approach to the development of software for the content monitoring of stop words to determine the author in Ukrainian-language style texts based on the Web Mining is proposed. The peculiarity of the approach

comprises the linguistic statistical analysis of lexemes adaptation to the morphological and syntactic constructions peculiarities of Ukrainian words/texts. The results of the proposed method experimental testing of determining the author in Ukrainian-language scientific texts of the technical profile in more than 300 individual scientific publications from issues of the Bulletin of the Lviv Polytechnic University of the "Information Systems and Networks" series for the period 2001–2021 were studied. It was found that for the selected experimental base with more than 300 works of the best results according to the density criterion achieve the article analysis method without initial mandatory information such as abstracts and keywords in different languages and the list of references. The reliability of scientific and practical results is confirmed by the relevant materials on the implementation of dissertation research, as well as by comparing the obtained practical results on different samples of the reliability of input data. CLS is realised on the platform http://victana.lviv.ua using CMS Joomla! (developing the site e-framework), PHP (implementation of text content processing methods), HTML (page markup), CSS (description of page styles), MySQL (storing data and dictionaries). An experimental study confirms the reliability of the method used to identify keywords and proves that for different source text processing algorithms, the average agreement of the identified keywords lists with the author's ones varies between 52.6 and 68.5%. The accuracy of keyword matching with the author's keywords ranges from 43.6% to 62.9%. The average match of meaningful keywords compared to all keywords found by the system varies between 38.9 and 75.8%, depending on the analysis stage of the article texts. The accuracy of matching keywords compared to all those found by the system varies between 34.3 and 71.9%, depending on the stage of analysis of the article texts.

The results of the thesis are the following: the CLS model for processing textual content in the Ukrainian language; information technology of intellectual analysis of text flow based on the processing of information resources; modules for processing Ukrainian-language textual content of decision-making support systems; methods and models of linguistic analysis and processing of Ukrainian-language textual content; a module for identifying keywords in Ukrainian-language texts and a method for

determining stable word combinations when identifying keywords in a Ukrainian-language text; module for determining the author of the Ukrainian-language text and determining the style of the author of the text. They were implemented during the development of the intelligent system for automatic monitoring of the condition of long-term structural elements for "Hydraulik-Partner" LLC (Lviv) during the execution of contract 265 "Business analysis of the activities of "Hydraulik Partner" LLC based on an ontological approach" and the digit recognition system module in voice based on neural networks for "A-Solutions Development" LLC (Lviv) during the execution of farm contract 677, as well as used during the implementation of scientific research works on state budget topics at the National University "Lviv Polytechnic", which is confirmed by the relevant acts of implementation.

The results of the work are included in the educational process of students of specialties 124 “System Analysis” and 126 “Information Technologies and Systems”, namely in the disciplines “Computer Linguistics”, “Language Recognition”, “Computational Methods and Data Visualization” through the use of educational aids: Pasichnyk V., Shcherbyna Y., Vysotska V., Shestakevich T. Mathematical linguistics. Combinatorial linguistics. Lviv: LPNU, 2019; Lytvyn V., Peleshchak R., VysotskaV. Depth learning.Lviv: LPNU, 2021; Andrunyk V., Vysotska V., Pasichnyk V., Chyrun L., Chyrun L. Numerical methods in computer sciences. Lviv: NovySvit, 2017. T. 1-2; Ryshkovets Y., Vysotska V. Algorithmization and programming. Part 1-2.Lviv: NovySvit, 2018; Vysotska V., Lytvyn V., Lozynska O. Discrete mathematics: practicum. Lviv: NovySvit, 2019; Vysotska V., Oborska O. Python: algorithmization and programming. Lviv: NovySvit, 2020.

Keywords: NLP, computational linguistics, text content, Ukrainian language, grapheme analysis, morphological analysis, lexical analysis, syntactic analysis, semantic analysis, structural analysis, pragmatic analysis, information technology, machine learning, natural language processing, information system, ontology, keywords, stable word phrases, author's style, author identification, psychological analysis of the text.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

Статті у періодичних виданнях, індексованих у Scopus та Web of Science

1. Lytvyn V., Pukach P., Vysotska V., Vovk M., Kholodna N. Identification and correction of grammatical errors in Ukrainian texts based on machine learning technology. Mathematics. 2023. Vol. 11. 904. ISSN 2227-7390. (квартиль Q2 відповідно до SCImago Journal).
2. Bisikalo O., Danylchuk O., Kovtun V., Kovtun O., Nikitenko O., Vysotska V. Modeling of operation of information system for critical use in the conditions of influence of a complex certain negative factor. International Journal of Control, Automation and Systems. 2022. Vol. 20. P. 904–1913. Print ISSN 1598-6446. (квартиль Q2 відповідно до SCImago Journal).
3. Bublyk M., Kowalska-Styczeń A., Lytvyn V., Vysotska V. The Ukrainian economy transformation into the circular based on fuzzy-logic cluster analysis. Energies. 2021. Vol. 14(18). Art. 5951. ISSN:1996-1073. (квартиль Q2 відповідно до SCImago Journal).
4. Lytvyn V., Vysotska V., Peleshchak I., Rishnyak I., Peleshchak R. Time dependence of the output signal morphology for nonlinear oscillator neuron based on Van der Pol model. International Journal of Intelligent Systems and Applications. 2018.Vol. 10(4). P. 8–17. ISSN: 2074-904X. (квартиль Q2 відповідно до SCImago Journal).
5. Висоцька В. Метод авторифікації тексту науково-технічних публікацій на основі лінгвістичного аналізу коефіцієнтів мовної різноманітності. Радіоелектроніка. Інформатика. Управління. 2020. № 1(52). С. 108–124.
6. Висоцька В. Інформаційна технологія просування інтернет-ресурсів в пошукових системах на основі контент-аналізу ключових слів web-сторінок. Радіоелектроніка, інформатика, управління. 2021 № 3 (58). С. 133-151.
7. Алєксєєва К. А., Берко А. Ю., Висоцька В. А. Технологія управління комерційним web-ресурсом на основі нечіткої логіки. Радіоелектроніка. Інформатика. Управління. 2015. № 3 (34). С. 71–79.
8. Бісікало О. В., Висоцька В. А. Виявлення ключових слів на основі методу контент-моніторингу україномовних текстів. Радіоелектроніка. Інформатика. Управління. 2016. № 1 (36). С. 74–83.
9. Бісікало О. В., Висоцька В. А. Застосування методу синтаксичного аналізу речень для визначення ключових слів україномовного тексту. Радіоелектроніка. Інформатика. Управління. 2016. № 3 (38). С. 54–65.
10. Lytvyn V., Pukach P., Bobyk I., Vysotska V. The method of formation of the status of personality understanding based on the content analysis. Eastern-European Journal of Enterprise Technologies. 2016. Vol. 5. P. 4–12.
11. Литвин В. В., Бобик І. О., Висоцька В. А. Застосування системи алгоритмічних алгебр для граматичного аналізу символьних обчислень виразів логіки висловлювань. Радіоелектроніка. Інформатика. Управління. 2016. № 4 (39). С. 77–89.

12. Lytvyn V., Vysotska V., Pukach P., Bobyk I., Pakholok B. A method for constructing recruitment rules based on the analysis of a specialist's competences. Eastern-European Journal of Enterprise Technologies. 2016. Vol. 6/2 (84). P. 4–14.
13. Lytvyn V., Vysotska V., Pukach P., Brodyak O., Ugryn D. Development of a method for determining the keywords in the Slavic language texts based on the technology of web mining. Eastern-European Journal of Enterprise Technologies. 2017. Vol. 2/2 (86). P. 14–23.
14. Lytvyn V., Vysotska V., Pukach P., Vovk M., Ugryn D. Method of functioning of intelligent agents, designed to solve action planning problems based on ontological approach. Eastern-European Journal of Enterprise Technologies. 2017. Vol. 3/2 (87). P. 11–17.
15. Lytvyn V., Vysotska V., Pukach P., Bobyk I., Uhryn D. Development of a method for the recognition of author's style in the Ukrainian language texts based on linguometry, stylemetry. Eastern-European Journal of Enterprise Technologies. 2017. Vol. 4/2 (88). P. 10–18.
16. Коробчинський М. В., Чирун Л. Б., Висоцька В. А., Нич М. О. Особливості прогнозування результатів матчів у кіберспорті. Радіоелектроніка. Інформатика. Управління. 2017. № 3. С. 95–105.
17. Коробчинський М. В., Чирун Л. Б., Висоцька В. А., Кондратьєв Є. О. Особливості формування та аналізу контенту інтернет-газети музичних новин. Радіоелектроніка. Інформатика. Управління. 2017. № 4. С. 139–150.
18. Lytvyn V., Vysotska V., Uhryn D., Hrendus M., Naum O. Analysis of statistical methods for stable combinations determination of keywords identification. Eastern-European Journal of Enterprise Technologies. 2018. Vol. 2/2 (92). P. 23–37.
19. Lytvyn V., Vysotska V., Maria H. Method of data expression from the Ukrainian content based on the ontological approach. Радіоелектроніка. Інформатика. Управління. 2018. № 3 (46). P. 144–157.
20. Lytvyn V., Vysotska V., Pukach P., Nytrebych Z., Demkiv I., Kovalchuk R., Huzyk N. Development of the linguometric method for automatic identification of the author of text content based on statistical analysis of language diversity coefficients. Eastern-European Journal of Enterprise Technologies. 2018. Vol. 5. P. 16–28.
21. Lytvyn V., Vysotska V., Kuchkovskiy V., Pelekh I., Bobyk I., Malanchuk O., Ryshkovets Y., Brodyak O., Bobrivetc V., Panasyuk V. Development of the system to integrate and generate content considering the cryptocurrent needs of users. Eastern-European Journal of Enterprise Technologies. 2019. Vol. 1/2. P. 18–39.
22. Lytvyn V., Vysotska V., Pukach P., Nytrebych Z., Demkiv I., Senyk A., Malanchuk O., Sachenko S., Kovalchuk R., Huzyk N. Analysis of the developed quantitative method for automatic attribution of scientific and technical text content written in Ukrainian. Eastern-European Journal of Enterprise Technologies. 2018. Vol. 6/2 (96). P. 19–31.
23. Berko A., Vysotska V., Lytvyn V., Naum O. Planning the activities of intellectual agents in the electronic commerce systems. Радіоелектроніка. Інформатика. Управління. 2018. № 4. С. 143–158.
24. Lytvyn V., Vysotska V., Demchuk A., Demkiv I., Ukhans'ka O., Hladun V., Kovalchuk R., Petruchenko O., Dzyubyk L., Sokulska N. Design of the architecture of an intelligent system for distributing commercial content

in the internet space based on SEO-technologies, neural networks, and machine learning. Eastern-European Journal of Enterprise Technologies. 2019. Vol. 2/2(98). P. 15–34.

25. Lytvyn V., Vysotska V., Shatskykh V., Kohut I., Petruchenko O., Dzyubyk L., Bobrivetc V., Panasyuk V., Sachenko S., Komar M. Design of a recommendation system based on collaborative filtering and machine learning considering personal needs of the user. Eastern-European Journal of Enterprise Technologies. 2019. Vol. 4/2 (100). P. 6–28.
26. Vysotska V., Demchuk A., Lytvyn V. Features of the architecture for Internet commercial content management system based on methods of Machine Learning, Web mining and SEO technologies. Радіоелектроніка. Інформатика. Управління. 2019. № 4. С. 121–135.
27. Lytvyn V., Vysotska V., Budz I., Pelekh Y., Sokulska N., Kovalchuk R., Dzyubyk L., Tereshchuk O., Komar M. Development of the quantitative method for automated text content authorship attribution based on the statistical analysis of N-grams distribution. Eastern-European Journal of Enterprise Technologies. 2019. Vol. 6/2 (102). P. 28–51.
28. Кравець П., Литвин В., Висоцька В. Ігрова модель онтологічної підтримки проектів. Радіоелектроніка, інформатика, управління. 2021. № 1(56). С. 172–183.
29. Литвин В. В., Бублик М. І., Висоцька В. А., Мацелюх Ю. Р. Технологія візуальної симуляції пасажиропотоків у сфері громадського транспорту smart city. Радіоелектроніка, інформатика, управління. 2021 № 4 (59). С. 106-121.
30. Кравець П. О., Литвин В. В., Висоцька В. А. Моделювання ігрової задачі призначення персоналу для виконання ІТ-проектів на основі онтологій. Радіоелектроніка, інформатика, управління. 2022. № 1 (60). С. 130–145.
31. Lytvyn V., Vysotska V., Veres O., Rishnyak I., Rishnyak H. Classification methods of text documents using ontology based approach. Advances in Intelligent Systems and Computing. 2017. Vol. 512. P. 229–240.
32. Shakhovska N., Vysotska V., Chyrun L. Intelligent systems design of distance learning realization for modern youth promotion and involvement in independent scientific researches. Advances in Intelligent Systems and Computing. 2017. Vol. 512. P. 175–198. ISSN 2194-5357.
33. Lytvyn V., Vysotska V., Burov Y., Veres O., Rishnyak I. The contextual search method based on domain thesaurus. Advances in Intelligent Systems and Computing. 2018. Vol. 689. P. 310–319.ISSN 2194-5357.
34. Kanishcheva O., Vysotska V., Chyrun L., Gozhyj A. Method of integration and content management of the information resources network. Advances in Intelligent Systems and Computing. 2018. Vol. 689. P. 204–216.
35. Vysotska V., Fernandes B. V., Emmerich M. Web content support method in electronic business systems. CEUR Workshop Proceedings. 2018. Vol. 2136. P. 20–41. E-ISSN: 1613-0073.
36. Lytvyn V., Vysotska V., Dosyn D., Y Burov. Method for ontology content and structure optimization, provided by a weighted conceptual graph. Webology. 2018. Vol. 15(2). P. 66–85. E-ISSN: 1735-188X.
37. Lytvyn V., Vysotska V., Osypov M., Slyusarchuk O., Y Slyusarchuk. Development of intellectual system for data de-duplication and distribution in cloud storage. Webology. 2019. Vol. 16(2). P. 1-42.

38. Vysotska V., Lytvyn V., Burov Y., Gozhyj A., Makara S. The consolidated information web-resource about pharmacy networks in city. CEUR Workshop Proceedings. 2018. Vol. 2255. P. 239–255.
39. Lytvyn V., Sharonova N., Hamon T., Vysotska V., Grabar N., Kowalska-Styczen A. Computational linguistics and intelligent systems. CEUR Workshop Proceedings. 2018. Vol. 2136. 390 p.
40. Rusyn B., Lytvyn V., Vysotska V., Emmerich M., Pohreliuk L. The virtual library system design and development. Advances in Intelligent Systems and Computing (AISC). 2019. Vol. 871. P. 328–349.
41. Vysotska V., Fernandes B. V., Lytvyn V., Emmerich M., Hirnyak M. Method for determining linguometric coefficient dynamics of Ukrainian text content authorship. Advances in Intelligent Systems and Computing (AISC). 2019. Vol. 871. P. 132–151. ISSN 2194-5357.
42. Gozhyj A., V Vysotska., Yevseyeva I., Kalinina I., Gozhyj V. Web resources management method based on intelligent technologies. Advances in Intelligent Systems and Computing. 2019. Vol. 871. P. 206–221.
43. Vysotska V., Lytvyn V., Burov Y., Berezin P., Emmerich M., Fernandes B. V. Development of information system for textual content categorizing based on ontology. CEUR Workshop Proceedings. 2019. V. 2362. P. 53–70.
44. Burov Y., Vysotska V., Kravets P. Ontological approach to plot analysis and modeling. CEUR Workshop Proceedings. 2019. Vol. 2362. P. 22–31. E-ISSN: 1613-0073.
45. Zdebskyi P., Vysotska V., Peleshchak R., Peleshchak I., Demchuk A., Krylyshyn M. An application development for recognizing of view in order to control the mouse pointer. CEUR Workshop Proceedings. 2019. Vol. 2386. P. 55–74. E-ISSN: 1613-0073.
46. Lytvyn V., Vysotska V., Rusyn B., Pohreliuk L., Berezin P., Naum O. Textual content categorizing technology development based on ontology. CEUR Workshop Proceedings. 2019. Vol. 2386. P. 234–254.
47. Lytvyn V., Vysotska V., Rzheuskyi A. Technology for the psychological portraits formation of social networks users for the IT specialists recruitment based on Big Five, NLP and Big Data. CEUR Workshop Proceedings. 2019. M2392. P. 147–171. E-ISSN: 1613-0073.
48. Vysotska V., Burov Y., Lytvyn V., Oleshek O. Automated monitoring of changes in web resources. Advances in Intelligent Systems and Computing. 2020. Vol. 1020. P. 348–363. ISSN 2194-5357, E-ISSN: 2194-5365.
49. Demchuk A., Lytvyn V., Vysotska V., Dilai M. Methods and means of web content personalization for commercial information products distribution. Advances in Intelligent Systems and Computing. 2020. Vol. 1020. P. 332–347. ISSN 2194-5357, E-ISSN: 2194-5365.
50. Lytvyn V., Vysotska V. Mykhailyshyn V. Rzheuskyi A. Semianchuk S. System development for video stream data analyzing. Advances in Intelligent Systems and Computing. 2020. Vol. 1020. P. 315–331.
51. Kravets P., Burov Y., Lytvyn V., Vysotska V. Gaming method of ontology clusterization. Webology. 2019. Vol. 16(1). P. 55–76. ISSN: 1735-188X.
52. Chyrun L, Leshchynskyy E., Lytvyn V., Rzheuskyi A., Vysotska V., Borzov Y. Intellectual analysis of making decisions tree in information systems of screening observation. CEUR Workshop Proceedings. 2019. Vol. 2488. P. 281–296. E-ISSN: 1613-0073.
53. Lytvyn V., Kowalska A., Peleshko D., Rak T., Voloshyn V., Noennig J., Vysotska V., Nykolyshyn L., Pryshchepa H. Aviation aircraft planning system project development. AISC. 2020. Vol. 1080. P. 315–348.

54. Lytvyn V., Burov Y., Kravets P., Vysotska V., Demchuk A., Berko A., Ryshkovets Y., Shcherbak S., Naum O. Methods and models of intellectual processing of texts for building ontologies of software for medical terms identification in content classification. CEUR Workshop Proceedings. 2019. Vol. 2488. P. 354–368.
55. Lytvyn V., Vysotska V., Shakhovska N., Mykhailyshyn V., Medykovskyy M., Peleshchak I., Fernandes V.B., Peleshchak R., Shcherbak S. A smart home system development. Advances in Intelligent Systems and Computing. 2020. Vol. 1080. P. 804–830. ISSN 2194-5357, E-ISSN: 2194-5365.
56. Lytvyn V., Gozhyj A., I Kalinina., Vysotska V., Shatskykh V., Chyrun L., Borzov Y. An intelligent system of the content relevance at the example of films according to user needs. CEUR Workshop Proceedings. 2019. Vol. 2516. P. 1–23. E-ISSN: 1613-0073.
57. Peleshko D., Rak T., Noennig J. R., Lytvyn V., Vysotska V. Drone monitoring system DROMOS of urban environmental dynamics. CEUR Workshop Proceedings. 2020. Vol. 2565. P. 178–193. E-ISSN: 1613-0073.
58. Krislata I., Katrenko A., Lytvyn V., Vysotska V., Burov Y. Traffic flows system development for smart city. CEUR Workshop Proceedings. 2020. Vol. 2565. P. 280–294. E-ISSN: 1613-0073.
59. Bisikalo O., Vysotska V. Linguistic analysis method of Ukrainian commercial textual content. CEUR Workshop Proceedings. 2020. Vol. 2608. P. 224–244. E-ISSN: 1613-0073.
60. Bisikalo O., Vysotska V., Kravets Y., Burov P. Conceptual model of process formation for the semantics of sentence in natural language. CEUR Workshop Proceedings. 2020. Vol. 2604. P. 151–177.
61. Vysotska V. Ukrainian participles formation by the generative grammars use. CEUR Workshop Proceedings. 2020. Vol. 2604. P. 407–427. E-ISSN: 1613-0073.
62. Batiuk T., Vysotska V., Lytvyn V. Intelligent system for socialization by personal interests on the basis of SEO technologies and methods of machine learning. CEUR Workshop Proceedings. 2020. V. 2604. P. 1237–1250.
63. Oliinyk V., Vysotska V., Burov Y., Mykich K., Basto-Fernandes V. Propaganda detection in text data based on NLP and machine learning. CEUR Workshop Proceedings. 2020. Vol. 2631. P. 132–144.
64. Kalinina I., Vysotska V., Sachenko S., Kovalchuk R., Gozhyj A. Qualitative and quantitative characteristics analysis for information security risk assessment in e-commerce systems. CEUR Workshop Proceedings. 2020. Vol. 2762. P. 177–190. E-ISSN: 1613-0073.
65. Lytvyn V., Hryhorovych A., Hryhorovych V., Vysotska V., Bublyk M., Chyrun L. Medical content processing in intelligent system of district therapist. CEUR Workshop Proceedings. 2020. V. 2753. P. 415–429.
66. Bublyk M., Lytvyn V., Vysotska V., Sokulska N., Chyrun L. Matseliukh Y. The decision tree usage for the results analysis of the psychophysiological testing. CEUR Workshop Proceedings. 2020. Vol. 2753. P. 458–472.
67. Vysotska V., Bublyk M., Korolenko O., Matseliukh Y., Kopach T. Network modelling of resource consumption intensities in human capital management in digital business enterprises. CEUR Workshop Proceedings. 2021. Vol. 2851. P. 366–380. E-ISSN: 1613-0073.
68. Kravets P., Lytvyn V., Vysotska V., Burov Y., Andrusyak I. Game task of ontological project coverage. CEUR Workshop Proceedings. 2021. Vol. 2851. P. 344–355. E-ISSN: 1613-0073.
69. Kravets P., Burov Y., Oborska O., Vysotska V., Dzyubyk L., Lytvyn V. Stochastic Game Model of Data Clustering. CEUR Workshop Proceedings. 2021. Vol. 2853. P. 198–213. E-ISSN: 1613-0073.

70. Bublyk M., Vysotska V., Panasyuk V., Brodyak O., Chyrun L. Assessing security risks method in e-commerce system for IT portfolio management. CEUR Workshop Proceedings. 2021. Vol. 2853. P. 462–479.
71. Vysotska V., Lytvyn V., Danylyk V. Vyshemyrska S. Lurie I. Luchkevych M. Detecting items with the biggest weight based on neural network and machine learning methods. Communications in Computer and Information Science.2020.V.1158.P.383-396.ISSN1865-0929.
72. Kalinina I., Vysotska V., Bidyuk P., Gozhyj A. Methods for forecasting nonlinear non-stationary processes in machine learning. Communications in Computer and Information Science. 2020. Vol. 1158. P. 470–485.
73. Matseliukh Y., Bublyk M., Vysotska V. Development of intelligent system for visual passenger flows simulation of public transport in Smart City based on neural network. CEUR Workshop Proceedings. 2021. Vol. 2870. P. 1087–1138. E-ISSN: 1613-0073.
74. Pashchetnyk O., Lytvyn V., Zhyvchuk V., Polishchuk L., Vysotska V., Rybchak Z., Pukach Y. The ontological decision support system composition and structure determination for commanders of Land Forces formations and units in Ukrainian Armed Force. CEUR Workshop Proceedings. 2021. Vol. 2870. P. 1077–1086.
75. Lytvyn V., Pashchetnyk O., Klymovych O., Polishchuk L., Kolb I., Burov Y., Vysotska V. Assessment of the hydro-meteorological conditions impact on the combat troops operations preparation and conduct in the geo-information subsystem of the automated battlefield system. CEUR Workshop Proceedings. 2021. Vol. 2870. P. 1063–1076. E-ISSN: 1613-0073.
76. Tymoshenko K., Vysotska V., Kovtun O., Holoshchuk R., Holoshchuk S. Real-time Ukrainian text recognition and voicing. CEUR Workshop Proceedings. 2021. Vol. 2870. P. 357–387. E-ISSN: 1613-0073.
77. Vysotska V., Holoshchuk S., Holoshchuk R. A comparative analysis for English and Ukrainian texts processing based on semantics and syntax approach. CEUR Workshop Proceedings. 2021. Vol. 2870. P. 311–356.
78. Dokhnyak B., Vysotska V. Intelligent Smart Home System Using Amazon Alexa Tools. CEUR Workshop Proceedings. 2021. Vol. 2917. P. 441-464. E-ISSN: 1613-0073.
79. Zdorenko Y., Lavrut O., T Lavrut., V Lytvyn., Burov Y., Vysotska V. Route Selection Method in Military Information and Telecommunication Networks Based on ANFIS. CEUR Workshop Proceedings. 2021. Vol. 2917. P. 514-524. E-ISSN: 1613-0073.
80. Balush I., Vysotska V., Albota S. Recommendation System Development Based on Intelligent Search, NLP and Machine Learning Methods. CEUR Workshop Proceedings. 2021. Vol. 2917. P. 584-617.
81. Kholodna N., Vysotska V., Albota S. A Machine Learning Model for Automatic Emotion Detection from Speech. CEUR Workshop Proceedings. 2021. Vol. 2917. P. 699-713. E-ISSN: 1613-0073.
82. Kravets P., Lytvyn V., Dobrotvor I., Sachenko O., Vysotska V., Sachenko A. Matrix Stochastic Game with Q-learning for Multi-agent Systems. Lecture Notes on Data Engineering and Communications Technologies. 2021. Vol. 83. P. 304–314. ISSN 23674512.
83. Kravets P., Burov Y., Lytvyn V., Vysotska V., Ryshkovets Y., Brodyak O., Vyshemyrska S. Markovian Learning Methods in Decision-Making Systems. Lecture Notes on Data Engineering and Communications Technologies. 2022. Vol. 77. P. 423-437. ISSN 23674512.

84. Vysotska V., Berko A., Lytvyn V., Kravets P., Dzyubyk L., Bardachov Y., Vyshemyrska S. Information resource management technology based on fuzzy logic. AISC. 2020. Vol. 1246. P. 164–182.
85. Kravets P., Lytvyn V., Vysotska V., Ryshkovets Y., Vyshemyrska S., Smailova S. Dynamic coordination of strategies for multi-agent systems. Advances in Intelligent Systems and Computing. 2020. Vol. 1246. P. 653–670.

Статті у наукових фахових виданнях України

86. Алєксєєва К. А., Берко А. Ю., Висоцька В. А. Управління Web-ресурсами за умов невизначеності. Технологічний аудит та резерви виробництва. 2015. № 2 (2). С. 4–7.
87. Висоцька В. А. Гопяк М. В., Козлов П. Ю. Особливості технології управління web-ресурсом. Інженерія програмного забезпечення. 2015. № 1 (21). С. 25–35.
88. Висоцька В. А., Чирун Л. В. Формальна модель опрацювання інформаційних ресурсів в системах електронної контент-комерції. Вісник НУ «Львівська політехніка». 2015. № 814. С. 44–54.
89. Берко А. Ю., Висоцька В. А., Чирун Л. В. Лінгвістичний аналіз текстового комерційного контенту. Вісник НУ «Львівська політехніка». 2015. № 814. С. 203–227.
90. Висоцька В. А. Особливості моделювання синтаксису речення слов'янських та германських мов за допомогою породжувальних контекстно-вільних грамматик. Вісник НУ «Львівська політехніка». 2015. № 814. С. 246–276.
91. Кісь Я. П., Висоцька В. А., Чирун Л. Б., Фольтович В. М. Застосування контент-аналізу для опрацювання текстових масивів даних. Вісник НУ «Львівська політехніка». 2015. № 814. С. 282–292.
92. Шестакевич Т. В., Висоцька В. А., Чирун Л. В., Чирун Л. Б. Моделювання семантики речення природною мовою за допомогою породжувальних граматик. Вісник НУ «Львівська політехніка». 2015. № 814. С. 335–352.
93. Висоцька В. А. Нога А. Ю., Козлов П. Ю. Управління Web-проектами електронного бізнесу для реалізації комерційного контенту. Вісник НУ «Львівська політехніка». 2015. № 814. С. 421–434.
94. Висоцька В. А., Чирун Л. В. Концептуальна модель опрацювання інформаційних ресурсів в системах електронної контент-комерції. Математичні машини і системи. 2015. № 3. С. 179–190.
95. Висоцька В. А., Чирун Л. В. Опрацювання інформаційних ресурсів у системах електронної контент-комерції. Відбір і обробка інформації. 2015. Вип. 42 (118). С. 84–92.
96. Алєксєєва К. А., Берко А. Ю., Висоцька В. А. Особливості процесу управління web-ресурсом комерційного контенту на основі нечіткої логіки. Вісник НУ «Львівська політехніка». 2015. № 826. С. 201–211.
97. Висоцька В. А. Особливості рубрикації текстового комерційного контенту. Вісник НУ «Львівська політехніка». 2015. № 826. С. 359–367.
98. Алєксєєва К. А., Берко А. Ю., Висоцька В. А. Інформаційна технологія управління Web-ресурсом на основі нечіткої логіки. Вісник НУ «Львівська політехніка». 2015. № 829. С. 7–28.
99. Висоцька В. А. Аналітичні методи опрацювання інформаційних ресурсів в системах електронної контент-комерції. Вісник НУ «Львівська політехніка». 2015. № 829. С. 76–101.

100. Гасько Р. В., Висоцька В. А., Чирун Л. Б. Інформаційна система аналізу психологічного стану особистості. Вісник НУ «Львівська політехніка». 2015. № 829. С. 102–128.
101. Бісікало О. В., Висоцька В. А. Експериментальне дослідження пошуку значущих ключових слів україномовного контенту. Вісник НУ «Львівська політехніка». 2015. № 829. С. 255–272.
102. Чирун Л. Б., Кучковський В. В., Висоцька В. А. Особливості методів контент-аналізу текстових масивів даних web-ресурсів в межах регіону контенту. Вісник НУ «Львівська політехніка». 2015. № 829. С. 296–320.
103. Андруник В. А., Висоцька В. А., Чирун Л. Б. Проект розроблення та впровадження системи електронної контент-комерції. Вісник НУ «Львівська політехніка». 2015. № 829. С. 321–348.
104. Козлов П. Ю., Висоцька В. А., Чирун Л. Б. Сучасні технології управління Web-ресурсами в інформаційній системі аналізу сервісу цифрової дистрибуції. Вісник НУ «Львівська політехніка». 015. № 832. С. 103–128.
105. Кучковський В. В., Висоцька В. А., Нитребич С. З., Оливко Р. М. Застосування методів Інтернет-маркетингу для аналізу Web-ресурсів в межах регіону. Вісник НУ «Львівська політехніка». 2015. № 832. С. 129–164.
106. Шаховська Н. Б., Висоцька В. А., Чирун Л. Б. Методи та засоби дистанційної освіти для заохочення і залучення сучасної молоді до проведення самостійних наукових досліджень. Вісник НУ «Львівська політехніка». 2015. № 832. С. 254–284.
107. Литвин В. В., Висоцька В. А., Досин Д. Г., Гірняк М. Г. Розроблення методів та засобів побудови інтелектуальних систем опрацювання інформаційних ресурсів з використанням онтологічного підходу. Вісник НУ «Львівська політехніка». 2015. № 832. С. 295–314.
108. Алєксєєва К. А., Берко А. Ю., Висоцька В. А. Аналіз процесу опрацювання web-ресурсу інформаційного продукту на основі нечіткої логіки та контент-аналізу. Вісник НУ «Львівська політехніка». 2016. № 843. С. 122–134.
109. Андруник В. А. Висоцька В. А., Чирун Л. В. Особливості формування електронних дайджестів. Вісник НУ «Львівська політехніка». 2016. № 843. С. 3–14.
110. Бісікало О. В., Висоцька В. А. Метод лінгвістичного аналізу україномовного комерційного контенту. Вісник НУ «Львівська політехніка». 2016. № 854. С. 185–204.
111. Вінтоняк С. М., Коробчинський М. В., Чирун Л. Б., Висоцька В. А. Аналіз особливостей Інтернет-порталу аматорських спортивних ігор. Вісник НУ «Львівська політехніка». 2016. № 854. С. 21–41.
112. Vysotska V., Chyrun L., Chyrun L. Online newspaper content analysis based on SEO technologies. Вісник НУ «Львівська політехніка». 2016. № 859. С. 3–16.
113. Литвин В. В., Ремешило-Рибчинська О. І., Висоцька В. А. Побудова онтології архітектурних термінів. Відбір і обробка інформації. 2017. Вип. 44 (120). С. 90–96.
114. Фольтович В. М., Коробчинський М. В., Чирун Л. Б., Висоцька В. А. Метод контент-аналізу текстової інформації Інтернет газети. Вісник НУ «Львівська політехніка». 2017. № 864. С. 7–19.

115. Гасько Р. В., Чирун Л. В., Висоцька В. А. Особливості контент-аналізу користувацької Інтернет-діяльності для формування зрізу психологічного стану особистості. Вісник НУ «Львівська політехніка». 2017. № 864. С. 221–238.
116. Lytvyn V., Vysotska V., Veres O., Brodyak O., Oryshchyn O. Big Data analytics ontology. Технологічний аудит та резерви виробництва. 2018. Vol. 1, № 2(39). С. 16–27.
117. Висоцька В. А., Наум О. М. Порівняння складності автоматичного опрацювання англійських та українських текстів з врахуванням семантики та синтаксису природних мов. Вісник НУ «Львівська політехніка». 2017. № 872. С. 149–162.
118. Шаховська Н. Б. Висоцька В. А., Скотар О. О. Розроблення архітектури інтелектуальної системи на основі інноваційних методів навчання студентів. Вісник НУ «Львівська політехніка». 2017. № 872. С. 220–229.
119. Русин Б., Висоцька В., Погрелюк Л. Особливості проектування та розроблення інформаційної системи Virtual Library. Оптико-електронні інформаційно-енергетичні технології. 2017. Т. 34, № 2. С. 18–33.
120. Литвин В. В., Висоцька В. А., Кучковський В. В., Дуткевич С. Ю., Наум О. Метод інтеграції та управління контентом мережі інформаційних ресурсів туризму згідно з потребами користувача. Вісник НУ «Львівська політехніка». 2018. № 901. С. 22–36.
121. Литвин В. В., Висоцька В. А., Кучковський В. В., Оливко Р. М. Архітектура інформаційної системи інтеграції та формування контенту про криптовалюти на основі аналізу бірж. Вісник НУ «Львівська політехніка». 2018. № 901. С. 43–60.
122. Русин Б. П., Погрелюк Л. В., Висоцька В. А., Осипов М. М., Варецький Я. Ю., Капшій О. В. Архітектура системи дедублікації та розподілу даних у хмарних сховищах під час резервного копіювання. Інформаційні технології та комп'ютерна інженерія. 2019. Т. 2, № 45. С. 40–63.
123. Литвин В. В., Наум О., Висоцька В. А., Дверій М. В. Архітектура системи онлайн-туризму для пошуку та планування подорожей із урахуванням потреб користувача. Вісник НУ «Львівська політехніка». 2019. Вип. 6. С. 13–29.
124. Русин Б., Погрелюк Л., Висоцька В., Осипов М. Метод дедублікації та розподілу даних у хмарних сховищах під час резервного копіювання даних Вісник НУ «Львівська політехніка». 2019. № 6. С. 1–12.
125. Пелещак Р. М., Литвин В. В., Пелещак І. Р., Висоцька В. А. Розробка штучної нейронної мережі з осциляторними нейронами для розпізнавання спектральних образів. Вісник НУ «Львівська політехніка». 2020. Вип. 7. С. 16–23.
126. Пелещак Р. М., Литвин В. В., Пелещак І. Р., Висоцька В. А., Черняк О. І. Побудова оптимізованої багатошарової нейронної мережі в межах нелінійної моделі узагальненої похибки. Вісник НУ «Львівська політехніка». 2021. Вип. 9. С. 53–60.
127. Батюк Т.М., Висоцька В. А. Розробка інтелектуальної системи підтримки соціалізації користувача за подібністю інтересівю Сучасний стан наукових досліджень та технологій в промисловості. 2022. Вип. 1(19). С. 13-26.

128. Batiuk T., Vysotska V. Decision-making support system to support of social networks users based similar common interests and preferences. Computer systems and information technologies. 2022. Vol. 1. P. 11-22.
129. Батюк Т.М., Висоцька В. А. Інформаційна підтримка процесів соціалізації особистості на основі інтересів. Вісник НУ «Львівська політехніка». 2022. №11. С. 56 – 86.
130. Олексів Н., Висоцька В. Мобільна інформаційна система контролю раціону харчування людини. Вісник НУ «Львівська політехніка». 2022. Вип. 11. С. 145 – 172.

Монографії

131. Lytvyn V., Vysotska V., Chyrun L., Dosyn D. Methods based on ontologies for information resources processing. Saarbrücken: LAP Lambert Academic Publishing, 2016. 324 p.
132. Литвин В. В., Висоцька В. А., Досин Д. Г. Методи та засоби опрацювання інформаційних ресурсів на основі онтологій. Львів: Піраміда, 2016. 404 с.
133. Висоцька В. А. Технології електронної комерції та Інтернет-маркетингу. Saarbrücken: LAP Lambert Academic Publishing, 2018. 285 с.
134. Vysotska V., Lytvyn V. Web resources processing based on ontologies. Saarbrücken: LAP Lambert Academic Publishing, 2018. 232 p.
135. Vysotska V., Shakhovska N. Information technologies of gamification for training and recruitment. Saarbrücken: LAP Lambert Academic Publishing, 2018. 248 p.
136. Vysotska V Internet systems design and development based on Web Mining and NLP. Saarbrücken: LAP Lambert Academic Publishing, 2018. 316 p.
137. Vysotska V. Computer linguistics for online marketing in information technology. Saarbrücken: LAP Lambert Academic Publishing, 2018. 396 p.
138. Висоцька В. А., Досин Д. Г., Микіч Х. І., Завущак І. І., Рибчак З. Л. Методи та засоби функціонування систем підтримки прийняття рішень на основі онтологій. Львів: Новий світ, 2019. 334 с.
139. Peleshchak R., Peleshchak I., Vysotska V. Methods for recognizing multispectral images based on neural networks. Beau Bassin: Lambert Academic Publishing, 2020. 153 p.

Статті у міжнародних виданнях

140. Vysotska V., Chyrun L. Linguistic analysis and modelling semantics of textual content for digest formation. MEST Journal. 2015. Vol. 3, № 1. P. 127–148. ISSN: 2334-7171.
141. Chyrun L., Vysotska V. Features of the content-analysis method for text categorization of commercial content in processing online newspaper articles. Applied Computer Science. 2015. Vol. 11, № 1. P. 15–30.
142. Chyrun L., Andrunyk V., Vysotska V. Electronic content commerce system development. MEST Journal. 2015. Vol. 3, № 2. P. 10–33. ISSN: 2334-7171.
143. Vysotska V., Chyrun L. The means structure of information resources processing in electronic content commerce systems. Journal of Information Sciences and Computing Technologies. 2015. Vol 3, № 3. P. 241–248.
144. Vysotska V., Chyrun L. Methods and means of processing information resources in electronic content commerce systems. Applied Computer Science. 2015. Vol. 11(2). P. 68–85. ISSN: 2353-6977.

145. Chyrun L. Vysotska V., Laba R. Information resources analysis in electronic content commerce systems. Applied Computer Science. 2016. Vol. 12(1). P. 48–66. ISSN: 2353-6977.
146. Vysotska V. Chyrun L., Kozlov P. Analysis of business processes in electronic content-commerce systems. Econtechmod. 2016. Vol. 5, № 1. P. 111–125. ISSN: 2084-5715.
147. Vysotska V., Chyrun L., Kozlov P. Design and analysis features of generalized electronic content-commerce systems architecture. Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska. 2016. № 6. P. 48–59.
148. Chyrun L., Vysotska V., Kozak I. Informational resources processing intellectual systems with textual commercial content linguistic analysis usage constructional means and tools development. Econtechmod. 2016. Vol. 5, № 2. P. 85–94. ISSN: 2084-5715.
149. Chyrun L., Andrunyk V., Vysotska V. Content analysis peculiarities of user internet activities for personality psychological state slice formation. MEST Journal. 2017. Vol. 6, № 2. P. 26–46. ISSN: 2334-7171.
150. Lytvyn V., Vysotska V., Veres O. Ontology of big data analytics. MEST Journal. 2018. Vol. 6(1). P. 41–60.
151. Lytvyn V., Vysotska V., Bublyk M., Nanivskyi R., Grudowski P., Matseliukh Y. Developing methods for building intelligent systems of information resources processing using an ontological approach. AISC. 2021. Vol. 1293. P. 345–370. ISSN 2194-5357.
152. Bisikalo O., V Vysotska., Lytvyn V., Brodyak O., Vyshemyrska S., Rozov Y. Experimental investigation of significant keywords search in Ukrainian content. AISC. 2021. Vol. 1293. P. 3–29.
153. Burov Y., Horodetska A., Bublyk M., Nashkerska M., Vysotska V. Intellectual Tourist Service with the Situation Context Processing. Advances in Social Science, Education and Humanities Research. 2021. Vol. 557. P. 233-243. ISSN (Online): 2352-5398.

Статті у міжнародних конференціях, які індексуються у Scopus та Web of Science

154. Alieksieieva K., Berko A., Vysotska V. Technology of commercial web-resource processing. CADSM : XIII Міжнар. наук.-техн. конф., 24–27 лют., Львів, Поляна, 2015. С. 340–344.
155. Andrunyk V., Chyrun L., Vysotska L. Electronic content commerce system development. CADSM : матер. XIII Міжнар. наук.-техн. конф., 24–27 лют., Львів, Поляна, 2015. С. 434–438.
156. Vysotska V. Chyrun L. Methods of information resources processing in electronic content commerce systems. CADSM: XIII Міжн. наук.-техн. конф., 24–27 лют., Львів, Поляна, 2015. С. 328–332.
157. Vysotska V., Chyrun L. Analysis features of information resources processing. CSIT : proc. of the Xth Intern. conf., 14–17 Sept., Lviv, Ukraine, 2015. P. 124–128.
158. Lytvyn V., Vysotska V. Designing architecture of electronic content commerce system. CSIT : proc. of the Xth Intern. conf., 14–17 Sept., Lviv, Ukraine, 2015. P. 115–119.
159. Vysotska V., Hasko R., Kuchkovskiy V. Process analysis in electronic content commerce system. CSIT: proc. of the X Intern. conf., 14–17 Sept., Lviv, Ukraine. 2015. P. 120–123.
160. Vysotska V. Linguistic analysis of textual commercial content for information resources processing. TCSET : proc. of the XIII Intern. conf, Feb. 23–26, Lviv, Slavske, Ukraine, 2016. P. 709–713.
161. Lytvyn V., Vysotska V., Veres O., Rishnyak I., Rishnyak H. Content linguistic analysis methods for textual documents classification. CSIT: proc. of the XIth Intern. conf., 6–10 Sept., Lviv. P. 190–192.

162. Vysotska V., Chyrun L., Chyrun L. The commercial content digest formation and distributional process. CSIT: proc. of the XIth Intern. conf. CSIT, 6–10 Sept., Lviv, Ukraine. 2016. P. 186–189.
163. Vysotska V., Chyrun L., Chyrun L. Information technology of processing information resources in electronic content commerce systems. Computer science and information technologies: proc. of the XIth Intern. conf., 6–10 Sept., Lviv, Ukraine. 2016. P. 212–222.
164. Shakhovska N., Vysotska V., Chyrun L. Features of e-learning realization using virtual research laboratory. CSIT: proc. of the XIth Intern. conf., 6–10 Sept., Ukraine. 2016. P. 143–148.
165. Lytvyn V., Vysotska V., Chyrun L., Chyrun L. Distance learning method for modern youth promotion and involvement in independent scientific researches. 1st IEEE International conference on data stream mining and processing, DSMP : proc. Aug. 23–27, Lviv, Ukraine. 2016. P. 269–274.
166. Lytvyn V., Vysotska V., Veres O., Rishnyak I., Rishnyak H. The risk management modelling in multi project environment. CSIT: proc. of the Intern. conf. 5–8 Sept., Lviv, Ukraine. 2017. P. 32–35.
167. Korobchinsky M., Vysotska V., Chyrun L., Chyrun L. Peculiarities of content forming and analysis in internet newspaper covering music news. Computer science and information technologies : proc. of the XIIth Intern. conf., 5–8 Sept., Lviv, Ukraine. 2017. P. 52–57.
168. Naum O., Chyrun L., Kanishcheva O., Vysotska V. Intellectual system design for content formation. CSIT: proc. of the XII Intern. conf., 5–8 Sept., Ukraine. 2016. P. 131–138.
169. Lytvyn V., Vysotska V., Dosyn D., Holoschuk R., Rybchak Z. Application of sentence parsing for determining keywords in Ukrainian texts. Computer science and information technologies : proc. of the XIIth Intern. conf., 5–8 Sept., Lviv, Ukraine. 2017. P. 326–331.
170. Su J., Vysotska V., Sachenko A., Lytvyn V., Burov Ye. Information resources processing using linguistic analysis of textual content. IDAACS : proc. conf., Bucharest, Sept. 21–23. 2017. P. 573–578.
171. Lytvyn V., Vysotska V., Burov Y., Demchuk A. Defining author's style for plagiarism detection in academic environment. DSMP : proc. of Intern. conf., Aug. 21–25, Lviv, Ukraine. 2018. P. 128–133.
172. Lytvyn V., Vysotska V., Lozynska O., Oborska O., Dosyn D. Methods of building intelligent decision support systems based on adaptive ontology. Data stream mining and processing : proc. of Intern. conf., Aug. 21–25, Lviv, Ukraine. 2018. P. 145–150.
173. Chyrun L., Vysotska V., Kis I., Chyrun L. Content analysis method for cut formation of human psychological stat. DSMP: proc. of Intern. conf., August 21–25, Lviv, Ukraine. 2018. P. 139–144.
174. Lytvyn V., Vysotska V., Burov Y., Bobyk I., Ohirko O. The linguometric approach for co-authoring author's style definition. IEEE IDAACS-SWS : proc., Lviv, 20–21 September 2018. P. 29–34.
175. Vysotska V., Lytvyn V., Hrendus M., Brodyak O., Kubinska S. Method of textual information authorship analysis based on stylometry. CSIT: proc. of Intern. conf., 11–14 вер., Львів. 2018. С. 9–16.
176. Vysotska V., Kanishcheva O., Hlavcheva Y. Authorship identification of the scientific text in Ukrainian with using the lingvometry methods. CSIT: proc. of the Intern. conf., 11 вересня 2018 р., Львів. 2018. С. 34–38.
177. Chyrun L., Kis I., Vysotska V., Chyrun L. Content monitoring method for cut formation of person psychological state in social scoring. CSIT: proc. of the Intern. conf., Львів, 11–14 вер. 2018 р., С. 106–112.

178. Su J., Lytvyn V., Vysotska V., Sachenko A., Dosyn D. Model of touristic information resources integration according to user needs. CSIT: proc. of Intern. conf., Львів, 11–14 вер., С. 113–116.
179. Rusyn B., Vysotska V., Pohreliuk L. Model and architecture for virtual library information system. CSIT: proc. of Intern. conf., Львів, 11–14 вер. 2018 р., С. 34–41.
180. Lytvyn V., Peleshchak I., Peleshchak R., Vysotska V. Satellite spectral information recognition based on the synthesis of modified dynamic neural networks and holographic data processing techniques. CSIT: proc. of Intern. conf., Львів,11–14 вер. 2018. Т. 1. С. 330–334.
181. Lytvyn V., Vysotska V., Burov Y., Demchuk A. Architectural ontology designed for intellectual analysis of e-tourism resources. CSIT: proc. of Intern. conf., Львів,11–14 вер. 2018. С. 335–338.
182. Gozhyj A., Kalinina I., Vysotska V., Gozhyj V. The method of web-resources management under conditions of uncertainty based on fuzzy logic. CSIT: proc. of the Intern. conf., Львів,11–14 вер. 2018. P. 343–346.
183. Lytvyn V., Kuchkovskiy V., Vysotska V., Markiv O., Pabyrivskyy V. Architecture of system for content integration and formation based on cryptographic consumer needs. Computer Sciences and Information Technologies (CSIT) : proc. of the IEEE Intern. conf., Львів,11–14 вер. 2018. С. 391–395.
184. Lytvyn V., Peleshchak I., Peleshchak R., Vysotska V. Information encryption based on the synthesis of a neural network and AES algorithm. Advanced information and communication technologies, AICT : proc. of the 3rd Intern. conf., Lviv, Ukraine, July 2–6 2019. P. 447–450.
185. Lytvyn V., Vysotska V., Mykhailyshyn V., Peleshchak I., Peleshchak R., Kohut I. Intelligent system of a smart house. AICT: proc. of the 3rd Inter. conf. (Lviv, Ukraine, July 2–6 2019. P. 282–287.
186. Vysotsky A., V Vysotska., Lytvyn V., Burov Y., Demchuk A., I Lyudkevych. Consolidated information web resource for online tourism based on data integration and geolocation. Computer Sciences and Information Technologies (CSIT): proc. of the IEEE Intern. conf., Львів, 17–20 вересня 2019. С. 15–20.
187. Lytvyn V., Vysotska V., Peleshchak I., Basyuk T., Kovalchuk V., Kubinska S., Chyrun L., Rusyn B., Pohreliuk L., Salo T. Identifying textual content based on thematic analysis of similar texts in big data . CSIT: proc. of Intern. conf., Львів, 17–20 вер. 2019. Т. 2. С. 84–91.
188. Vysotsky A., Lytvyn V., Vysotska V., Dosyn D., Lyudkevych I., Antonyuk N., Naum O., Vysotskyi A., Chyrun L., Slyusarchuk O. Online tourism system for proposals formation to user based on data integration from various sources. Computer Sciences and Information Technologies (CSIT) : proc. of the IEEE Intern. conf., Львів, 17–20 вер. 2019. С. 92–97.
189. Vysotska V., Lytvyn V., Kovalchuk V., Kubinska S., Dilai M., Rusyn B., Pohreliuk L., Chyrun L., Chyrun S., Brodyak O. Method of similar textual content selection based on thematic information retrieval. CSIT: proc. of Intern. conf., Львів, 17–20 вересня. 2019. Т. 3. С. 1–6.
190. Rzheuskyi A., Kutyuk O., Vysotska V., Burov Y., Lytvyn V., Chyrun L. The architecture of distant competencies analyzing system for IT recruitment. Computer Sciences and Information Technologies (CSIT): proc. of the IEEE Intern. conf., Львів, 17–20 вересня. 2019. Т. 3. С. 254–261.
191. Gozhyj A., Kalinina I., Gozhyj V., Vysotska V. Web service interaction modeling with colored petri nets. 10th IEEE IDAACS : proc., September 18–21, Metz, France. 2019. P. 319–323.

192. Shu C., Dosyn D., Lytvyn V., Vysotska V., Sachenko A., Jun S. Building of the predicate recognition system for the NLP ontology learning module. IDAACS : proc., September 18–21, Metz, France. 2019. P. 802–808.
193. Kalinina I., Vysotska V., Bidyuk P., Gozhyj A., Vasilev M., Malets R. Forecasting nonlinear nonstationary processes in machine learning task. DSMP: proc. of the 3rd inter. conf., Lviv, Ukraine. 2020. P. 28–32.
194. Lytvyn V., Vysotska V., Burov Y., Hryhorovych V. Knowledge novelty assessment during the automatic development of ontologies. DSMP: proc. of the Intern. conf., Lviv, Ukraine. 2020. P. 372–377.
195. Lytvyn V., Dosyn D., Vysotska V., Hryhorovych A. Method of ontology use in OODA. DSMP: proc. of the IEEE 3rd Intern. conf., Lviv, Ukraine. 2020. P. 409–413.
196. Vysotska V., Lytvyn V., Bublyk M., Demchuk A., Demkiv L., Shpak Y. Method of ontology quality assessment for knowledge base in intellectual systems based on ISO\IEC 25012. CSIT: proc. of Intern. conf., Збараж, 23–26 вересня, 2020. P. 109–113.
197. Vysotska V., Berko A., Bublyk M., Chyrun L., Vysotsky A., Doroshkevych K. Methods and tools for web resources processing in e-commercial content systems. Computer Sciences and Information Technologies (CSIT): proc. of the IEEE Intern. conf., Збараж, 23–26 вересня, 2020. P. 114–118.
198. Lytvyn V., Dosyn D., Vysotska V., Demchuk A., Demkiv L., Lytvyn I. Intellectual agent construction method based on the subject field ontology. Computer Sciences and Information Technologies (CSIT): proc. of the IEEE Intern. conf., Збараж, 23–26 вересня, 2020. P. 40-46.
199. Lytvyn V., Vysotska V., Burov Y., Brodyak O. Approach to automatic construction of interpretation functions during ontology learling. CSIT: proc. of Int. conf., Збараж, 23–26 вер., 2020. P. 267–271.
200. Burov Y., Lytvyn V., Vysotska V., I Shakleina. The basic ontology development process automation based on text resources analysis. CSIT: proc. of Int. conf., Збараж, 23–26 вер, 2020. P. 280–284.
201. Lytvyn V., Vovnyanka R., Oborska O., Dosyn D., Vysotska V., Panasyuk V. Intelligent agent behavior simulation based on reinforcement learning. Computer Sciences and Information Technologies (CSIT): proc. of the IEEE Intern. conf., Збараж, 23–26 вересня, 2020. P. 285–290.
202. Peleshchak R., Lytvyn V., Peleshchak I., Vysotska V. Stochastic Pseudo-Spin Neural Network with Tridiagonal Synaptic Connections. SIST, 28-30 April 2021, Nur-Sultan, Kazakhstan. Art. 9465998.
203. Lytvyn V., Bublyk M., Vysotska V., Panasyuk V., Brodyak O., Luchkevych M. Modelling of the Intelligent Agent's Behavior Scheduler Based on Petri Nets and Ontological Approach. SIST, 28-30 Apr. 2021, Nur-Sultan, Kazakhstan. Art. 9465994.
204. Lytvyn V., Y Burov., Vysotska V., Pukach Y., Tereshchuk O., Shakleina I. Abstracting Text Content Based on Weighing the TF-IDF Measure by the Subject Area Ontology. SIST, 28-30 April 2021, Nur-Sultan, Kazakhstan. Art. 9465978.
205. Tchynetskyi S., Peleshchak R., Peleshchak I., Vysotska V. A Neural Network Development for Multispectral Images Recognition. CSIT : proc. of the Intern. conf., 22-25 Sept., Lviv, Ukraine. 2021. P. 278–284.
206. Dmytriv A., Vysotska V., Bublyk M. The Speech Parts Identification for Ukrainian Words Based on VESUM and Horokh Using. Computer Sciences and Information Technologies (CSIT) : proc. of the IEEE 16th Intern. conf., 22-25 Sept., Lviv, Ukraine. 2021. Vol. 2. P. 21–33.

207. Kubinska S., Vysotska V., Matseliukh Y. User Mood Recognition and Further Dialog Support. CSIT: proc. of Intern. conf., 22-25 Sept., Lviv, Ukraine. 2021. Vol. 2. P. 34–39.
208. Ivanchyshyn D., Vysotska V., Albota S. The Film Script Generation Analysis Based on the Fiction Book Text Using Machine Learning. Computer Sciences and Information Technologies (CSIT) : proc. of the IEEE 16th Intern. conf., 22-25 Sept., Lviv, Ukraine. 2021. Vol. 2. P. 68–80.
209. Sartiukova A., Peleshchak R., Peleshchak I., Vysotska V. The Multiclass Classification of Objects Based on Multispectral Images Recognition. Computer Sciences and Information Technologies (CSIT): proc. of the IEEE 16th Intern. conf., 22-25 Sept., Lviv, Ukraine. 2021.Vol. 1. P. 52–60.
210. Voloshynskyi O., Vysotska V., Bublyk M. Cardiovascular Disease Prediction Based on Machine Learning Technology. CSIT: proc. of Intern. conf., 22-25 Sept., Lviv, Ukraine. 2021. Vol. 1. P. 69–75.
211. Aksonov D., Gozhyj A., Kalinina I., Vysotska V. Question-Answering Systems Development Based on Big Data Analysis. CSIT: proc. of Intern. conf., 22-25 Sept., Lviv, Ukraine. 2021. P. 113–118.
212. Mykytiuk A., Vysotska V., Albota S. Spam Filtration System with the Use of Machine Learning Technology. CSIT: proc. of Intern. conf., 22-25 Sept., Lviv, Ukraine. 2021. Vol. 1. P. 124–130.
213. Zanchak M., Vysotska V., Albota S. The Sarcasm Detection in News Headlines Based on Machine Learning Technology. CSIT: proc. of Intern. conf., 22-25 Sept., Lviv, Ukraine. 2021. P. 131–137.
214. Voloshyn S., Peleshchak R., Peleshchak I., Vysotska V. Big Data Analysis for Multispectral Images Recognition Based on Deep Learning. Computer Sciences and Information Technologies (CSIT): proc. of the IEEE 16th Intern. conf., 22-25 Sept., Lviv, Ukraine. 2021. Vol. 1. P. 160–170.
215. Lytvyn V., Vysotska V., Bublyk M., Gozhyj A., Schuchmann V. Solving Scheduling Issues Methods Analysis in Computational Grid. CSIT, 22-25 Sept., Lviv, Ukraine. 2021. Vol. 1. P. 267–273.
216. Kravets P., Lytvyn V., Burov Y., Vysotska V., Chyrun L., Panasyuk V. Making Optimal Decisions with Learning Method Based on Fuzzy Logic. Advanced Information and Communication Technologies (AICT): proc. of the IEEE 4th Intern. conf., 21-25 Sept., Lviv, Ukraine. 2021. P. 183–188.
217. Gozhyj A., Kalinina I., Nechakhin V., Gozhyj V., Vysotska V. Modeling an Intelligent Solar Power Plant Control System Using Colored Petri Nets. IDAACS, 22-25 Sept., Cracow, Poland. 2021. P. 626–631.
218. Peleshchak R., Lytvyn V., Kholodna N., Peleshchak I., Vysotska V. Two-Stage AES Encryption Method Based on Stochastic Error of a Neural Network. TCSET, Lviv-Slavske, Ukraine, Feb. 22 - 26, 2022.

Статті та тези доповідей у збірниках праць конференцій

219. Козлов П., Висоцька В. Особливості технології управління web-ресурсом. V Міжн. наук.-практ. конф. «Обробка сигналів і негаусівських процесів», 20-22 травня, 2015, Черкаси. С. 38–40.
220. Козлов П., Висоцька В. Аналіз процесу управління комерційним контентом. Міжн. наук. конф. ISDMCI, Залізний Порт, Україна, 25–28 трав. 2015. С. 36–38.
221. Висоцька В. А., Чирун Л. Б., Чирун Л. В. Контент-моніторинг текстової інформації Web-ресурсів. Міжнар. наук. конф. ISDMCI, Залізний Порт, Україна, 25–28 трав. 2015. С. 36–38.
222. Козлов П., Висоцька В. Технологія управління комерційними контентом в системах електронного бізнесу. Міжнар. наук. конф. ІКС, 20–23 трав. 2015, Львів, Славське. С. 48–49.

223. Кондратєв Є., Висоцька В Контент-аналіз текстових масивів даних. 4 Міжн. наукова конференція ІКС, 20–23 трав. 2015, Україна, Львів, Славське. С. 170–171.
224. Литвин В. В. Висоцька В.А., Оливко Р. М. Метод визначення семантичної метрики на основі тезаурусу предметної області. ІСПЛ, Харків, 14 квіт. 2016 р. С. 10–12.
225. Chyrun L., Vysotska V., Lytvyn V. Specifics informational resources processing for textual content linguistic analysis. Proceeding of MEMSTECH 2016, 20–24 Apr., 2016, Polyana, 2016. P. 214–219.
226. Литвин В. В., Висоцька В. А., Оливко Р. М., Черна Т. М. Особливості рубрикації текстових документів з використанням онтології. ISDMIT, Україна, 25–28 трав. 2016. С. 292–295.
227. Висоцька В. А. Чирун Л. Б., Чирун Л. В. Аналіз процесу супроводу текстового комерційного контенту. Міжнар. наук. конф. ISDMIT, Залізний Порт, Україна, 25–28 трав. 2016. С. 42–44.
228. А Берко. Ю., Висоцька В. А., Чирун Л. В., Чирун Л. Б. Аналітичний метод супроводу текстового контенту інформаційних ресурсів. Математика. Інформаційні технології. Освіта. Східноєвропейський НУ ім. Лесі Українки. Луцьк, 2016. С. 11–20.
229. Висоцька В. А., Козлов П. Ю. Управління Web-ресурсом.. Математика. Інформаційні технології. Освіта. V Міжн. наук.-практ. конф., 5–7 черв. 2016 р., Луцьк. С. 62–63.
230. Берко А. Ю., Висоцька В. А., Чирун Л. В., Чирун Л. Б. Особливості формування критеріїв оцінювання знань студентів згідно їх компетентності у ІТ-сфері. Математика. Інформаційні технології. Освіта. V Міжн. наук.-практ. конф., 5–7 черв. 2016 р., Луцьк. С. 117–118.
231. Каніщева О., Главчева Ю., Висоцька В. Визначення стилю автора для виявлення плагіату в академічному середовищі. SAIT 2017, May 22–25, 2017, Kyiv. P. 78–79.
232. Lytvyn V., Vysotska V., Chyrun L., Smolarz A., Naum O. Intelligent system structure for web resources processing and analysis. 1st Intern. conf., COLINS, 21 Apr. 2017, Kharkiv. P. 56–74.
233. Lytvyn V., Vysotska V., Wojcik W., Dosyn D. A method of construction of automated basic ontology. 1st Intern. conf., COLINS, 21 Apr. 2017, Kharkiv. P. 75–83.
234. Висоцька В. А. Методика аналізу компетентностей для рекрутингу. International scientific and practical conf. on Scientific Research Priorities., 22–23 June 2017, Nowy Sanz, Poland. P. 60–62.
235. Литвин В. В., Наум О. М., Висоцька В. А. Метод інтеграції та управління контентом мережі інформаційних ресурсів туризму згідно потреб користувача. Міжнар. наук. конф. ISDMCI, 22–26 трав. 2017, Залізний Порт. С. 78–80.
236. Висоцька В. А., Чирун Л. Б., Чирун Л. В. Інтернет-портал аматорських спортивних ігор. Міжнар. наук. конф. ISDMCI, 22–26 трав. 2017 Залізний Порт. С. 45–47.
237. Литвин В. В., Оборська О. В., Висоцька В. А., Бобик І. О. Метод аналізу авторства тексту на основі стилеметрії. Міжнар. наук. конф. ISDMCI, 21–27 трав. 2018 р., Залізний Порт. С. 240–243.
238. Чирун Л. Б., Чирун Л. В., Висоцька В. А. Метод визначення авторства текстового україномовного контенту. ISDMCI, 21–27 трав. 2018 р., Залізний Порт, Україна. С. 287–289.
239. Русин Б. П., Висоцька В. А., Погрелюк Л. В. Модель інформаційної системи Virtual Library. Міжнар. наук. конф. ISDMCI, 21 трав. 2018 р., Залізний Порт, Україна. С. 100–102.

240. Kovalchuk V., Lytvyn V., Vysotska V., Hrendus M., Naum O. The information system for identification of content set based on analysis of similar texts. Computational Linguistics and Intelligent Systems. Proceedings. Vol. 2: Proc. of the 2nd Intern. Conf. COLINS 2018. P. 122–127. ISSN 2523-4013.
241. Lytvyn V. Vysotska V., Chyrun L., Hrendus M., Naum O. Content analysis of text-based information in E-commerce systems. COLINS. Vol. 2: Proc. of the 2nd Intern. Conf. 2018. P. 81–94.
242. Rusyn B., Vysotska V., Pohreliuk L. Methods of information resources processing in virtual library. COLINS. Vol. 2 : Proc. of the 2nd Int. Conf., 2018. P. 28–39. ISSN 2523-4013.
243. Lytvyn V., Oborska O., Vysotska V., Dosyn D., Demchuk A. Ontology using for decision making in a competitive environment . COLINS. Vol. 2 : Proc. of the 2nd Intern. Conf. 2018. P. 17–27.
244. Chyrun L., Vysotska V., Chyrun L., Gozhyj A., Kalinina I. SEO technology for web resource processing. COLINS. Proceedings. Vol. 2 : Proc. of the 2nd Intern. Conf., 2018. P. 40–52.
245. Досин Д. Г., Висоцька В. А., Литвин В. В. Побудова системи підтримки прийняття рішень на базі адаптивної онтології. Обчислювальні методи і системи перетворення інформації: зб. пр. V-ї наук.-техн. конф., Львів, 4–5 жовт. 2018. С. 135–138.
246. Висоцька В. А., Литвин В. В., Олешек О. І. Автоматизований моніторинг змін у Web-ресурсах. Міжнар. наук. конф. ISDMCI, Залізний Порт, 21–25 трав. 2019. С. 30–32.
247. Литвин В. В., Висоцька В. А., Михайлишин В. Ю., Сем'янчук С. О. Розроблення інформаційної системи аналізу даних відеопотоку. ISDMCI, Україна, 21–25 трав. 2019. С. 94–97.
248. Демчук А. Б., Литвин В. В., Висоцька В. А. Технологія персоналізованого поширення комерційного контенту через Web-ресурс Е-комерції. ISDMCI, Україна, 21 трав. 2019. С. 49–51.
249. Lytvyn V., Oborska O., Vysotska V., Dosyn D., Demchuk A., Burov Y., Kravets P., Oleksiv N. Problems of ontology structure and meaning optimization and theirs solution methods. COLINS. Proceedings of the 4th Intern. Conf., Lviv, Ukraine; June 23-24, 2020. Vol. II. P. 21–40.
250. Kutyuk O., Lytvyn V., Oborska O., Vysotska V., Dosyn D., Demchuk A., Burov Y., Kravets P. Intelligent system development of distant matrix analysis for recruitment in the IT sector. COLINS. Proceedings of the 4th Intern. Conf., Lviv, Ukraine; June 23-24, 2020. Vol. II. P. 41–78.
251. Tymoshenko K., Vysotska V. Algorithm of text recognizing in Ukrainian on the video mode. COLINS. Proceedings of the 4th Intern. Conf., Lviv, Ukraine; June 23-24, 2020. Vol. II. P. 81–89.
252. Висоцька В. Суб'єктивізм трактування академічної доброчесності в межах наукової діяльності видавництв. Академічна доброчесність: виклики сучасності. Варшава, 06.11.2020. С. 31-35.
253. Bublyk M., Zahreva Y., Vysotska V., Matseliukh Y., Chyrun L., Korolenko O. Information system development for recording offenses in smart city based on cloud technologies and social networks. Webology. 2022. Vol. 19, No. 2. P. 1870–1898.
254. Bublyk M., Kalynii T., Varava L., Vysotska V., Chyrun L., Matseliukh Y. Decision support system design for low voice emergency medical calls at smart city based on chatbot management in social networks. Webology. 2022. Vol. 19, No. 2. P. 2135–2178.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

БД – база даних;

БЗ – база знань;

ГА – графемний аналіз;

ЗМІ – засоби масової інформації;

ІАТПК – інтелектуальний аналіз текстових потоків контенту;

ІІП – інтелектуальний інформаційний пошук;

ІІПМ – інтелектуальна інформаційно-пошукова мова;

ІППС – інтелектуальна інформаційно-пошукова система;

ІІПТ – інтелектуальний інформаційно-пошуковий тезаурус;

ІС – інформаційна система (англ. Information Technology);

ІТ – інформаційна технологія (англ. Information Technology);

КЛ – когнітивна лінгвістика;

КЛС – комп’ютерна лінгвістична система

ЛА – лексичний аналіз;

ЛС – лінгвістична система;

ЛР – лінгвістичний ресурс;

МА – морфологічний аналіз;

МН – машинне навчання;

ОІР – опрацювання інформаційних ресурсів;

ПА – прагматичний аналіз;

ПО – предметна область;

ПОК – пошуковий образ контенту;

ПСУМ – правила синтаксису української мови;

СА – синтаксичний аналіз;

СЕА – семантичний аналіз;

СД – сховище даних;

СППР – система підтримки прийняття рішень;

СШІ – система штучного інтелекту;

ТЕ – тематичний елемент;

ШІ – штучний інтелект;

ФА – фонологічний аналіз;

AI – штучний інтелект (англ. Artificial Intelligence);

AL – прикладна лінгвістика (англ. Applied Linguistics);

CL комп'ютерна лінгвістика (англ. Computational Linguistics);

CPC – плату за клік (англ. Cost Per Click);

QA-система – питально-відповідна система (Question-Answering System);

IVR – інтерактивне голосове меню (англ. Interactive Voice Response) для маршрутизації дзвінків всередині call-центру або через тональний набір;

k-NN – метод k найближчих сусідів (англ. k-nearest neighbors algorithm);

KPI – ключовий показник ефективності (англ. Key Performance Indicators):

ML – машинне навчання (англ. Machine Learning);

Naive Bayes classifiers – наївний Баєсів класифікатор;

NLP – опрацювання природної мови (англ. Natural-Language Processing);

NN – нейрона мережа;

OCR – оптичне розпізнавання тексту (англ. optical character recognition);

POS – частина мови (англ. Part-of-speech);

POST – розмічування частин мови (англ. Part-of-speech tagging);

SEO – пошукова оптимізація сайту (англ. Search Engine Optimization):

SEM – пошукове просування (англ. Search Engine Marketing);

SSI – інтерфейс безмовного доступу (англ. Silent speech interfaces) через мовленнєві сигнали на ранній стадії артикуляції;

SVM – метод опорних векторів (англ. Support vector machines);

RIA – насичений Web-додаток (англ. Rich Internet application, installable Internet application, або Rich web application) з підтримкою інтерактивних функцій користувацького дружнього інтерфейсу незалежно від браузера користувача;

TCLC – життєвий цикл контенту (англ. Text Content Life Cycle);

TF-IDF – статистичний показник, що використовується для оцінки важливості слів у контексті (англ. TF — term frequency, IDF — inverse document frequency);

URL – уніфікована адреса ресурсу (англ. Uniform Resource Locator).

ВСТУП

Актуальність теми. На сьогодення активний розвиток інформаційних технологій (ІТ) знаходиться на перетині глобалізації та інформатизації. Швидкі темпи зростання інформатизації суспільства напряму пов’язані з темпами розвитку та впровадженням комп’ютерних лінгвістичних систем (КЛС), розроблення яких базується на моделях та методах опрацювання природної мови (ОПМ). Складність розроблення моделей, методів та засобів ОПМ полягає не лише в розв’язку не типових задач ОПМ, але й в адаптації цих моделей, методів та засобів для конкретної природної мови. Кожна природня мова є унікальною, зі своїм колоритом правил, історії, граматики, виключень та особливостей генерування лінгвістичних одиниць для передачі сенсу, що ускладнює процес розроблення КЛС.

Зазвичай кожний успішний проект розроблення КЛС призначений під конкретну задачу (наприклад, машинний переклад, ідентифікація плагіату/рерайту, рубрикація тексту, аналіз атрибуції тексту, інформаційний пошук, реферування, голосові помічники, інтелектуальні чат-боти тощо) та одночасно є одноразовим та закритим (наприклад, Amazon Alexa, Google Assistant, Facebook, Voice Mate, Bixby, Siri, Abby Lingvo, Microsoft Cortana, Microsoft Word, Grammarly, Google Translation, PROMT, CuneiForm, Trados, OmegaT, Wordfast, Dragon, IBM via voice, Speereo, Finereader, Tesseract, OCRopus тощо) без можливості ознайомитися з вмістом бажаючим ІТ-фахівцям/спеціалістам. Рідкісні випадки, коли до таких проектів КЛС розробники надають відкритий доступ та можливість ознайомитися з їх структурою та змістом. Створення будь-якого прикладного додатку ОПМ для довільної природної мови із понад 7000 мов та діалектів базується на дослідженні великих текстових одномовних/паралельних корпусів цієї мови, який містить понад сотень мільйонів слів та лінгвістичних ресурсів. Лише близько для 20 природних мов (англійська, китайська, західноєвропейські мови, японська тощо) відомі результати досліджень таких корпусів, що дає змогу для цих мов розробляти КЛС різної складності. Нажаль в сучасних реаліях українська мова

вважається в міжнародному науковому суспільстві екзотичною мовою з низьким показником ресурсності, тобто не має достатньо навчальних, дослідницьких та опрацьованих даних для розроблення сучасних прикладних додатків ОПМ. Такі прикладні додатки використовуються для побудови КЛС в кібербезпеці (виявлення фейків та пропаганди, так званих тролів/ботів в соціальних мережах), соціології (аналіз динаміки зміни громадської думки на тематичні питання), філології (автоматичне дослідження великих масивів даних різного тематичного спрямування та різних часових періодів), психології (аналіз психологічного портрету особи, ідентифікація посттравматичного стресового розладу учасників бойових дій або окупації), національній безпеці (інформаційна війна), юриспруденції (криміналістика та судова справа), соціальних комунікаціях (аналіз дописів спільнот в соціальних мережах) та в інших важливих галузях сучасної України. Означене обумовлює актуальність теми дисертаційного дослідження.

Наукові дослідження N. Chomsky, В.М. Глушкова, А.В. Гладкого, Д.В. Ланде, В.А. Широкова, Н.В. Шаронової, Н.Ф. Хайрової, О.П. Левченко, О.В. Бісікала, С.Н. Бук, Н.П. Дарчук, З.В. Партика, А.В. Анісімова, Ю.Д. Апресяна, О.О. Марченка, І.М. Кульчицького, А.О. Никоненка, М. Гросса, А. Лантена, V.H. Yngve, S. Sharoff, Ю.А. Шрейдера, D. Jurafsky, B. Bengfort, J.H. Martin, L. Tesniere, T. Ojeda, P.M. Postal, D.G. Hays, T.A. van Dijk, S. Marcus, J. Lyons, L.W. Tosh, Y. Bar-Hillel, D.G. Bobrow, G. Lakoff, R. Bilbro, N. Kotsyba, А.Ю. Берка, Ю.М. Щербини, В.Ю. Величка, В.Ф. Старка та багатьох інших дають змогу зрозуміти основні принципи лінгвістичного опрацювання тексту в залежності від особливостей конкретної природньої мови. Більше 80% таких досліджень стосуються опрацювання англомовних текстів. Суттєво менше досліджень для слов'янських мов, зокрема, для низькоресурсної української мови. Зокрема відсутні публікації щодо рекомендацій розроблення, функціональних вимог, загальної структури та типової архітектури КЛС опрацювання україномовного текстового контенту. Напряму застосувати моделі, методи, алгоритми та ІТ опрацювання англійської мови для україномовного

текстового контенту не приводить до позитивних результатів. Вже на рівні морфологічного аналізу виникає суттєвий конфлікт між розробленими методами для англомовного тексту та їх використанням для україномовного тексту. Наприклад, для простого алгоритму Портера (стемінг) без відповідної модифікації не коректним є відокремлення основи слова від флексії, що призводить до неточності ідентифікації ключових слів, що, в свою чергу, впливає на розв’язок будь-якої задачі ОПМ, де необхідно швидко ідентифікувати множину ключових слів (рубрикація, пошук, анотування тощо). Визначення основних особливостей та процесів лінгвістичного аналізу українськомовних текстів значно полегшить етапи опрацювання текстового потоку інформації як інтеграція, супровід та управління контентом. В свою чергу, адаптація процесів інтелектуального аналізу текстового контенту з ідентифікацією функціональних вимог до відповідних модулів КЛС призведе до можливості розробити її типову архітектуру на принципі модульності (додавання компонентів в залежності від змісту задачі ОПМ та призначення КЛС).

Наведене свідчить про актуальність досліджень під час вирішення важливої науково-прикладної проблеми аналізу та синтезу КЛС для розв’язання різних задач опрацювання україномовного текстового контенту, що дасть змогу підвищити рівень ресурсності природної української мови на основі розроблення нових та удосконаленні відомих моделей, методів та засобів ОПМ.

Зв'язок роботи з науковими програмами, планами, темами. Тема дисертації відповідає науковому напряму «Дослідження, розроблення і впровадження інтелектуальних розподілених інформаційних технологій та систем на основі ресурсів баз даних, сховищ даних, пристроїв даних та знань з метою прискореного формування інформаційного суспільства» кафедри інформаційних систем та мереж Національного університету «Львівська політехніка». Дисертація виконана в межах науково-дослідної роботи цієї кафедри «Розроблення інтелектуальних розподілених систем на основі онтологічного підходу з метою інтеграції інформаційних ресурсів» (№ 0115U004228, терміни: 05.2015–12.2017 рр.), держбюджетної науково-дослідної

роботи «Методи та засоби функціонування систем підтримки прийняття рішень на основі онтологій» (№ 0118U000269, терміни: 01.2018–12.2019 рр.), а також держбюджетної науково-дослідної роботи «Система підтримки прийняття рішень розпізнавання мультиспектральних образів на основі технологій машинного навчання та онтологічного підходу» (№ 0120U102203, терміни: 04.2020–12.2021 рр.).

Мета і завдання дослідження. Метою роботи є розроблення моделей, методів, засобів аналізу та синтезу комп'ютерних лінгвістичних систем на базі нових та удосконалення відомих методів опрацювання україномовного текстового контенту для розв'язання задач опрацювання природньої мови. Метою дисертаційної роботи визначено необхідність виконання таких завдань:

1) провести аналіз специфіки побудови КЛС шляхом систематизації процесів їх реалізації та функціонування, що забезпечить можливість виділити клас систем, функціональні властивості яких дозволяють виконувати кількісне оцінювання очікуваних ефектів впровадження типової КЛС опрацювання україномовного текстового контенту для розв'язку різних задач ОПМ;
2) розробити інформаційну технологію побудови КЛС опрацювання україномовного тексту, що дасть змогу визначити їх базову структуру, функціональні вимоги, послідовність налаштування та навчання системи, загальні засади проектування;
3) запропонувати ІТ опрацювання інформаційних ресурсів як інтеграція, управління та супровід українськомовного контенту на основі вдосконалення лінгвістичного аналізу текстового контенту для розроблення метрик оцінювання ефективності функціонування КЛС для розв'язку різних задач ОПМ;
4) розробити методи опрацювання україномовного текстового контенту для розв'язку різних задач ОПМ для підвищення точності отриманих результатів;
5) розробити методи та засоби інтелектуального аналізу текстового контенту для підвищення ефективності розв'язку різних задач ОПМ;

6) створити програмні модулі опрацювання україномовного текстового контенту для розв’язку різних задач ОПМ та проведення експериментів;
7) провести апробацію отриманих результатів шляхом побудови та впровадження прикладних КЛС опрацювання україномовного текстового контенту.

Об’єктом дослідження є процеси аналізу та синтезу комп’ютерних лінгвістичних систем опрацювання україномовного текстового контенту.

Предметом дослідження є моделі, методи та засоби опрацювання україномовного текстового контенту для розв’язку різних задач ОПМ.

Методи дослідження. Для досягнення поставленої мети використано: теорію формальних граматик та автоматів, теорію множин, теорію моделей даних та знань, теорію ймовірності і математичної статистики, теорію моделей, алгоритмів та логіко-лінгвістичних числень, теорію інформації, теорію графів та методи подання знань для моделювання процесів опрацювання україномовного текстового контенту та розроблення модулів машинного навчання; моделі та методи опрацювання та аналізу текстового контенту для реалізації процесів розв’язку різних задач ОПМ; методи об’єктно-орієнтованого та системного аналізу і проектування – для проектування та розроблення КЛС; теорію реляційних баз даних, методи штучного інтелекту, об’єктно-орієнтоване програмування – для програмної реалізації КЛС опрацювання україномовного текстового контенту для розв’язку різних задач ОПМ.

Наукова новизна одержаних результатів полягає у вирішенні важливої науково-прикладної проблеми аналізу та синтезу КЛС для розв’язання різних задач опрацювання україномовного текстового контенту на основі розроблення нових та удосконалення відомих моделей, методів та засобів ОПМ. Отримано такі нові наукові результати:

вперше

– розроблено метод ідентифікації ключових слів в україномовних текстах на основі графемного та морфологічного аналізу основ слів через регулярні

вирази та N-грами, що дало змогу підвищити точність пошуку ключових слів, здійснити пошук стійких словосполучень та рубрикацію контенту;

– розроблено метод визначення стилю автора тематичного україномовного текстового контенту на основі аналізу ключових слів, стійких словосполучень, N-грам, лінгвометрії та стилеметрії, що дало змогу визначити стилістичний вклад кожного з авторів та підвищити точність атрибуції науково-технічної публікації;
– розроблено метод обчислення ступеня верифікації автора україномовного тексту із множини можливих на основі порівняльного аналізу стилів потенційних авторів, що дало змогу підвищити точність класифікації за подібністю стилю;
– розроблено методи аналізу та синтезу КЛС на основі створення загальної типової структури системи опрацювання текстового контенту українською мовою через підтримку модульності, моделювання взаємодії основних процесів і компонентів, що дало можливість розширити колекцію розв’язків різних типових задач ОПМ шляхом реалізації типового програмного забезпечення таких систем;

одержали подальший розвиток

– методи опрацювання інформаційних ресурсів, такі як інтеграція, управління та супровід контенту, які на відмінну від існуючих адаптовані для опрацювання україномовного тексту та враховують потреби постійної цільової аудиторії на основі аналізу історії діяльності цільової аудиторії на веб-ресурсі КЛС, що дало можливість сформувати множину метрик та показників ефективності функціонування КЛС для розв’язку різних задач ОПМ;
– модель лінгвістичного опрацювання текстового контенту на основі вдосконалення графемного, морфологічного, лексичного та синтаксичного аналізів, які на відмінну до існуючих адаптовані для опрацювання україномовного тексту через регулярні вирази та машинне навчання, дала змогу адаптувати процеси опрацювання україномовного текстового контенту

та підвищити точність отриманих результатів в залежності від конкретної задачі ОПМ;

удосконалено

– методи ОПМ, які на відмінну від існуючих реалізовані на основі розроблених регулярних виразів графемного та морфологічного аналізу україномовного тексту та модифікованого алгоритму стемінгу Портера як ефективного способу ідентифікації афіксів лем для можливості розмічування аналізованого слова, що дало змогу оптимізувати процес та покращити точність сегментації/нормування українського слова/речення;
– методи токенізації та нормалізації тексту, які на від міну від існуючих використовують каскади простих підстановок розроблених регулярних виразів узгодження з шаблонами на основі продукційних правил, скінченних автоматів та онтологічної моделі правил синтаксису української мови, що дало змогу адаптувати алгоритми лексичного та синтаксичного аналізів для опрацювання україномовного контенту;
– модель інтелектуального аналізу текстового потоку, яка на відмінну від існуючої базується на процесах опрацювання інформаційних ресурсів та машинного навчання, що дало змогу адаптувати типові структури модулів інтеграції, управління та супроводу контенту, розробити конвеєр опрацювання україномовного тексту та підвищити ефективність функціонування КЛС в залежності від розв’язку конкретної задачі ОПМ.

Практичне значення одержаних результатів полягає у тому, що їх можна використати для побудови прикладних КЛС опрацювання україномовного текстового контенту. Зокрема, практично цінними є такі результати:

– застосування методу ідентифікації стійких словосполучень при визначенні ключових слів в україномовних наукових текстах технічного профілю дозволяє підвищити точність пошуку ключових слів на 6-9% та виділити з тексту тематичні терми для подальшої рубрикації публікації;

– розроблення формального підходу до проектування модуля контент-моніторингу для ідентифікації ключових слів в україномовних текстах на основі видобування веб-даних, ОПМ та лексичного аналізу визначених слів текстового контенту, що дозволило розробити загальну структуру типових КЛС та підвищити ефективність функціонування КЛС на 6-9% в залежності від розв’язку конкретної задачі ОПМ;
– застосування методу обчислення ступеня верифікації автора україномовного тексту на основі аналізу стилів потенційних авторів дозволило підвищити точність ідентифікації на 6-12% та провести декомпозицію методу через дослідження коефіцієнтів стилістики як зв’язність мовлення, ступінь синтаксичної складності, лексична різноманітність, індекси концентрації та винятковості тексту;
– розроблення модуля контент-моніторингу для ідентифікації потенційного автора тексту із множини можливих на основі порівняння результатів аналізу шаблонного авторського тексту з досліджуваним для зменшення обсягу відповідної множини до [9;34]% із загальної кількості учасників проекту в залежності від тематики та часового діапазону написання науково-технічної публікацій, а також частоти публікацій цього автора в цей проміжок на конкретну тематику;
– експериментальна апробація методу ідентифікації стилю автора в україномовних текстах на основі видобування веб-даних та лексичного аналізу визначених стопових слів, що дозволяє виділити множину потенційно подібного за стилем контенту з множини потенційних авторських публікацій.

Особистий внесок здобувача. Усі наукові результати, подані у дисертації, одержані здобувачем особисто. Роботи [209, 404, 471, 472, 473, 476, 937] опубліковано без співавторів. У друкованих працях, опублікованих у співавторстві, особисто здобувачу належать такі результати: [136, 804, 805, 958] – метод класифікації текстових документів; [38, 56, 56, 58, 59, 144, 588, 984] – вдосконалений метод управління контентом інформаційних ресурсів; [40, 850, 929] – метод інтеграції контенту інформаційних ресурсів; [480, 946] – аналіз

часової залежності морфології вихідного сигналу на основі машинного навчання та нейронних мереж; [103, 132, 348, 811, 883, 936] – метод супроводу контенту інформаційного ресурсу; [19, 37, 112, 142, 875] – вдосконалені лінгвістичні методи опрацювання тексту; [84, 86, 86, 182, 960, 961, 962, 976] – метод визначення автора текстового україномовного контенту; [54, 813, 814, 954, 955] – аналіз ігрових методів опрацювання інформації; [20, 21, 535] – вдосконалений метод семантичного аналізу текстового україномовного контенту; [41, 61, 350, 493, 856, 948, 496, 497, 989, 995, 997] – аналіз процесів опрацювання контенту в різних предметних областях на основі машинного навчання та аналізу великих даних; [20, 162, 292, 535] – вдосконалений метод морфологічного аналізу текстового україномовного контенту; [60, 145, 401, 402] – метод інтелектуального пошуку текстового контенту; [114, 160, 256, 257, 407, 959] – метод визначення ключових слів текстового україномовного контенту; [987, 988, 1002] – метод опрацювання службового контенту; [474, 475, 477, 477, 903, 904] – метод опрацювання інформаційних ресурсів; [113, 803, 809, 808, 812, 816, 996] – онтологічний підхід для опрацювання текстового контенту; [163, 535] – вдосконалений метод синтаксичного аналізу текстового україномовного контенту; [141, 405, 406, 885] – метод контент-аналізу для опрацювання текстових масивів даних; [39, 310, 403, 849, 884] – метод аналізу психологічного стану особистості на основі ОПМ.

Апробація результатів дисертації. Основні результати дисертаційної роботи доповідалися на міжнародних, українських та міжвузівських конференціях та семінарах, зокрема: Міжнародна конференція «Computational Linguistics and Intelligent Systems» (CoLInS, Lviv-Kharkiv, 2017-2021); Міжнародна конференція «Modern Machine Learning Technology» (MoMLeT, Shatsk, 2019-2021); IEEE Міжнародна конференція «Smart Information Systems and Technologies» (SIST, Nur-Sultan, 2021); Міжнародна конференція «Intelligent data acquisition and advanced computing systems: technology and applications» (IDAACS, Бухарест, 2017; Metz, 2019; Cracow, 2021); IEEE Міжнародна конференція «Advanced information and communication technologies» (AICT, Lviv,

2019, 2021); Міжнародна конференція «Академічна доброчесність: виклики сучасності» (Warszawa, 2020); IEEE International Conference: Modern problems of radio engineering, telecommunications and computer science TCSET (Lviv-Slavske, 2016, 2022); Міжнародна конференція «Computer Science and Information Technologies» (CSIT, Lviv, 2015-2021); Міжнародна наукова конференція "Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту ISDMIT (Залізний Порт, 2015-2019); International scientific and practical conference "Scientific Research Priorities: theoretical and practical value" (Nowy Sanz, 2017); Міжнародний симпозіум «Intelligent data acquisition and advanced computing systems» (IDAACS-SWS, Львів, 2018); Науково-технічна конференція «Обчислювальні методи і системи перетворення інформації» (Львів, 2018); International conference of System analysis and information technology SAIT (Kyiv, 2017); IEEE Міжнародна конференція «Data Stream Mining and Processing» (DSMP, Lviv, 2016, 2018, 2020); Міжнародна конференція «Математика. Інформаційні технології. Освіта» (Луцьк, 2016); International conference of perspective technologies and methods in MEMS Design MEMSTECH (Lviv-Polyana, 2016); Всеукраїнська науково-практична конференція «Інтелектуальні системи та прикладна лінгвістика» (Харків, 2016); IEEE International Conference: The Experience of Designing and Application of CAD Systems in Microelectronics CADSM (Lviv-Polyana, 2015); Міжнародна конференція «Обробка сигналів і негаусівських процесів» (Черкаси, 2015); Міжнародна конференція «Інформація, комунікація, суспільство 2015» (Славське, 2015). Результати дисертаційних досліджень регулярно доповідалися на наукових семінарах кафедри «Інформаційні системи та мережі» Національного університету «Львівська політехніка» (2015 р. – 2022 р.).

Публікації. Основні результати дисертаційного дослідження опубліковано у 254 наукових публікаціях, серед яких 71 стаття у наукових фахових виданнях України (зокрема, 26 із них включено до Scopus або Web of Science), 72 статті у наукових періодичних виданнях інших держав (зокрема, 59 із них включено до Scopus або Web of Science, з них 4 статті опубліковано в

журналах з квартилем Q2), 100 тез доповідей та матеріалів конференцій (зокрема, 64 із них включено до Scopus або Web of Science), 9 монографій та 2 розділи монографії, які включено до міжнародних наукометричних баз. Зокрема 50 статей у фахових наукових виданнях України та 31 стаття у наукових періодичних виданнях інших держав відповідають вимозі МОН України щодо публікації в одному виданні.

Структура та обсяг роботи. Дисертаційна робота складається з анотацій, вступу, шести розділів, висновків, списку використаних джерел з 1044 назв на 52 сторінках та 6 додатків на 82 сторінках. Загальний обсяг дисертації – 480 сторінок, з них: 306 сторінок основного тексту, 179 рисунків, 62 таблиці.

РОЗДІЛ 1
СУЧАСНИЙ СТАН ТА ПЕРСПЕКТИВИ РОЗВИТКУ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ ОПРАЦЮВАННЯ ПРИРОДНОЇ МОВИ КОНТЕНТУ

1.1. Аналіз відомих комп'ютерних лінгвістичних систем

1.1.1. Поняття комп'ютерних лінгвістичних систем

Сучасний розвиток інформаційних технологій (ІТ) знаходиться на перетині глобалізації та інформатизації [1]. Швидкі темпи зростання інформатизації суспільства напряму пов'язані із темпами розвитку та впровадженням ІТ опрацювання природної мови (Natural-Language Processing, NLP) [2], основними інструментами яких є комп'ютерні лінгвістичні системи (КЛС). Згідно з [3] для існуючого терміну лінгвістична система (ЛС) існує два різних трактування:

1. Множина лінгвістичних одиниць відповідного мовленнєвого рівня (фонологія, морфологія, синтаксис тощо) з врахуванням їх єдності та взаємозв'язку; типи лінгвістичних одиниць і правила їх формування, перетворення та поєднання. Ідентифікація мови як ЛС приписується F. de Saussure [4] і ґрунтується на працях W. von Humboldt [5] та I. A. Baudouin de Courtenay [6].
2. Множина опозицій (фактів) на відповідному лінгвістичному рівні з використанням для опису та ідентифікації метамови [7-8].

В [9-10] лінгвістична інформаційна система або ЛС визначена як система, яку індивід застосовує для своєї мовленнєвої діяльності.

Згідно з тлумаченням Стенфордської енциклопедії [11] комп'ютерна лінгвістика (Computational Linguistics, CL) це наукова та інженерна дисципліна для знаходження підходів розуміння письмової та усної мови комп'ютерними засобами, а також створення методів опрацювання природної мови. Оскільки мова є дзеркалом розуму, комп'ютерне розуміння мови також сприяє розуміння процесу мислення та змісту інтелекту [12]. Якщо природна мова є найприроднішим і універсальним засобом спілкування, то відповідне програмне забезпечення (ПЗ) з лінгвістичною компетентністю мають значно полегшувати

людську взаємодію через комп’ютери між собою та інформаційними системами (ІС) усіх видів для задоволення повсякденних потреб, наприклад, при ІІП/аналізі величезних текстових масивів даних та інших Website [13-15]. Відповідно КЛС призначена для розв’язування NLP-задач відповідно до потреб користувача [13-24]. Головними ознаками КЛС є застосування методів штучного інтелекту (Artificial Intelligence, AI) [25-26], прикладної лінгвістики (Applied Linguistics, AL) [27-29], системного аналізу [220-227, 237, 1034-1038] та ІТ [30-31] для розуміння природної інформації при розв’язані різних NLP-задач [32-34] як в повсякденному житті людини, так і в сучасних дослідженнях спеціалізованого наукового спрямування (Рис. 1.1) [35-42].

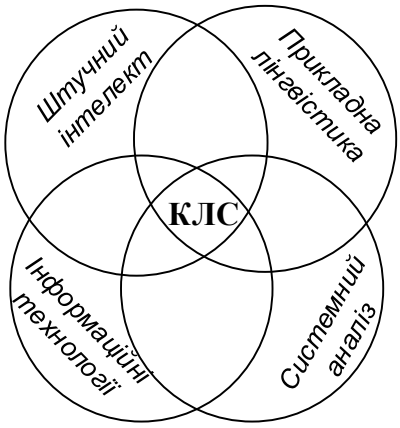


Рис. 1.1. Основні сучасні напрями для синтезу КЛС

Головним об’єктам комп’ютерної лінгвістики є контент – довільна напівструктурована та частково формалізована інформація, подана усно промовою, письмово текстом, візуально/емоційно мімікою та жестами, графічно смайликами/зображеннями і/або будь-яким іншим способом передачі. Контент – колекція різнотипових даних (текстових, звукових, службових, комерційних, додаткових тощо), які формують відповідну множину мета-моделей (опис структури та особливостей функціонування моделі) та моделей-шаблонів, реалізованих в межах конкретної інформаційної системи (ISO/IEC/IEEE 24765:2010(E), 3.1398, ISO/IEC 15474-1:2002, Information technology) [585-587].

Публікацій та досліджень розв’язання різних NLP-задач є досить багато для опрацювання англомовних текстів. Суттєво менше є досліджень для

слов’янських мов, зокрема, для української. І взагалі, відсутні публікації щодо рекомендацій розроблення, функціональних вимог, загальної структури та типової архітектури КЛС. Зазвичай кожний успішний проект розроблення КЛС призначений під конкретну задачу та одночасно є одноразовим та закритим (наприклад, Siri, Amazon Alexa, Google Assistant, Grammarly, Abby Lingvo, Facebook, Voice Mate, Bixby, Microsoft Cortana, Microsoft Word, Google Translation, PROMT, CuneiForm, Trados, OmegaT, Wordfast, Dragon, IBM via voice, Speereo, Finereader, Tesseract, OCRopus тощо) без можливості ознайомитися з вмістом бажаючим ІТ-фахівцям/спеціалістам. Досить рідкі випадки, коли до таких проектів надають відкритий доступ та можливість ознайомитися з його структурою та іншим змістом. Відповідно дослідження в напрямі аналізу та синтезу КЛС, зокрема, для опрацювання україномовних текстів на сьогодення є актуальним та перспективним [43-50].

1.1.2. Загальна класифікація комп’ютерних лінгвістичних систем

Сьогодні напрям CL стрімко розвивається, але більшість проектів є комерціалізованими та одноразовими. Тому немає єдиного однозначного підходу, типових загальних рекомендацій, порад та вимог щодо проектування, аналізу, розроблення та синтезу відповідних КЛС. Також немає єдиної думки щодо типізації, категоризації та класифікації КЛС. Це значно ускладнює зорієнтуватися в хаосі публікацій та досліджень, які методи та інструменти необхідно застосувати для ефективного отримання бажаних результатів, зокрема, при розв’язанні конкретної NLP-задачі опрацювання україномовних текстів. Наприклад, за класифікацією від Grammarly існують лише три основні типи КЛС [51-53]: аналітичні, трансформаційні та комбіновані (Рис. 1.2). Різновидів та можливостей КЛС набагато більше [54-58], ніж описано в [51-53]. Цей список необхідно доповнити рекомендаційними системами [59-73], ІС засобів масової інформації, системами аналізу психологічного стану особи (наприклад, IBM Watson™ Personality Insights) [74-78], системами ідентифікації плагіату (копірайт/рерайт) [79-83], системами визначення авторського стилю

мовлення [84-88], інтерфейсами безмовного доступу [90-91], системами розпізнавання жестової мови [92-99] тощо.

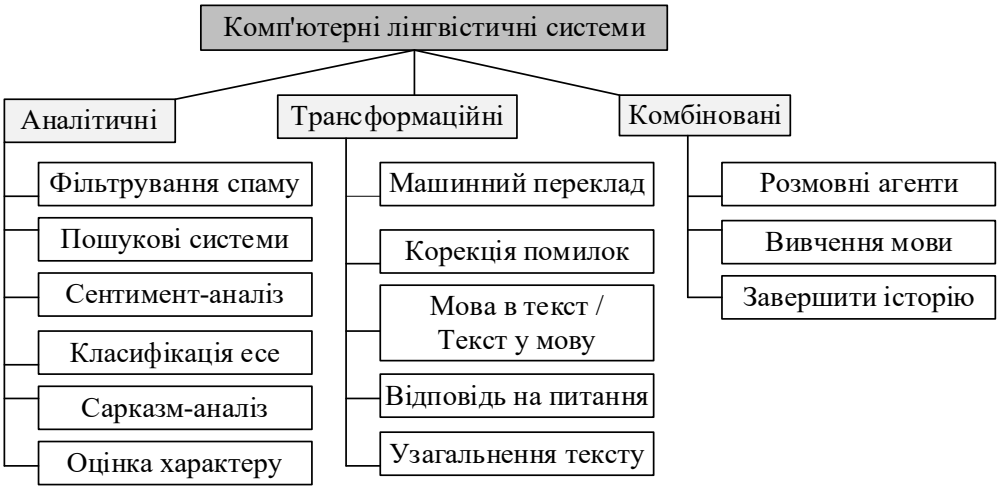


Рис. 1.2. Класифікація комп’ютерних лінгвістичних систем за Grammarly

Стівен Хокінг – однин із найвідоміших людей – застосовував мовленнєвий комп’ютер для спілкування [100-102]. Послуга IBM Watson ™ Personality Insights надає API для збирання статистичних даних та корпоративної інформації із соціальних мереж та інших е-інструментів комунікацій [74-78]. Служба використовує лінгвістичну аналітику для формування висновку про внутрішні характеристики особистості людей за допомогою е-інструментів комунікацій, таких як електронна пошта, текстові повідомлення, твіти та повідомлення на форумі [74-78].

1.1.3. Основні NLP-задачі комп’ютерних лінгвістичних систем

Основним критерієм розвитку ринку та частоти використання КЛС є мотивація застосування інтелектуального ПЗ, хмарних рішень/додатків на основі NLP, які покращують обслуговування різних клієнтів всіх можливих напрямів діяльності людини та суттєво збільшують потенційну аудиторію користувачів сучасних ІТ без вимог володіння спеціальними навичками та знаннями для їх використання [103-108]. На це здійснив вплив спектр задач, які мають розв’язувати різного типу/призначення КЛС (Рис. 1.3) [109-114]. Основними напрями розв’язку задач для КЛС є аналіз і/або генерування текстів природньою мовою, розпізнавання та синтез природнього мовлення [115-116]. Частина

актуальних задач одночасно відносять до деяких напрямів [109-116], наприклад, діалогові системи спираються на такі NLP-інструменти як розпізнавання мови, виділення змісту та контексту, ідентифікація намірів, а потім вибудовування діалогу, виходячи з вищезгаданого (в ідеалі – шляхом синтезу мовлення) [68-73]. Так розумний асистент має розв'язувати задачі розпізнавання мови, аналіз текстів, генерування текстів та відповідно синтез мови [60-63, 68-73, 113-114, 117-120]. А машинний переклад розв'язує задачі аналізу текстів, синтез мовлення та генерування текстів [121-129]. Для QA-систем (питально-відповідні) достатньо розв'язати задачі аналізу текстів [60-63, 68-73].

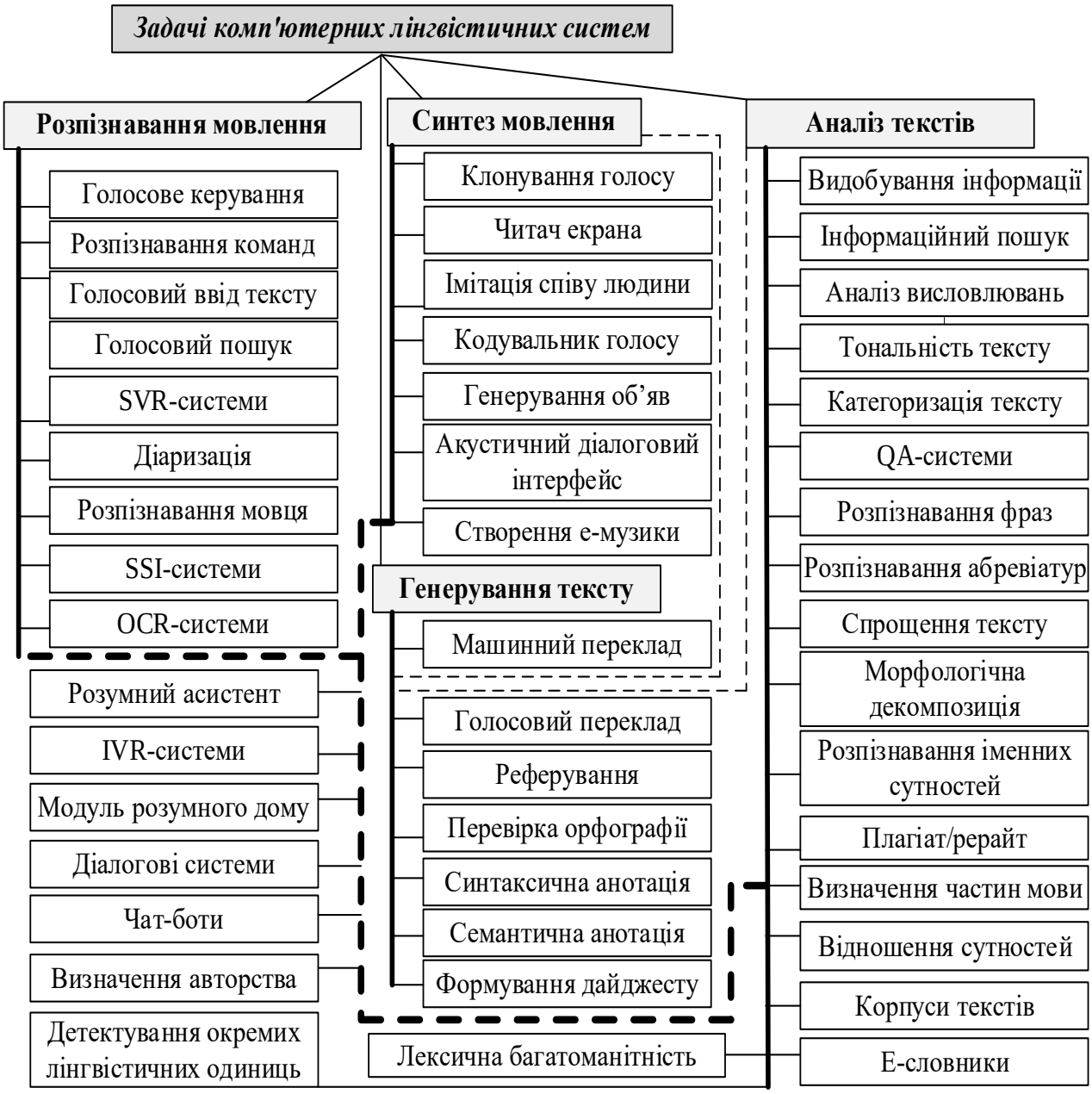


Рис. 1.3. Класифікація задач комп'ютерних лінгвістичних систем

Але це лише умовні припущення у зв'язку з тим, що кожна реалізація конкретної КЛС є зазвичай закритим комерційним проектом, що не дає можливості ІТ-фахівцям ознайомитися із детальною структурою відповідних систем та реалізованими NLP-алгоритмами.

1.1.4. Приклади та порівняльний аналіз відомих сучасних КЛС

На сьогодення при бурхливому розвитку ШІ, прискореному зростанні великих обсягів даних і знань та швидких темпах інформатизації суспільства розроблено та впроваджено безліч КЛС (ПЗ/Web-сервіси) різного призначення для розв'язку відповідного типу NLP-задач згідно потреб користувачів [130-147]. В цьому напрямку працюють провідні світові компанії як Google, Apple, IBM, Microsoft, тощо. Поряд з ними над різними типами КЛС працюють інші менш відомі компанії, в тому числі і українські. Ці КЛС мають власні переваги та недоліки. Розглянемо порівняємо лише найпопулярніші світові та міжнародні проекти КЛС відповідно з кожного класу NLP-задач (Таблиця 1.1). В напряму комп'ютерної лінгвістики працюють такі компанії як НАСА, IBM, Apple, Amazon, Microsoft, Google, Yamaha, Grammarly, DARPA, Yahoo, тощо [148-159].

Таблиця 1.1

Відомі інструменти для розв'язку відповідної NLP-задачі [148-159]

NLP-задача	Інструмент розпізнавання мовлення
Розпізнавання мовлення	
Голосове керування та розпізнавання команд	компонента ОС Microsoft Windows (Vista, v. 7-11), OS/2 Warp 4 та Mac OS X, а також Voice Access, IBM ViaVoice, Microsoft Voice Command, Yandex SpeechKit, Dragon NaturallySpeaking, Speereo Speech Engine, Lexy, linguatec Voice Pro, Speech (Apple Macintosh), тощо;
Голосовий ввід (набір) даних	VoiceNavigator, Dragon Naturally Speaking MSpeech (Google Voice API), Voco, Dictate, SpeechPad (Chrome), VoiceNote II (Chrome), TalkTyper (Chrome, 37 мов, free), SpeechTexter (Google Play), Google Docs (Gmail), Voice Notepad (Chrome, 120 мов), Odrey (українська розробка), VoiceTypist (введення тексту голосом українською мовою) тощо;
Аналіз промови	IBM via voice, Dragon,;
Голосовий ІІП	Google, Yandex SpeechKit Mobile SDK, МедиаИнсайт тощо;
Субвокальне розпізнавання (англ. Subvocal recognition, SVR) промови в процесі мовчання особи на основі аналізу електроміограм	технологія НАСА (Ames Research Laboratory) з дослідницького центра Еймса в Маунтін-Вью (Каліфорнія), під керівництвом Charles Jorgensen тощо;
Діаризація спікера для ідентифікації приналежності частини промови та її	National Institute of Standards and Technology (NIST), тощо;

NLP-задача	Інструмент розпізнавання мовлення
змісту конкретній особі із множини діалогу	
Розпізнавання мовця (ідентифікація людини залежно від особливостей голосу для поведінкової біометрії	GoVivace Inc., Barclays, Barclays Wealth, CSELT, тощо;
SSI (інтерфейс безмовного доступу) як допоміжний інструмент для створення звукової фонації аудіозного мовлення або спілкування при наявності фонового шуму	AlterEgo (Arnav Kapur, MIT Researcher), SpeakUP (Varun Chandrashekhar), NTT DoCoMo;
OCR (оптичне розпізнавання тексту)	ABBYY FineReader, CuneiForm, Brainware, COCR2, ExperVision TypeReader & RTK, Tesseract, FineReaderOnline.ru, FreeOCR, GOCR, HOCR, img2txt.com, Microsoft Office OneNote 2007, Microsoft Office Document Imaging, OCRopus, Kirtas Technologies Arabic OCR, NovoDynamics VERUS, NewOCR.com, OnlineOCR.ru, Ocrad, OmniPage, Persian Reader, Readiris, ReadSoft, RelayFax Network Fax Manager, Scantron Cognition, SILVERCODERS OCR Server, SimpleOCR, SmartScore, Tesseract, ViewWise, WeOCR, Zonal, тощо.
Синтез мовлення	
Клонування голосу (англ. voice cloning, voice changing)	CereVoice Me, iSpeech, Voice Anonymizer, LyreBird, Resemble AI, Voice changer, Morphvox, SDK пакети, Voice Cloning Toolkit for Festival and HTS (Mac, Дослідницький Центр Мовленнєвих Технологій, Junichi Yamagishi із Університету Единбургу) тощо;
Імітація співаючої людини за технологією вокалоїд (англ. Vocaloid) від Yamaha Corporation	SONiKA, LEON, CUL, MAYU, IA, UNI, AVANNA, MEW, DAINA, KAITO, DEX, OLIVER, YANHE, MEIKO, MEIKO V3, LUMi, MATCHA, AZUKI, MAIKA, FUKASE, MIRIAM, LOLA, KAITO V3, CYBER DIVA, Ken, Kaori, Chris, Amy, Haruno Sora, KAITO V5, YANHE V5, CYBER SONGMAN II, CYBER DIVA II, VY1v5, VY2v5, Yuezheng Ling V5, ARSloid, Sachiko, SeeU, Kaai Yuki, Yuzuki Yukari, Luo Tianyi, GUMI Native (Megpoid), Clara, Bruno, Ryuto (Gachapoid), GUMIv3 (Megpoid), VY1 (VocaloWitter), eVY1, VY1 (Mizki), VY1v3 (Mizki), VY1 (i-VOCALOID), VY2 (Yuma), VY2v3 (Yuma), VY1V4, YOHIOloid, ZOLA PROJECT, IA ROCKS, CYBER SONGMAN, V flower, v4 flower, Gachapoid V3, Megpoid V4, Rana V4, Hiyama Kiyoteru V4, Kaai Yuki V4, SF-A2 miki V4, Tohoku Zunko V4, Hatsune Miku V4X, Gackpoid V4, Kagamine Rin and Len V4X, Macne Nana V4, Tone Rion V4, Nekomura Iroha V4, Luo Tianyi V4, Megurine Luka V4X, Xin Hua V4, Xin Hua V4 Japanese, Hatsune Miku V4 Chinese, Yuzuki Yukari V4, galaco NEO, Kokone, Ruby, Tohoku Zunko, Tone Rion, Chika, Merli, Lily, Rana, Xin Hua, Tonio, Macne Nana, Aoki Lapis, SF-A2 miki, Megpoid English, anon & kanon, Yuezheng Ling, Hatsune Miku V3 English, Big-AL, Hiyama Kiyoteru, Vocaloid Keyboard, Pocket Miku, Megpoid GUMI, Miku Append, Megurine Luka, Utatane Piko, Nekomura Iroha, Prima, Kamui Gakupo (Gackpoid), Kagamine Rin and Len, Hatsune Miku, Sweet ANN, Kagamine Rin and Len Append, Mo Qingxian, Zhiyu Moke, Mirai Komachi, Kizuna Akari, Yuezheng Longya, Yumemi Nemu, Otomachi Una, Xingchen (Stardust), тощо;
Кодувальник голосу за технологією вокодера (англ. voice encoder) та VST-plugins	Cylonix, Darkoder, Lpc-vocoder, Formulator, AC vocoder, Voctopus, Akai DC Vocoder, Fruity Vocoder, Steinberg Vocoder, FL Studio, Steinberg Cubase, Cakewalk Sonar, Buzz Composer, Max MSP, NI Reactor/Generator;
Читач екрана (англ. screen reader) або синтез промови з тексту	VoiceOver (Apple Inc Mac OS X), Агафон, Narrator (MS Windows), Microsoft Agent, GNOME, NVDA, VS Robotics, Window-eyes, JAWS, Festival, AT&T Natural Voices, Gnuspeech, ESpeak, pVoice (Perl), RSS To Speech, Read Words Eng 4, Digalo (Acapela ELAN TTS), Nuance RealSpeak (ScanSoft), Sakrament TTS Engine, Govorilka, Speak Aloud, ToM Reader, CoolReader, Linguatec, Acapela, Oddcast, iSpeech, Google Translate, 2yxa, Балаболка, Microsoft Speech Api 4.0, тощо; україномовними є Lesya

NLP-задача	Інструмент розпізнавання мовлення
	(KobaVision/KobaSpeech, Code Factory, NextUp, Nuance Vocalizer), Розмовлялка, WaveNet, UkrVox, RHVoive, CyberMova/VymovaPlus/VymovaPro, тощо;
Генерування повідомлень/об'яв	HKUST Xunfei, VS Robotics;
Акустичний діалоговий інтерфейс на основі технології Partner-assisted scanning (voice output communication aids або Speech-generating devices, SGDs)	Equalizer для Stephen Hawking (Walter Woltosz, CEO of Words Plus), також Tony Proudfoot, Roger Ebert, Pete Frates (засновник ALS Ice Bucket Challenge), тощо;
Створення електронної музики	Adobe Audition, Final Cut Pro (Apple Pro, Mac), MainStage, Logic Pro, Compressor (Apple Qmaster, Apple Qadministrator, Share Monitor), Motion, тощо;
Аналіз текстових масивів даних	
Інформаційний пошук або інформаційно-пошукові системи	Google, Yahoo!, Yongzin (Китай), AltaVista, A9.com (Amazon), LightStorage, Ask.com (Teoma), Alltheweb FAST-Engine, ALLhave, Search engine site ABC Engine, ZipLocal (США, Канада), Neti (Естонія), Яндекс (Росія, Білорусь, Туреччина, Казахстан), Yahoo Japan (Японія), Walla! (Ізраїль), Seznam (Чехія), Sesam (Норвегія, Швеція), Search.ch (Швейцарія), Rediff (Індія), Rambler (Росія), Pipilika (Бангладеш), Naver (Корея), Najdi.si (Словенія), Miner.hu (Угорщина), Maktoob (Арабський світ), Leit.is (Ісландія), Goo (Японія), Fireball (Німеччина), Egerin (Курдистан), Daum (Корея), Biglobe (Японія), Accoona (Китай, США), Youdao, Yippy, WebCrawler, Swisscows, Startpage.com, Soso, Sogou, Searx, Qwant, Mojeek, MetaCrawler, Lycos, HotBot, Гігабласт, Excite, Exalead, Ecosia, DuckDuckGo, Dogpile, Бінг, Байду, Voilà, Nomade, Locace, Francité, Ez2find, Abacho, Wseeker, Пошук AOL, SAPO (Португалія, Мозамбік, Кабо-Верде, Ангола), Google Scholar, Scirus, ArXiv.org, ScienceDirect, PubMed, PDF Search System (PDFSS), LightStorage (медіа файли), GlobalFileSearch (файли), Tineye (зображення), тощо; серед українських пошукових систем необхідно виділити наступні: Мета, Шукалка, Bigmir, I.ua, Online.ua, TopPING, UAport, Ukr.net, search.com.ua;
Аналіз висловлювань або контент-аналіз	якісний (Kwalitan, MAXQDA, тощо) та кількісний (TextQuest, Textanz, тощо), WebAnalyst (Megaputer Intelligence), Autonomy Knowledge Server, Text Miner (SAS Institute), InfoStream (Елвіста, українська розробка), Lithium Community Platform, Meltwater Buzz, тощо;
Розроблення е-словників	Abby Lingvo, ForceMem, dict, Stardict, GoldenDict, WordNet (семантичний англомовний словник), ConceptNet, Мультитран, Викисловарь, WordNet-Affect, SenticNet, SentiWordNet, тощо;
Текстова аналітика (англ. text mining), видобування інформації або інтелектуальний аналіз тексту	Intelligent Miner for text (IBM), SAS Text Analytics, WebAnalyst (Megaputer Intelligence), Autonomy Knowledge Server, SemioMap (Entrieva), TextAnalyst (Megaputer Intelligence), Text Miner (SAS Institute), Apache OpenNLP (Java), OpenCalais (Thomson Reuters), Natural Language Toolkit (Python), Galaktika-ZOOM, InfoStream (Елвіста, українська розробка), Russian Context Optimizer (RCO), Lithium, Ontos (TAIS Ontos, Ontos SOA, LightOntos for Workgroups, OntosMiner), Paai's text utilities тощо;
Аналіз тональності тексту	SAS Sentiment Analysis, Lithium Social Media Monitoring, InfoStream (Елвіста, українська розробка), OpinionEQ, Radian6, OpenAmplify SocialView (Visual Intelligence), Meltwater Buzz, LIQUID CAMPAIGN Opinion Mining, Social Mention, Tweetfeel, Twittratr, тощо;
Виявлення ключових слів та стійких словосполучень (колокацій, collocations)	Feature extraction tool (Intelligent Miner for text, IBM), SemioMap (Entrieva), VICTANA (українська платформа), Oracle Text, InterMedia Text, Galaktika-ZOOM, тощо;
Категоризація тексту	Categorisation tool (Intelligent Miner for text, IBM), SemioMap (Entrieva), Autonomy Knowledge Server, TextAnalyst (Megaputer Intelligence), RCO, AskNet, тощо;

NLP-задача	Інструмент розпізнавання мовлення
Кластеризація текстів	Clusterisation tool (Intelligent Miner for text, IBM), SemioMap (Entrieva), TextAnalyst (Megaputer Intelligence), Vivisimo Nigma, Quintura Searchcrystal, тощо;
Питально-відповідні системи (QA-системи)	ELIZA, Watson (IBM), DrQA (Facebook Research), тощо;
Розпізнавання фраз	WebAnalyst (Megaputer Intelligence), Oracle Text, InterMedia Text, Google Translate, TextGrabber, Translate.Ru, Яндекс Translate, Microsoft Translate, Translator Foto - Voice, Text & File Scanner, iA Writer, TextExpander, Odrey (українська розробка), Apache OpenNLP, тощо;
Морфологічна декомпозиція	Oracle Text, InterMedia Text, iA Writer, TextExpander, Odrey (українська розробка), RCO, Apache OpenNLP;
Розпізнавання іменних сутностей, колокацій (англ. collocations), ідіом, фразеологізмів та крилатих фраз	OpenNLP, SpaCy, GATE, SemioMap (Entrieva), Autonomy Knowledge Server, Oracle Text, InterMedia Text, Galaktika-ZOOM, DBpedia Spotlight, Apache OpenNLP, тощо;
Визначення частин мови слів	Oracle Text, InterMedia Text, Apache OpenNLP, тощо;
Ідентифікація мови	Language identification tool (Intelligent Miner for text, IBM), тощо;
Розпізнавання абревіатур	OpenNLP, SpaCy, GATE, VICTANA (українська платформа), тощо;
Спрощення тексту	WebAnalyst (Megaputer Intelligence), Poetica, Test-the-Text, HamingwayApp, Readability, Стоп-слов нет, Типографська розкладка Іллі Бірмана (Gagadget, українська версія), Типограф Артемія Лєбєдєва, LeoBilingua, Forson,тощо;
Ідентифікація плагіату/рерайту або дублювання тексту	Unplag/Unichek, Etxt Antiplagiat, Advego Plagiatus, Plag.com.ua, Plagiarisma, Content-watch, StrikePlagiarism.com, TEXT.RU, Edu-Birde, InfoStream (Елвіста, українська розробка), тощо;
Визначення відношень між сутностями	Text Miner (SAS Institute), SemioMap (Entrieva), Autonomy Knowledge Server, Galaktika-ZOOM, InfoStream (Елвіста, українська розробка), Link Grammar Parser, Mystem, LingSoft, Cíbola/Oleada CLR, StarLing, MCR DLL, SyTech;
Розв’язок лексичної багатоманітності	Oracle Text, InterMedia Text;
Кореферентний аналіз (англ. Coreference) – визначення множини термінологічних іменних сутностей, що мають відношення до одного об'єкта, суб’єкта, явища або події	TextAnalyst (Megaputer Intelligence), Customer Intelligence Center;
Статистичний аналіз тексту	WordStat, netXtract Relevant, URS, FRQDictW, Лемматизатор Мультитран, Textarc, Ngram Statistics Package (NSP), Rhymes, WordTabulator, тощо;
Детектування окремих лінгвістичних одиниць	Autonomy Knowledge Server, SemioMap (Entrieva), Galaktika-ZOOM, тощо;
Розмічення та маркування текстів для формування лінгвістичних корпусів текстів	GenCode, TeX, LaTeX, Scribe, GML, SGML, HTML, XML, XHTML, Textual Analysis Computing Tools (TACT),тощо;
Генерування сценаріїв/сюжетів для вистав/телепрограм/фільмів	Final Draft, тощо;
Редактор для концентрації уваги	FocusWriter, iA Writer, тощо;
Автоматичний HTML-редактор	Реформатор, тощо;
Створення конкордасів	WordSmith Tools, MonoConc, Textual Analysis Computing Tools (TACT), ParaConc, WordSmith Tools Mike Scott, Concordance 2.0.0 R.J.C. Watt, тощо.
Генерування текстових масивів даних на основі розпізнавання/синтезу мовлення та аналізу текстів	
Машинний переклад	Google Translate, Microsoft Translator, PROMT, SYSTRAN, Yandex.Translate, TIDES, Babylon translator, MT@EC, Trados, OmegaT, Apertium, SDL Trados, STAR Transitfr, Déjà Vu, SDLX, Abby Lingvo, Socrat, Across Language Server, Crowdin, GlobalSight, gtranslator, MateCat, memoQ, MetaTexis, Open Language Tools, Phrase, Poedit, Pootle, Babylon, SDL Trados Studio, SmartCAT, UNMIN, Virtaal, Wordfast, Anusaaraka, DeepL, GramTrans, IBM

NLP-задача	Інструмент розпізнавання мовлення
	Watson, IdiomaX, Moses, Moses for Mere Mortals, NiuTrans, OpenLogos, тощо; українська розробка: Trident Software, Pragma, L-Master 98, Language Master; утілити GoogleTalk, Facebook, MSN Messenger, Skype, тощо;
Ідентифікація пошукового спаму (англ. Spamdexing)	Google, Yahoo!, AIRWeb, тощо;
Створення рерайту	BIPOD, ReWrite Suite, SeoGenerator, korektoronline.pl, Програма для рерайту (plati.ru), тощо;
Перевірка орфографії та граматики (англ. spell checker)	Grammarly (розробка українських програмістів), Microsoft Word, myspell, aspell, ispell, Орфо, SPELL (Ralph Gorin ,Stanford Artificial Intelligence Laboratory), WordPerfect, WordStar, Firefox, GNU Aspell, Mac OS X, Pidgin, Kmail, Opera, Konqueror, Google, Online Corrector (українська розробка), Draft, Google Docs, Орфограммка, Language Tool (українська орфорграфія), тощо;
Голосовий переклад	Speereo Voice Translator, тощо;
Реферування	Annotation tool (Intelligent Miner for text, IBM), Oracle Text, InterMedia Text, RCO, тощо;
Синтаксична анотація	WebAnalyst (Megaputer Intelligence), RCO, LeoBilingua, Forson,тощо;
Семантична анотація	TextAnalyst (Megaputer Intelligence), RCO, Customer Intelligence Center, Ontos, тощо;
Формування дайджестів	InfoStream (Елвіста, українська розробка), тощо;
Латентно-семантичний аналіз (ЛСА)	патент від Lynn Streeter, Karen Lochbaum, Thomas Landauer, Richard Harshman, George Furnas, Susan Dumais, Scott Deerwester як латентно-семантичне індексування (англ. Latent Semantic Indexing, LSI);
Фільтрація спаму та маршрутизація е-поштою	Kaspersky Anti-Spam, Apache SpamAssassin, AntispamSniper (The Bat!), тощо;
Стилеметрія (класифікація за стилем та жанром)	Emma, VICTANA (українська платформа), тощо;
Лінгвометрія	netXtract Relevant, WordTabulator, Ngram Statistics Package, Rhymes, Langsoft, VICTANA (українська платформа);
Глотохронологія	VICTANA (українська платформа), тощо;
Оцінка читабельності	WebFX Readability Test Tool, Automatic Readability Checker, Readability Calculator, Perry Marshall, StoryToolz, Readability Checker, Word Counter, Joe's Web Tools, progaonline.com, ru.readability.io, copywritely.com, glvrd.ru, Advego, turgenev.ashmanov.com, тощо.
Змішаний напряму розпізнавання/синтезу мовлення та аналізу текстів	
Розумний асистент	Google Assistant, Siri (Apple), Amazon Alexa, Яндекс Алиса, Robin (Audioburst), Vani Dialer (Bolo International Limited), Ассистент Дуся (UseYoVoice), Маруся (VK.com), Окей Блокнотик (Dmitriy V. Lozenko), МИРИ (BlueTo), Cortana (Windows 10), Горыныч, AGGREGATE, Typle (Windows), тощо;
IVR-системи (англ. Interactive voice response, інтерактивна голосова відповідь, система голосових меню)	VoiceNavigator, VoiceKey.IVR, (Customer Engagement Platfrom), тощо;
Модуль розумного дому	Apple Siri, Google Home, Facebook M, Xiao Ai, Amazon Alexa, Microsoft Cortana, Sonos One, Яндекс Алиса, тощо;
Діалогові системи (англ. Dialogue system або розмовний агент, speech agent, SA)	GUS system, CSLU, NLUI, LinguaSys, модулі в сучасних іграх, Olympus, Nextnova, Quack.com, NADIA, тощо;
Створення чат-ботів	сервіси SendPulse, Flow XO, ManyChat, Chatfuel, MobileMonkey, ChatbotsBuilder, Botmother, ChatBot.com;
Визначення авторства текстів	Emma, Лингвоанализатор, Атрибутор, СМАЛТ, Антиплагиат, Fresh Eye, тощо;
Аналіз психологічного профіля автора	IBM Watson™ Personality Insights, Авторовед, тощо;
Аналіз наукової літератури (визначення новизни та актуальності, семантичний пошук, ідентифікація омонімів тощо)	NaCTeM сервіси (National Centre for Text Mining, Манчестерський та Токійській університети), BioText (School of Information, Каліфорнійський університет, Берклі, США), TAPoR (Альбертський університет, Едмонт, Канада), тощо.

Більшість проектів КЛС є закритими, одноразовими та комерційними. Лише окремі ентузіасти відкривають секрети своїх проектів та надають користувачам та IT-фахівцям доступ до свої розробок. Крім того більшість із розроблених КЛС працюють з англомовними текстами, або множиною європейських та азіатських мов, в список яких не входить українська мова. КЛС ELIZA (Рис. 1.4) є однією із перших прикладів розв'язку NLP-задачі ведення діалогу комп'ютера з користувачем, наслідуючи відповідь Роджерського психотерапевта [68-70].

Рис. 1.4. Приклади діалогів в КЛС ELIZA

Нажаль цей діалог є обмежений за своєю структурою та призначений лише для англомовних учасників [71-73]. ELIZA є класичною КЛС, яка застосовує шаблон для ідентифікації англомовних фраз вигляду *Ви є X* та трансформує у типові питання типу *Що змушує вас думати, що я X?* (де *X* довільний ланцюжок слів англійською мовою від користувача). Це є імітація діалогу без семантичного аналізу для усвідомлення змісту запитань від користувача. В роботі [69] автор зазначив, що ELIZA реалізує один з кількох діалогових жанрів, де слухачі можуть діяти так, ніби вони нічого не знають про навколишній світ. На початках більшість користувачів ELIZA прийшли до думки, що система дійсно розуміла їх та їхні проблеми навіть після публікації з поясненнями в [70]. Це одна з перших спроб реалізації чат-ботів (ChatBot), якими зараз наповнені сучасні соціальні мережі, сервіси надання послуг та системи е-комерції.

Звичайно, сучасні розмовні агенти – це набагато більше, ніж розвага; вони можуть відповісти на запитання, забронювати рейс або знайти ресторани, функції, на які вони покладаються, набагато складніші для розуміння, ніж наміри

користувача [68]. Тим не менше, прості, засновані на моделях, методи, які використовують ELIZA та інші чат-боти, відіграють вирішальну роль для розв'язку сучасних NLP-задач. Але для чат-ботів українською мовою звичайне застосування шаблонів унеможливлює процес імітації спілкування із-за наявності словозмін (за відмінком, часом, множиною тощо) в залежності від контексту. Без простого морфологічного, лексичного та синтаксичного аналізу побудувати відповідь або запитання українською в таких КЛС є не прийнятним результатом. Крім того, для текстових шаблонів необхідно застосовувати регулярні вирази, нормалізацію тексту, токенізацію, лематизацію, стемінг, сегментацію та розрахунок редакційної відстані [73]. Регулярні вирази використовують для ідентифікації послідовності символів, які необхідно витягнути з попередніх запитів користувача. Для цього використовують токенізацію слів як відокремлення їх від основного тексту (простої ідентифікації меж слів за наявністю пробілів та знаків пунктуації є недостанім процесом для вилучення словосполучень типу *проспект Червона Калина, Улан-Уде, Алма-Ата*, Південна Корея, *Івано-Франківськ, село Залізний Порт, місто Гола Пристань, місто Кривий Ріг, Кам'янець-Подільська фортеця* тощо або скорочення типу і т.п., т.д., англ., грн., та різні абревіатури, наприклад КЛС, NLP). Також сьогодні при токенізації треба враховувати різні знаки пунктуації для передачі емоцій у вигляді смайликів, наприклад, :), :(, :)))) тощо або хеш-теги, такі як #lpnu, #друзі. Наявність стилістичних помилок як відсутність пробілів між словами або відповідних знаків пунктуації ускладнює процес токенізації. Нормалізація тексту є трансформуванням його до зручної стандартної форми для сприйняття користувачем КЛС з врахуванням та узгодженням всіх флексій для всіх слів в реченні.

Нормалізація тексту також використовує сегментацію або парсинг [160-168] – розбиття тексту на окремі речення, використовуючи розділові знаки як сигнали. При лематизації ідентифікують однакові слова за аналізом їх коренів, , незважаючи на їх різницю, наприклад, слова *біжать, бігли, забігли* тощо – це форми дієслова *бігати*. Стемінг є простішою версією лематизації, де скорочують

слова до основи простим відкиданням суфіксів та/або флексій [169-170]. Для швидкості опрацювання текстів українською мовою краще застосовувати стемінг (ІП за ключовими словами та стійкими словосполученнями), а для точності отриманого результати – лематизацію (ідентифікація плагіату та рерайту) [171-183]. Для порівняння слів з розпарсеними ланцюгами символів використовують метрику – редакційну відстань [184-191], яка визначає ступень подібності аналізованих лінгвістичних одиниць на основі кількості необхідних редагувань (вставки, видалення, заміни) для заміни однієї послідовності символів на іншу. Застосовують найчастіше при ідентифікації та виправленні помилок, визначення рівня плагіату-рерайту, ІП, генерування рерайту тексту, розпізнавання мови/мовлення конкретної особи та машинному перекладі. Розрізняють такі категорії систем машинного перекладу: статистичні (Statistical Machine Translation, SMT) [192-194], на основі граматичних правил (Rule-Based Machine Translation, RBMT) [195-198], та гібридні системи. Але в кожній із них застосовують методи аналізу обчислювальної семантики для передачі конкретного змісту тексту при перекладі ні іншу мову [199] – різні способи моделювання значень слів, фраз, речень, фрагментів текстів. Обчислювальну семантику поділяють на дистрибутивну, онтологічну, формальну, операційну, та традиційну. Зокрема, дистрибутивну семантику застосовують для визначення значень лінгвістичної одиниці на основі статистики поєднання слів у великих текстових корпусах певної тематики [200-202]. В онтологічній семантиці розраховують семантичні залежності лінгвістичних одиниць контексту для формування множини знань [203-205]. Формальну семантику застосовують для опису значень виразів через математичну логіку та булеву алгебру [206-213]. Операційна семантика дозволяє описати множину речень тексту як множину команд керування деяким процесом генерування подій або функціонування виконавчого пристрою [214]. Традиційна семантика описує значення лінгвістичних одиниць тексту за допомогою спеціальних мов тлумачень [215]. Кожна із них має свої переваги та недоліки, особливо для синтаксичних природних мов як українська [199-215].

1.2. Основна загальна схема процесу лінгвістичного аналізу тексту природньою мовою засобами КЛС

1.2.1. Структурна схема лінгвістичного аналізу текстового контенту

Будь-який текст природною мовою містить суттєвий обсяг абстрактних неформалізованих неструктурованих змістовних даних [216-219]. Це є змістовний ланцюг символьних (лінгвістичних) одиниць з набором відповідних властивостей p_j для розв'язку певних лінгвістичних задач (Рис. 1.5) [51-53], як:

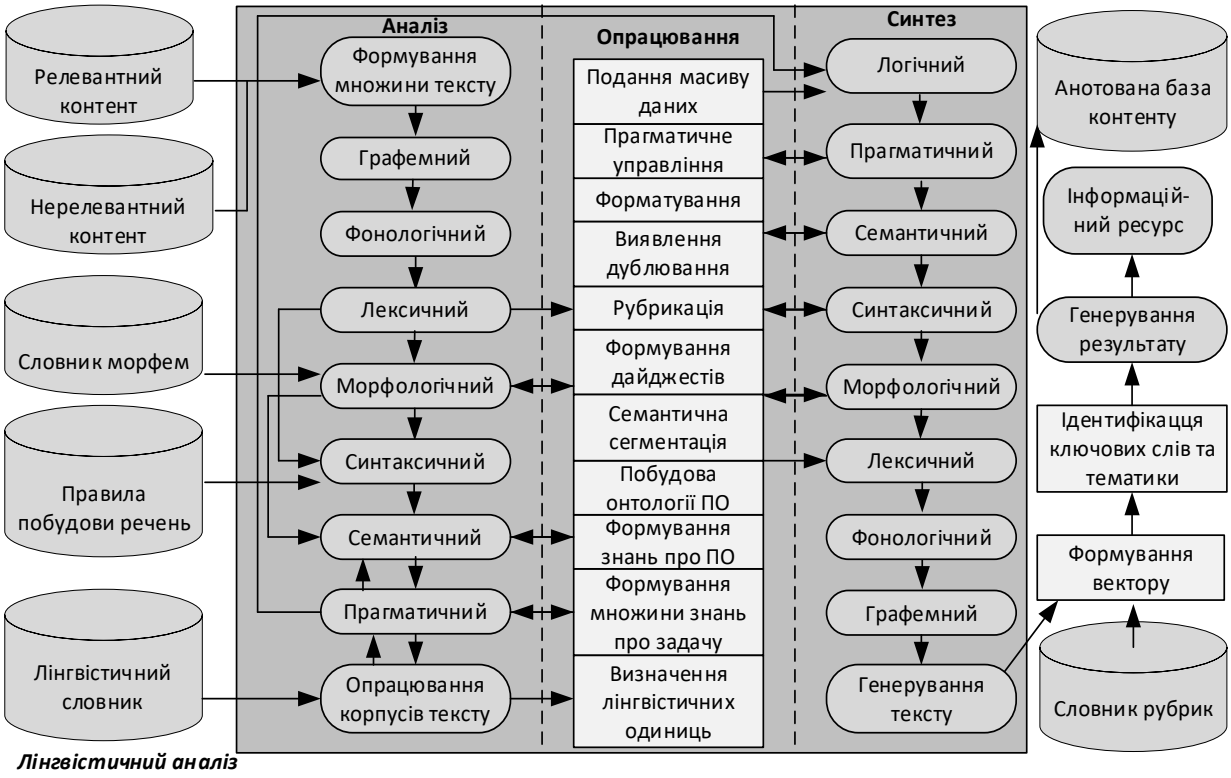


Рис. 1.5. Структурна схема лінгвістичного аналізу текстового контенту [211]

- кількість речень, слів, слів на речення тощо;
- розмір та розміщення абзаців;
- довжина слова та місцерозташування слова в реченні;
- кількість складів у слові та кількість змісту слова;
- співвідношення приголосних та голосних;
- глибина слова в дереві залежностей речення;
- N-grams та морфеми: афікси, корені, закінчення;
- чи є слово з великої літери / з дефісом / складеним;

- граматичні категорії різних POS тощо.

При лінгвістичному аналізі в КЛС застосовують різні рівні аналізу тексту природною мовою для розв'язку конкретних задач [51-53, 211]:

- сегментація – кортежі лінійних ланцюгів символів, розмежованих відповідними розділовими знаками;
- стемінг – множини лінійних ланцюгів морфологічних структур;
- токенізація – лінійна послідовність ланцюгів символів (слів тощо);
- парсинг – мережа взаємопов'язаних структурних єдностей в цих реченнях (категорії граматичні – лексичні – фонологічні).

1.2.2. Стани та властивості комп'ютерних лінгвістичних систем

Будь-який стан КЛС визначається кортежом головних властивостей в конкретний момент часу або активності відповідного NLP-процесу [220-227]:

$$s_i = (p_{i1}, p_{i2}, \dots, p_{im}),\ i = \overline{1,n}, \tag{1.1}$$

де s_i – відповідний i-тий стан в конкретний момент часу t_l з множини з потужністю $|S|=n$, p_{ij} – відповідна ij-та властивість стану з множини з потужністю $|P|=m$, яка визначає поведінку КЛС:

$$p_j = (r_{ij1}, r_{ij2}, \dots, r_{ijv}),\ j = \overline{1,m}, \tag{1.2}$$

де r_{ijk} – відповідний параметр конкретної властивості p_{ij} для стану s_i.

Для будь-якої КЛС станом s_i можуть бути один із процесів опрацювання природної мови, наприклад, ідентифікація ключових слів і/або стійких словосполучень для наступного стану s_{i+1} системи як рубрикація текстового масиву даних. Відповідно властивостями для стану s_i є морфологічна p_{i1}, лексична p_{i2} та синтаксична p_{i3}, в окремих випадках для точності аналізу може бути і семантична тощо. Тоді для властивості p_j буде визначається множина параметрів для відповідного аналізу тексту в залежності від конкретної NLP-задачі. За цими параметрами уточнюють стратегію функціонування КЛС в біжучий момент часу [220-227]. Наприклад, для:

- морфологічної властивості p_{i1} параметрами є N-grams та морфеми: корені r_{i11}, закінчення r_{i12}, афікси r_{i13}; граматичні категорії різних POS r_{i14}, довжина слова r_{i15}, місцерозташування слова в реченні 6, кількість складів у слові r_{i17}, кількість змісту слова r_{i18}, співвідношення приголосних та голосних r_{i19}, тощо;
- лексичної властивості p_{i2} параметрами є місцезнаходження речення в тесті r_{i21}, місцезнаходження слова в реченні r_{i22}, вага слова r_{i23}, вага речення r_{i24}, основа слова r_{i25}, флексія слова r_{i26} тощо;
- синтаксичної властивості p_{i3} параметрами є глибина слова в дереві залежностей речення r_{i31}, місцерозташування слова в реченні r_{i32}, кількість змісту слова r_{i33}, кількість слів на речення r_{i34}, кількість слів r_{i35} та речень r_{i36}, чи є слово з великої літери r_{i37}/ з дефісом r_{i38}/ складеним r_{i39} тощо;
- семантичної властивості p_{i4} параметрами є кількість змісту слова r_{i41}, глибина слова в дереві залежностей речення r_{i42}, розмір абзаців r_{i43}, розміщення абзаців r_{i45} тощо;

В залежності від кортежу властивостей p_j визначається поведінка КЛС, тобто реалізація множини правил (активація дій або подій) реалізації конкретного NLP-процесу для досягнення певної мети в залежності від вхідних текстових даних [228-230]. Відповідно подією o_l є зміна однієї властивості на іншу $p_{ij} \rightarrow p_{ik}$ або $o_l: p_i \rightarrow p_j$ згідно виконання певних умов U для вхідного аналізованого тексту X та проміжного опрацьованого тексту C:

$$p_i = o_l(p_j, U, X, C), \tag{1.3}$$

Дія d_g є процесом активації події o_l іншою подією o_v в КЛС [228-230]:

$$C' = d_g(o_l \circ o_v). \tag{1.4}$$

Чим складніша мова (морфологія, синтаксис, тощо), тим складніше автоматизувати опрацювання відповідних текстів природною мовою. Крім того, для таких мов як українська не стандартизовані правила та словники опрацювання тестів природною мовою для розв’язку відповідних NLP-задач.

Багато наукових лінгвістичних шкіл [231-255] та ІТ-фахівців [256-268] працюють над створенням україномовних словників [269-276] та правил для опрацювання українських текстів. Але зазвичай це лінгвісти та філологи [231-255], які не ознайомлені з особливостями конкретних сучасних інструментів, як мови програмування, методи машинного навчання, BigData аналізу тощо. Існує колосальна прогалина між результатами дослідження філологів та прикладних лінгвістів з одного боку [231-255], та ІТ-фахівцями з іншого [256-268] для опрацювання україномовних тестів. Крім того, сьогодні досить мало реалізовано та впроваджено для загально доступу NLP-інструментів для української мови.

1.2.3. Класифікація та особливості основних властивостей станів комп'ютерної лінгвістичної системи

Кожний стан s_i КЛС для розв'язку конкретної NLP-задачі використовує декілька або всі рівні NLP-процесів для формування кортежу головних властивостей $s_i = (p_{i1}, p_{i2}, \dots, p_{im})$, $i = \overline{1,n}$ [256-259]. Зміст мовлення будь-якої людини, незалежно від конкретного подання інформації (письмове чи звукове) передається кожною із шести властивістю p_{ij} NLP-процесу або рівнем аналізу людської мови незалежно від походження (Рис. 1.6) [277-282]:

$$S_{LA} = d(p_{\text{I}}, p_{\text{II}}, p_{\text{III}}, p_{\text{IV}}, p_{\text{V}}, p_{\text{VI}}). \quad (1.5)$$

I. Фонологічний рівень	Організація та інтерпретація звуків мовлення
II. Морфологічний рівень	Ідентифікація та аналіз структури та форми слів
III. Лексичний рівень	Поділ на розділи ,абзаци, речення та слова
IV. Синтаксичний рівень	Аналіз слів як граматичної структури речення
V. Семантичний рівень	Визначення змісту речення у контексті тексту
VI. Прагматичний рівень	Інтерпретація речень у відповідних контекстах

Рис. 1.6. Класифікація основних підпроцесів опрацювання природної мови

Основні NLP-задачі тісно взаємопов'язані. Тому частина процесів в КЛС для розв'язку різних NLP-задач є подібними або навіть частково однаковими. Наприклад для NLP-задач машинного перекладу, виправлення граматичних помилок, визначення ключових слів, рубрикація тексту тощо точно необхідно

застосовати морфологічний та синтаксичний аналізи тексту. А процес рубрикації тексту обов'язково включає в себе процес визначення ключових слів. Процеси реферування та семантичного анотування включають в себе не лише морфологічний та семантичний аналізи, але і семантичний аналіз та визначення ключових слів. Будь-який аналіз тексту має включати в себе лексичний рівень. Крім того кожний рівень аналізу тексту для розв'язку конкретної NLP-задачі може складатися з різних послідовностей кроків та їх кількості. Та і NLP-методи застосовують для відповідних аналізів в межах розв'язку конкретної NLP-задачі різні. Для зручності основні підпроцеси NLP розділяють на лінгвістичні категорії [68-87, 283-297], які розв'язують певними методами (Рис. 1.7).

Правила опрацювання текстового контенту генерують згідно цих методів. Для ефективнішого NLP контенту необхідно та достатньо, щоб в КЛС була реалізована максимально можлива кількість модулів відповідних мовленнєвих рівнів для розв'язку конкретної NLP-задачі. Кожний із відповідних рівнів аналізу текстового контенту має власну множину методів для досягнення ефективних конкретних результатів в залежності від мови тексту [68-87, 283-297].

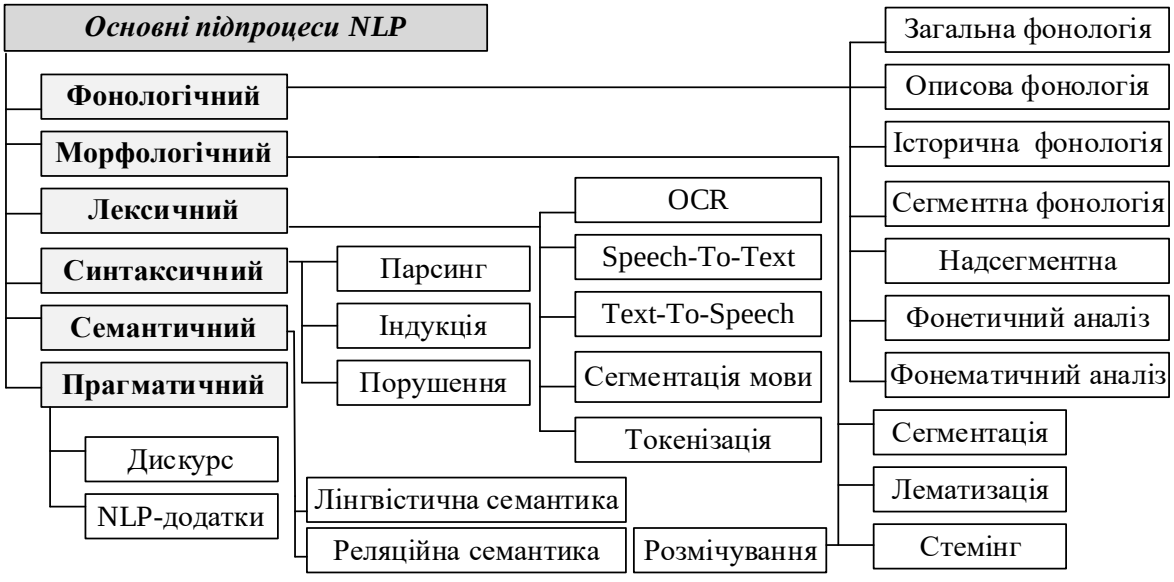


Рис. 1.7. Класифікація основних методів опрацювання природної мови

Наприклад, при визначенні множини ключових слів в англійському тексті використовують парсинг, стемінг та різні статистичні методи для аналізу частоти вживання слів іменникової групи та їх розподіл по тексту [164-170]. Відповідно

для україномовних текстів при визначенні ключових слів простий алгоритм стемінгу необхідно замінити на модифікований алгоритм стемінгу із-за наявності великої кількості різноманіття флексій в аналізованому тексті для ідентифікації іменної групи [19-22, 169-170, 209-213]. Крім того, порівнювати треба між собою не слова, а основи слів іменної групи [209-213, 256-259], так як в українській мові часто ключові слова визначаються не лише послідовністю слів, але можлива їх взаємна перестановка в різних відмінках (наприклад, *пошук інформації* – основи *пошук інформац*, *інформаційний пошук* – основи *інформац пошук*, *пошуку інформації* – основи *пошук інформац*, тощо, тобто відсікання не лише флексій, але і суфіксів для приведення до основи слова) [209-213, 256-259].

1.3. Класичні підходи та напрями опрацювання природної мови

1.3.1. Класифікація основних NLP-підходів

Для опрацювання природної мови використовують індукцію, дедукцію, метод гіпотез, аналіз та синтез, спостереження, ідеалізацію, моделювання та формалізацію [298-303]. Крім того, застосовують спеціалізовані підходи для дослідження явищ та закономірностей конкретної природньої мови як об’єкта комп’ютерної лінгвістики (Рис. 1.8) [212, 298-306]. Ці підходи дозволяють визначити множину процедур та алгоритмів аналізу мовленнєвих явищ для вирішення конкретної задачі та відповідно перевірки отриманих результатів при експериментальній апробації [305]. Зазвичай для конкретної задачі NLP використовують гібридний підхід як комбінування декількох різних підходів (Рис. 1.8) [212, 298-306]. Наприклад методи статистичного аналізу, ймовірнісного моделювання та МН поряд з лінгвістичним підходом використовують для визначення авторства тексту, стилістики окремого автора, у дешифруванні, стенографії, лінгво-дидактиці, реферуванні, зняття полісемії та ІП [305]. Статистичні методи застосовують в контент-аналізі для ідентифікації стану соціальної свідомості або емоційного забарвлення для просування відповідної політичної та/або комерційної реклами в соціальних мережах [305].

Рис. 1.8. Класифікація основних підходів опрацювання природної мови

При лінгвістичному моніторингу окрім переліченої множини методів використовують регулярні вирази та мішок слів при дослідженні функціонування мови у конкретному науковому, політичному або ЗМІ дискурсі. Метою моніторингу є також ідентифікація іншомовних запозичень, плагіату/рерайту, граматичних/стилістичних помилок, лексики емоцій/почуттів, тематичної/просторової/часової лексики тощо.

1.3.2. Загальна класифікація напрямів дослідження для NLP-задач

Відповідні підходи використовують для розв’язку конкретних NLP-задач у типових КЛС-системах при відповідних напрямах дослідження (Рис. 1.9) [212, 298-309]. Але при розв’язку кожної NLP-задачі при застосуванні конкретних підходів в залежності від мови дослідження використовують різну множину інструментів для успішного ефективного досягнення поставленої мети.

Наприклад, аналіз та ідентифікації психологічних ефектів, закладених автором текстового контенту залежить від наявності персоналізованого словника автора та сентимент-словника цього регіону (не всі слова мають ті ж самі емоційні забарвлення і різних мовах та в різних регіонах ще і різних людей

конкретних людей – простий переклад не допоможе отримати реальний опис психологічного стану особи) [310-313].

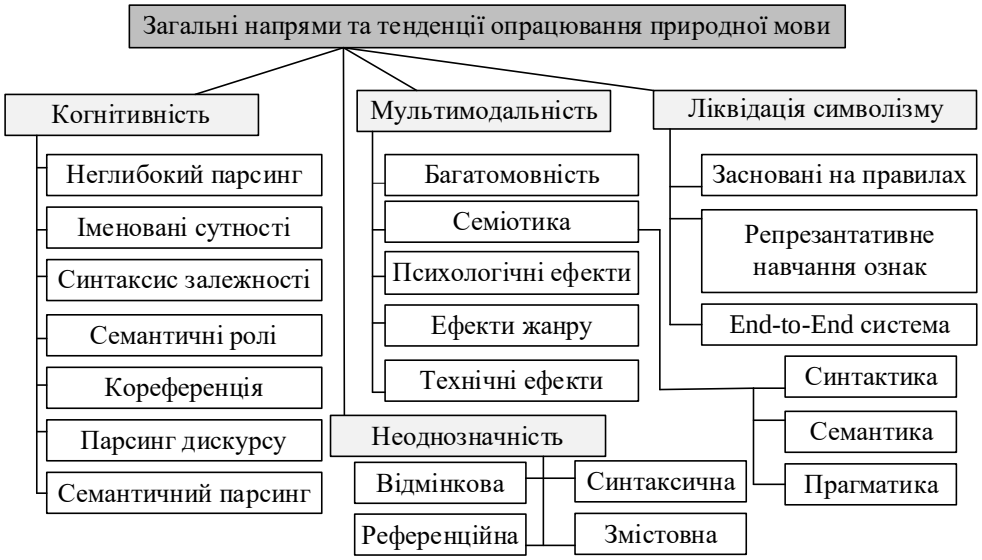


Рис. 1.9. Загальна класифікація напрямів дослідження для NLP-задач

Наприклад, за BigFive-моделлю визначають 5 показників психологічного стану особи за його коментарями в соціальних мережах за певний період часу, зокрема, рівні Extraversion екстраверсії/інтроверсії або амбіверсії/, доброзичливості Agreeableness, відкритості досвіду Openness, невротизму Neuroticism та сумлінності Conscientiousness [314-322]. Для аналізу рівня Extraversion досліджують лексичні міри у вигляді аналізу використаних в текстах множини слів-маркерів, які відповідно відображають риси конкретного типажу. Одна множина маркерів класифікуються як активний, товариський, балакучий, компанійський, комунікабельний, а інша множина як стриманий, тихий, пасивний, задумливий, тощо. Маркерами можуть бути не лише прикметники та іменні групи, але і дієслівні групи в певному часовому відмінку як опис дій в часі (активних або відповідно пасивних). І на цьому етапі виникає складність в синтаксичному та семантичному аналізі в залежності від мови автора тексту. В англомовному тексті, особливо в розмовних діалогах є чіткий порядок груп слів (іменникова, дієслівна), порівняно з українськими текстами. Крім того, середня довжина речень значно менша в англомовному тексті. Тому для них легше побудувати синтаксичне дерево залежностей для аналізу змістовності маркерів, а не лише їх наявність (як у відомій цитаті з казки – де кому поставити у фразі

казнить нельзя помиловать; наявність у відповідному місці знаку пунктуації визначить рівень доброзичливості Agreeableness автора крилатої фрази).

1.3.3. Додаткові методи лінгвістичного дослідження для NLP-задач

Для вищого рівня NLP-додатків застосовують додаткові методи (Рис. 1.10) [323-325]. Когнітивно-ономасіологічний аналіз ідентифікує мотиваторів та мотиваційну базу фразеологічних одиниць для інтерпретації та моделювання структури знань ПО текстової інформації та семантичної залежності між мотиватором та фразеологічною одиницею української мови [326-331].

Рис. 1.10. Методики лінгвістичного дослідження для NLP-задач

Описовий метод використовують як частину ФА, ЛА, МА та СА як інвентаризації l_1, сегментації l_2, таксономії l_3 та інтерпретації l_4:

$$L = l_4(l_{41}, l_{42}) \circ l_3 \circ l_2 \circ l_1. \quad (1.6)$$

Внутрішня інтерпретація l_{41} групує лінгвістичні одиниці за множиною критерій [332-333]. Зовнішня інтерпретація l_{42} ілюструє зв'язки лінгвістичної одиниці із змістовним явищем, об'єктами, суб'єктами та змодельованими подіями [332-333] конкретних текстових потоків контенту відповідної мови.

Порівняльно-історичний метод застосовують для аналізу спорідненості мов на основі зовнішньої (залучення даних) і внутрішньої (співвідношення явищ) реконструкції, лінгвостатистики та лінгвогеографії [334-338].

Зіставний метод застосовують для ідентифікації специфічних та спільних характеристик аналізованих текстів мовлення у граматичній словниковій і звуковій системах [339-341] на основі порівняння для формування критерію як еталона зіставлення внутрішньої форми, ономасіологічної структури, та словотвірних типів на словотворчому рівні, компонентний склад значень порівнювальних еквівалентів на лексичному рівні для систем машинного перекладу [342-347], діалогових систем [68-73, 348-353] та чатботів [354-359].

Типологічний метод застосовують для ідентифікації та групування основних лінгвістичних ознак і закономірностей мовлень (кластеризація) на основі розбіжності та подібності лінгвістичних характеристик [360-362]. Для дослідження використовують мову-еталон, наприклад, синтаксичні, морфологічні та фонетичні моделі, семантичне поле, граматичні правила, лінгвістична категорія, конкретна мова, штучна мова тощо [360-362].

Структурний метод використовують для дослідження структури мовлення в методиках трансформаційній, безпосередніх складників, дистрибутивній і компонентного аналізу [363-365]. Аналізують синтагматичні, парадигматичні, епідигматичні відношення між реченнями, грамемами, лексемами, морфемами, фонемами [323]. Дистрибутивний аналіз ідентифікує ознаки та функціональні характеристики лінгвістичних одиниць з врахуванням середовища (дистрибуції) [366]. Аналіз за безпосередніми складниками ґрунтується на почерговому поділу лінгвістичної одиниці (речення → словосполучення → слова) на складові до того моменту, коли отримаємо неподільні частини [367-379]. Трансформаційний аналіз ідентифікує семантичні та синтаксичні відмінності та подібності між лінгвістичними одиницями через ознаки в множинах їх трансформацій при дослідженні лексичної семантики, словотвору, морфології та синтаксису [363-365, 380]. Компонентний аналіз застосовують для визначення лексичного значення слова як семи (еталонно організована множина елементарних змістовних лексичних одиниць) для формування тлумачних словників [381-382].

Також при дослідженні та опрацювання природної мови застосовують методи математичні (статистика та закономірності) [212, 304-306, 383-393],

психолінгвістичні (асоціативний експеримент, велика п’ятірка) [310-322, 394-395], соціолінгвістичні (аналіз через опитувальники) [396-399], тощо [400-407]. Досить цікаві результати дають методи психолінгвістичного аналізу як асоціативний експеримент (вільний, спрямований або ланцюговий) через семантичний диференціал [408-410]. Останній є якісним та кількісним індексуванням змісту слова через двох-полюсні шкали із градацією парою антонімічних прикметників [411-415].

1.3.4. Методи дослідження когнітивної лінгвістики

Когнітивна лінгвістика поєднує знання та дослідження з психології та лінгвістики засобами ІТ та методами ШІ. Підходами КЛ є генеративна граматика (Generative grammar, автор Avram Noam Chomsky) [416-430], когнітивна лінгвістика (Cognitive Linguistics або linguistics framework, автор George Philip Lakoff) [431-432] та інтегративна когнітивна лінгвістика (Integrative cognitive linguistics або cognitive semantics) [433-437]. За George Philip Lakoff КЛ поділяється на теорію концептуальної метафори (Conceptual metaphor theory або аналіз метафор) та когнітивну і конструктивну граматику (Cognitive and construction grammar або аналіз конструкцій як при форма-значення із порівнянням з мемами як з одиницями мовної/мовленнєвої еволюції) [433-440]. Джордж Лакофф пропонує методологію побудови алгоритмів NLP з точки зору когнітивної науки, разом із виведеними когнітивної лінгвістики, із 2 аспектами:

1. Застосуйте теорію концептуальної метафори для розуміння одного змісту лінгвістичної одинці (слова, фрази, речення чи фрагмента тексту) на основі іншого для ідентифікації наміру автора [431-440].
2. Призначте відносні міри значення аналізованій лінгвістичній одинці на основі інформації, представленої до та після фрагмента тексту, що аналізується, наприклад, за допомогою імовірнісної безконтекстної граматики (PCFG). Математичне рівняння для таких алгоритмів подано в патенті США 9269353 [441]:

$$w(t_N) = p(t_N) \times \frac{1}{2d} (\sum_{i=-d}^{d} (p(t_{N-1}) \times f(t_N, t_{N-1}))_i) \quad (1.7)$$

де w – відносна міра значення; t – токен, будь-який блок тексту, речення, фрази чи слова; N – кількість аналізованих токенів; p – це ймовірна міра значення, заснована на корпусах; d – розташування маркера вздовж послідовності токенів $N-1$; f –функція ймовірності, специфічна для мови.

За класифікацією Л.А. Ковбасюк КЛ поділяють на такі напрями [442-443]:

1. Когнітивна поетика – дослідження когнітивних процесів на базі яких продукується, сприймається та інтерпретується текстовий масив даних [449];
2. Фреймова семантика досліджує когнітивні моделі та ментальні простори (фрейми) [444-453];
3. Концептуальна метафора та концептуальна метонімія [449, 454-456] (аналіз, ідентифікація та інтерпретація змісту на основі іншого);
4. Теорія семантичних прототипів (структурування категорії та ідентифікація складових на основі заданого прототипу, наприклад прототип *собак – вівчарка* або *маламут, котів – Шотландський висловухий, птахів – орел* тощо) [449-451].

Підходи для розроблення когнітивних моделей застосовують в когнітивній [444-456], функціональній [457-459] та конструктивній граматиці [460-465], комп'ютерній психолінгвістиці та когнітивній нейронауці (наприклад, ACT-R або Adaptive Control of Thought-Rational – адаптивний контроль думки-раціональності, автори Christian Lebiere and John Robert Anderson з Університету Карнегі-Меллона) [466-470]. Напрями дослідження когнітивного NLP є частиною підходу когнітивний ШІ [471-479], в тому числі на основі нейронних моделей для мультимодальних NLP [480-484].

1.4. Основні методи та методики опрацювання природної мови засобами машинного навчання

1.4.1. Класифікація основних ML-методів для NLP-процесів

Кластеризація та класифікація великих текстових масивів даних зазвичай здійснюється на основі ML-методів (Machine Learning) [480-484], аналізу

великих даних (Big data analysis) [485-499]. Для побудови таких методів використовують засоби теорії графів, теорії ймовірностей, методів оптимізації, математичного аналізу, чисельних методів, математичної статистики, різні техніки роботи з даними в е-формі [212, 304-306, 383-393, 500-511]. КЛС основі машинного навчання складається з основних частин як NLP, кластеризація та класифікація (Рис. 1.11) [490-514].

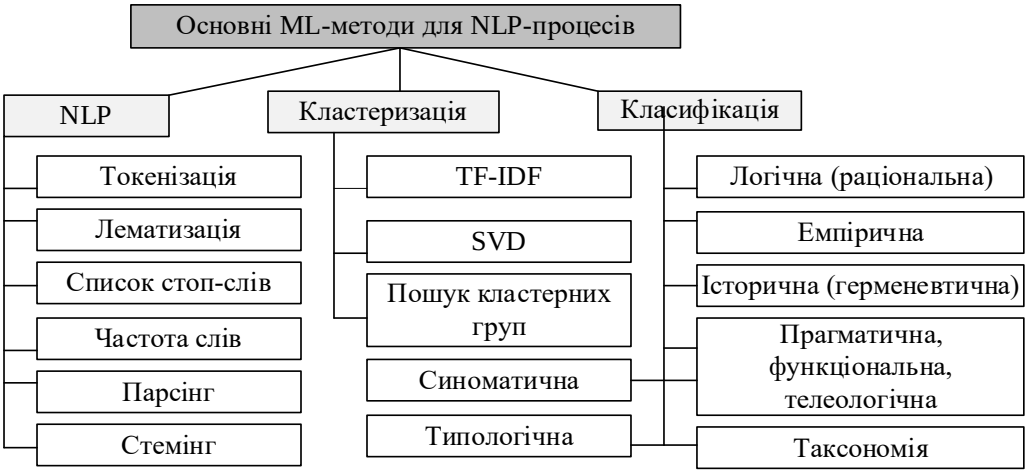


Рис. 1.11. Класифікація основних ML-методів для NLP-процесів

Дослідження текстів є однією з найскладніших задач для програмістів через неоднозначність значення слів. Деякі компанії, наприклад Alchemy і Thomson Reuters розробили NLP-служби та ML-алгоритми для ідентифікації змісту тексту [515-521]. Компанія Aylien запропонувала власний інструментарій API для аналізу тексту для можливості створення різних NLP-сервісів [522-525].

API дає можливість швидко ідентифікувати в документі заголовки і основний текст, виділити зміст і основні концепції, скласти реферат або анотацію [522-525]. Виділені з тексту дані зберігаються в форматі JSON, для забезпечення доступу до них використовується Mashape [526-529]. Нажаль інструмент працює лише з англійською та німецькою мовами [522, 530].

Основні NLP-задачі є розроблення алгоритмів вилучення та аналізу ознак лінгвістичних одиниць виміру з мови та застосувати їх для вирішення більш широкого кола задач КЛ [68, 212, 305, 506]. Такими ознаками, зокрема, є [51-53]:

- кількість речень, слів, слів в реченнях тощо;
- розмір та місцерозташування абзаців;

- позиція слова в реченні та довжина слова;
- співвідношення голосних та приголосних;
- кількість складів у слові та значень слова;
- глибина слова в дереві залежностей речення;
- склад морфем: афікси, корені, закінчення;
- N-grams та граматичні категорії різних POS;
- слово з великої літери / з переносом / дефісом / складене.

1.4.2. Основні проблеми опрацювання україномовних текстів

Основними проблемами для розроблення КЛС опрацювання української мови є розщеплення (Splitting) лінгвістичних одиниць [531-535], розмічування частин мови (POS tagging) [169, 273, 534-541], парсинг (Parsing) [157-170] та прагматика (Pragmatics) [542-546], тобто як контекст впливає на зміст [547-550]. Прагматика досліджує такі ознаки як імпликатура [547-550] (неоднозначності висловлювань, натяки, здогадки), мовленнєві дії, актуальність та розмова [551].

Відповідно Workflow (потоки робіт) для NLP для типової задачі є [552-556]:

1. Дослідити наявні дані та NLP-алгоритми;
2. Підготувати тестовий набір та базову лінію та визначити метрики;
3. Розробити NLP-алгоритм: дизайн ознак; NLP-pipeline (налагодження потоку/джерела інформації як конвеєр); NLP-ресурси; обрати підхід заснований на правилах / статистичний / ML;
4. Впровадити та тестувати рішення;
5. Контролювати виконання.

Цікавою NLP-задачею є ідентифікація або генерування вірусних заголовків новин в соціальних мережах або онлайн-газетах. Є декілька ознак, якими має володіти потенційний вірусний заголовок [51]:

- унікальність (uniqueness) назви – відсутність аналогу;
- сусідство або близькість (proximity) – присутність посилання на країну/місто/заклад/район джерела новини;

- надзвичайність (superlativeness) – із вказанням масштабу/розмаху або сили/якості впливу на явище/суб’єкт/об’єкт/середовище;
- емоція (sentiment) – використання емоційного забарвлення мови;
- несподіванка (surprise) – застосування незвичних фраз/зворотів;
- відомість (prominence) – присутність посилання на видатних осіб (людей, місцезнаходження, звання) або події/дії цих осіб.

Семантичний аналіз тексту є однією з ключових NLP-проблем як теорії створення КЛС. Результати семантичного аналізу використовують для вирішення проблем у таких областях, як, наприклад: системи автоматичного перекладу (Google translation); ІІПС (Google повністю заснована на семантичному аналізі); філологія (аналіз авторських текстів); торгівля (аналіз попиту на певні товари на основі коментарів до цього товару); політологія (прогнозування результатів виборів); психіатрія (для діагностики пацієнтів) тощо. Візуалізація результатів семантичного аналізу є важливим етапом його реалізації, оскільки вона може забезпечити швидке та ефективне прийняття рішень на основі результатів аналізу. Аналіз публікацій у мережі з прихованого семантичного аналізу (LSA) показує, що візуалізація результатів аналізу здійснюється у вигляді двокоординатного семантичного просторового графіка із нанесеними словами та документами координат [557-560]. Така візуалізація не дозволяє однозначно визначити групи суміжних документів та оцінити рівень їх семантичного зв'язку за словами в тексті. Для груп слів та документів без візуалізації визначали лише мітки кластерів та координати центроїдів.

1.5. Огляд відомих інформаційних технологій розроблення комп’ютерних лінгвістичних систем

1.5.1. Особливості інтелектуального аналізу потоку контенту

В роботах [561-572] акцентована увага на актуальності та перспективності інтеграції інформаційних потоків на основі прагматичних методів text mining для розв’язку низки ТДЗ-задач, зокрема, реферування, аналізу інформаційних

портретів, контент-аналізі текстів, формування дайджестів, ІІП тощо. Це досить інформативна робота, але не розкриває особливостей опрацювання текстів описаними різними мовами. Так для словосполучення *content analysis* майже не зустрінеш в англійському тесті його інший варіант *analysis of content*. В українському тексті для ключового слова контент-аналіз існують та часто вживанні такі еквіваленти як *контентний аналіз* та *аналіз контенту* та їх аналоги *змістовний аналіз*, *аналіз змісту*. Це ускладнює процес семантичного аналізу для NLP-задач видобування інформації з текстового контенту, що дає не точні результуючі дані. Процес ідентифікації/маркування ключових слів/термінів, ІІП за ключовими словами, інтеграція потоків інформації значно ускладняється, так як в текстовому контенті українською мовою можуть набувати різні форми із-за відмінювання із зміною флексій в залежності від роду/множини іменника та прикметника, наявності суфіксів, чергування літер при словозміні тощо.

В [13-15, 573-583] доводиться для ефективності ІІП краще застосовувати онтологічний підхід для англомовного контенту. Це досить ефективно при видобуванні знань з українських текстів лише в тому випадку, якщо перед цим проведені детальні та коректні морфологічні, графемні, лексичні та семантичні аналізи. Побудувати онтологію можливо лише із коректним визначенням та відповідним маркуванням всіх зав'язків між всіма сутностями із подальшим їх збереженням в формах (для дієслів в неозначеній формі - інфінітив, для іменників у формі називного відмінка однини та прикметників у формі називного відмінка чоловічого роду). Тобто для речення – *комп'ютерна лінгвістична система розв'язує конкурентну задачу опрацювання природної мови* відповідний аналог буде в граматичному дереві залежностей у вигляді листя за синтаксичним аналізом *комп'ютерний лінгвістичний система розв'язувати конкурентний задача опрацювання природний мова*. Без аналізу цього дерева неможливо ідентифікувати залежності слів та їх підпорядкування в реченнях для побудови відповідної онтології автоматично на основі пресингу.

В [561-572] акцентована увага на актуальності та перспективності аналізу змін в текстових потоках контенту на основі лінгвістичного аналізу, в тому числі

для ІІП інформації у вигляді текстового контенту. Опрацювання текстової інформації подано як набір операторів формування, управління та супровід множини текстового контенту *C*. Аналогічно як і в попередніх роботах, всі методи наведені для опрацювання контенту без прив'язки до особливостей конкретної мови. В більшості публікаціях про особливості ІІП інформації рекомендують розглядати множину аналізованого контенту *C* як сукупність підмножин релевантного C_{rt} та нерелевантного C_{rf} контенту, або знайденого C_{st} та незнайденого C_{sf} контенту, або корисного контенту C_{ut} для кінцевого користувача та некорисного C_{uf}, або часто відвідуваного контенту користувачами C_{vt} та рідко відвідуваного C_{vf}, або час ознайомлення з контентом більше певного значення C_{pt} або менше C_{pf}:

$$C = C_{rt} \cup C_{rf} = C_{st} \cup C_{sf} = C_{ut} \cup C_{uf} = C_{vt} \cup C_{vf} = C_{pt} \cup C_{pf}. \quad (1.8)$$

Результат ІІП текстового контенту оцінюють за відповідними критеріями як степінь релевантності k_1, актуальності k_2, популярності k_3, достовірності k_4, унікальності k_5, тощо. В табл. Таблиця 1.2 подано формули визначення критеріїв ефективності ІІП. Кожний з перелічених критеріїв має свою шкалу оцінювання для формування рейтингу результату ІІП [584-594]. Точний розрахунок кожного з критеріїв в конкретному з результатів ІІП не покращує його якість. Покращення значення одного із критеріїв призведе до погіршення іншого. Знаходження балансу значень критеріїв ІІП трудомісткий процес та не дає жодних позитивних результатів. А ось знати, які із показників краще використовувати при конкретній меті ІІП значно полегшить отримання очікуваного результату.

Таблиця 1.2

Критерії формування результату ІІП текстового контенту [584-594]

k_i	Назва	Зміст	Формула
k_1	Релевантність	Відповідність кількості ключових слів n знайденого контенту до кількості ключових слів N запиту ІІП та M знайденого контенту, або відношення корисного користувачу до загального	$\frac{n}{N}$, $\frac{n}{M}$, або $\frac{n^2}{MN}$ для конкретного контенту, а для всього знайденого тоді $\frac{\lvert C_{ut} \cap C_{st} \rvert}{\lvert C_{uf} \cap C_{sf} \rvert + \lvert C_{ut} \cap C_{st} \rvert}$
k_2	Популярність	Відношення часто відвідуваного та більше певного часу переглянутого контенту до всього знайденого релевантного контенту	$\frac{\lvert C_{vt} \cap C_{pt} \rvert}{\lvert C_{rt} \cap C_{st} \rvert + \lvert C_{vt} \cap C_{pt} \rvert}$

k_i	Назва	Зміст	Формула
k_3	Актуальність	Відношення часто відвідуваного корисного користувачу контенту до всього знайденого релевантного контенту	$\frac{\lvert C_{vt} \cap C_{ut} \rvert}{\lvert C_{vf} \cap C_{st} \rvert + \lvert C_{vt} \cap C_{ut} \rvert}$
k_4	Достовірність	Відповідність контенту реальним значенням та надійності джерела p_s	$\frac{k_1}{p_s}, p_s = [0;1]$
k_5	Унікальність	Показник оригінальності авторських даних контенту по відношенню до знайдених	$\frac{k_{1i} \cdot \lvert C_{st} \rvert}{\sum_{i=1}^{\lvert C_{st} \rvert} k_{1i}}$
k_6	Автентичність	Показник відповідності джерелу походження та авторству контенту	$\frac{k_{1i} \cdot \lvert C_{st} \rvert}{\sum_{i=1}^{\lvert C_{st} \rvert} k_{1i}} \cdot p_s, p_s = [0;1]$
k_7	Прибутковість	Відношення кількості повернень $\lvert C_{rp} \rvert$ до загальної кількості перегляду контенту	$\frac{\lvert C_{rp} \cap C_{st} \rvert}{\lvert C_{vf} \cap C_{st} \rvert + \lvert C_{vt} \cap C_{st} \rvert}$
k_8	Повнота	Відношення знайденого релевантного контенту до всього можливого релевантного контенту	$\frac{\lvert C_{rt} \cap C_{st} \rvert}{\lvert C_{rt} \cap C_{st} \rvert + \lvert C_{rt} \cap C_{sf} \rvert} = 1 - k_{13}$
k_9	Точність	Відношення знайденого релевантного контенту до всього знайденого контенту	$\frac{\lvert C_{rt} \cap C_{st} \rvert}{\lvert C_{rt} \cap C_{st} \rvert + \lvert C_{rf} \cap C_{st} \rvert} = 1 - k_{10}$
k_{10}	Шум	Відношення знайденого нерелевантного контенту до всього знайденого контенту	$\frac{\lvert C_{rf} \cap C_{st} \rvert}{\lvert C_{rt} \cap C_{st} \rvert + \lvert C_{rf} \cap C_{st} \rvert} = 1 - k_9$
k_{11}	Осад	Відношення знайденого нерелевантного контенту до всього нерелевантного контенту	$\frac{\lvert C_{rf} \cap C_{st} \rvert}{\lvert C_{rf} \cap C_{sf} \rvert + \lvert C_{rf} \cap C_{st} \rvert} = 1 - k_{12}$
k_{12}	Специфічність або селективність	Відношення незнайденого нерелевантного контенту до всього нерелевантного контенту	$\frac{\lvert C_{rf} \cap C_{sf} \rvert}{\lvert C_{rf} \cap C_{sf} \rvert + \lvert C_{rf} \cap C_{st} \rvert} = 1 - k_{11}$
k_{13}	Залишок, втрата або мовчання	Відношення незнайденого релевантного контенту до всього можливого релевантного контенту	$\frac{\lvert C_{rt} \cap C_{sf} \rvert}{\lvert C_{rt} \cap C_{st} \rvert + \lvert C_{rt} \cap C_{sf} \rvert} = 1 - k_8$
k_{14}	Невизначеність	Відношення незнайденого релевантного контенту до всього можливого нерелевантного контенту	$\frac{\lvert C_{rt} \cap C_{sf} \rvert}{\lvert C_{rf} \cap C_{sf} \rvert + \lvert C_{rt} \cap C_{sf} \rvert} = 1 - k_{15}$
k_{15}	Неоднозначність	Відношення незнайденого нерелевантного контенту до всього можливого нерелевантного контенту	$\frac{\lvert C_{rf} \cap C_{sf} \rvert}{\lvert C_{rf} \cap C_{sf} \rvert + \lvert C_{rt} \cap C_{sf} \rvert} = 1 - k_{14}$
k_{16}	Пертинентність	Відношення обсягу корисного контенту згідно потреб користувача до загального обсягу знайденого контенту	$\frac{\lvert C_{ut} \rvert}{\lvert C_{st} \rvert}$
k_{16}	Відповідність	Відношення релевантності корисного контенту до знайденого контенту	$\frac{\sum_{i=1}^{\lvert C_{st} \rvert} k_{1i}}{\lvert C_{st} \rvert \cdot \lvert C_{ut} \rvert}$
k_{17}	Задоволеність	Показник ефективності для інтегральної оцінки ІІП	$k_8 + k_9$
k_{18}	Функціональність	Показник ефективності оцінки якості ІІП	$k_8 \cdot k_9$
k_{19}	Уточнення	Відношення коефіцієнта точності до ймовірності релевантності випадкового контенту p_{rc} в масиві	k_9 / p_{rc}
k_{20}	Конверсійність	Відношення відмов $\lvert C_{cv} \rvert$ до загальної кількості перегляду контенту	$\frac{\lvert C_{cv} \cap C_{st} \rvert}{\lvert C_{vf} \cap C_{st} \rvert + \lvert C_{vt} \cap C_{st} \rvert}$
k_{21}	Новизна	Показник ефективності для оцінки якості знайденого релевантного контенту	$k_3 \cdot k_5$

Якщо результатом є задовільнити кінцевого користувача в конкретній ІППС – це одна множина критеріїв, причому навіть в ній є більш важливіші критерії та менш важливіші (показник релевантності може переважати за

показник унікальності, а для розрахунку релевантності використовують критерії точності та повноти з врахування часу читання та частоти відвідування конфетну попередніми відвідувачами) [584-594]. Якщо ж треба ідентифікувати множину Website, де використані деякі сірі/чорні SEO-методи, то застосовують іншу множину критеріїв, зокрема як шум та осад. Стовідсоткова результативність ІІП неможлива із-за суб'єктивізму авторського контенту, наявності шуму завдяки застосуванню сірого/чорного SEO-технологій для просування Website та некоректності створення пошукових образів контенту (ПОК) із складності лінгвістичного опрацювання мов, зокрема української.

1.5.2. Технології інтелектуального аналізу текстового потоку

Однією із розповсюджених ІТ аналізу потоку текстового контенту є інтеграції даних з різних джерел (Рис. 1.12) [561-572]. Зазвичай інтегрують даних зі достовірних джерел за аналізом тегів. Але біль складний процес інтеграції є на основі видобування інформації або даних з різних джерела контенту з виростанням NLP-методів [584-594]. Якісне генерування нового текстового контенту з множини різних за природою, але подібних за змістом даних з різних джерел є на сьогодення однією із актуальних та перспективних NLP-задач, наприклад, для успішного ведення е-бізнесу. Етапи генерування та застосування множини текстового контенту визначають методологію збору, фільтрації, індексації, форматування, структурування інформації з відповідних джерел, та подальше збереження, опрацювання, супровід, формування, управління тощо, тобто основні етапи інтелектуального аналізу потоку текстового контенту (Рис. 1.13) [561-594]. Процес інтелектуального аналізу потоку тексту складається з:

1) інтеграції контенту на основі розпізнавання та аналізу тексту;
2) управління контентом на основі аналізу та опрацювання тексту;
3) супроводу контенту на основі аналізу та синтезу інформації.

Процес інтеграції контенту забезпечує формування контенту на основі методів контент-моніторингу, контент-аналізу, видобування інформації та

розпізнавання мовлення з різних джерел згідно інформаційних потреб постійної/потенційної аудиторії (Рис. 1.14) [561-594], зокрема:

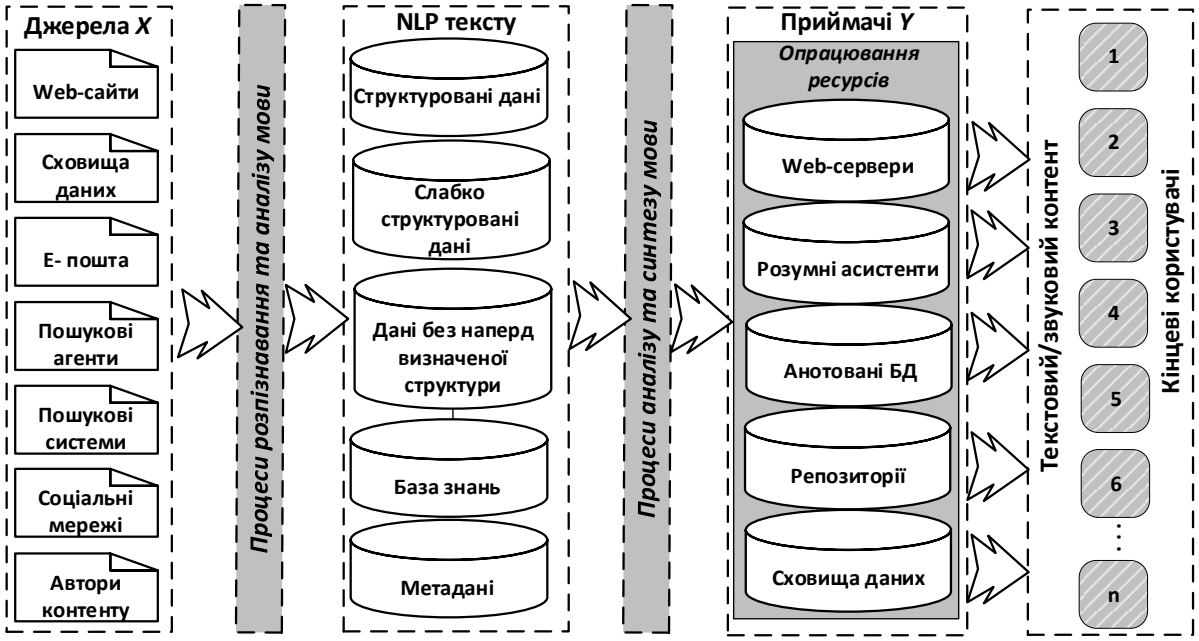


Рис. 1.12. Процес інтеграції тестових даних з джерел *X* в приймач *Y*

Рис. 1.13. Загальна схема інтелектуального аналізу текстового потоку

адміністратор → *правила інтеграції* → *база знань* → *формування множини параметрів ІПП* → *база даних контенту* → *ІПП за параметрами* → *база кешованих даних* → *парсинг джерела* → *видобування інформації* → *анотована база даних контенту* → *формування контенту* → *база даних контенту* → *розподіл контенту* → ***модератор***

модератор → *правила розпізнавання та аналізу тексту* → *база знань* → *формування контенту з інтегрованих даних* → *база даних контенту* → *систематизація контенту* →

база даних контенту → розподіл контенту → ***редактор*** *→ публікація контенту →* ***Website/Webpage***

інтегровані дані *→ збирання контенту → база кешованих даних → форматування контенту → база даних контенту → перевірка та вилучення плагіату → база даних контенту → вилучення дублів → виправлення помилок → база даних контенту → визначення ключових слів → анотована база даних контенту → анотування контенту → анотована база даних контенту → реферування → база даних контенту → рубрикація → база даних контенту → сентимент-аналіз → база даних контенту →формування дайджесту → база даних контенту → поширення контенту →* ***потенційна/постійна аудиторія***

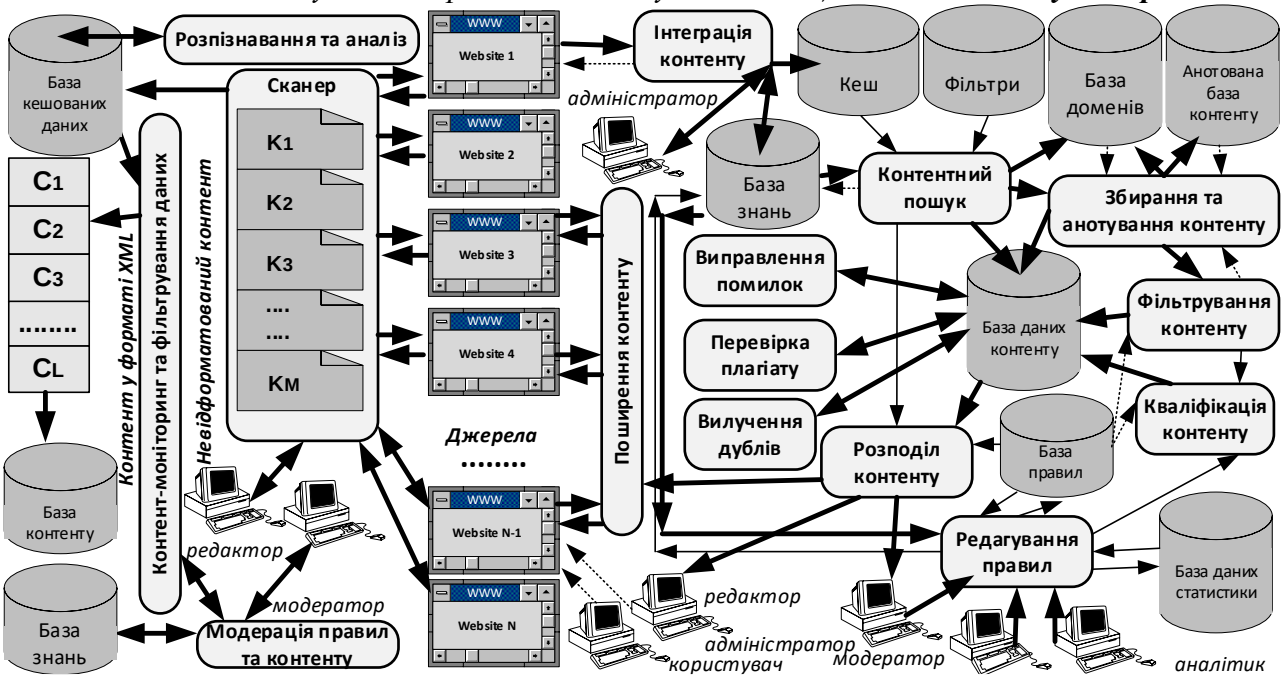


Рис. 1.14. Схема інтеграції текстового контенту зрізних джерел

Процес управління контентом описують наступними відношеннями:

Користувач *→ опрацювання запиту → база даних фільтрів→ контент-моніторинг → база даних контенту → контент-аналіз → форматування контенту → подання контенту →* ***Website/Webpage***

Процес управління контентом Website класифікують за відповідними критеріями формування відповіді на запит користувача (Рис. 1.15) [561-594]:

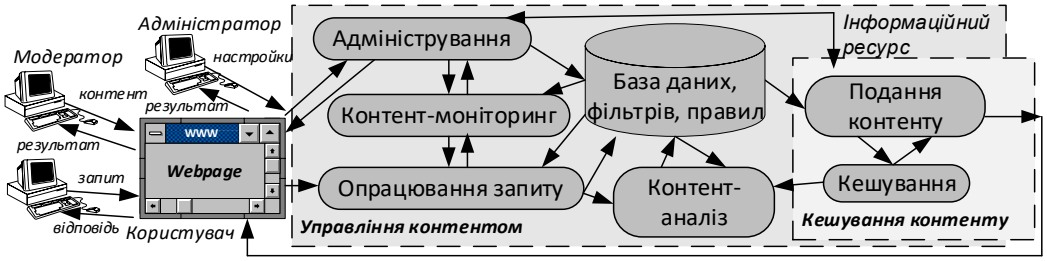
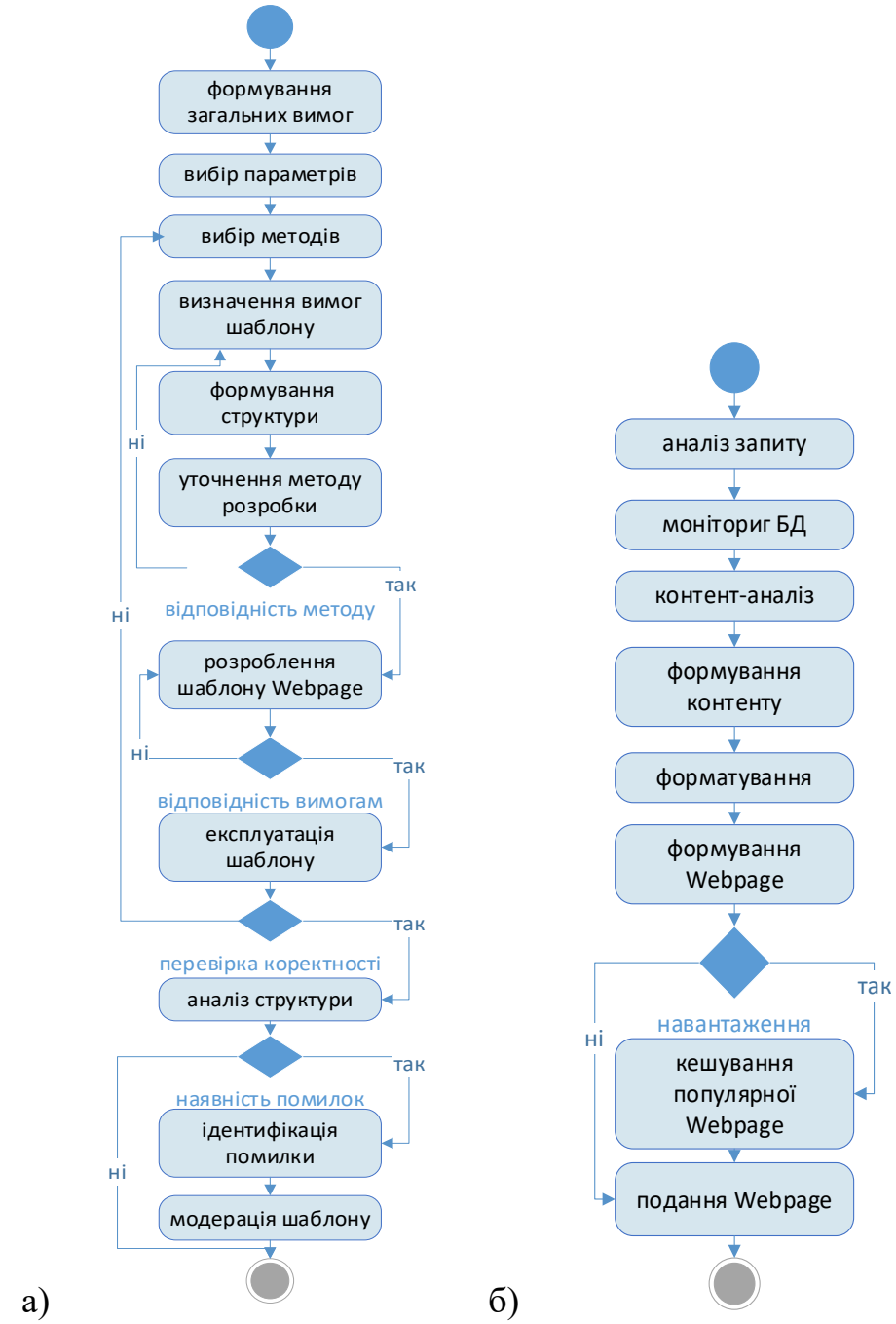


Рис. 1.15. Процес управління контентом за користувацьким запитом

1) Формування змісту Webpage за конкретним персоналізованим запитом користувача з БД в певний момент часу (Рис. 1.15-Рис. 1.16).



а) б)

Рис. 1.16. Генерація а) шаблону та б) Webpage за запитом користувача

Формування Webpage залежить від конкретного запиту кожного користувача постійної аудиторії. Це призводить до значного зростання навантаження на Webserver при кожному користувацькому запиті постійної аудиторії відповідного Website. Навантаження зменшують за рахунок кешування часто запитуваної інформації в певний проміжок часу згідно попереднього статистичного аналізу динаміки запитів [561-594].

2) Формування статичних Webpage при редагуванні модератором Website (Рис. 1.17-Рис. 1.18) [561-594].

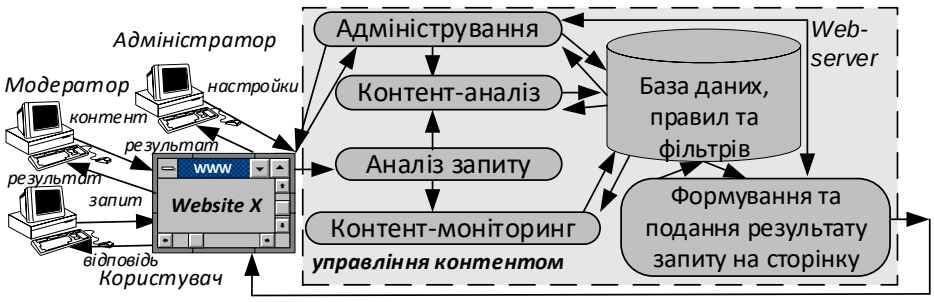
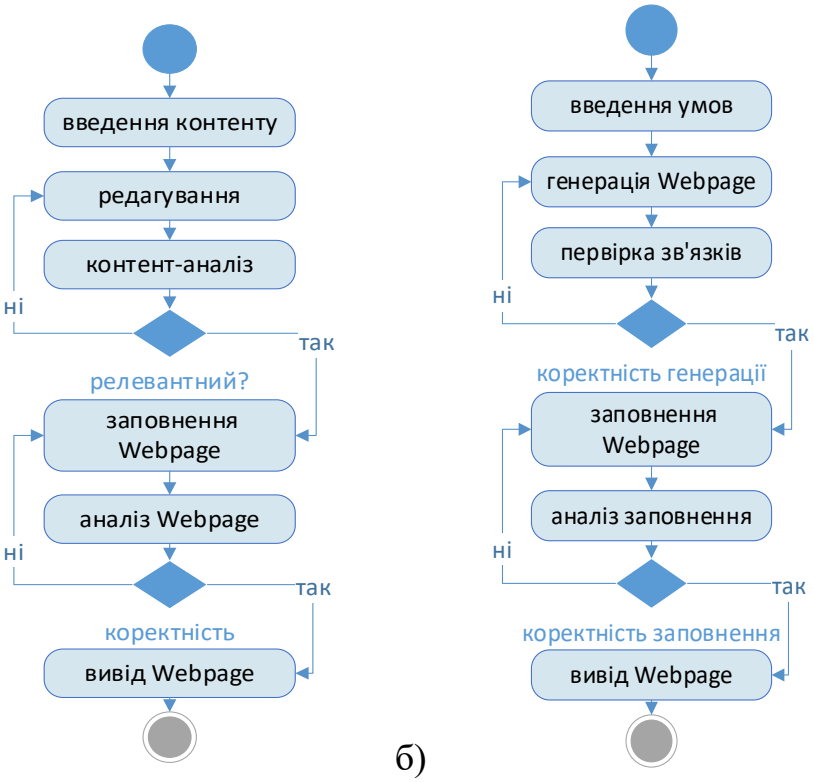


Рис. 1.17. Формування статичних Webpage при модерації



а) б)

Рис. 1.18. а) Генерація та б) заповнення Webpage при модерації контенту

Повнотекстовий контент-моніторинг у великих базах/сховищах даних є неефективним. Проблему оперативності та точності контент-моніторингу вирішує ІІП в анотованих БД. Ефективно застосовувати контент-моніторинг для ІІП тексту за ПОК (шаблонами, анотаціями) із зваженими ключовими словами та стійкими словосполученнями з найбільшими ваговими значеннями [561-594]. Проблему відсутності інтерактивного діалогу між користувачем та Website забезпечує не лише наявність кешування часто запитуваних даних, але і аналіз статистики запитів цього клієнта за певний/весь період часу.

3) Кешування Webpage згідно аналітики запитів (останніх подібних запитів) користувачів та переходів з ІІПС з досягненням конверсії відвідування (Рис. 1.19-Рис. 1.21) [561-594].

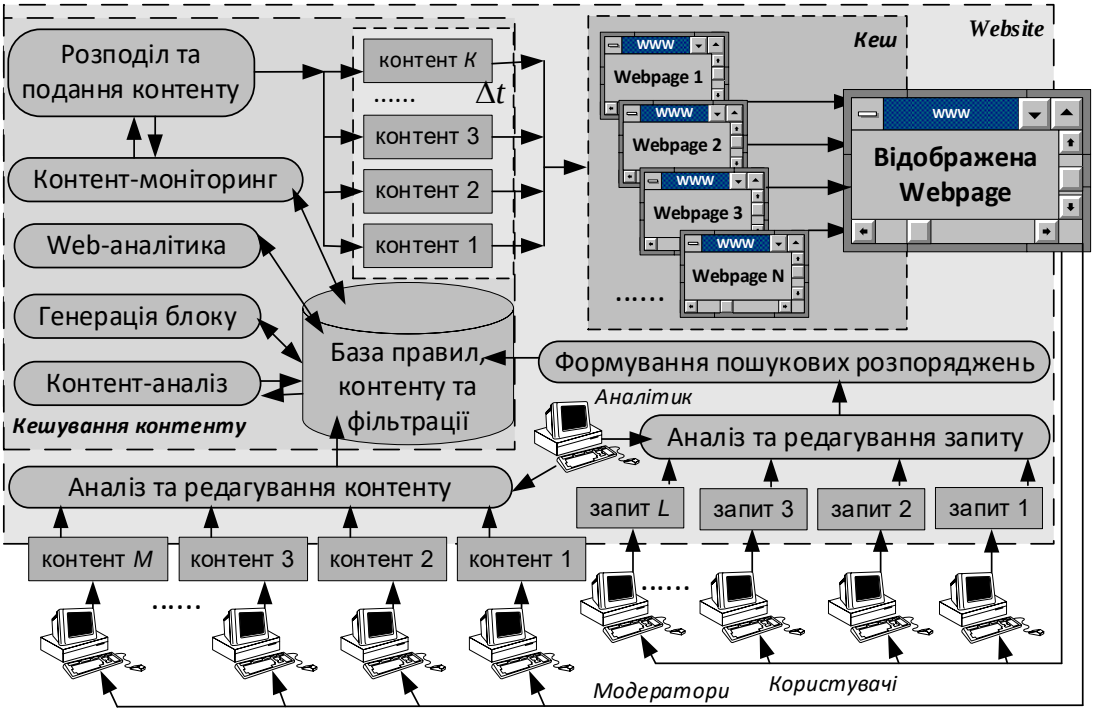
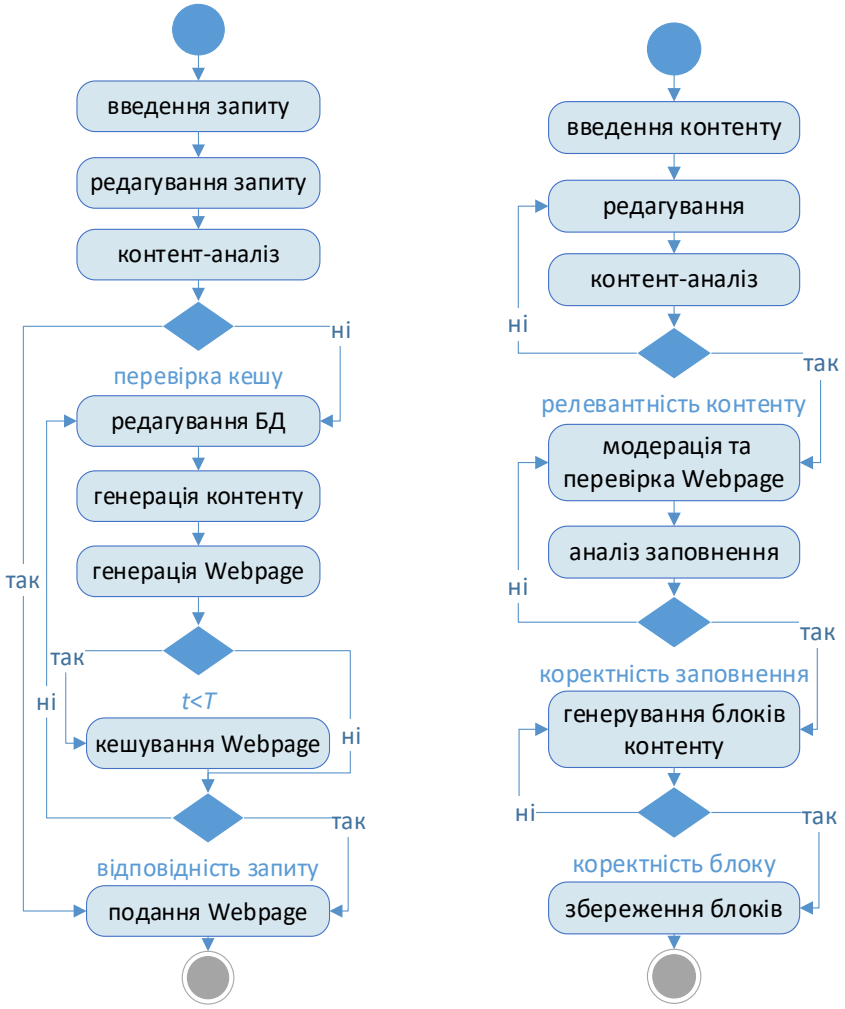


Рис. 1.19. Кешування згенерованих Webpage згідно аналітики запитів

КЛС генерує Webpage один раз в певний момент та зберігає її образ в БД. Webpage в кеші збережена на протязі часу Δt (поки зміст контенту Webpage за певний період $\Delta' t$ буде не затребувана іншими користувачами). Множина Webpage в кеші оновлюється згідно історії запитів постійної аудиторії. До таких Webpage швидше мають доступ користувачі, ніж очікувати заповнення нової Webpage. Кеш оновлюється періодично вручну/автоматично: по закінченню терміну Δt, або Webpage не буде певний час затребувана користувачами (Рис. 1.20), або суттєвій модифікації Website/контенту із змістом цих Webpage.

Аналіз зміни динаміки та періоду часу звернень до відповідних кешовних даних визначає множину тематичних зацікавлень постійної аудиторії (Рис. 1.21). Це також визначає швидкість розвитку потреб кінцевих користувачів до відповідних оперативних тематичних напрямів контенту Website. Відповідний своєчасний аналіз такої динаміки змін запитів та часу зацікавленості аудиторії дозволяє скорегувати наповнення Website відповідним актуальним контентом.



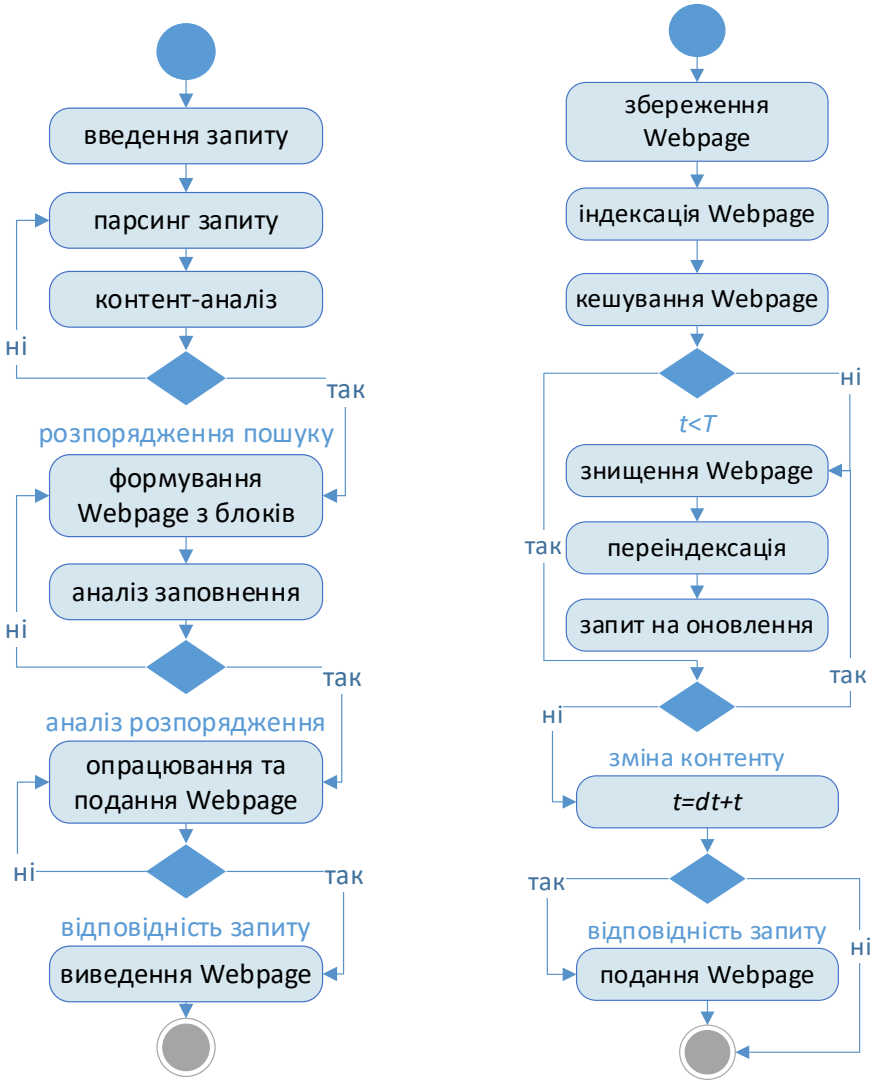
а) б)

Рис. 1.20. Етапи а) генерації Webpage та б) генерування інформаційних блоків для генерування та кешування Webpage згідно аналітики запитів

Статистичний аналіз зав’язків між контентом текстового потоку дозволяє визначити тематичну кореляцію в певний проміжок часу та ефективність посилань для досягнення конверсії відвідувань користувачів (Рис. 1.22) [561-594]. Застосування методів кластерного аналізу дозволяє кількісно оцінити ваги тематичних зав’язків в текстових потоках в певний проміжок часу для передбачення популярності тематики контенту серед кожної групи постійної аудиторії. Це дозволить скорегувати за пріоритетністю кешування відповідних тематичних інформаційних блоків в певний проміжок часу.

Інтелектуальний аналіз українського текстового потоку Website з врахуванням статистичної Web-аналітики досягнення конверсії відвідувань користувачів складніше успішно впровадити з врахуванням оперативного

взаємодії постійного користувача через інтерактивний гнучкий діалоговий інтерфейс для надання доступу до актуального релевантного контенту без надлишку та шуму даних (Рис. 1.23) [561-594].



а) б)

Рис. 1.21. Етапи а) подання Webpage з кешованих блоків та б) кешування Webpage згідно аналітики запитів та переходів з ІІПС

Якісний, ефективний, оперативний та своєчасний аналіз аналітики запитів постійних користувачів, анонімних відвідувачів, переходів з соціальних та ІІПС за множиною тематичних ключових слів, час затримки та дій на конкретній цільовій Webpage, досягнення конверсії для певних тематичних Webpage та відмов для інших тощо значно прискорить процес аналізу певного тематичного текстового потоку контенту для формування інформаційних блоків, їх

кешування та подальшого контент-моніторингу згідно запитів користувачів (Рис. 1.24) [561-594].

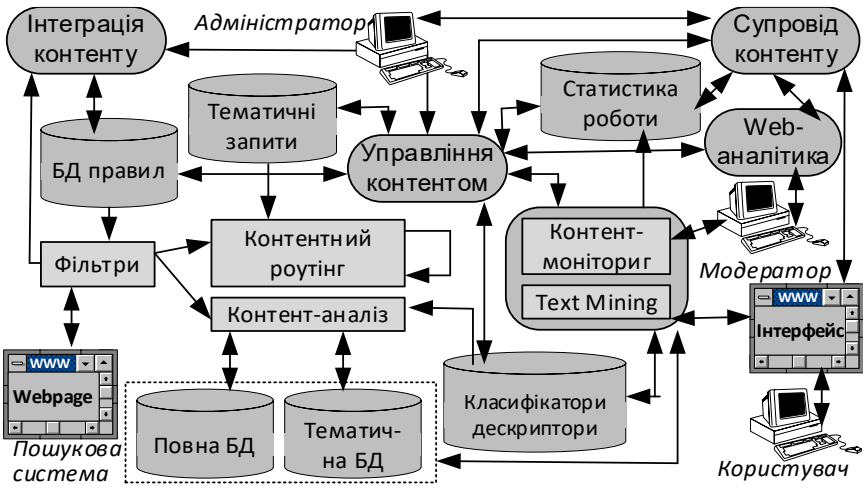


Рис. 1.22. Функціональна схема інтелектуального аналізу тексту

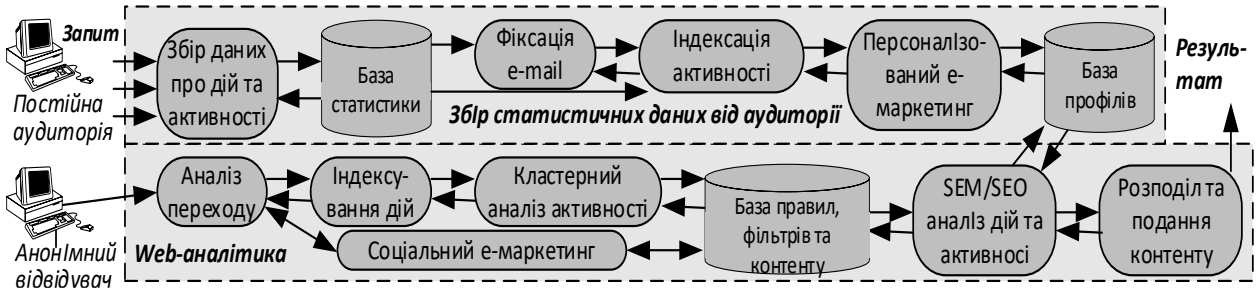


Рис. 1.23. Схема процесу аналізу аналітики запитів користувачів

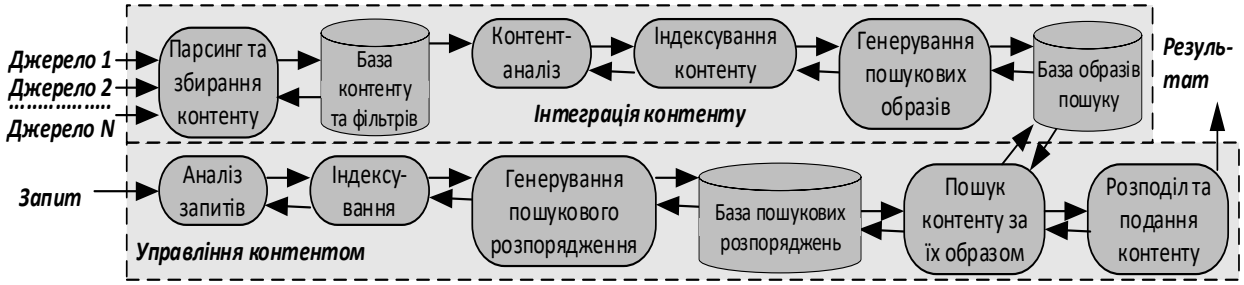


Рис. 1.24. Структурна схема процесу контент-моніторингу тексту

Суттєве зростання обсягу контенту на Website та змінна динаміки, релевантності/точності/тематичності/актуальності потоків текстового контенту (оперативне систематичне оновлення) сприяє зростанню контентного надлишку/шуму/осаду, дублювання, плагіату, рерайту та надмірності запитів/результатів ІІП [561-594]. Інтеграція та контент-моніторинг, контент-аналіз та узагальнення великого обсягу оперативних динамічних потоків

текстових контенту з Internet-джерел як Website вимагає впровадження нових ефективних ІТ ІІП/аналітики тексту.

1.6. Критерії оцінки ефективності КЛС на основі технології машинного навчання та аналізу великих даних

1.6.1. ML-методи аналізу великих даних з множини текстових потоків контенту

На сьогодення IT/NLP-фахівці активно ПЗ опрацювання природної мови на основі машинного навчання [506-514]. Деякі сучасні КЛС на основі ML видобувають/інтегрують/генерують актуальний/релевантний/корисний контент з неопрацьованої/неструктурованої інформації [595]. Такі КЛС не лише аналізують природний текст, але підтримують інтерактивний діалог з користувачем та адаптацію до оперативних змін навколишнього середовища. Результати функціонування КЛС мають бути повними/точними/значущими, використані методи – якісними/ефективними/інтелектуальними, а відповідно організація/структура/архітектура ІС – простою/адаптованою в реалізації [586]. Ці особливості розкривають основну методологію розроблення КЛС, засновану на аналізі природної мови: кластеризація подібного тексту в значущі групи або класифікація тексту на основі конкретних міток, тобто, ML без/з вчителем.

Яскравим прикладом слугує КЛС фільтрації відгуків Yelpy Insights [596] на основі сентимент аналізу, ідентифікації стійких фраз/виразів/словосполучень та методів ІІП для класифікації ресторанів щодо персональних смаків/дієт конкретного користувача [597]. Ще одним цікавим та актуальним прикладом є КЛС на основі супроводу рекомендаційних тегів (мета-дані про фрагменти контенту) [598-601], реалізованих такими компаніями, як YouTube [602], Facebook [603], Amazon [604], Netflix [598-604], Stack Overflow [605-607]. Теги важливі для ІІП та генерування рекомендацій, а також при визначенні семантичного вмісту контенту згідно зацікавлень конкретного користувача тощо [506-514]. Теги ідентифікують ознаки описаного ними контенту, використовуються для кластеризації/класифікації подібних фрагментів та пропонують тематичні назви для відповідних кластерів [506].

Google Smart Reply підтримує генерування інтелектуальних відповідей на користувацькі е-листи [608-610]. Голосові віртуальні помічники як Siri [611], Alex [612-613], Google Assistant [614] та Cortana [615] здатні аналізувати мову і давати найймовірніші релевантні відповіді. Siri та Netflix підтримують українську [616]. Textra [617-618], iMessage [619] та інші ПЗ обміну миттєвими повідомленнями роблять прогнози щодо майбутнього користувацького тексту на основі введеного, а функція автоматичного коректури виправить орфографічні помилки. Reverb підтримує персоналізований RSS-агрегатор (агрегатор новин) [620] на основі словника Wordnik [621]. ChatBot Slack супроводжує діалог з ідентифікацією контексту [622].

Лінгвістичні ознаки, які роблять природну мову унікальним інструментом спілкування, ускладнюють її аналіз на основі детермінованих правил. Гнучкість людської інтерпретації при понад 50 тисяч символічних уявлень пояснює перевершення середньостатистичною людиною будь-якого комп'ютера в миттєвому розумінні мовлення. Тому для реалізації КЛС необхідні нечіткі гнучкі чутливі обчислювальні NLP-методи на основі машинного навчання.

Основною метою ML є припасування існуючих даних під деяку модель формування подання реального світу, яка допомагає приймати рішення або генерувати прогнози на основі нових даних через пошук закономірностей в них. Тобто це вибір множини моделей для визначення взаємозв'язків між цільовими та вхідними даними, задання форми/шаблону з параметрами/функціями та мінімізація помилки моделі на начальних даних на основі відповідної процедури оптимізації. Потім навченій моделі передають нові дані для побудови прогнозу та повернення маркерів, ймовірності, ознак належності або значення. Задача полягає у знаходженні балансу між здібностями з високою точністю знаходити закономірності у відомих даних та узагальнення для аналізу невідомих даних.

Більшість ПЗ для аналізу природної мови побудовані на декількох ML-моделях, які взаємодіють між собою та впливають один на одну. ML-моделі повторно навчаються на нових даних, використовуючи нові простори рішень і налаштовуючись на конкретного користувача для безперервного розвитку по

ходу поступлення нового контенту та зміни різних аспектів КЛС з часом. В КЛС ML-моделі рангують, старіють та щезають (замінюють на нові або модифікують). Тобто ML-модулі КЛС реалізують життєві цикли контенту/процесів, які забезпечують відповідність динаміки розвитку та регіональних особливостей природної мови з робочим процесом КЛС ля підтримки/супроводу/аналізу/моніторингу текстового контенту.

ML застосовують для аналізу великих даних з множини текстових потоків за певними ознаками як різноманітність, частота вживання, унікальність, закономірність, обсяг, швидкість, надійність, час тощо для розв'язку конкретної NLP-задачі, в тому числі виправлення помилок. Застосування кластеризації дозволяє згрупувати в множини лінгвістичні ознаки або типові помилки за відповідними подібними характеристиками. Це є неконтрольоване ML за відповідним алгоритмом/методом [506-514]: k -середніх (англ. k-means method) [623-625]; DBSCAN (англ. density-based spatial clustering of applications with noise, просторова кластеризація на основі щільності для додатків із шумами) [626-627]; OPTICS (англ. Ordering points to identify the clustering structure, впорядкування точок для знаходження кластерної структури) [628]; PCA (англ. principal component analysis, метод головних компонент) [629] тощо. Але найкращими методами є TF-IDF (англ. term frequency – inverse document frequency) [630], сингулярний розклад матриці (англ. singular-value decomposition SVD) [631] та знаходження кластерних груп [632]. Загальновідомими методами класифікації тексту є TF-IDF [630], k-NN метод [633]; naive Bayes classifiers [634], SVM-метод [635], латентно-семантичний аналіз (англ. Latent semantic analysis, LSA) [557], EM алгоритм (англ. Expectation-maximization algorithm) [636], дерева рішень (англ. decision trees) як алгоритм ID3/C4.5 (англ. decision trees) [637], штучна нейронна мережа (англ. artificial neural networks, ANN) [638], добування даних (англ. data mining) [639], глибинний аналіз понять (англ. Concept mining) [640], класифікація на основі м'яких (англ. Soft set theory) [641] або грубих множин (англ. Rough set theory) [642], навчання за множиною зразків (англ. multiple-instance learning, MIL) [643] та інших ML-методів опрацювання природної мови.

Останнім часом популярності набули методи глибинного навчання (англ. deep learning) [644].

1.6.2. Кластеризація текстового контенту при неконтрольованому ML

В кластеризації при неконтрольованому ML алгоритм шукає приховані зв’язки між вхідними даними текстового контенту на основі моделі прихованих (латентних) змінних, яка включає: EM алгоритм [636]; латентно-семантични аналіз [557]; PCA-алгоритм [629]; аналіз незалежних компонентів (англ. independent component analysis, ICA) [645]; BSS метод (англ. blind signal separation) [646]; метод моментів знаходження оцінок (англ. Method of moments) [647]; розклад невід'ємних матриць (англ. Non-negative matrix factorization, NMF) [648]; ієрархічна кластеризація (англ. hierarchical cluster analysis, HCA) [649] або таксономія (англ. Taxonomy) [650]; сингулярний розклад матриці (англ. singular-value decomposition, SVD) [631] тощо. Метриками аналізу лінгвістичних одиниць зазвичай бувають індекс Ренда (англ. Rand index) [651], F-міра (англ. F-measure) [652], індекс Жаккара (англ. Jaccard index) [653], індекс Соренса (англ. Dice index) [654] та індекс Фаулкса-Маллоуса (англ. Fowlkes-Mallows index) [655].

Індекс Ренда розраховує наскільки кластери (повернені алгоритмом кластеризації) подібні до еталонних класифікацій [651]:

$$I_R = \frac{k_{TP}+k_{TN}}{k_{TP}+k_{TN}+k_{FP}+k_{FN}}, \tag{1.9}$$

де k_{TP} – число істино позитивних (true positives) результатів, k_{TN} – число істинно негативних (true negatives) результатів; k_{FP} – кількість хибно позитивних (false positives) результатів; k_{FN} – число хибно негативних (false negatives) результатів.

F-міру застосовують для збалансування хибно негативних результатів шляхом зважування повноти (recall) параметром $\varepsilon \geq 0$ [652]:

$$I_{FP} = \frac{k_{TP}}{k_{TP}+k_{FP}},\ I_{FR} = \frac{k_{TP}}{k_{TP}+k_{FN}}, \tag{1.10}$$

де I_{FP} – швидкість точності або влучності (precision) та I_{FR} – швидкість повноти (чутливість). В ІІП [652]:

$$I_{FP} = \frac{|\{a_i\}\cap\{b_j\}|}{|\{b_j\}|},\ I_{FR} = \frac{|\{a_i\}\cap\{b_j\}|}{|\{a_i\}\}}, \tag{1.11}$$

де $\{a_i\}$ – множина релевантного контенту, $\{b_j\}$ – множина знайденого контенту. Обчислюють F-міру за такою формулою [652]:

$$I_{F\varepsilon} = \frac{(\varepsilon^2+1)I_{FP}I_{FR}}{\varepsilon^2 I_{FP}+I_{FR}}, \tag{1.12}$$

коли $\varepsilon = 0$, $I_{F\varepsilon} = I_{FP}$, тобто I_{FR} не впливає на F-міру $I_{F\varepsilon}$ при $\varepsilon = 0$, а зростання ε виділяє все більшу кількість ваги для I_{FR}в остаточній F- мірі.

Індекс Жаккара застосовують для кількісної оцінки подібності між двома наборами даних X та Y (приймає значення від 0 до 1) та визначається як [653]:

$$I_{Jcr} = \frac{|X\cap Y|}{|X\cup Y|} = \frac{k_{TP}}{k_{TP}+k_{FP}+k_{FN}}, \tag{1.13}$$

Це кількість унікальних елементів, спільних для обох наборів, поділена на загальну кількість унікальних елементів в обох наборах. Індекс 1 означає, що два набори даних ідентичні, а індекс 0 вказує на те, що набори даних не мають спільних елементів [653].

Індекс Соренса подвоює вагу k_{TP} при цьому ігноруючи k_{TN} [654]:

$$I_{DSC} = \frac{2k_{TP}}{2k_{TP}+k_{FP}+k_{FN}}. \tag{1.14}$$

Індекс Фаулкса-Маллоуса обчислює подібність між поверненими кластерами та еталонними класифікаціями. Чим вище значення I_{FM}, тим подібніші кластери та еталонні класифікації [655]:

$$I_{FM} = \sqrt{\frac{k_{TP}}{k_{TP}+k_{FP}}\frac{k_{TP}}{k_{TP}+k_{FN}}}, \tag{1.15}$$

де k_{TP} – кількість справжніх позитивних результатів, k_{FP} – кількість хибнопозитивних, k_{FN} – кількість помилково негативних. Індекс I_{FM} – це середнє геометричне значення точності I_{FP} та повноти I_{FR}, і відомий як G-міра, а F-міра – це їх гармонічне значення [655]. Крім того, I_{FP} і I_{FR} відомі як індекси Уоллеса (англ. Wallace's indices) I' і I'' [656]. Нормалізовані вибірки I_{FP}, I_{FR}та G-виміри відповідають індексу інформованості I_{YJs} (Youden's index або Youden's J statistic) [657], індексу маркованості I_M (англ. markedness) [658], коефіцієнту кореляції Метьюза I_{MCC} (англ. Matthews correlation coefficient (MCC) або phi coefficient) [659] та сильно пов’язані коефіцієнтом Каппа Коена $I_{CK\kappa}$ (англ. Cohen’s kappa

coefficient, κ) [660]. Індекс Уоллеса I_{YJS} фіксує ефективність дихотомічного діагностичного експерименту через аналіз чутливості та специфічності [656]:

$$I_{YJS} = \frac{k_{TP}}{k_{TP}+k_{FN}} + \frac{k_{TN}}{k_{TN}+k_{FP}} - 1. \quad (1.16)$$

Інформованість – узагальнення I_{YJS} на багатокласовий випадок і оцінює ймовірність інформованого рішення [657]. Коефіцієнт кореляції Метьюза I_{MCC} розраховують як коефіцієнт phi Пірсона [661]:

$$I_{MCC} = \frac{k_{TP}k_{TN}-k_{FP}k_{FN}}{\sqrt{(k_{TP}+k_{FP})(k_{TP}+k_{FN})(k_{TN}+k_{FP})(k_{TN}+k_{FN})}}. \quad (1.17)$$

Коефіцієнт Каппа Коена $I_{CK\kappa}$ - вимірювання надійності між оцінниками та надійності внутрішньої оцінки для якісних/категоріальних елементів [660]:

$$I_{CK\kappa} = \frac{2(k_{TP}k_{TN}-k_{FP}k_{FN})}{(k_{TP}+k_{FP})(k_{FP}+k_{TN})+(k_{TP}+k_{FN})(k_{FN}+k_{TN})}. \quad (1.18)$$

Існують суперечки навколо коефіцієнта Каппа Коена $I_{CK\kappa}$ через складності в інтерпретації показників узгодженості. Деякі дослідники припускають, що концептуально простіше оцінити розбіжності між елементами [660].

1.7. Основні напрями дослідження

На сьогодення є багато комп’ютерних лінгвістичних систем різного призначення, навіть для опрацювання україномовного текстового контенту. Але це зазвичай комерційні проекти закритого типу (немає ні публікацій ні доступу до адміністративної частини) та найчастіше це є іноземні проекти. Публікацій ніби багато для розуміння як в загальному відбувається процес опрацювання природної мови, особливо для англійських текстів. Але застосувати ці моделі, методи, алгоритми та технології напряму для україномовного текстового контенту не приводить майже ні до якого позитивного результату. Вже саме на рівні морфологічного аналізу виникає суттєвий конфлікт між розробленими методами та вхідним українським текстом – на виході не коректний результат. Наприклад для простого алгоритму Потрера (стемінг) без відповідної модифікації не коректне буде відокремлення основи слова від флексії, що призведе до некоректної ідентифікації ключових слів текстів, щ в свою чергу

впливає на будь-яку NLP-задачу, де необхідно швидко ідентифікувати множину ключових слів (рубрикація, пошук, анотування тощо). Визначення основних процесів та особливостей лінгвістичного аналізу україномовних текстів значно полегшить етапи опрацювання текстового потоку контенту як інтеграція, супровід та управління контентом (Рис. 1.25). В свою чергу адаптація процесів інтелектуального аналізу текстового контенту з ідентифікацією функціональних вимог до відповідних модулів КЛС призведе до можливості розробити типову архітектуру подібних систем на принципі модульності (додавання компонентів в залежності від змісту NLP-задачі та призначення КЛС).

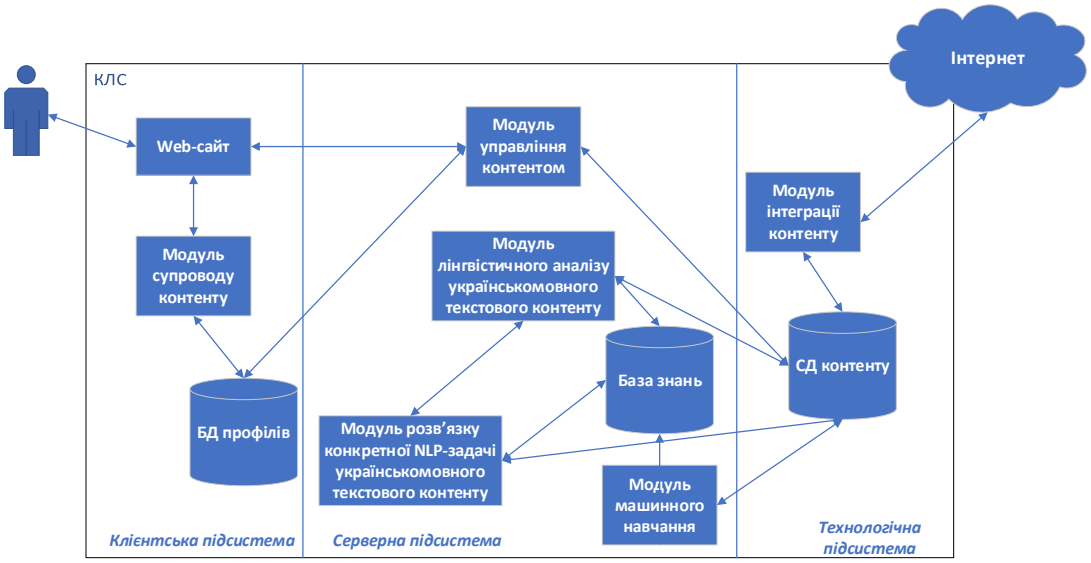


Рис. 1.25. Узагальнена архітектура комп'ютерної лінгвістичної системи

Застосування вказаних технологій/методів/моделей в типовій архітектурі КЛС, адаптованих для будь-якої NLP-задачі опрацювання україномовного текстового контенту, є необхідною передумовою успішної реалізації проекту комп'ютерної лінгвістичної системи для розв'язку конкретної NLP-задачі, який вимагає застосування відповідної множини стандартних бібліотек, утиліт та ПЗ з відкритим кодом, що вирішуватимуть спеціалізовані задачі проекту згідно потреб кінцевого користувача.

1.8. Основні результати та висновки розділу

Аналіз та синтез КЛС базується на застосуванні лінгвістичного аналізу україномовного текстового контенту, інтелектуальному опрацювання

текстового потоку контенту, машинному навчанні системи на достовірних даних та статистичному аналізі для знаходження закономірностей появи лінгвістичних подій. Проведено аналіз сучасного стану та перспективи розвитку ІТ опрацювання природної мови, що дало змогу визначити проблему та задачі дослідження, а також сформувати загальні напрями дослідження при відсутності некомерційних КЛС з відкритим кодом для опрацювання україномовного текстового контенту та стандартизованого підходу проектування. Визначено поняття КЛС та наведена загальна їх класифікація. Проведений детальний аналіз відомих КЛС, що дало можливість вдосконалити загальну класифікацію відповідних ІС. Визначені основні NLP-задачі комп'ютерних лінгвістичних систем, на основі яких наведені приклади та порівняльний аналіз відомих сучасних КЛС. Це дало можливість сформувати загальні напрями дослідження. Описана та проаналізована основна загальна схема процесу лінгвістичного аналізу тексту природньою мовою засобами КЛС. Визначені основні стани та властивості КЛС, їх класифікація та особливості. Проаналізовано відомі класичні підходи та напрями опрацювання природної мови. Наведена загальна класифікація основних NLP-підходів, напрямів та додаткових методів лінгвістичного дослідження для NLP-задач. Проведено аналіз існуючих основних методів та методики опрацювання природної мови засобами машинного навчання. Проведена їх класифікація та визначені типові проблеми ML-методів опрацювання україномовних текстів. Зроблений огляд відомих ІТ розроблення КЛС на основі особливостей та технологій інтелектуального аналізу потоку україномовного контенту. Визначені основні вимоги до оцінювання ефективності КЛС на основі технології ML та аналізу великих даних. Розглянуті основні ML для аналізу великих даних з множини текстових потоків контенту. Визначені вимоги до кластеризації текстового контенту при неконтрольованому ML. Основні результати розділу опубліковані у роботах [19-23, 35-42, 48-49, 54-61, 84, 86-87, 103, 112-114, 130, 132, 135-147, 160, 162-163, 182-183, 209-213, 219, 256-259, 287-297, 304-305, 310-313, 350, 354-355, 400-407, 471-484, 490-505, 534-535, 573-574, 585-592].

РОЗДІЛ 2
ОСОБЛИВОСТІ ПРОЕКТУВАННЯ ТА РОЗРОБЛЕННЯ КОМП'ЮТЕРНИХ ЛІНГВІСТИЧНИХ СИСТЕМ

2.1. Основні етапи лінгвістичного аналізу текстового потоку

2.1.1. Особливості аналізу україномовного текстового потоку

Будь-який лінгвістичний аналіз тексту включає основні NLP-підпроцеси (NLP-рівні) лінгвістичного аналізу (Рис. 2.1) [210-212, 561-565, 585-586].

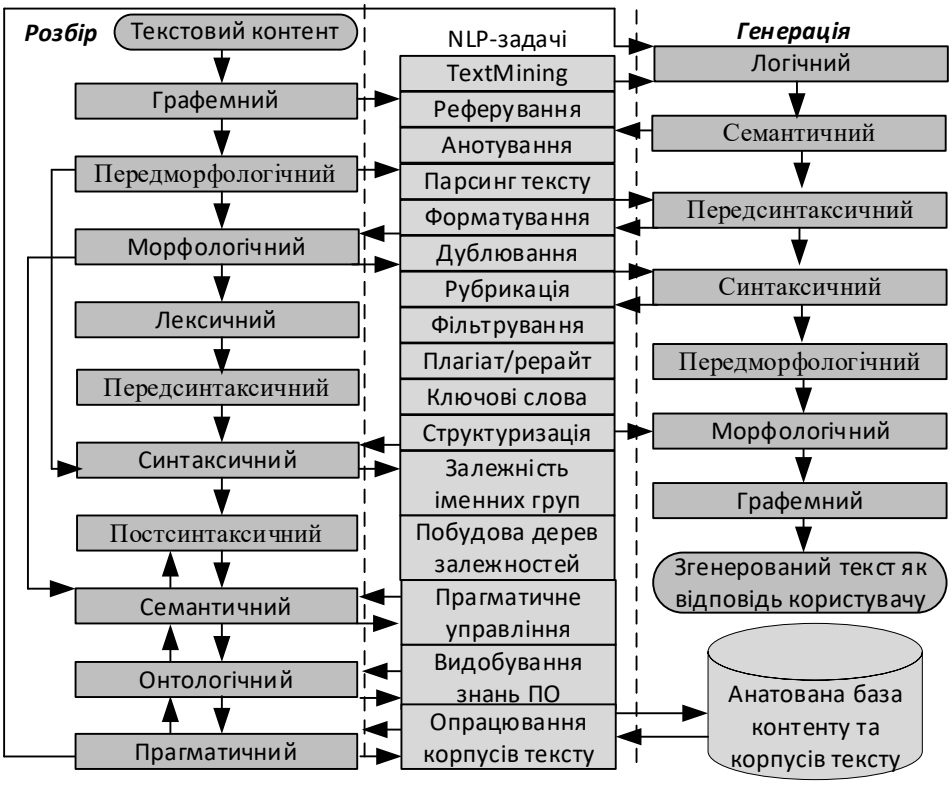


Рис. 2.1. Структурно-лінгвістична схема лінгвістичного аналізу тексту

Для кожної мови складність є в реалізації синтаксичного аналізу, але є мови як українська, складність ще полягає в реалізації морфологічного аналізу, від якого залежать інші NLP-рівні лінгвістичного аналізу (Таблиця 2.1). Розроблення повноцінних детальних словників ПО, основ слів та їх особливостей відмінювання в залежності від частини мови та їх особливостей (роду, часу, множини/однини) з врахуванням чергування літер значно полегшить проведення МА тексту української мови. Це дозволить провести точніший синтаксичний (структури речень) та семантичний (використаних понять) аналіз

для підготовки витягування знань з відповідного тексту через прагматичний аналіз (коректність мети використання понять).

Таблиця 2.1

Етапи лінгвістичного аналізу текстової інформації [211-212]

№	Аналіз	Пояснення
1	Графематичний або графемний	Виділення/об'єднання синтаксичних (заголовків, основного тексту, вставок, зносок, коментарів тощо) і/або структурних (абзаци, речення, окремі слова і розділові знаки) одиниць текстового контенту з подальшим фільтруванням
2	Передморфологічний	Виділення/об'єднання нерозривних незмінних стійких словосполучень в одну лінгвістичну одиницю: *_Залізний_Порт_* (місто), *_Червона_Калина_* (проспект), *_Нью_-_Йорк_, _Івано_-_Франківськ_, _і_так_далі_, _яким_-_небудь_, _таким_чином_, _будь_-_хто_*, тощо
3	Морфологічний	Визначення нормальної форми словоформи, і навпаки генерування словоформи з нормальної форми з врахуванням місцерозташування в синтаксичному дереві залежності для узгодження слів у реченні.
4	Передсинтаксичний	Об'єднання окремих лексичних одиниць в одну синтаксичну як стійкі словосполучення (наприклад, фразеологізми та метафори як *бити байдики*), поділ на окремі (наприклад, *словоформа*, *криптовалюта*, *відеомонтаж*, але не *качкодзьоб, водогін, зорепад, чорнозем*) та сегментація.
5	Синтаксичний	Розгортання синтаксичних дерев залежності речень із узгодженням слів. Перетворення дерева на лінійний порядок слів із врахуванням параметрів.
6	Постсинтаксичний	Нормалізація синтаксичних дерев залежності речень із врахуванням та уточненням параметрів слів та їх змістовного навантаження у виразі.
7	Семантичний	Уточнення взаємозв'язків слів в дереві для видобування знань або генерація відповіді з врахуванням семантичних ролей іменних груп та їх дій/подій.

Логічне виведення у вигляді множини природнього текстового контенту на основі лінгвістичного аналізу вхідного тексту є розповсюдженим явищем для будь-якого статистичного аналізу тексту (сентимент-аналіз, аналіз тональності, контент-аналіз тощо), діалогових систем, QA-систем, систем генерування рефератів/анотацій/дайджестів тощо. Природній текст зазвичай є частково структурованою та формалізованою інформацією з наявністю натяків, замовчуванням, скороченням, неповнотою, шумом, неточностями, заплутаністю тощо, особливо це притаманно для синтаксичних груп мов, як слов'янські мови. Ідентифікувати та опрацювати такі конструкції є складним процесом (наприклад, українською *пташка сидить на столі*, хоча може *стояти*, *кішка* може *сидіти*, *лежати* та *стояти*, *стакан стоїть на столі*, а *тарілка лежить на столі* тощо, в свою чергу, в англійській мові зазвичай використовують для всіх перелічених випадків дієслово *є – is*). Також цікавими конструкціями розмовної української мови є *шмигати носом*, *зробити ноги*, *говорити абсурдні речі*, *дати прочухана*, *золота молодь*, *зробити ляпсус*, *пам'ять як у рибки*, *піти по воду*, *стригти*

купони, *вештатися містом*, *зелена капуста*, *теревенити* (*базікати*) *по телефону*, *тримати ніс за вітром*, *кмітливий пуцьвірінок*, *дати телефон*, тощо.

2.1.2. Графемний аналіз і синтез україномовного тексту

Основною будь-якого графемного аналізу тексту [211-212] є ідентифікація розділових знаків, скорочень, абревіатур, великих літер у власних назвах тощо.

Апостроф в українській та англійській мовах не є розділовим знаком, хоча існує подібний розділовий знак – одинарна лапка для відокремлення цитат. В англійській мові простіше – апостроф зустрічається в кінці іменника (визначення приналежності) або множини окремих символів біля літери s або для скорочення деяких дієслів форм. В українській мові зазвичай апостроф зустрічається в коренях слів та їх варіаціях, зокрема, після губних приголосних (*б*, *в*, *м*, *п*, *ф*) в коренях деяких слів, після *р* в кінці складу, після префіксів перед твердою приголосною початку кореня та після першої частини деяких складних слів.

Наявність подвійних лапок свідчить або про власну назву, або про цитату, або про сарказм. Кожний із перелічених лінгвістичних одиниць несе своє змістовне навантаження та є різним рушієм для генерування синтаксичного дерева розбору та визначення ключових слів як стійких словосполучень (*кінотеатр* «*Зірка*», *зірка на кінотеатрі*, *зірки готелю* або «*золота рибка*» як риса людини або *золота рибк*а як рибка в акваріумі тощо). Інколи власні назви співпадають із загальновживаними словами (група *Мертвий півень*, студенти *Тарас Оксана*, *Сергій Семен*, *Лема Тарас*, *Сало Михайло* (*немає сьогодні Михайла Сала*) та *Тесля Софія*, співачки *Катя Чилі* та *Вінницька Альона*, актор *Девід Духовний*, проспект *Червоної Калини* або *Свободи*, вулиця *Перемоги* тощо), але мають різний сенс. Існують лексеми, які не підпорядковані граматиці (1979, 12%, кг, млн, км тощо). Тому графемний аналіз дозволяє маркувати та класифікувати лексеми, які виходять за межі стандартного лінгвістичного аналізу граматики. Наперед визначити та поповнити словники такими лексемами не можливо. Лише використання правил попереднього аналізу тексту з використанням методів машинного навчання з вчителем. Підтримувати

словники все можливих імен, географічних назв, скорочень, числових значень тощо неможливо. При ідентифікації лексеми як графеми з невідомих наперед невизначених елементів формується проміжний словник, який має переглянути модератор та відповідно їх маркувати. Але простіше підтримувати набір правил для ідентифікації нестандартних лексем на основі аналізу графем з частковим використанням словників часто вживаних розповсюджених виключень.

Додатковими пунктами графемного аналізу є ідентифікації графем у вигляді спеціальних знаків, як кінець абзацу, наявність рисунків, таблиць, формул тощо, наявність символів алфавіту іншої конкретної мови, HTML-тегів, елементів форматування як заголовок, вирівнювання, смайликів і т.п. [210-212]. Результатом графемного аналізу є розбудова графемної структури тексту із класифікованих множин графемних ланцюгів та зав'язків між ними.

$$C_\alpha = \alpha(D_\alpha, R_\alpha, X), \tag{2.1}$$

де X – вхідний текст; C_α – опис графемної структури вхідного тексту; α – оператор графемного аналізу (ідентифікація, класифікація та маркування графеми); D_α – словники розділових знаків, скорочень, абревіатур, географічних назв тощо; R_α – правила графемного аналізу, в тому числі регулярні вирази.

2.1.3. Морфологічний аналіз і синтез україномовного тексту

Основною метою МА є ідентифікація нормальної форми слова $f_{\beta i}^{n}$ для будь-якої словоформи $w_{\beta i}^{t}$ у вхідному тексті, та відповідного кортежу описових критеріїв та параметрів $c_{\beta i}$ (частина мови, рід, число, відмінок тощо) [313]:

$$C_\beta = \beta(C_\alpha, D_\alpha, R_\alpha, X),\ C_\beta = \{c_{\beta 1}, c_{\beta 2}, \dots, c_{\beta n},\}, \tag{2.2}$$

$$c_{\beta i} = (w_{\beta i}^{t}, r_{\beta i}^{w}, f_{\beta i}^{n}, r_{\beta i}^{f}, p_{\beta i}^{w}),\ p_{\beta i}^{w} =< n_{\beta i}^{p}, v_{\beta i}^{p} >, \tag{2.3}$$

де X – вхідний текст; C_β – множина кортежів описових критеріїв та параметрів для кожного слова $w_{\beta i}^{t}$ вхідного тексту; $c_{\beta i}$ – кортеж описових критеріїв та параметрів для i слова вхідного тексту; β – оператор морфологічного аналізу; C_α – результат ГА; D_α – словники слів в нормальній формі або основ слів з описовими параметрами; R_α – правила МА; $r_{\beta i}^{w}$ – частина мови слова $w_{\beta i}^{t}$ вхідного

тексту; $f_{\beta i}^{n}$ – нормальна форма слова $w_{\beta i}^{t}$ вхідного тексту; $r_{\beta i}^{f}$ – частина мови нормальної форми $f_{\beta i}^{n}$ (наприклад, для дієприслівника як форми дієслова); $p_{\beta i}^{w}$ – колекція морфологічних параметрів та критеріїв $w_{\beta i}^{t}$; $n_{\beta i}^{p}$ – назва морфологічного параметра слова (відмінювання, час, число, рід, стисліть форми прикметника та інші параметри слів відповідної природної мови); $v_{\beta i}^{p}$ – конкретне значення морфологічного параметра слова вхідного тексту відповідної природної мови.

Різноманітність залежності у різних мовах місця розташування конкретної форми слова із відповідною частиною мови значно ускладнює лінгвістичний аналіз тексту. Попереднє опрацювання слів вхідного тексту через МА скорочує список слів, з якими необхідно працювати на наступному етапі (наприклад, слово *інформація*, а не всі варіанти, утворенні при відмінюванні та зміні числа). Так для іменників обирають запис виду *слово* → <*частина мови*, *рід*, *відмінок*, *істота*, *число*> згідно різної методики, наприклад для *донька* записують [567]:

1) 1593 → < 01 0202 0301 0601 0901 >;

2) донька → < і, рід = ж, число = од, відмінок = нз, істота = і >;

3) донька → < ім, ж, од, наз, іст >.

Для першого пункту кожне слово має свій номер в словнику або його перетворюють на число за відповідністю символів в ASCII-таблицях (наприклад, слово *донька* в словнику має номер 1593) [567]. Частині мови іменник відповідає значення 01, параметру рід – 02, а жіночий рід – також 02, тому отримуємо 0202. Іменники не змінюють рід, але сформовані з них дієслова та прикметники в українські мові можуть змінювати рід в залежності від змісту [567]. Тому одній словоформі можуть приписуватися декілька кортежів (омонімія), наприклад:

1) *доньки* → донька → < ім, ж, **од**, **род**, іст >
доньки → донька → < ім, **мн**, **наз**, іст >
2) *мати* → мати → < **ім**, ж, од, наз, іст >
мати → мати → < **д**, перехідне, 1 дієв, недок >
3) *опали* → опал (камінь) → < **ім**, ч, мн, наз, іст >
опали → опасти → < **д**, мин, мн, 3 ос, док. >
4) *ягуари* → ягуар (тварина) → < ім, ч, од, наз, **іст** >
ягуар → ягуар (машина) → < ім, ч, од, наз, **неіст** >
5) *замок* → замок (будівля) → < ім, ч, од, наз, неіст >

замок (інструмент) → < ім, ч, од, наз, неіст >
6) *дракон* → дракон (тварина) → < ім, ч, од, наз, **іст** >
дракон (корабель) → < ім, ч, од, наз, **неіст** >
6) *кішки* → кішка (тварина) → < ім, ж, **од**, **род**, **іст** > або < ім, **мн**, **наз**, **неіст** >
кішки → кішка (частина взуття) → < ім, ж, **од**, **род**, **іст** > або < ім, **мн**, **наз**, **неіст** >
але *кишки* → кишка (частина тіла) → < ім, ж, **од**, **род**, **іст** > або < ім, **мн**, **наз**, **неіст** >
6) *коси* → коса (зачіска) → < ім, **мн**, **наз**, **неіст** > або < ім, ж, **од**, **род**, **іст** >
коси → коса (мілина) → < ім, **мн**, **наз**, **неіст** > або < ім, ж, **од**, **род**, **іст** >
коси → коса (інструмент) → < ім, **мн**, **наз**, **неіст** > або < ім, ж, **од**, **род**, **іст** >
коси → коса (селезіка) → < ім, **мн**, **наз**, **неіст** > або < ім, ж, **од**, **род**, **іст** >

Зазвичай застосовують словниковий морфологічний аналіз (Рис. 2.2) [313], тобто зберігають повний словник слів [567]. Недоліками є [567]:

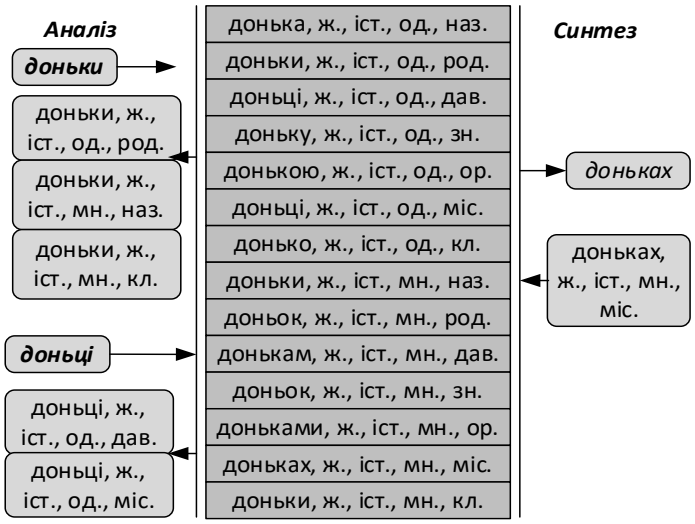


Рис. 2.2. Структурно-лінгвістична схема прикладу подання слова в словнику

1) неможливо опрацювати слова, яких немає в словнику;

2) громіздкість (багато переборів та порівнянь) та надлишок (наявність декількох варіантів результатів ІІП) даних для опрацювання слів в тексті.

Сучасна українська мова має понад 256 тисяч слів [662]. Іменник має 7 відмінків, тобто приймає 14 форм, а прикметники – 24, тобто наявність різних флексій та в деяких випадках чергування літер. Є безліч синонімів, наприклад горизонт має 12. Кількість словоформ дієприслівників і дієприкметників як форм дієслова сягає 300 (біля 25 форм на парадигму). Це все ускладнює МА [313].

Перехід до дерева частково вирішує цю проблему (Рис. 2.3) [313, 567]. Зазвичай МА провадять посимвольно з кореня дерева. Цей спосіб є складним для реалізації – треба врахувати всі можливі варіанти з всіх можливих слів [567]. Тому кращим способом є поєднання двох цих двох способів з парсингом

символів з кінця слова (ідентифікація флексій за деревом всіх можливих закінчень для визначення частини мови, відокремлення кореня та ідентифікація кореня в словнику) [313]. В [663] побудоване статичне дерево закінчень для слів з бази Aspell (біля 1,4 млн форм українських слів) в межах 1-11 символів.

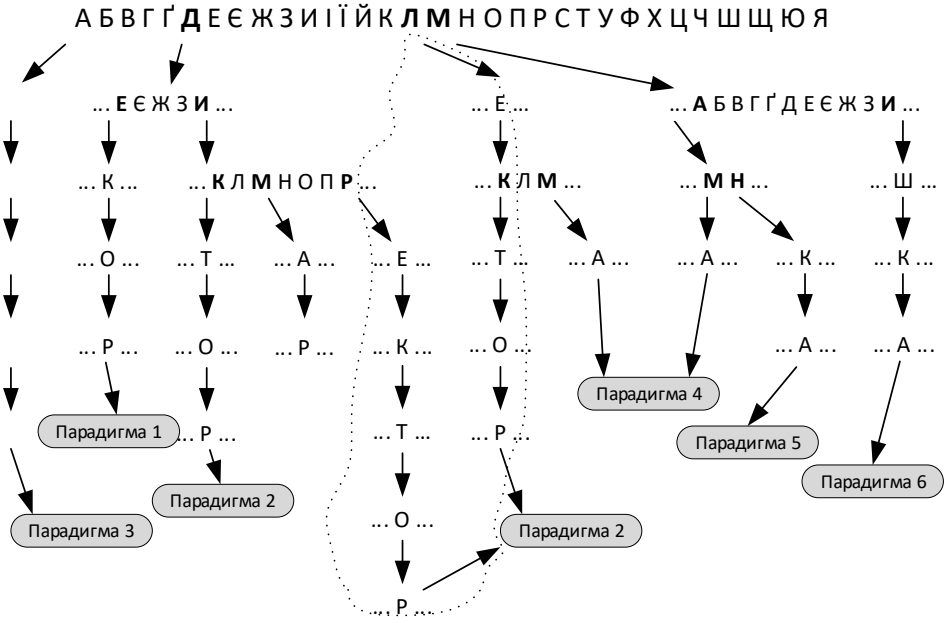


Рис. 2.3. Структурно-лінгвістична схема прикладу побудови дерева префіксів

Завдяки дослідженням автора [169] можна ранжувати флексії за частотою використання та відокремити по блоках за належністю до частин мови (Таблиця 2.2) [663]. Більшість флексій з сумарна питома вагою використання меншою за 1% належать в більшості випадків іменникам, зокрема, г (4) – *ґирлиґ*, *дзиґ*, *зиґзаґ*, *меґ* [663]. Аналогічно це відноситься до флексій ц (34), щ (110), ф (214), б (281), п (341), ж (353), з (581), г (636), л (754), с (914), ч (959), д (1038), н (2531), р (2709).

Таблиця 2.2

Статична таблиця розповсюджених українських флексій [663]

Флексія	*ням* (9434)	*лись* (10337)	*є* (11466)	*ним* (19093)	*ною* (20280)	*им* (31166)	*х* (61506)
т (2980)	*ня* (9765)	*лися* (10338)	*ку* (11624)	*ної* (19098)	*мося* (20532)	*ім* (31343)	*ми* (62080)
к (7299)	*ями* (9844)	*тися* (10379)	*шся* (11775)	*теся* (19103)	*мось* (20536)	*ого* (31389)	*е* (66988)
кою (7497)	*ях* (9855)	*ало* (10465)	*ті* (12596)	*тесь* (19105)	*на* (21328)	*ої* (31421)	*а* (68134)
істю (7598)	*нні* (9909)	*ав* (10547)	*ям* (15717)	*еся* (19105)	*ися* (21940)	*го* (31445)	*й* (77109)
ість (7606)	*всь* (10016)	*ала* (10610)	*ів* (15898)	*ному* (19112)	*ві* (22543)	*ому* (31585)	*ю* (80877)
стю (7648)	*ню* (10075)	*али* (10666)	*ом* (17018)	*есь* (19114)	*ись* (22656)	*ні* (31679)	*і* (90275)
ості (7636)	*вся* (10076)	*ати* (10819)	*ові* (17191)	*ш* (19163)	*ну* (23125)	*те* (32651)	*о* (90454)
сть (7688)	*лась* (10229)	*ка* (11029)	*ло* (17238)	*нім* (19333)	*ться* (25036)	*в* (32681)	*у* (94504)
ося (8044)	*лася* (10230)	*ємо* (11136)	*ли* (17711)	*ній* (19549)	*ься* (25211)	*ть* (33055)	*сь* (111459)
ось (8047)	*лось* (10231)	*єте* (11137)	*ла* (17945)	*ах* (20023)	*ося* (30769)	*ій* (33241)	*м* (119779)
сті (8731)	*лося* (10233)	*єш* (11138)	*ний* (19042)	*ти* (20025)	*ось* (30788)	*мо* (33568)	*и* (123402)
нням (8975)	*ася* (10235)	*ють* (11222)	*ними* (19089)	*ами* (20106)	*ими* (31121)	*ї* (34702)	*ся* (148160)
ння (9001)	*ась* (10239)	*ймо* (11229)	*ного* (19090)	*ам* (20154)	*их* (31127)	*му* (35023)	*ь* (151355)
нню (9054)	*тись* (10366)	*йте* (11230)	*них* (19092)	*не* (20257)	*ий* (31136)	*ою* (39616)	*я* (164062)

Слова групують за парадигмами (множини всіх постфіксів на основі додатку В [567] і морфологічних параметрів для всіх словоформ відповідного слова, наприклад, слова *лектор* і *професор*). Тоді зберігають єдину стрічку у дереві постфіксів. Групування за парадигма залежить від особливостей слів, їх морфологічних параметрів та NLP-задачі. Так, слова *лектор* і *вектор* не входять в одну парадигму через різні флексії в знахідному відмінку. Але слова *мама* і *лема* можуть увійти в одну парадигму, якщо розглядати для конкретної NLP-задачі значення істота/неістота (для рубрикації не треба, а ось для ПА необхідно).

Посимвольний аналіз словоформи з кореня дерева вимагає збереження масиву вказівників на наступну вершину – конкретну літеру [313, 567]. Необхідно у кожній вершині зберігати алфавіт мови. Але для української мови для збереження всіх ланцюгів із 8 літер треба понад 46 млрд вказівників. Частина з них відсікається (наприклад, немає слів на м’який знак). Тому масиви літер алфавіту у вершинах заповнюються щільно біля кореня дерева, а ближче до листя – розріджені. Крім того, частина постфіксів слів унікальні для частин піддерев, тому їх зберігають як стрічку. Але це все не дозволить врахувати всі можливі варіанти піддререв та викликає зайві навантаження на МА-процес– в деревах зберігають зазвичай всі слова в нормальній формі. При збережені всіх варіантів відмінювання слів в таких деревах з врахуванням чергувань символів призводить до зростання надлишку збереження даних. Збереження дерев постфіксів та морфологічний аналіз з кінця слів за їх флексіями/постфіксами зменшить кількість операцій [313, 567]. Наприклад, для визначення ключових слів достатньо розглядати слова лише з іменникової групи (без займенників), тоді всі закінчення (постфікси) притаманні для дієслів суттєво зменшать кількість слів, які необхідно аналізувати. Для рубрикації вхідних текстів тоді достатньо ідентифікувати іменникові групи та провести відповідний МА.

2.1.4. Лексичний аналіз україномовного тексту

Процес лексичного аналізу полягає в аналітичному розборі (сегментації) вхідного масиву тексту після детального морфологічного аналізу на формування

колекцій токенів (послідовностей символів за відповідними шаблонами) як лексем з подальшою ідентифікацією їх типів [407]. Лексемою зазвичай є слово, словоформа, або словосполучення як змістовна лексична одиниця виразу/речення [407]. Сегментація речення є ще одним важливим кроком у опрацюванні тексту [567]. Модулем ЛА є сканер, токенізатор або лексичний аналізатор в залежності від мети NLP-задачі. Не всі токени є лексемами, наприклад число 13, математичний вираз, знак пунктуації тощо. Найбільш корисними символами для сегментації тексту в речення є пунктуація, наприклад, крапки, знаки питання, знаки оклику тощо. Знаки питання/оклику відносно однозначні маркери меж речень. Крапки, з іншого боку, більш неоднозначні, наприклад між маркером речення та маркером абревіатур, таких як млн. або р. Останнє скорочення проілюструвало складний випадок цієї неоднозначності, в якій позначена крапка р. є як скороченням слова *рік*, так і маркером межі речення. З цієї причини токенізація речення та слова мають провадитися паралельно та одночасно. Загалом, методи токенізації речення працюють шляхом побудови бінарного класифікатора (на основі послідовності правил або на машинному навчанні), який вирішує, чи крапка є частиною слова, або маркером межі речення. При прийнятті цього рішення він допомагає з'ясувати, чи крапка належить до загальноприйнятої абревіатури; таким чином, корисним є словник скорочень. Найсучасніші методи токенізації речення ґрунтуються на використанні машинного навчання. Ідентифікацію лексем через класифікацію за типами токенів у контексті певної граматики/мови. Якщо лексему як токен мови не можливо ідентифікувати згідно з відповідною граматикою, тоді перевіряють із словником спец символів, математичних знаків тощо. Якщо і в цьому випадку не можна ідентифікувати, то маркують як спеціальний токен-помилку. Токен – це є структура за певним шаблоном із ідентифікатором типу/класу. Ідентифікація відбувається у два етапи у вигляді скінченного автомата – сканування за регулярними виразами та оцінювання для подальшого класифікації за типом та передачі на вхід до синтаксичного аналізатора. Інколи для спрощення синтаксичний аналізатор поєднають із лексичним для деяких NLP-задач.

Тоді синтаксичні аналізатори провадять аналіз як парсинг тексту в два етапи (Рис. 2.4) [407, 567]: ідентифікують змістовні лексеми (ЛА) та генерується дерево розбору речення (залежності ідентифікованих лексем).

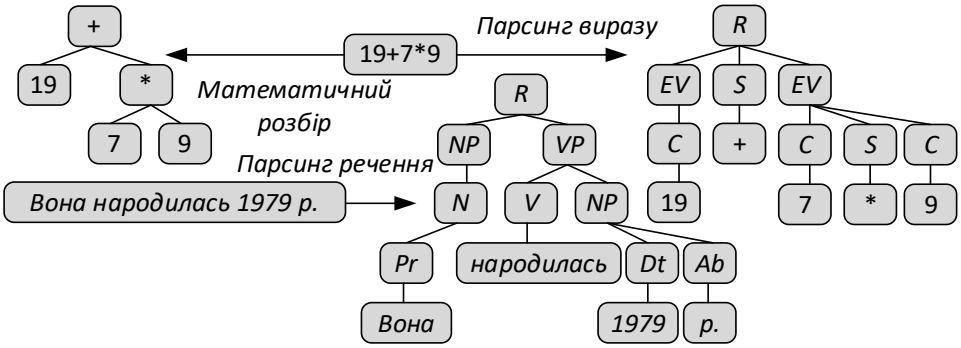


Рис. 2.4. Приклади розбору виразів та генерування дерева залежності

Токен є атомарним змістовним об’єктом із послідовності в межах [1,N] символів [407, 567]. Ідентифікують токени на основі регулярних виразів та за місцерозташуванням у наборі символів/реченні та контексті. Це не графемний аналіз як відокремлення групи символів між розділовими знаками. Токени ідентифікуються правилами лексера з врахуванням вже граматичних ознак з попереднього кроку МА відповідно природньої мови вхідного тексту, зокрема:

- Маркування набору вхідних символів тексту у набір токенів;
- Ідентифікація окремого токена як логічної лінгвістичної одиниці тексту (слово, математичний знак, число, знак пунктуації тощо).
- Встановлення відношення між токеном і лексемою – конкретний текст токена (“для”, “1979”, “+”, “змінна”, “.”, “р.”, “;” тощо).
- Ідентифікація додаткових атрибутів токена (наприклад, крапка як межа речення чи частина скорочення).
- Формування кортежу токенів як вхідної інформації для СА.

Лексичний аналізатор не перевіряє коректність зав’язків в кортежі токенів, але лише їх ідентифікує, маркує та класифікує (Рис. 2.5) [407, 567]. Лексичний аналізатор розпізнає дужки, знаки пунктуації та математичні знаки як знаки, але не перевіряє, чи кожному знаку “(” відповідає інший – “)”, а кожний математичний знак знаходиться між конкретними двома числами [407, 567].

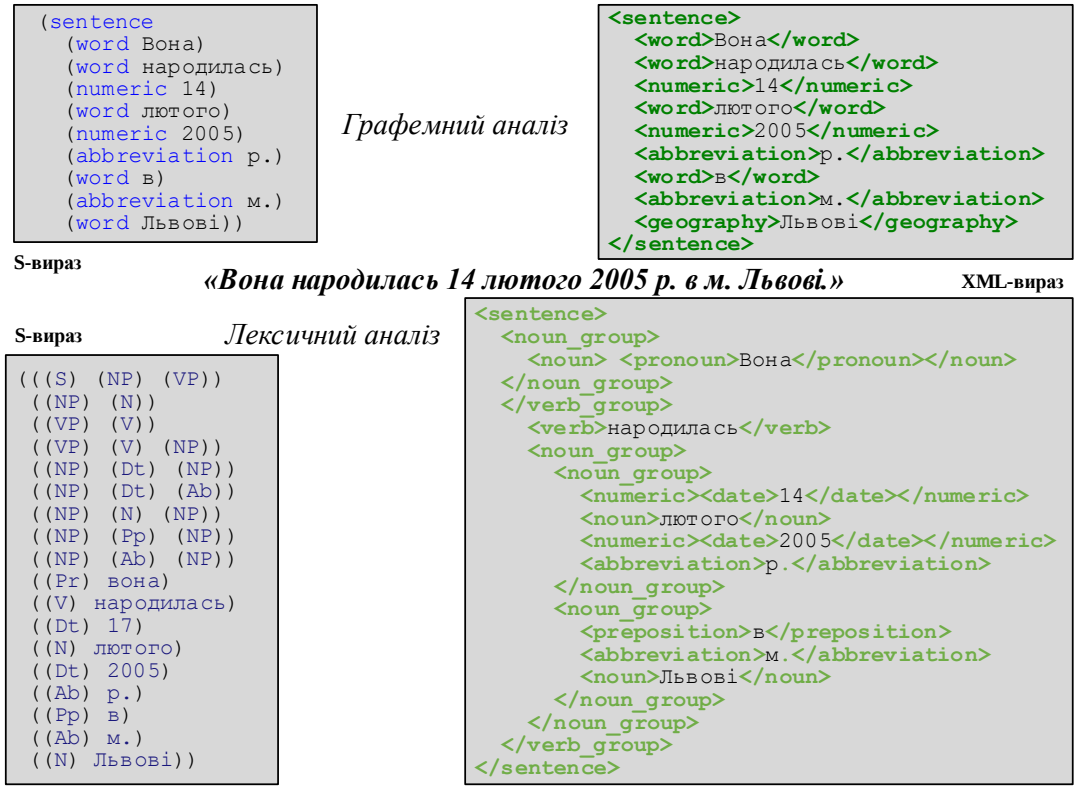


Рис. 2.5. Приклади результатів S/XML-виразів для графемного та лексичних аналізів речення «Вона народилась 14 лютого 2005 р. в м. Львові.»

Такі функції притаманні для синтаксичного/семантичного парсера/аналізатора у відповідних NLP-задачах.

2.1.5. Синтаксичний аналіз та парсинг україномовного тексту

Для аналізу синтаксису тексту застосовують зазвичай граматики Н. Хомські [416-430], системні граматики Холідея [664-673], дерева підпорядкування та системи складових дослідника А.В. Гладкого [367-379], розширення мережі переходів Петрі [821, 858, 914, 674-679]. Ефективним інструментом англомовного синтаксичного моделювання (правила утворення речень зі словоформ) є генеративна граматика (Generative grammar), започаткована у працях американського лінгвіста Н. Хомські [416-430]. За його теорією словоформи позначають термінальними символами, синтаксичні категорії – нетермінальними символам [219], а правила виведення речень (синтаксична структура) – продукційними правилами та подають в термінах безпосередніх складових [963]. Науковець застосував формальний аналіз схеми

речень для виокремлення синтаксичної схеми виразу незалежно від значення [416-430]. Дослідження Н. Хомські продовжив лінгвіст А.В. Гладкий [367-379], який застосовував систему складових та синтаксичні дерева залежностей для аналізу речень природньою мовою [367-379]. Науковець розробив основи моделювання синтаксису на основі синтаксичних груп для ідентифікації складових словосполучень як одиниць генерування дерева залежності [367-379]. Такий підхід дозволив поєднати переваги дерев залежностей і безпосередніх складових для опрацювання та аналізу слов'янських мов [367-379].

Лінгвістичні дослідження Н. Хомські [416-430], А.В. Гладкого [367-379], Є.І. Большакової, Е.С. Клишинського, Д.В. Ланде, А.А. Носкова, О.В. Пескової та Є.В. Ягунової [567], А.Є. Пентуса та М.Р. Пентуса [680], В.С. Фомічева [681], Е.В. Попова [682], Б.К. Мартиненко [683], А.С. Герасимова [684], І.А. Волкової, Т.В. Руденка [685], Н.Ц. Більгаєвої [686], Ю.Д. Апресяна [687-688], А.В. Анісімова [689] та О.О. Марченко, А.О. Никоненко [690], М. Гросса та А. Лантена [691], Н.Г. Арсент'єва [692], В. Інгве [693-696], Е.В. Падучевої [697] С.А. Шарова [698], Ю.А. Шрейдера [699], L. Tesniere [700], P.M. Postal [701], D.G. Hays [702], L.W. Tosh [703], Y. Bar-Hillel [704] та D. G. Bobrow [705] дозволяють зрозуміти основні принципи синтаксичного аналізу текстових масивів даних в залежності від особливостей конкретної мови, в тому числі і для української мови на основі відповідних досліджень українських фахівців [704-724]. Під час СА кожне речення формалізують та перетворюють у структуру даних у вигляді дерева синтаксису та залежності лексем (Рис. 2.6) [257, 963].

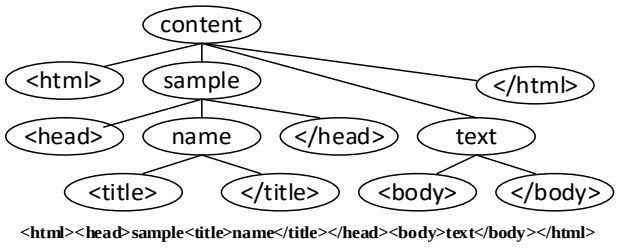


Рис. 2.6. Приклад розбору виразу в дерево залежності лексем

Синтаксис речень є множиною правил конкретної мови для формування залежності лінгвістичних одиниць для визначення семантичних ролей та

відповідності між собою сутностей/об’єктів/явищ/подій/дій в контексті тексту на основі операцій логіки висловлювань. Тоді синтаксичний парсинг є процесом розбору маркованої на попередніх рівнях вхідної інформації для ідентифікації граматичної структури відповідно формальної граматики відповідної мови з подальшою побудовою дерева залежності. Це є досить складним процесом для синтетичного типу флективних мов як українська, де лексичне значення синтезується з граматичним в межах лексеми на основі суплетивізму (генерування граматичних форм слів від різних основ, наприклад, *сказати - говорити*, *взяти – брати* тощо), чергування звуків, формотворчих афіксів (частина слова, яка змінює значення основи, наприклад, *заїхати*, *пароплав*, *лісостеп*, *заморський* тощо) та флексій. Дієвідмінювання дієслів та відмінки іменникових груп визначають способи зміни лексем для опису взаємовідносин лексем між собою в межах конструкції речення для передачі змісту. Тому речення у синтетичних мовах як українська засновані на словозміні для опису структури відношень лексем, і не залежать від місцерозташування лексем у реченні за винятком лише декількох моментів (наприклад, частка *не* завжди попереду лексеми-заперечення, будь-який прийменник завжди попереду лексем типу іменникової групи або іменника і не зустрічаються перед дієсловом).

Аналітичні ж мови як англійська та німецька відносно обмежені у морфології, зокрема відмінках, флексіях та дієвідмінюванні, але розвинуті у застосуванні різноманіття прийменників та артиклів (без них в таких мовах речення розсипаються у контексті). Тобто синтетичні мови передають контекст через відношення лексем на основі словозміни в межах речення, а аналітичні мови застосовують прийменники для утворення цих відношень. Речення флективних мов програмно проаналізувати складно. У складі природньої мови часто присутні неоднозначності (лексеми, які передають багато варіантів змісту, але лише одне для конкретного контексту). Коректний вибір значення часто залежить від змісту речення/тексту, та передбачити всі можливі варіанти є недоречним. Складно реалізувати структуровані правила реалізації неформальних подій, але за рахунок ідентифікації контексту та побудови дерева

залежності можна суттєво звузити список варіантів до мінімуму. Результатом синтаксичного аналізу є синтаксична структура речення у вигляді дерева розбору/синтаксису та залежності лексем. Синтаксичне дерево є графічне подання етапів складових/залежності розбору вхідного тексту згідно контексту.

2.1.6. Семантичний та онтологічний аналіз україномовного тексту

Семантичний аналіз формує структуру змісту тексту на основі уточнення зв'язку лексем на СА та визначення семантичних ролей суб'єктів/об'єктів тексту [219]. Також СЕА відфільтровує некоректні значення лексем та семантичну незв'язність. Для семантичного аналізу тексту використовують як фреймові моделі Мінського та семантичні мережі, так і на основі онтології, референційного та структурного аналізу для формування множини міжфразових єдностей. Результатом СЕА є розуміння змісту та контексту вхідного тексту. Н.М. Леонтьєва виділяє такі типи семантичних структур: лінгвістичні структури речень тексту (локальне розуміння), семантичні мережі цілого тексту (глобальне розмите розуміння), інформаційні структури цілого тесту (глобальне узагальнене розуміння) та структури баз даних та знань (вибіркове спеціальне розуміння). Для СЕА речень запропоновано в [567] відмінкові граматики і семантики (здатність лексеми приєднувати інші лексеми відповідним синтаксичним способом), завдяки яким семантику фраз описують через відношення головного слова з його семантичними відмінками. Наприклад, головне слово *надіслати* описують семантичними відмінками відправника, адресата та об'єкта пересилання. Для аналізу семантики тексту застосовують предикати (продукційні правила) та семантичні мережі (розмічені графи, де вузли є означенням, а орієнтовані ребра – відношення з-поміж них) [567]. У рамках генеративного підходу валентності слів (насамперед дієслів) описують у вигляді спеціальних фреймів (subcategorization frames) [567], а в межах підходу, заснованого на деревах залежності – моделі управління. Теорія дискурсу та прагматики (опрацювання окремих фраз та текстів) заснована на дослідженнях

Ван Дейка [725-734]. Для дискурсивного синтезу зв'язних текстів аналізують анафоричні посилання та інші явища дискурсу [567].

В семантичній моделі типу *зміст↔текст* [567] розглядають особливий перетворювач заданого змісту (інваріанти усіх синонімічних перетворень тексту) в текст і навпаки [567]. Зміст зв'язного фрагмента без розчленування на фрази/словоформи подають у вигляді спеціальної семантичної структури з двох компонентів: семантичного графа та відомостей про комунікативну організацію сенсу через семеми (sememe, семантична одиниця) та семи (seme, змістовна одиниця; атом семеми) [567]. Лексема складається із семеми (лексико-семантичний варіант – різні значення) та семи (формальний варіант), зміна семантичних значень якої відбувається через розширення (збільшення значень), звуження (конкретизація значення) та зміщення (перевизначення значення). Термін семи введений Eric Buyssens [735-736] і досліджений Bernard Pottier [737].

Семеми є фундаментом розбудови онтологій ПО. Подібним до семеми згідно досліджень Leonard Bloomfield [738] та Kenneth Pike [739] є епісемема як одиниця значення тагмеми (найменший функціональний елемент у граматичній структурі мови). Це аналог морфеми, визначеної як найменша значуща одиниця лексичної форми. Процес ідентифікації сем у значенні слів є компонентним аналізом (розщеплення значення лексеми на компоненти як семеми, маркери або семантичні множники) на основі вибудовування бінарних опозицій. Класичними опозиціями є еквіполентні (класифікація за якісною відмінністю), градуальні (класифікація за різною градацією ознаки) та приватні (дихотомічна класифікація елементів за наявністю/відсутністю диференціальної ознаки).

Marcus Solomon запропонував ще диз'юктивні (відсутність подібності) та нульові (тотожні) опозиції [740-747]. Nikolai Trubetzkoy в [748-750] на противагу класичним опозиціям між членами запропонував по відношенню до системи: багатомірні (відношення яких покриває інші опозиції), ізольовані (відсутність іншої опозиції з подібним відношенням) та пропорційні (тотожність відношень між членами двох опозицій, тобто наявність кореляції для ідентифікації певної мовленнєвої закономірності).

Основи компонентного аналізу складової структурної семантики (Structural semantics або structuralist semantics) заклали Bernard Pottier [751] та Algirdas Julien Greimas [752-754] на основі методики Nikolai Trubetzkoy [748-750] – опозитивного фонологчногої аналізу через порівняння фонем з ідентифікацією їх ознак. Компонентний аналіз напряму пов'язаний із теорією семантичного поля на основі досліджень Roman Jakobson [755-759], Louis Trolle Hjelmslev [760-765] та інших лінгвістів з акцентом перенесення принципів фонології Nikolai Trubetzkoy [748-750] в граматику (опис відмінкових значень) та семантику (опис семантичного поля). У порівнянні із фонологією тут число диференційних ознак значно зростає та є неоднорідними за ступенем узагальнення (чим більш узагальнені семантичні ознаки, тим менші їх кількість, і навпаки, конкретизовані семантичні ознаки, тим більша їх кількість. Предметно-логічний аналіз є надлишковим та неефективним.

Синтагматичний (дистрибутивний) та парадигматичний аналіз на сьогодення більш надійні на основі дослідження семантичного поля (множина слів та їх значень з парадигматичними відношеннями на основі семантичної інтегральної ознаки та відмінні принаймні за однією диференціальною ознакою). Слова та ознаки семантичного поля утворюють ієрархічно організовані структури як онтології, наприклад, на основі інтегральної ознаки спорідненості та такими диференційними ознаками як ступінь, наслідування, покоління тощо. Семантична ознака в різних семантичних полях має різний ієрархічний статус (від елемента ознаки категорії до диференціальної ознаки).

Структурна семантика започаткована дослідженнями Ferdinand de Saussure [766-767] та продовжена в теорії лексичного поля [768-771], реляційної семантики від John Lyons [772-776], компонентного аналізу (Eugenio Coseriu [777-780], Bernard Pottier [751] та Algirdas Greimas [752-754]), генеративної лінгвістики від Noam Chomsky [416-430]. Ferdinand de Saussure стверджує, що мова є системою взаємопов'язаних одиниць і структур і що кожна одиниця мови пов'язана з іншими в межах однієї системи [766-767]. Відомими розробниками структурної семантики були Horst Geckeler [778, 781], Kurt Baldinger [782-784],

Klaus Heger [785-786], Émile Benveniste [787-788], Louis Hjelmslev [760-765]. Carl Hempel [789-793], Willard van Orman Quine [794-799] та Karl Popper [800-802] активно провадили дослідження зв'язків між значеннями термінів у реченні та того, як значення може складатися з менших елементів.

Структуралізм є дуже ефективним аспектом семантики, пояснює узгодженість у значенні певних слів і висловлювань. Концепція змістовних відношень як засобу семантичної інтерпретації є відгалуженням цієї теорії. Структуралізм змінив семантику до її теперішнього стану, а також допомагає правильному розумінню інших аспектів лінгвістики. Наслідковими сферами структуралізму в лінгвістиці є змістовні відношення (лексичні і фразові).

Зміст зв'язного фрагмента тексту без розчленування на фрази і словоформи подається у вигляді спеціального семантичного уявлення (онтології) [803-809], що складається з двох компонентів: семантичного графа та значень про комунікативну організацію змісту. Особливості теорії: орієнтація на синтез текстів (здатність породжувати змістовно-коректні тексти) [803-809]; багаторівневість та модульність, зокрема наявні рівні як глибинний (семантичний) та поверхневий (чистий) синтаксис; інтегральність; збереження кожної рівневої інформації відповідним модулем з переходом на наступний рівень; спеціальні засоби опису синтаксису (правил з'єднання одиниць) на кожному з рівнів на основі множини лексичних функцій через сформульовані правила синтаксичного перифразування; акцент на словник, а не на граматику (збереження інформації різних рівнів мови, зокрема, для синтаксичного аналізу застосовують моделі управління слів, що описують їх синтаксичні та семантичні валентності). Семантична модель типу *зміст↔текст* [567] спирається на тлумачно-комбінаторний словник, у словниковій статті якого крім морфологічної, синтаксичної та семантичної інформації (синтаксичні та семантичні валентності) подані відомості про лексичну сполучність цього слова.

Також застосують словники синонімів, паронімів (зовні подібних слів, які різняться за змістом), основ типових словосполучень, тезауруси (семантичний словник із змістовними відношеннями слів як синоніми, відношення рід-вид,

частина-ціле, асоціації тощо) та онтології (сукупності семантично залежних понять відповідно до множини продукційних правил) [811-822]. Онтології розробляють на основі лексики (лінгвістичні, наприклад, WordNet, EuroWordNet) та граматики (множина правил виразу загальних синтаксичних властивостей слів та груп слів) природньої мови, тип яких залежить від моделі синтаксису [803-809]. Через наявність неоднозначності на глибших рівнях аналіз тексту природною мовою в рамках одного із NLP-етапів часто не може бути однозначно та коректно виконаний семантичний аналіз [823-848]. Тоді в таких ситуаціях найкращим варіантом є генерування множини найбільш ймовірних результатів аналізу на основі інтелектуальних методів опрацювання даних [849-861]. Однак, але застосування такого підходу призводить до значних обчислювальних навантажень, а оптимізація через відкидання частини результатів призводить до ймовірності втрати актуальної інформації та відсутності допустимих інтерпретацій на наступних етапах семантичного аналізу. Іншим підходом є застосування специфікованих структур, де інформацію подають в неповній формі на кожному NLP-етапі для уникнення вибору між різними варіантами. Використання структур ознак дозволяє подати інформацію в специфічній формі при наявності ознак без значень для відповідних змінних. Але неоднозначність у вигляді чи в одну структуру ознак вкладена інша або навпаки. Рішенням є застосування семантики мінімальної рекурсії (Minimal recursion semantics, MRS) [823-825] як перетворення вкладеної структури ознак (або предикатів) на плоску – множину структур, об’єднаних кон'юнкціями. Семантика мінімальної рекурсії є основою для комп’ютерної семантики та реалізований у формалізмах типізованої структури ознак/фраз (feature structure) [826-829], таких як граматика структури фрази, керованої головою (Head-driven phrase structure grammar, HPSG) [830-834], і лексична функціональна граматика (Lexical functional grammar, LFG) [835-839]. Розвинута Ivan Sag [840-845], Carl Pollard [830, 840-841, 845], Dan Flickinger [841-842], Ann Copestake [823-825, 829, 841-842] для розбору мови обчислень і генерації природної мови. Дозволяє формулювати граматичні обмеження для лексичної та

фразової семантики, включаючи принципи семантичної композиції, наприклад, при машинному перекладі. Формалізм RMRS (Robust Minimal Recursion Semantic) є розвитком MRS [825, 846-848], відмінність якої є в розбитті структури з кількох ознак (багатоаргументні предикати) на однознакові (бінарні предикати). Якщо структури ознак подати як орієнтовані графи через множини ребер, для кожного з яких вказано початкову і кінцеву вершину, причому такі вказівники подають як константи/змінні. У поданні можуть бути задані додаткові обмеження, наприклад, вимоги щодо різниці значення деяких змінних.

2.2. Постановка проблеми опрацювання україномовного тексту

2.2.1. Загальний аналіз проблеми аналізу україномовного тексту

Кожна природна мова має особливу структуру та унікальну колекцію лінгвістичних одиниць для генерування змістовного контенту (Таблиця 2.3), що в свою чергу суттєво ускладнює/унеможливлює процес адаптації NLP-алгоритмів однієї мови для іншої для розв'язку конкретної NLP-задачі. Для розроблення нових NLP-методів для конкретної мови при розв'язку конкретної NLP-задачі необхідно багато ресурсів, зусиль та часу, що приводить до не конкурентоспроможності відповідних проектів. Але основна складність полягає зазвичай у відсутності в таких проектах носіїв мови як фахівців на перетині галузей IT, AI та CL, бо не носій мови в мислені обмежений структурою та особливостями власної природної мови. Наприклад, в українській мові є лінгвістичні звороти незрозумілі для більшості іноземців [716], зокрема, *на столі стакан стоїть, а виделка лежить*. Але якщо *вткнути* цю ж *виделку* в стіл, вона буде *стояти*. Ніби просто – горизонтальні речі лежать, а вертикальні – стоять. Але це не так – *пательня та тарілка стоять на столі*, але *тарілка лежить в пательні*. *Кіт на столі* може *лежати, сидіти* або *стояти*, а *жива пташка* – лише *сидіти*, а *іграшка пташки – лежати, опудало пташки – стояти*. *Чобіт – сидить на нозі*, але *стоїть/лежить біля столу*. *Сукня/спідниця гарно сидить на дівчині*. Для не носія мови тут взагалі немає логіки. В англійській мові все просто

– об’єкт/суб’єкт *є на/біля/під* тощо об’єкті/суб’єкті. Це одна із основних причин, чому для не носіїв мови українська мова є досить складною та незрозумілою.

Таблиця 2.3

Типова структура природної мови [862]

Лінгвістичний аналіз текстового контенту	NLP-процеси				
	Структурний			Аналітичний	
	Графемний	Морфологічний	Лексичний	Синтаксичний	Семантичний
Письмо (орфографія)	літера	склад	слово	речення	корпус
Усне мовлення (фонетика)	звук				

Для повною мірою використання закодованих мовою даних необхідно і достатньо розглядати будь-яку природню мову не як зрозумілу і природну, а як необмежену та неоднозначну. Лінгвістичною одиницею аналізу текстового контенту є *лексема* (послідовність закодованих символів/байтів). *Слова* є ширшим значенням лексем, зокрема змістовною послідовністю символами у вигляді словесної конструкції зображення/звуку. Лексеми не є словами. Слова не мають універсального фіксованого значення, незалежного від контекстів культури/мови. Англійці та німці застосовують адаптивні форми слів з суфіксами і префіксами, які змінюють час, стать тощо [862-866]. Китайці ж розпізнають множину піктографічних зображень, де значення ідентифікують через порядок послідовності. Українці на відміну від англійців для зв’язки між самостійними лінгвістичними одиницями застосовують зміну закінчень, звуків у коренях, формотворчих афіксів та суплетивізму (додаток А, таблиця А.1) [716, 862-864]. Англійці ж застосовують для цього порядок лінгвістичних одиниць у поєднанні із службовими словами (артиклів, часток, прийменників). Синтетичні мови у порівнянні з аналітичними є більш архаїчними та мають більш розвинуту морфологію, тому складнішу семантику. Надмірність, неоднозначність та візуальні асоціації визначають природні мови як динамічні, здатні швидко/оперативно розвиватися і передавати досвід сьогодення. Наприклад, сучасний розвиток смайликів (емограм) дозволяє перекласти коротко/змістовно дитячу/ підліткову художню шкільну літературу. Коли ж буде розроблена формальна граматика та визначені граматичні/синтаксичні правила застосування смайликів, то ця мова ще більше зміниться, адаптувавшись до потреб сьогодення

та розвитку ІТ (зміна або збільшення змісту конкретних смайликів, поява нових та перетворення на архаїзми інших, тощо). Під час написання цієї дисертаційної роботи українська мова набула деяких трансформацій. Зокрема, 22 травня 2019 року Кабмін ухвалив нову редакцію Українського правопису (процес зміни тривав із червня 2015 р., громадське обговорення – з 2018 р.). Перехідний етап складатиме 5 років – до 2024 р. Але кожна нова зміна впливає на правила опрацювання україномовного текстового контенту КЛС. Це додавання нових не лише символів/слів і структур для адаптації мови на сьогодення, але і визначень/контекстів/методів вживання. Для чіткої ідентифікації значення слів необхідно більше розрахунків та аналізу, ніж простий пошук в словнику КЛС.

2.2.2. Основні проблеми опрацювання україномовного тексту

Україномовний текстовий контент незалежно від стилю зазвичай містить значний обсяг неструктурованої абстрактної інформації. Це змістовний ланцюг лінгвістичних одиниць з наперед визначеною структурою, цілісністю та зв’язністю. Коректний, оперативний та повноцінний контент-аналіз відповідного україномовного тексту дозволяє розв’язати багато сучасних NLP-задач. Розбір текстового україномовного контенту на лексеми на основі скінчених автоматів та граматики Хомські є класичним підходом. Але він не вирішує основні проблеми опрацювання україномовного текстового контенту, зокрема:

1) Коректне узгодження всіх словоформ в реченні, особливо при використанні/генеруванні дієприкметників в складних реченнях [404]. Середня довжина слова для англійської мови складає біля 4,3-4,4 літери (3,5 фонеми), а для української мови – біля 4,9-5,2 фонеми/літери (залежить від жанру). Але середня довжина англомовного речення більша за україномовного із-за наявності артиклів та службового слова *of*. Якщо ж не враховувати артиклі (вважати, що артиклі є невід’ємною частиною більшості іменних груп) та *of* (це лише зв'язок всередині іменної групи), то середня кількість слів в англомовному реченні значно зменшиться. ІІП відбувається по ключових словах без врахування артиклів та *of,* хоча останнє значно впливає на результат семантичного аналізу.

В україномовних тестах таких підказок простих немає – там треба враховувати флексії по відношенню до основ слів та місцерозташування цих слів по відношенню один до одного, враховуючи знаки пунктуації та інші службові слова. Згідно з [867-869] середня кількість знаків в україномовному реченні 72,4, в англомовному - 83,5; літер в україномовному реченні 67,7, а в англомовному – 79,2; слів 13,1 та 18,2 відповідно. Якщо не враховувати артиклі та *of* – в англомовному реченні середня довжина слів 10-11. Але якщо розглядати лише розмовний текст (діалог), то розрив відповідних значень зростає між цими показниками [870-871]. Це спрощує опрацювання англомовних текстів, і майже не спрощує опрацювання україномовних текстових діалогів.

2) Наявність та узгодженість складних речень. Згідно з [870-873] застосування складносурядних речень в англомовних текстах приблизно 11%, відповідно в україномовних текстах – 15%. Відповідно застосування складнопідрядних речень – 89% та 85% серед 300 зразків для кожної відповідної мови серед всіх складних речень. Але автор навіть в наведених ним прикладах не врахував, що в україномовних складних реченнях часто присутнє більше двох речень, порівняно з англомовними варіантами. Крім того речень з поєднанням сурядного та підрядного зв’язку в україномовних більше не лише за чисельністю, але за варіаціями 12% та 9% відповідно для цих двох мов. Значить правил опрацювання має бути більше, що впливає на складність аналізу. В загалом частка використання складних речень коливається в межах 10-40%, в залежності від стилю автора та жанру текстового контенту.

3) Аналіз номінативних речень (буттєвих, оцінних та вказівних) та їх особливостей (без використання дієслівних груп в реченні/висловлюваннях, головний член – іменникова група) в діалогових текстах. Згідно з [874] в проаналізованих текстах кожної з двох мов для англомовних текстів буттєві речення 25%, для україномовних – 41%. Відповідно для оцінних речень 75% та 55% відповідно. Вказівні в більшості характерні для україномовних текстів – 4%. Проблема лише в тому, що аналіз автор провадив лише серед номінативних речень відповідних мов без врахування часто вживання цих типів речень серед

інших в загальних текстових масивах даних. Зазвичай такі речення використовую в поезії. І на противагу англійській мові, в україномовних текстах, особливо в діалогах часто застосовують номінативні речення – *Зараз тепло. Сьогодні холодно. А ти весела людина. Посміхайся! Знову в школу. Вже кінець літа. Десь попереду.*

4) Відсутність чіткої структури речення на відмінну від англійської мови, яка має у реченні твердий (прямий) порядок лінгвістичних одиниць (*підмет* → *присудок* → *додаток* як ядро речення – Рис. 2.7) [862]. Наприклад, для одного англійського речення *Teenagers like music* українською існує 6 варіантів, зокрема:

Підліткам подобається музика. → *Музика подобається підліткам.* → *Підліткам музика подобається.* → *Музика підліткам подобається.* → *Подобається підліткам музика.* → *Подобається музика підліткам.*

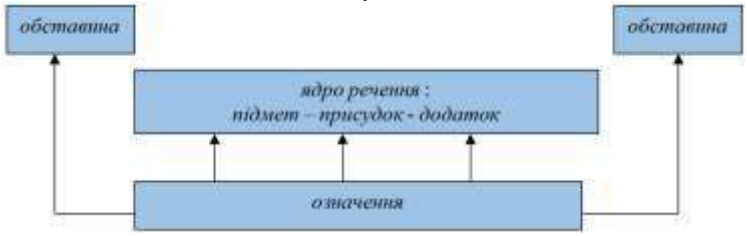


Рис. 2.7. Правила побудови англомовного речення

Помінявши місцями слова в англійській мові *teenagers* та *music* спонукає розуміти речення так, що *музиці подобаються тинейджери* (відсутність сенсу) 862-864]. Але для речення *teenagers like singer* (молоді подобається співак) перестановка слів як *singer like teenagers* (співаку подобається молодь) призведе до утворення нового змісту тексту. В українській завдяки відповідності флексій припустима перестановка слів з уникнення утворення нового сенсу/нісенітниці. Але це ускладнює значно реалізацію POST-процесу для ідентифікації сенсу.

5) Складність ідентифікації іменної групи, яка може виконувати різні функції, зокрема: підмета речення, додатка, обставини одночасно з прийменником, означення або іменної частини складного присудка в тому числі з прикметником, займенник, власна назва або скорочення без відповідного відображення в словнику. Іменна група визначається множиною відповідної змістовної лексики мовця з врахуванням його суб'єктивізму, зокрема, слова або словосполучення відносяться до однієї із категорій як:

1. Прямі однозначні визначення, незалежно від контексту тексту.
2. Зміст залежить від конкретного контексту тексту (багатозначні) або мають значення, відмінне від їх словотворчих складників.
3. Новоутворені, запозичені або вузькоспеціалізовані, яких немає в загальнодоступних словниках та зміст яких неоднозначний.

6) Складність ідентифікації прикметника (якість, ознака або властивість іменника) в іменній групі (не лише по його закінченню в українській мові та місцерозташуванню – зазвичай перед іменником або іншим прикметником). Розрізняють якісні, відносні та присвійні прикметники, а також за простою або складною формою вищого/найвищого ступеня порівняння.

7) Складні ідентифікації дієслівної групи в залежності від можливих складових цієї групи (дієслів, іменних груп як обставини, дієприкметників, дієприслівників тощо) та словозмін в залежності від часу (майбутній, минулий, давноминулий та теперішній), форми дієслова (інфінітив, особова, дієприкметник, безособова, зворотна та дієприслівник), виду (недоконаний, доконаний), перехідністю/неперехідністю (наявність/відсутність прямого додатка), дієвідмінювання (I/II), способів (дійсний, умовний та наказовий) та стану (активний або пасивний). Для відповідних утворень дієслів застосовують префікси, суфікси, чергування звуків/літер, наголосу та різних основ.

8) Складність полягає ще у наявності великого діапазону синонімів для опису явищ/подій тощо, морфологічного аналізу українського тексту та змісту конкретної NLP-задачі. Наприклад, для рубрикації україномовного тексту або визначення авторства статті ускладняється процесом ідентифікації множини ключових слів (наявність синонімів та складність МА) та стійких словосполучень (із-за нестрогого порядку слів, наявності декількох варіантів слів за одним змістом та різноманіття стійких словосполучень). Задача реферування україномовного тексту ускладняється всіма NLP-етапами аналізів від графемного до прагматичного.

9) Побудова е-словників, тезаурусів та граматик є об’ємним та складним процесом, ніж розробка лінгвістичної моделі та відповідного NLP-модуля.

Автоматизація побудови лінгвістичних ресурсів або віртуальних бібліотек [875-878] є одним із перспективних напрямів досліджень комп’ютерної лінгвістики, але напряму пов'язаний з коректо побудованим попередньо описаними рівнями аналізу природньої мови як морфологічний та синтаксичний [567]. Е-словники зазвичай генерують конвертацією звичайних текстових словників, проте для їх коректної побудови додатково використовують колекції та корпуси текстів відповідної ПО, зібрані за певним принципом категоризації (за жанром, авторської належності тощо) та відповідним чинно розмічені/марковані (анотовані) – акцентно, морфологічно, синтаксично тощо [269-277, 879-882].

Зазвичай розмічені корпуси створюють лінгвісти та їх застосовують для різних лінгвістичних досліджень та налаштування КЛС на основі математичних методів машинного навчання, наприклад, для ІІП, машинного перекладу, виправлення помилок, аналізу анафоричних посилань, розпізнавання/синтезу мовлення, розв’язання лексичної неоднозначності тощо. Корпуси текстів завжди обмежені за поданням в них мовленнєвих явищ і це є суттєвим недоліком. Тому найкращим варіантом є використання як лінгвістичного ресурсу текстові потоки конкретної мови в Internet як бази корпусів текстів з достовірних джерел [586-587]. Але це потребує розроблення спеціальних ІТ та відповідних КЛС.

10) Відсутність загальних правил та стандартів структур типових КЛС та етапів розроблення в свою чергу приносить свої недоліки для побудови таких систем. Тому необхідно розробити NLP-моделі/NLP-методи та загальну структуру типової КЛС. Також полегшить роботу визначення функціональних вимог, типової архітектури та рекомендації розроблення відповідних КЛС на основі сучасних методів ML.

2.3. Проект типової комп’ютерної лінгвістичної системи

2.3.1. Основні характеристики комп’ютерної лінгвістичної системи

Метою типової КЛС є реалізація методів та апробація ІТ інтелектуального аналізу текстового потоку для розв’язку конкретної NLP-задачі. Проектування загальної структурної схеми КЛС спричиняє конкретизації/типізації ІТ

інтелектуального аналізу текстового потоку в КЛС через основні етапи інтеграції/управління/супроводу для оптимальності/якості/ефективності розв'язку конкретної NLP-задачі для спеціалізованої ПО [883-887]. Застосування таких КЛС зменшує загальний час опрацювання/аналізу інтегрованих текстових потоків інформаційних ресурсів [888-905], статистики/динаміки життєвого циклу текстового контенту (TCLC) [586-587], діяльності постійних/потенційних користувачів, та функціонування КЛС (Рис. 2.8) [906-911] та зростання обсягів функціональних можливостей КЛС та постійної/потенційної цільової аудиторії.

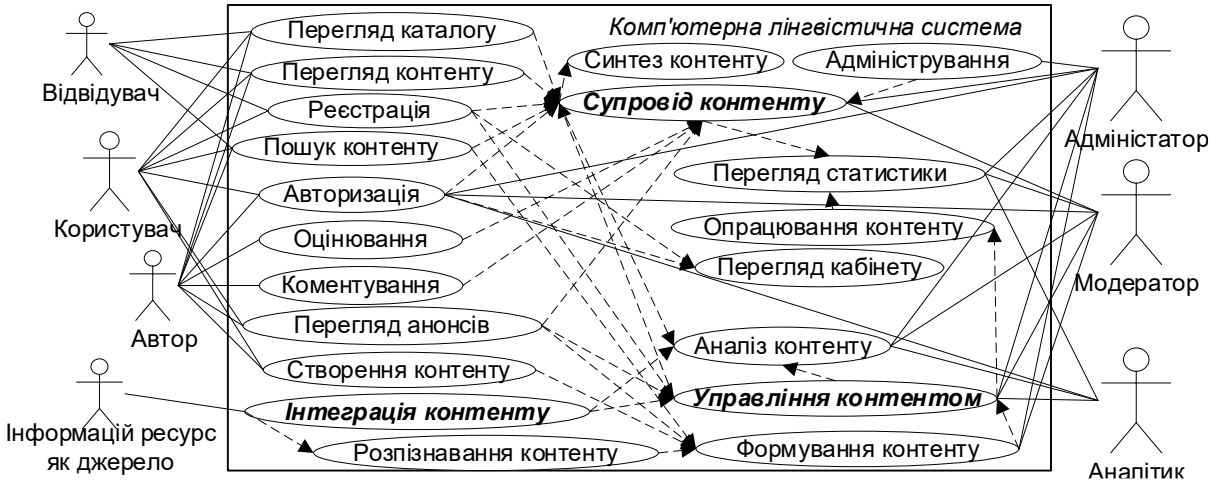


Рис. 2.8. Діаграма use case проекту типової КЛС

Процес інтелектуального аналізу текстового потоку в КЛС складається з:

1) інтеграції контенту на основі розпізнавання та аналізу тексту [586-587, 912-917] (збирання/створення/формування текстового контенту з різних джерел, фільтрування/збереження, форматування, структурування, сортування/анотування, кластеризація та класифікація, формування/генерування відповідних правил фільтрування/ІІП/інтеграції/розпізнавання/аналізу);
2) управління контентом на основі аналізу та опрацювання тексту [586-587, 917-935] (заповнення БД/СД/БЗ; кешування популярних інформаційних блоків/Webpage/результатів ІІП; збір/аналіз статистичних даних динаміки функціонування КЛС, конверсії відвідувань користувачів та історії переходів між контентом; формування звітів згідно запитів користувачів; персоналізація

діяльності користувачів; генерування Webpage/форм згідно запитів користувачів для розв'язку конкретної NLP-задачі; підтримка ІІП контенту; підтримка інтерактивної взаємодії з постійною аудиторією; генерування та постійне оновлення контенту Website; збереження відгуків, коментарів, голосування постійної аудиторії);

3) супровід контенту на основі аналізу та синтезу інформації [586-587, 936-943] (генерування та оновлення інформаційних зрізів/портретів відносно часових проміжків потоку контенту [888-905], потенційних/постійних персоналізованих користувачів та цільової аудиторії; ідентифікація та оновлення сюжетів/сценаріїв класифікованого контенту відносно часових проміжків; генерування семантичних схем відношень контенту; ранжування контенту/аналітиків/модераторів/авторів; ідентифікація, кластеризація та класифікація конверсії/дій постійних користувачів/відвідувачів та відповідно подій в потоках контенту).

Коефіцієнти конверсії K_{wcv} (досягнення мети користувачами відповідно до всіх дій відповідного змісту) для КЛС розраховують наступним чином [888-905]:

$$K_{wcv} = \frac{N_{wcv}}{N_{vrb}};\ K_{wcv} = \frac{N_{wcv}}{N_{vtb}};\ K_{wcv} = \frac{N_{wcv}}{N_{wvr}};\ K_{wcv} = \frac{N_{wcv}}{N_{wvt}};\quad (2.4)$$

де N_{wcv} – число конверсії КЛС, N_{vrb} – загальна число користувачів Website, коли досягнута відповідна конверсія (успішна конверсія), N_{vtb} – загальна кількість відвідувань Website, коли досягнута відповідна конверсія, N_{wvr} – загальна кількість користувачів Website, N_{wvt} – загальна кількість відвідувань Website.

КЛС застосовують для розв'язку конкретної NLP-задачі згідно відповідних вимог/потреб кінцевого користувача або потенційної аудиторії, наприклад, для реалізації е-бізнесу інформаційного обслуговування на основі ІТ, машинного навчання та основних етапів NLP. КЛС застосовують для надання інформаційних послуг у відповідних сферах діяльності постійного користувача та цільової аудиторії, наприклад, для продажу контенту через Internet-магазин,

Internet-видання, Internet-журнал, Internet-видавництво, Internet-газета, Internet-маркетинг, надання консалтингових або SEO послуг тощо [944-953].

КЛС використовують як додаткову підсистему системи електронної комерції для просування інформаційних послуг/товарів, наприклад, через інформаційні агентства, освітні заклади, журнали, компанії розроблення ПЗ, газети, видавництва, тощо [944-953]. Необхідність застосовувати КЛС для розв'язання різних NLP-задач пов'язано із прискореними оперативними темпами збільшення обсягів/масштабів текстового потоку контенту в Internet/е-бізнесі та зростання/розповсюдження доступу до різноманітних джерел інформації, збільшення множини функціональних можливостей КЛС та автоматизації розв'язку різноманіття NLP-задач, збільшенням попиту/потреб на актуальну/релевантну/оперативну інформацію, розробленням/впровадженням ІТ/ПЗ опрацювання текстів відповідної природньої мови та зростанням кількості ПО застосування NLP-технологій для досягнення поставленої мети кінцевого користувача або цільової аудиторії комп'ютерної лінгвістичної системи.

2.3.2. Обґрунтування реалізації проекту типової КЛС

Відсутність стандартизованих загальновідомих та некомерціалізованих ІТ розроблення типової КЛС та основних модулів інтелектуального аналізу текстових потоків контенту спонукає до збільшення множини проблем проектування загальної структури ІС розв'язку конкретної NLP-задачі залежно від самої природної мови. Із-за відсутності загальноприйнятої стандартної та детальної типізації КЛС та NLP-задач є проблемним процес розроблення спеціалізованих ІТ/ІС/ПЗ інтелектуального аналізу текстових потоків контенту. З цього випливає проблематичність стандартизації основних процесів/модулів КЛС як супровід/інтеграція/управління текстовим контентом конкретної мови.

Згідно застосувань на Website S_{wtm} КЛС є модуль розв'язку конкретної NLP-задачі M_{dis}, модуль супроводу контенту M_{dmr}, модуль інтеграції контенту M_{dcp} для підтримки написання якісного/ефективного актуального унікального контенту копірайтерами контенту Website, журналістами, авторами тощо та

M_{dvm} модуль управління контентом. Для кожного розраховують власний ключовий показник ефективності KPI (англ. Key Performance Indicators) [937]:

$$S_{wtm} =< M_{dis}, M_{dmr}, M_{dcp}, M_{dvm} >. \quad (2.5)$$

Загально-розповсюджені та популярні сучасні КЛС функціонують на основі невідомих методів для більшості NLP-фахівців у зв'язку з цим, що ці КЛС є комерційними проектами закритого типу. При розроблені нових КЛС NLP-фахівці створюють заново або модифікують методи/засоби/модулі інтелектуального аналізу текстових потоків контенту та супроводу TCLC. Досить багато матеріалів є в загальному доступі про ІТ на основі комп'ютерної лінгвістики. Але в більшості випадків вони несуть суто теоретичне навантаження і майже не відображають практичних рекомендацій навчання фахівців з опрацювання конкретної мови. В більшості ці матеріали присвячені опрацюванню саме англійської мови. І майже відсутні для української мови.

Відсутні в широкому доступі публікації щодо якості/ефективності впливу наявності реалізованих етапів TCLC на динаміку роботи КЛС для інтелектуального аналізу текстових потоків контенту українською мовою. Дослідження динаміки роботи КЛС практично відсутні із-за неможливості організації доступу широкого кола дослідників до адміністративних панелей підсистем сучасних популярних КЛС із-за їх комерціалізації.

Актуальність реалізації КЛС-проекту полягає у розробленні основної структури, уніфікованих методів/модулів/ІТ/ПЗ побудови КЛС та основних TCLC-етапів. Впровадження основних модулів інтеграції/управління/супроводу контенту в КЛС спричиняє зменшення етапів/часу генерування результатів згідно запитів постійних користувачів та відповідно інтелектуального аналізу текстових потоків контенту українською мовою. Паралельно це спонукає до зростання обсягу потенційної/постійної цільової аудиторії користувачів КЛС, що дозволяє накопичувати статистичні дані функціонування КЛС для подальшого машинного навчання на основі аналізі зібраних великих даних. Це призводить до активного та оперативного зростання/адаптації функціональних можливостей відповідних КЛС. Розроблення загальних основних рекомендацій щодо

проектування та розроблення архітектури КЛС на основі основних TCLC-етапів та модулів інтелектуального аналізу текстових потоків контенту дасть можливість ефективно/якісно/своєчасно/оперативно підтримати життєвий цикл побудови відповідних КЛС на декількох рівнях. Зокрема на рівні розробника це спричиняє скорочення обсягів часу/ресурсів на реалізацію та збільшення якості/ефективності функціонування КЛС, а також уніфікація/стандартизація процесів інтелектуального аналізу контенту (Рис. 2.9-Рис. 2.10). На рівні власника – збільшення рентабельності та зацікавленості постійної аудиторії. На рівні користувача – збільшення обсягів вибору функціональних можливостей КЛС, підтримка/спрощення інтерфейсу, результативність/зрозумілість).

	Название задачи	Тривалість	Початок	Завершення	Попередники
0	**Розроблення КЛС**	**36 днів**	**Пн 12.04.21**	**Пн 31.05.21**	
1	Збір/уточнення даних з ПО конкрентної NLP-задачі	2 днів	Пн 12.04.21	Вт 13.04.21	
2	Формування множини специфікацій ПО відповідної NLP-задачі	1 день	Вт 13.04.21	Вт 13.04.21	1
3	Технічне завдання проекту КЛС конкрентної NLP-задачі	3 днів	Ср 14.04.21	Пт 16.04.21	2;1
4	Уточнення ТЗ через взаємодію із потенційною аудиторією	2 днів	Пт 16.04.21	Пн 19.04.21	3;1
5	Ідентифікація та аналіз множини функціональних вимог до КЛС	3 днів	Чт 15.04.21	Пн 19.04.21	1;2;3;4
6	Аналіз та уточнення архітектури інформаційного ресурсу КЛС	1 день	Вт 20.04.21	Вт 20.04.21	5;2
7	Ідентифікація та аналіз множини нефункціональних вимог	2 днів	Ср 21.04.21	Чт 22.04.21	6;1
8	Розроблення шаблону інформаційного ресурсу КЛС	2 днів	Чт 22.04.21	Пт 23.04.21	4;7
9	Розроблення множини типових шаблонів текстового контенту	2 днів	Пт 23.04.21	Пн 26.04.21	8;4
10	Створення та тестування інформаційног ресурсу КЛС	4 днів	Пн 26.04.21	Чт 29.04.21	9;3;5
11	Аналіз та уточнення архітектури КЛС відповідної NLP-задачі	6 днів	Чт 22.04.21	Чт 29.04.21	5;3
12	Аналіз архітектури підсистеми управління контентом КЛС	3 днів	Ср 28.04.21	Пт 30.04.21	11;5
13	Створення підсистеми управління текстовим контентом	6 днів	Пт 30.04.21	Пт 07.05.21	12;10
14	Розроблення сховища даних текстового контенту КЛС	2 днів	Чт 29.04.21	Пт 30.04.21	11
15	Створення бази знань для управління текстовим контентом	2 днів	Пт 30.04.21	Пн 03.05.21	14
16	Наповнення сховища даних текстового контенту КЛС	5 днів	Нд 02.05.21	Пт 07.05.21	15
17	Розроблення КЛС для розв'язку конкрteної NLP-задачі	7 днів	Чт 06.05.21	Пт 14.05.21	16;13
18	Аналіз архітектури інтеграції контенту з різних джерел	3 днів	Чт 13.05.21	Пн 17.05.21	5;10;13;16
19	Створення підсистеми інтеграції тектового контенту	5 днів	Сб 15.05.21	Пт 21.05.21	18
20	Створення сховища даних джерел текстовго контенту	3 днів	Пн 17.05.21	Ср 19.05.21	19
21	Створення бази знань для фільтрів текстового контенту	3 днів	Вт 18.05.21	Чт 20.05.21	20;17
22	Аналіз архітектури підсистеми супроводу контенту КЛС	3 днів	Вт 18.05.21	Чт 20.05.21	5;19
23	Створення підсистеми супроводу текстового контенту	5 днів	Ср 19.05.21	Вт 25.05.21	22;21
24	Тестування підсистеми управління текстовим контентом	6 днів	Пт 07.05.21	Пт 14.05.21	13
25	Тестування КЛС розв'язку конкрентної NLP-задачі	3 днів	Пт 14.05.21	Вт 18.05.21	17;24
26	Тестування підсистеми інтеграції текстового контенту	3 днів	Пн 17.05.21	Ср 19.05.21	25;19
27	Усунення недоліків роботи підсистеми управління контентом	6 днів	Пн 10.05.21	Пн 17.05.21	24
28	Тестування підсистеми супроводу текстового контенту	3 днів	Пт 21.05.21	Вт 25.05.21	23;26;27
29	Усунення недоліків функціонування КЛС	4 днів	Вт 18.05.21	Пт 21.05.21	25;27
30	Усунення недоліків роботи підсистеми інтеграції контенту	3 днів	Пт 21.05.21	Вт 25.05.21	26;29
31	Усунення недоліків роботи супроводу текстового контенту	4 днів	Пн 24.05.21	Чт 27.05.21	27;28;29;30
32	Підготовка технічної документації розробленої КЛС	5 днів	Вт 25.05.21	Пн 31.05.21	24;25;26;27;28;29;30;31

Рис. 2.9. Орієнтований графік проектування та реалізації типової КЛС

На Рис. 2.9 подано загальний орієнтований план розроблення типової КЛС на основі реалізації основних етапів інтелектуального аналізу текстових потоків контенту українською мовою [586-587] для спрощення аналізу/оцінювання фінансових/часових/ресурсних затрат. Це скорочує часові затрати на реалізацію

КЛС-проекту, зменшує число NLP-фахівців та чітко описує регламент розроблення на основі аналізу обсягу використаного часу на відповідні етапи.

На Рис. 2.10 подано діаграму Ганта проектування та реалізації типової КЛС, що дозволяє проаналізувати чіткий та детальний регламент розроблення типової КЛС як етапів виконання інтелектуального аналізу текстових потоків контенту українською мовою та залучення відповідних NLP-фахівців на цих етапах [586-587]. Результати етапу 1 активізують етапи 2-5 і 7, а етапу 4 – 5, 8-9, що дозволяє завчасно перерозподіляти задачі між відповідними NLP-фахівцями в часі та учасників між командами тощо. Для активації етапу 5 необхідні вихідні результати з етапів 1-4, а результати етапу 5 активізують етапи 12, 18 та 22.

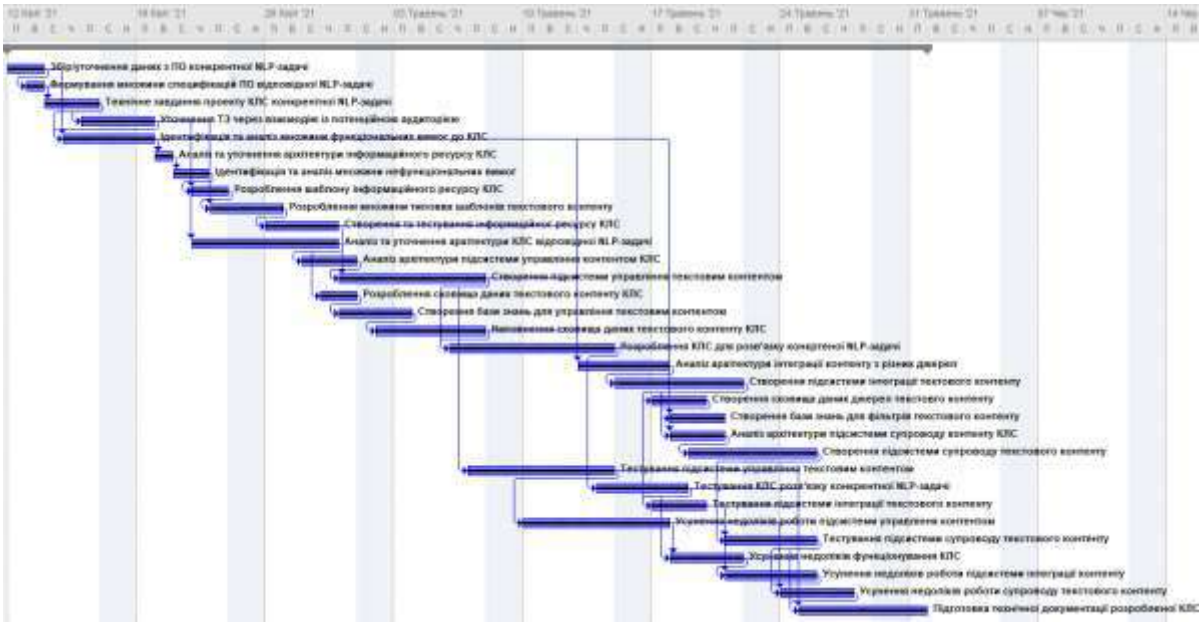


Рис. 2.10. Діаграма Ганта проектування та реалізації типової КЛС

Несвоєчасне реалізація етапу 10 призводить до одночасної затримки реалізації етапів 13 та 18. Скорочення часу реалізації етапів 11, 13, 16, 17, 19, 24 та 27 дозволить завчасно реалізувати КЛС-проект, але призведе до збільшення виникнення додаткових помилок, які зазвичай усувають на етапах 24-31.

2.3.3. Очікувані ефекти реалізації проекту типової КЛС

Прогнозований економічний ефект розв’язку конкретної NLP-задачі залежить від скорочення затрат на створення проекту, й загальної архітектури типової КЛС, застосування додаткових фахівців/спеціалістів/експертів/ресурсів

та наявності чіткого регламенту реалізації відповідних модулів інтелектуального аналізу текстових потоків контенту українською згідно таких факторів:

1. Наявність модуля розв'язку конкретної NLP-задачі [888-905, 937] на основі лінгвістичного опрацювання текстів українською мовою формує множину унікальної цільової аудиторії для подальшого аналізу та фіксації потреб користувачів та відповідне коригування цілей е-бізнесу для збільшення прибутку (не лише фінансового, але інформаційного/ресурсного).

У КЛС з модулем розв'язку конкретної NLP-задачі, ймовірно, більше значення K_{PI} (KPI), так як це є зазвичай основна мета кінцевого користувача за переходам з ІІПС, соціальних мереж, інших Website/банерів та прямого відвідування Website. Оцінювати достатньо за показниками звітів з Google Analytics, наприклад, кількості відвідувачів/користувачів N_{wvr} (але деякі K_{PI} для уточнення необхідно витягнути з інших модулів) [888-905, 937], а також:

$$M_{dis} =< N_{wvr}, S_{gcc}, S_{gco}, S_{gcv}, S_{gro}, P_{wnv}, I_{wnv} >, \quad (2.6)$$

де S_{gcc} – середній коефіцієнт конверсії згідно розрахунків Google Analytics, S_{gco} – середня вартість замовлень згідно розрахунків Google Analytics, S_{gcv} – середня вартість на відвідування (корисність відвідування згідно даних транзакцій електронної комерції) або середня корисність мети відвідування (на основі корисності цілей) згідно розрахунків Google Analytics, S_{gro} – середня P_{ROI} або середнє повернення на інвестиції згідно розрахунків Google Analytics та AdWords, P_{wiv} – відсоток прибутку від нових відвідувачів Website КЛС, I_{wnv} – індекс нових покупців/замовників при першому відвідуванні Website КЛС.

Показник ефективності для загального валового прибутку P_{ROI} [888]:

$$P_{\text{ROI}} = \frac{N_{Inc} - N_{Exp}}{N_{Exp}}, \quad (2.7)$$

де N_{Exp} – витрати, N_{Inc} – прибуток. Якщо P_{ROI}<0, то затрати на залучення користувачів цільової аудиторії більше за прибуток. P_{ROI} не враховує прибуток від надання послуг та кількість користувачів чи транзакцій.

Норми прибутку згідно розрахунків Google Analytics та AdWords [888]:

$$P_{\text{RR}} = \frac{N_{Inc} - N_{Exp}}{N_{Inc}}. \quad (2.8)$$

Число відвідувань, які необхідні для переконання здійснити замовлення, впливають на розрахунок P_{wiv}. Тому ймовірність перетворення нового відвідувача на постійного користувача при першому відвідуванні [888-905, 937]:

$$I_{wnv} = \frac{P_{wtv}}{P_{wnv}}, \quad (2.9)$$

де P_{wnv} – відсоток нових користувачів Website, P_{wtv} – відсоток транзакцій від нових користувачів Website. При I_{nv}=1 новий та повторний користувач з однаковою ймовірністю стануть постійними користувачами. При I_{nv}<1 новий користувач стане постійним з меншою ймовірністю, ніж повторний. І навпаки, при I_{nv}>1 новий стане постійним з більшою ймовірністю, ніж повторний.

2. Наявність модуля супроводу текстового контенту скорочує затрати на модераторів/аналітиків, які здійснюють збір/аналіз статистичних даних динаміки функціонування КЛС, активності постійної цільової аудиторії як реакції на зміни контенту Website/Webpage, формування правил аналізу інформаційних портретів користувачів та тематичних сюжетів контенту.

Для ідентифікації найкращого трафіку досліджують отриманий прибуток та P_{ROI}, затрати на компанію, коефіцієнт конверсії K_{wcv}. Тому K_{PI} модуля супроводу контенту змістовно перетинається з K_{PI} модуля розв'язку конкретної NLP-задачі на основі даних з AdWords. Відмінність у акцентуванні не лише на коефіцієнті конверсії замовлень, але цілей для аналізу/розвитку відношень з користувачами/відвідувачами, які потенційно здійснять замовлення, зокрема:

$$M_{dmr} =< I_{gyk}, K_{gvb}, P_{wap}, P_{wvk}, S_{grk}, I_{gck}, P_{wck}, P_{wvk}, K_{wcz}, P_{wvz} >, \quad (2.10)$$

де I_{gyk} – індекс якості рекламної кампанії згідно AdWords; K_{gvb} – коефіцієнт впізнання бренду; P_{wap} – відсоток нових/повторних замовників; P_{wvk} – відсоток нових/повторних користувачів; S_{grk} – середній P_{ROI} за типом рекламної кампанії; I_{gck} – індекс конверсії цілей за типом рекламної кампанії; P_{wck} – відсоток конверсії цілей за типом рекламної кампанії; P_{wvk} – відсоток відвідувань за

типом рекламної кампанії; K_{wcz} – коефіцієнт конверсії цілей за типом засобу; P_{wvz} – відсоток відвідувань за типом засобу [888-905, 937].

Індекс якості рекламної кампанії I_{gyk} пов'язаний з якісністю/ефективністю та результативністю таргетингу рекламної кампанії (залучення цільового трафіку на Website КЛС [888-905, 937].

$$I_{gyk}(w) = \frac{P_{wcv}(w)}{P_{wvk}(w)}, \tag{2.11}$$

де $P_{wvk}(w)$ – функція визначення відсотку відвідувань від рекламної кампанії w; $P_{wcv}(w)$ – функція визначення відсотку конверсії цілей для відвідувань від кампанії w; $I_{gyk}(w)$ – функція визначення індексу якості рекламної кампанії w. Якщо P_{vk}=50% користувачів переходять з AdWords, але цьому джерелу рекламної кампанії x відповідає лише P_{cv}=20% конверсії, то це неефективний таргетинг. Інша рекламна кампанія y генерує теж 50% трафіку та їй відповідає 80% конверсії, то це ефективний таргетинг. Значення індексу I_{gyk}=1.0 означає, що замовник з даної кампанії здійснить конверсію з тою ймовірністю, як і замовник з будь-якої іншої кампанії. Значення I_{gyk}<1.0 означає відповідно, що замовник з даної кампанії здійснить конверсію з меншою ймовірністю, ніж замовник з будь-якої іншої кампанії. При I_{gyk}>1.0 – замовник здійснить конверсію з більшою ймовірністю, ніж замовник з будь-якої іншої кампанії.

Коефіцієнт впізнання бренду [888-905, 937]:

$$K_{gvb} = \frac{N_{ubq}+N_{utv}}{N_{uaq}+N_{utv}}, \tag{2.12}$$

де N_{uaq} – загальне число користувацьких запитів ІІП (ключові слова); N_{utv} – число прямих відвідувань Website; N_{ubq} – число запитів ІІП із назвою бренду.

3. Наявність модуля інтеграції текстового контенту скорочує затрати на КЛС-модераторів та авторів контенту, автоматизуючи/реалізуючи деякі їх роботи/функції як збір контенту з множини різних достовірних джерел, його розпізнавання, фільтрування, збереження, форматування, аналіз, анотування, кластеризація, класифікація тощо [888-905, 937].

Для розробників КЛС головна мета – максимальне залучення постійної цільової аудиторії, основними ключовими показниками якого є обсяг часу/частота/Webpage на ознайомлення із контентом Website та збільшення зацікавленості користувачів. Для КЛС важливий KPI є обсяг відвідувань/замовлень за визначений період часу. Для аналізу часових показників згідно з повторним відвідуванням обирають найкращі для конкретної моделі КЛС часові проміжки $t_1<t_2$. Тому для модуля інтеграції [888-905, 937]:

$$M_{dcp} =< P_{glt}, P_{gst}, P_{ght}, K_{gvb}, K_{uzv}, P_{uav}, P_{uzv}, S_{gnc}, P_{wvv}, S_{gpv}, S_{gtp} >, \quad (2.13)$$

де P_{glt} – відсоток повторних відвідувань користувача з попереднього відвідування $>t_2$ днів згідно Google Analytics; P_{gst} – відсоток повторних відвідувань користувача з попереднього відвідування в межах $[t_1; t_2]$ днів при $t_1<t_2$ згідно Google Analytics; P_{ght} – відсоток повторних відвідувань користувача з попереднього відвідування $<t_1$ днів згідно Google Analytics; K_{gvb} – коефіцієнт впізнавання бренду; P_{uav} – відсотки нових/повторних відвідувачів згідно Google Analytics; P_{uzv} – відсоток зацікавленості відвідувачів; S_{gnc} – середнє число кліків на рекламі за N_{wvr} відвідувань; P_{wvv} – показник відмов для Webpage P_{vvp}; S_{gpv} – середнє число переглядів Webpage за відвідування згідно Google Analytics; S_{gtp} – середня тривалість перебування на Webpage через AdWords.

Показник відмов для однієї Webpage на основі даних з Google Analytics:

$$P_{vvp} = \frac{N_{vnp}}{N_{inp}}, \quad (2.14)$$

де N_{inp} – число відвідувань користувачами цієї Webpage напряму; N_{vnp} – число односторінкових відвідувань для цієї Webpage через Google Analytics.

Середня кількість кліків на рекламі за N_{vr} відвідувань [888-905, 937]:

$$S_{gnc} = \frac{N_{wcr}}{N_{wav}} \cdot N_{wvr}, \quad (2.15)$$

де N_{wvr} – число відвідувань для аналізу (часто N_{vr}=1000 згідно CPM – Cost Per Mille); N_{wav} – загальне число відвідувань згідно Google Analytics; N_{wcr} – середнє число кліків на рекламі згідно AdWords.

Показник зацікавленості відвідувачів [888-905, 937]:

$$K_{uzv} = \frac{N_{wad}}{N_{wav}}, \quad (2.16)$$

де N_{wav} – загальне число відвідувань згідно Google Analytics; N_{wad} – загальне число дій на Website згідно AdWords.

Відсоток зацікавленості відвідувачів [888-905, 937]:

$$P_{uzv} = \frac{N_{wzv}}{N_{wvk}}, \quad (2.17)$$

де N_{wvk} – загальне число користувачів згідно Google Analytics; N_{wzv} – загальне число зацікавлених користувачів згідно AdWords.

При ефективному ідеальному впровадженні/використанні КЛС [888, 937]:

$$P_{ght} >> P_{gst} >> P_{glt}. \quad (2.18)$$

При періодичному аналіз таких показників ідентифікують закономірності для корегування контенту для підтримки хоча б такого співвідношення.

$$P_{ght} \geq P_{gst} \geq P_{glt}. \quad (2.19)$$

4. Наявність модуля управління текстовим контентом скорочує затрати на модераторів/адміністраторів [888, 937], які оновлюють Website/Webpage та формують правила кешування/ІІП популярних інформаційних блоків.

Модуль управління контентом відповідають за безперервне та ефективне функціонування Website, контролюючи навантаження на сервери (очікуване число звертань користувачів), частота використання типових браузерів/мови:

$$M_{dvm} =< K_{wis}, P_{wep}, P_{gum}, P_{gup}, P_{gur}, P_{gus}, P_{gub}, P_{gul}, P_{wep}, K_{wdu}, S_{wdu} >, \quad (2.20)$$

де K_{wis} – показник внутрішнього ІІП; P_{wep} – відсоток видання Webpage з помилкою; P_{gum} – відсоток мобільних користувачів згідно Google Analytics; P_{gup} – відсоток користувачів з високошвидкісним підключенням до Internet; P_{gur} – відсоток користувачів з низькою/середньою/високою роздільною здатністю дисплею; P_{gus} – відросток користувачів з конкретною операційною системою; P_{gub} – відсоток користувачів з конкретним браузером згідно Google Analytics; P_{gul} – відсоток користувачів з підтримкою англійської/української мови; K_{wdu} – показник кількості користувачів, переглядів та відвідувань Webpage. Показник S_{wdu} є базовим модуля керування контентом згідно Google Analytics [888, 937]:

$$S_{wdu} =< N_{svt}, N_{sut}, N_{spt}, N_{spv} >, \quad (2.21)$$

де N_{spv} – середнє число переглядів Webpage за відвідування; N_{spt} – середнє число переглядів Webpage за конкретний Δt час; N_{sut} – середнє число унікальних користувачів за конкретний Δt час; N_{svt} – середнє число відвідувань за конкретний Δt час.

Відсоток генерування Webpage з помилкою (необхідно мінімізувати):

$$P_{wep} = \frac{N_{wep}}{N_{wpp}}, \quad (2.22)$$

де N_{wpp} – загальне число переглянутих Webpage; N_{wep} – загальне число виданих Webpage з помилкою [888-905, 937].

Показник внутрішнього ІІП згідно Google Analytics [888-905, 937]:

$$K_{wis} =< N_{nns}, P_{uts}, P_{ksp}, P_{bus}, P_{cuss}, P_{pop}, P_{ucs}, S_{vrs}, P_{uos}, P_{uns}, P_{unr}, \quad (2.23)$$
$$P_{uur}, S_{nup}, T_{svs}, P_{uis}, P_{nrp}, K_{wps} >,$$

де N_{nns} – число нульових результатів ІІП по Website; P_{uts} – відсоток користувачів, що перебували $> t$ часу на Website після здійсненого ІІП; P_{ksp} – відсоток користувачів, що переглянути $> k$ Webpage після здійсненого ІІП; P_{bus} – відсоток здійснених покупок серед користувачів, що використовують ІІП по Website; P_{cus} – відсоток покупців серед користувачів, що використовують ІІП по Website; P_{pop} – відсоток відмов після відвідування однієї Webpage як результату ІІП; P_{ucs} – відсоток конверсії від користувачів, що використовують ІІП по Website; P_{unr} – відсоток користувачів, які не використовують ІІП по Website; P_{uur} – відсоток відвідувачів, які використовують ІІП по Website; S_{nup} – середнє число Webpage, переглянутих відвідувачами після ІІП; T_{svs} – середній час перебування на Website для відвідування після ІІП; P_{uns} – відсоток відвідувачів, які проводять декілька ІІП по Website на протязі відвідування (враховуючи декілька ІІП для одного і того ж ключового слова); P_{uos} – відсоток відвідувачів, які покинули Website після перегляду результатів ІІП; S_{vrs} – середнє число результатів ІІП, переглянутих після ІІП; P_{uis} – відсоток відвідувань, в яких використовують ІІП по Website; P_{nrp} – відсоток нульових результатів ІІП по Website, зокрема,

$$P_{nrp} = \frac{N_{nps}}{N_{vps}}, \quad (2.24)$$

де N_{vps} – загальне число переглянутих Webpage ІІП; N_{nps} – загальне число нульових результатів ІІП Webpage [888-905, 937].

Показник використання ІІП K_{wps} по Website як залежності відвідувань:

$$K_{wps} = \frac{N_{wsv}}{N_{wns}}, \quad (2.25)$$

де N_{wns} – відвідування без ІІП по Website; N_{wsv} – відвідування із ІІП по Website. При сучасному поступовому зростанні числа Website КЛС на основі RIA-технології збільшується потреба в розрахунку відповідних K_{PI} [888-905, 937].

5. Наявність підсистем інтелектуального аналізу текстових потоків контенту скорочує час/затрати/персонал/ресурси на своєчасне оперативне отримання релевантного унікального актуального текстового контенту, що призводить до зростання обсягів цільової аудиторії КЛС, зокрема сприяє зростанню економічного ефекту від впровадження КЛС на кілька пунктів.

Аналітики важливі статистичні дані не лише про перегляди Webpage K_{wdu}, а динаміки множини постійних/потенційних/повторних подій/дій K_{was} від замовників/відвідувачів/користувачів на основі взаємодії з Website, зокрема,

$$K_{was} =< S_{wcc}, S_{wtv}, S_{wnv}, P_{wuv}, P_{wnv} >, \quad (2.26)$$

де S_{wcc} – середній коефіцієнт конверсії; S_{wtv} – середня тривалість відвідування; S_{wnv} – середнє число переглядів за відвідування; P_{wuv} – відсоток унікальних замовників/відвідувачів/користувачів; P_{wnv} – відсоток нових замовників Website.

Згідно відстеження подій K_{as} та взаємодії з Website K_{du} аналізують:

$$K_{usa} = \alpha(K_{wdu}, K_{was}) =< P_{vcu}, P_{sau}, P_{siu}, I_{wdx} >, \quad (2.27)$$

де P_{siu} – відсоток взаємодії з Website (наприклад, коментування, голосування, реєстрація, авторизація, підписка тощо); P_{sau} – відсоток користувачів, які активізують різні події (наприклад, клік на рекламу, запуск функції, пауза тощо); P_{vcu} – відсоток користувачів, взаємодіючих з різними типами подання контенту (перегляд наступного спілкування, панорамування, масштабування, тощо); I_{wdx} – значення міри корисності відповідно Webpage/Website/КЛС/контенту [888].

Розрахунок множини різних K_{PI} спонукає звернути увагу на онлайнові стратегії, найефективніші для генерації звернень, залучення користувачів, збільшення конверсії/прибутків е-бізнесу. Це надає можливість оптимізувати загальну структуру Website при розв'язку конкретної NLP-задачі для зростання ефективності/якості його застосування та обсягів постійних користувачів та замовників. Також можна ідентифікувати множину неефективних Webpage.

На основі аналізу даних про постійних користувачів/замовників оптимізують Webpage для Website при розв'язку конкретної NLP-задачі для ефективності/якості відвідування/перебування на ньому. Зазвичай покращують структуру Website через зміни URL-адрес Webpage входження для відповідного зручного/ефективного відвідування замовниками/користувачами конкретних Webpage, виправлення непрацюючих посилань, або коригування відповідного контенту Webpage для розміщення необхідного рекламного блоку. Алгоритм ідентифікації проблемних місць структури Website для подальшої оптимізації:

1. Формування набору популярних Webpage входження на основі аналізу показників відмов від користувачів/замовників.
2. Формування набору неефективних Webpage на основі аналізу ступеня корисності та ефективності/якості щодо функціональності.
3. Аналіз джерел входження (прямі входження згідно URL-адреси з історії попередніх відвідувань або перше пряме відвідування, посилання на/з інших Website, посилання в е-пошті, платна реклама, ІІПС, перехід з соціальних мереж або пошукових систем тощо).
4. Аналіз ключових слів входження відносно джерел/частоти/часу.
5. Візуалізація переходів по Website від користувача для досягнення мети/конверсії та результативності/ефективносі/якості ІІП.
6. Дослідження та аналіз результативності успішності ІІП по Website.

Формування набору неефективних Webpage засобами Web-аналітики провадять через аналіз множини відповідних показників, зокрема [888-905, 937]:

– дерево візуалізації залежних послідовностей (англ. Funnel Visualization);

– набір популярних Webpage входу/виходу (Top Landing and Exit Pages);

– значення міри корисності Webpage I_{wdx}, яке ідентифікують як [888]:

$$I_{wdx} = \frac{R_{wcv}+R_{wec}}{N_{upv}}, \quad (2.28)$$

де N_{upv} – число унікальних переглядів Webpage; R_{wec} – прибуток від е-бізнесу; R_{wcv} – значення міри корисності відвідування користувачів (на основі транзакцій е-бізнесу) та мети відвідування користувачів (на основі корисності цілей).

Якщо Webpage a_i відвідують замовники/користувачі із досягненням мети b_j, то її корисність впливає на зростання значення корисності Webpage a_i. При зростанні частоти відвідування Webpage a_i користувачами із досягненням мети b_j і чим більше значення корисності мети, тим швидше зростає міра корисності Webpage I_{wdx} (результат не має стосунок до конверсії та цілей). Рейтингування Webpage згідно I_{wdx} впливає на послідовність їх оптимізації. Неочікувані Webpage в наборі аналізованих (не мають стосунок до цілей) вказують на проблему контенту та структури Website (множини відповідних Webpage).

Показник відмов при дослідженні набору популярних є основним. Якщо користувачі відвідують Webpage c_k через відповідну точку входження і зразу покидають Website, то це характеристика низької залученості замовників Website е-бізнесу при розв'язку конкретної NLP-задачі. Якщо у Webpage входження c_k високе значення відмов, то контент Webpage c_k не відповідає очікуванням та зацікавленості замовників/користувачів/відвідувачів. Тоді аналізують джерела переходів як з інших джерел, так і в Website між Webpage по відношенню до Webpage c_k. Аналіз та дослідження статистики низьких значень цих переходів та їх закономірностей спонукає виконати відповідні конкретні дії, зокрема: вдосконалення рекламної політики та підтримка Webpage/Website у відповідних соцмережах серед типової цільової аудиторії, впровадження/підтримка відповідних of-line/on-line маркетингових заходів, активація рекламних та інших кампанії з оплачуваними результатами ІІП, підтримка оптимізації ІІП (SEO).

Через детальний аналіз ключових слів входження визначають основні цілі користувачів відповідно до змісту очікувань та сподівань від результатів ІІП при відвідуванні Webpage/Website КЛС. Демонстрація переходів користувача між

Webpage по Website КЛС для досягнення кінцевої мети спонукає оцінити проблемні частини структури Website як складні/незрозумілі/некоректні кроки здійснення замовлення. Часто користувачі/замовники застосовують ІІП по Website я внутрішню техніку, замінюючи меню/навігацію/катало по Website. Для Website з великою множиною Webpage ІІП є найкращим рішенням для користувачів для швидкого та оперативного знаходження шуканого текстового контенту. Для такого ІІП зазвичай застосовують таку ж структуру/ техніку, як і для ІІПС як Google. Аналіз успішності/результативності/оперативності ІІП по Website полягає в розрахунку множини показників, зокрема[888-905, 937]:

$$K_{iip} =< P_{wuv}, R_{ecc}, S_{wcv}, P_{wip}, P_{wcv}, N_{wvt}, R_{wcv}, R_{wec}, N_{wtr}, N_{wcv}, I_{ssp} >, \quad (2.29)$$

– значення корисності відвідування P_{wuv} Website/Webpage КЛС [888, 937]:

$$P_{wuv} = \frac{R_{wcv}+R_{wec}}{N_{wvt}}, \quad (2.30)$$

де N_{wvt} – число відвідувань; R_{wec} – корисність е-бізнесу; R_{wcv} – корисність мети.

– рейтинг конверсії в е-бізнесі R_{ecc} для КЛС відповідної NLP-задачі[937]:

$$R_{ecc} = \frac{N_{wtr}}{N_{wvt}} \cdot 100\%, \quad (2.31)$$

де N_{wvt} – число відвідувань; N_{wtr} – число транзакцій.

– значення середньої корисності S_{wcv} [888-905, 937]:

$$S_{wcv} = \frac{R_{wcv}+R_{wec}}{N_{wcv}+N_{wtr}}, \quad (2.32)$$

де N_{wtr} – число транзакцій; N_{wcv} – число конверсії; R_{wec} – корисність від е-бізнесу, R_{wcv} – корисність мети.

– значення прибутку е-бізнесу P_{wip} для КЛС відповідної NLP-задачі [937]:

$$P_{wip} = R_{wcv} + R_{wec}, \quad (2.33)$$

де R_{wec} – корисність від е-бізнесу; R_{wcv} – корисність мети відвідування.

– значення досягнутої конверсії P_{cv} відвідувань Website/Webpage КЛС:

$$P_{wcv} = \frac{N_{wcv}}{N_{wvt}} \cdot 100\%, \quad (2.34)$$

де N_{wvt} – число відвідувань; N_{wcv} – число конверсії [888-905, 937].

Застосовуючи для досягнення мети ІІП по Website користувач/замовник в декілька разів корисніший за інших. Звідси створення/впровадження сервісу ІІП

по Website ефективно/якісно/результативно впливає на показники відвідування Website для залучення нових відвідувачів та зростання обсягів постійної цільової аудиторії. Для цього застосовують розрахунок впливу на прибуток ІІП I_{ssp}:

$$I_{ssp} = (R_{ssv} - R_{snv}) \cdot N_{ssv}, \quad (2.35)$$

де N_{ssv} – число відвідувань з ІІП по Website; R_{snv} – корисність відвідування без ІІП по Website; R_{ssv} – корисність відвідування з ІІП по Website [888-905, 937].

Показник I_{ssp} регулює стратегії/плани щодо подальших інвестицій в розвиток сервісу ІІП по Website та КЛС в цілому для розв'язку конкретної NP-задачі та має бути понад 80% місячного доходу для Website КЛС.

Процес оптимізації заходів з маркетингу ІІПС (SEM) [888-905, 937]:

1. Дослідження ключових слів (для оплачуваних/неоплачуваних ІІП).
    a. Користувачі відвідали згідно до природних результатів ІІП.
    b. Користувачі застосовують внутрішній ІІП по Website.
2. Оптимізація ІІПП/Webpage-входження (SEO) (для всіх результатів ІІП).
3. Оптимізація рекламної кампанії (оплачувані результати ІІП).
4. Оптимізація оголошень AdWords (оплачувані результати ІІП), тобто:
    a. Позицій за відвідуванням Webpage згідно середньої тривалості перебування на Webpage/Website за певний час.
    b. Позицій за відсотком нових відвідувань (показник досягнутих переходів для мети 1 [для цілей 2-4], показник відмов, показник досягнутої конверсії [середня корисність, корисність відвідування, транзакції, прибутку, рейтинг конверсії в е-бізнесі]).
    c. Позицій за часом доби/сезону/місяця/тижня в AdWords.
    d. Позицій за корисністю відвідування Webpage/Website.
5. Оптимізація версій оголошень AdWords (оплачувані результати ІІП).
6. Наявність коректно-реалізованих модулів лінгвістичного опрацювання контенту українською мовою для ефективного/якісного аналізу тексту при розв'язку конкретної NLP-задачі відповідною КЛС з підтримкою TCLC.

Тематика множини ключових слів є одним із основних показником ІІП для ідентифікації конкретного контенту Webpage. Наявність на Webpage цих слів або їх частини при ІІП не є достатнім для додавання цієї Webpage до результатів пошуку за конкретним запитом користувача. Коректно-визначена тематика ключових слів для ІІП значно покращує якість/ефективність відвідування користувачами КЛС як результат ІІП. Зазвичай теми містять 5-10 стійких словосполучень/виразів/фраз на Webpage, ключові слова в яких перетинаються. Чим більше таких виразів, тим складніше визначити тематику, що значно зменшує рейтинг/ефективність/якість Webpage при ІІП. Краще Webpage розбити на декілька згідно ідентифікованих тематичних підмножин ключових слів.

Для множин ключових слів, що збільшують значення конверсії, оптимізують інвестиції, збільшуючи CPC в AdWords. Значення повернення на інвестиції (P_{ROI}) має бути позитивним ($N_{Inc} > N_{Exp}$) [888-905, 937], тобто:

$$P_{\text{ROI}} = \frac{N_{Inc} - N_{Exp}}{N_{Exp}} \cdot 100\% > 0, \tag{2.36}$$

де N_{Exp}– витрати; N_{Inc} – прибуток. Тоді P_{ROI} для валового прибутку [888, 937]:

$$P_{\text{ROI}\,vp} = \frac{(N_{Inc} \cdot A_{Inc})/100 - N_{Exp}}{N_{Exp}} \cdot 100\%, \tag{2.37}$$

де A_{Inc} – розмір прибутку. Тоді знаходять на скільки >q% коштів без ризику отримати P_{ROI}<0 можна затратити на конкретне ключове слово в AdWords.

Для розрахунку обсягів коштів на залучення користувачів застосовують:

$$C_{amax} = \frac{\frac{N_{Inc} \cdot A_{Inc}}{100}}{\frac{P_{\text{ROI}\,vp}}{100} + 1}. \tag{2.38}$$

Для розрахунку обсягів коштів на CPC для даного ключового слова на основі коефіцієнтів конверсії для кожного ключового слова застосовують:

$$C_{cmax} = C_{amax} \cdot \frac{R_{ecc}}{100}. \tag{2.39}$$

Тоді не треба переплачувати за ключові слова AdWords. Основні вимоги [888]:

1. Завжди враховувати зацікавлення користувачів Website для КЛС.

2. Для рекламних/маркетингових кампаній застосовувати для користувачів спеціальні Webpage входження згідно неоплачених/оплачених результатів ІІП.

3. Webpage входження як результат ІПП є завжди поряд із закликом до дії.

4. Тематичні ключові слова розмістити в HTML-тегах <title>.

5. Контент Webpage формувати навколо конкретної тематики із 5-10 подібних ключових слів для коректності та результативності ІПП.

6. Не зловживати/спамити ключовими словами для ІІПС.

7. Тематичні ключові слова розмістити змістовно в HTML-тегах <a>.

8. Насичений ключовими словами контент розмістити вверху Webpage.

9. Контролювати через файл robots.txt список Webpage індексування ІІПС.

10. Не розміщувати актуальний текст в рисунках/анімаціях тощо.

SEV-алгоритм для Website та визначення його ефективності/якості:

1. Формулювання та ідентифікація корисності відповідно до цілей.
2. Активація звітів е-бізнесу для КЛС згідно конкретної NLP-задачі:
    a. Визначити необмежене число цілей (≈4 цілі на кожний профіль).
    b. Ідентифікувати оптимальний обсяг відвідувань/часу кінцевого користувача/замовника для успішної конверсії.
    c. Проаналізувати обсяги вкладу кожної цілі в загальний прибуток.
    d. Поєднати цілі за категоріями/напрямами/видами.
    e. Сформувати окремі множини транзакцій як відповідних цілям.
3. Підтримка of-line актуальних маркетингових кампаній/замовників:
    a. На основі ІІП – орієнтація на послугу/ціну/зручність тощо.
    b. Кодовані URL – відома популярна NLP-послуга.
    c. Престижні URL – розміщати все на центральному домені.
4. Підтримка опрацювання службового контенту Website як складників е-бізнесу(завантаження/збереження фото, файлів pdf/txt/xls, тощо)

Впровадження КЛС спричиняє *збільшенню продуктивності роботи* NLP-фахівців, обсягів потенційної/постійної аудиторії користувачів системи, якості та ефективності інтелектуального аналізу текстових потоків контенту [586-587]. Паралельно відбувається скорочення обсягів часових/фінансових/ресурсних затрат на реалізацію КЛС та оперативне/своєчасне отримання доступу до унікального актуального релевантного текстового контенту згідно факторів:

1. Збільшення продуктивності роботи спричинене застосуванням автоматизації інтеграції/управління/супроводу контенту на основі інтелектуального аналізу текстових потоків та результатів роботи додаткових спеціальних ресурсів як Google Analytics та NLP-фахівців, зокрема, аналітиків, програмістів, лінгвістів, адміністраторів, модераторів та зворотного зв'язку від цільової постійної аудиторії.
2. Аналіз статистики/динаміки збільшення продуктивності роботи спричиняє формування множини факторів впливу на зростання якості та ефективність інтеграції/управління/супроводу контенту і зменшення часу/ресурсів/фінансів на реалізацію КЛС та оперативного отримання контенту цільовою аудиторією як результат успішності конверсії.
3. Зростання якості інтелектуального аналізу текстових потоків контенту спричинене ефективністю аналізу статистики/динаміки та основних показників функціонування КЛС за визначений період часу як кількість унікальних відвідувачів, кількість переглядів Webpage за відвідування, джерело трафіка та число переходів, нові відвідування, число перегляду Website/Webpage, динаміка контенту, досягнута мета ІІП, показник відмовлень, середній час перебування на Website/Webpage, число відвідувань, ступінь конверсії, тор ключових слів ІІП тощо.
4. Скорочення часових/фінансових/ресурсних затрат на реалізацію КЛС та оперативне отримання доступу до унікального релевантного актуального текстового контенту прямо пропорційне зростанню якості та ефективності прийняття рішень відповідними NLP-фахівцями для інтелектуального аналізу тексту при розв'язку конкретної NLP-задачі:
    a) адміністраторами для своєчасного оперативного адміністрування Website та КЛС та формування запитів контролю за транзакціями;
    b) модераторами на генерування актуальних правил інтеграції, розпізнавання, аналізу, опрацювання та синтезу контенту, зокрема, управління, супроводу, форматування, фільтрування, кластеризації, класифікації, кешування контенту тощо;

c) модераторами на формування списку адрес та правил для інтеграції актуальних оперативних даних з достовірних джерел;
d) авторами для генерування унікального релевантного актуального текстового контенту згідно рейтингового списку актуальних запитів від цільової аудиторії відповідно до актуальної тематики;
e) аналітиками для аналізу статистики/динаміки функціонування КЛС, генерування правил ідентифікації сюжетів, персоналізації роботи з постійною аудиторією, ранжування контенту.

Організаційний ефект спричинений множиною таких факторів [586-587]:

1. Зменшенням числа NLP-фахівців (аналітиків 1-3, адміністраторів 1-2, програмістів 1-2, лінгвістів 1-2, авторів 1-10, модераторів 1-3, експертів ПО 1-2, наприклад психологів), задіяних на етапах розроблення та впровадження КЛС для розв'язку конкретної NLP-задачі;
2. Зміна/фіксація організаційної структури проекту (функціональний розподіл між NLP-фахівцями проекту, тобто лінгвіст не виконує роботу аналітика, а експерт – модератора тощо, але можливе об'єднання функцій в деяких простих NLP-задачах або взаємозаміна);
3. Зменшення числа функцій NLP-фахівців КЛС-проекту (часткова автоматизація на основі інтелектуального аналізу текстових потоків);
4. Підтримка регламенту інтелектуального аналізу текстових потоків контенту для реалізації функцій прийняття рішень на основі модулів інтеграції/управління/супроводу контенту (інтеграція інформації для користувачів/авторів, фіксація/аналіз результатів/статистики/динаміки запитів/дій цільової аудиторії та інших статистичних даних для модераторів/аналітиків/адміністраторів/лінгвістів/експертів).

Технологічний ефект спричинений скороченням/вивільнення ресурсів як NLP-фахівців, якісним/ефективним застосуванням модулів інтелектуального аналізу текстових потоків контенту в КЛС, відносно фіксованим розподілом функцій між NLP-фахівцями проекту, а також реалізацією нових ІТ як

інтеграції/управління/супроводу контенту та організацією/аналізом зворотного зв’язку з постійною/потенційною цільовою аудиторією [586-587].

Соціальний ефект сприяє на основі регулювання змісту/тематики Website зростанню обсягів цільової аудиторії, числа унікальних/постійних користувачів Website, доступності до релевантного контенту/Webpage/Website, охоплення широкого кола соціальної аудиторії тощо. Супровід тематично актуального та подібного текстового контенту, інтеграція оперативного унікального тексту та відповідне управління ним через Website регулює межі обсягу соціальної цільової постійної аудиторії КЛС та сприяє прогнозувати/регулювати ці зміни.

Рекламний ефект на основі застосування шаблонів для Website, Webpage/контенту та інтеграції/генерації/створення унікального актуального контенту сприяє зростанню числа відвідування користувачів з ІІПС та є своєрідною саморекламою Website КЛС, множини послуг/контенту Webpage. Застосування результатів роботи Google Analytics/AdWords суттєво полегшують аналіз показників е-бізнесу, реклами та функціонування Website/КЛС [586-587].

Психологічний ефект сприяє організації/реалізації підтримки дружнього користувацького інтерактивного інтерфейсу для кожного NLP-фахівця [586-587], користувача та замовника Website КЛС на основі динамічного зворотного зв’язку. Це суттєво полегшує виконання обов’язків для лінгвістів, аналітиків, адміністраторів, модераторів, авторів, а також збір/аналіз психологічних показників постійних користувачів/замовників/відвідувачів КЛС на основі персоналізації роботи з ними через дружній інтерактивний інтерфейс Website.

Ергономічний ефект сприяє зростанню впливу результатів функціонування КЛС та модулів інтелектуального аналізу текстових потоків контенту через супровід/управляння/інтеграцію текстового контенту на основі розрахунків/аналізу числа джерел трафіка у %, абсолютно унікальних відвідувачів, нових відвідувань (%),переглядів Webpage/Webpage за всі/одне відвідування, досягнутої конверсії ІІП, відмов (%), відвідувань, а також динаміки опрацювання контенту (%), середнього часу відвідування на Website (хв:с) тощо.

2.3.4. Вхідний потік контенту комп’ютерної лінгвістичної системи

Класифікований список вхідного потоку контенту з множиною відповідних властивостей/ознак/параметрів сприяє розмежовувати учасників проекту через їх типізацію та обмеження прав доступу в залежності від контенту: постійні користувачі, потенційні відвідувачі, лінгвісти, аналітики статистики, адміністратори Website, модератори контенту/правил, автори унікального контенту, інформаційний ресурс як джерело контенту тощо [586-587]. Типізована структура шаблону вхідного потоку контенту з множиною відповідних властивостей/ознак/параметрів сприяє визначити основні функціональні вимоги до Website/КЛС та її типової структури та чітко окреслити нефункціональні можливості, класифікувати джерела, розрахувати частоти інтеграції та відповідні обмеження/умови інтеграції з типового джерела. Вхідними потоками контенту до КЛС є типові компоненти:

$$X =< X_a, X_s, X_q\ , X_f, X_s, X_w, X_b, X_d, X_k, X_v, X_u, X_r, X_t, X_o >, \quad (2.40)$$

– X_a – URL-адреси Website джерел для БД фільтрів КЛС;

– X_s – контент як результат інтеграції з різних за наперед визначеним списком URL-адрес джерел X_a без наперед визначеної структури у HTML/XML-форматі згідно релевантних тематичних запитів;

– X_q – тематичні запити відвідувачів/користувачів Website КЛС у вигляді множини ключових слів або стійких словосполучень;

– X_f – фактичні дані постійних користувачів/профілів та множина правил дозволених дій в межах відповідного типу користувача КЛС;

– X_s – статистичні дані дій/подій/явищ суб’єктів/об’єктів КЛС розв’язку відповідної NLP-задачі та правила збору/збереження/аналізу статистики в певних проміжках часу функціонування КЛС;

– X_w – статистичні дані функціонування Website КЛС, зібрані із заданою періодичністю з Google Analytics у вигляді XML-таблиць;

– X_b – вміст баз даних контенту/правила/фільтрів/анотацій тощо КЛС;

– X_d – різного виду лінгвістичні словники в залежності від призначення КЛС для розв'язку конкретної NLP-задачі;

– X_k – множина персоналізованих/анонімних відгуків/коментарів відвідувачів/користувачів до відповідного контенту Website КЛС;

– X_v – кортеж результатів персоналізованих/анонімних голосувань постійних/потенційних відвідувачів/користувачів щодо контенту КЛС;

– X_u –статистичні персоналізовані індивідуальні дії користувачів КЛС;

– X_r – множина зовнішньої/внутрішньої реклами тематичного контенту;

– X_t – тематичні стікери розважального/інформаційного контенту (курси валют, анонси, дайджести, погода, анекдоти, гороскоп тощо);

– X_o – кортеж опцій налаштування та зміни конфігурацій КЛС/Website.

2.3.5. Вихідний потік контенту комп'ютерної лінгвістичної системи

Наповнення кортежу вихідного опрацьованого тексту згідно призначення КЛС для розв'язку конкретної NLP-задачі напряму залежить від змісту вхідного класифікованого потоку контенту з наперед визначеною множиною відповідних властивостей/ознак/параметрів в залежності від взаємодії Website відповідних типів учасників проекту (постійні користувачі, потенційні відвідувачі, лінгвісти, аналітики статистики, адміністратори Website, модератори контенту/правил, автори унікального контенту, інформаційний ресурс як джерело контенту тощо):

$$Y =< Y_c, Y_q, Y_a, Y_v, Y_s, Y_p, Y_t, Y_r, Y_o, Y_k >, \quad (2.41)$$

– Y_c – текстовий контент як інформаційний продукт або результат надання відповідної інформаційної послуги розв'язку конкретної NLP-задачі на Website;

– Y_q – множина змістовно згенерованих/кешованих Webpage як результат тематичних запитів/ІІП користувачів/відвідувачів Website КЛС;

– Y_a – анотації/дайджести/реферати на текстовий тематичний контент;

– Y_v – кортеж статистики взаємодії користувачів/відвідувачів з Website;

– Y_s – кортеж змісту профілів постійних користувачів КЛС згідно персоналізованої статистики Y_v для відповідного генерування індивідуального портрету користувача/аудиторії в певні проміжки часу;

– Y_p – кортеж змістовного рекомендованого контенту Webpage Website, персоналізованого під конкретного постійного користувача згідно профіля/дій/взаємозв'язку із КЛС в певні проміжки часу;

– Y_t – множина тем/рубрик контенту з можливістю поновлення згідно результатів останніх ІІП/запитів від постійних користувачів Website;

– Y_o – схема взаємозв'язків текстового тематичного контенту за відповідною класифікацією (актуального, релевантного, авторського, застарілого, популярного, подібного, останньо-переглянутого, часто-переглянутого, послідовно за певним найчастіше переглянутого, довше переглянутого, найчастіше переглянутого з пошукових систем або внутрішнього ІІП, переглянутого типовою групою користувачів тощо);

– Y_r – множина результату рейтингування контенту за наперед визначеною шкалою у межах відповідної класифікації ранжування;

– Y_k – множина маркованого оцінювання/рейтингування коментарів користувачів як ступінь дозволу опублікування на Website/Webpage при необхідності з позначкою заборони для конкретного дописувача писати подальші коментарі та ранжування за ступенем довіри всіх дописувачів.

Список вихідного потоку контенту, його основні ознаки та відповідна класифікація, ІТ генерування/підтримка/аналіз сприяє визначити чіткі загальні функціональні вимоги реалізації КЛС для розв'язку будь-якої NLP-задачі.

2.4. Функціональні вимоги до проекту типової КЛС

2.4.1. Вимоги до програмних модулів типової КЛС

Функціональні/нефункціональні вимоги до типової КЛС основним складником для проектування та розроблення ПЗ для розв'язку конкретної NLP-задачі. Функціональні вимоги (англ. Functional Requirements) формують напрям розроблення та реалізації типової КЛС, але в більшості випадків їх неможливо розрахувати та виміряти (вимірюють як множина входів в КЛС та множини виходів які перевіряються) [586-587]. Нефункціональні вимоги (англ. Non-

Functional Requirements) дозволяють виміряти якість розроблення та ефективність впровадження КЛС на основі зворотного зв’язку від постійної аудиторії та темпів зростання обсягів постійних користувачів та конверсії їх дій. Функціональні вимоги до типової КЛС є множиною описових інструкцій щодо внутрішнього функціонування ІС та зміни динаміки її поведінки в залежності від станів системи через визначення набору специфічних функцій/модулів для розв’язку конкретної NLP-задачі, зокрема, опрацювання/модифікація контенту, маніпулювання/оперування даними, інтегрування/калькулювання даних тощо. Основними типовими вимогами до КЛС є відповідність стандартам, точність/правильність вихідних даних по відношенню до вхідних, безпечність ПЗ та сумісність з різними модулями/ПЗ/ІС. Загальні типові вимоги до КЛС:

- підтримка динамічного керування транзакціями КЛС/Website;
- підтримка швидкого впровадження WebOLTP-застосувань для КЛС;
- оперативний ефективний взаємозв’язок браузера і back-end БД;
- продуктивність/масштабність та якість/ефективність функціонування при великих обсягах транзакцій, сесій, користувачів/відвідувачів та одночасного доступу баз/сховищ контенту/правил тощо.

Для підтримки керування основними типовими транзакціями при функціонуванні КЛС/Website використовують наступні вбудовані ПЗ [586-587]:

- виклики розподілених елементів для своєчасної оперативної якісної підтримки взаємозв’язку в багаторівневій структурі КЛС/Website;
- сервіси ефективного оперативного запуску/керування сервлетами;
- Web-сервіси якісного керування транзакціями КЛС/Website/Webpage;
- інструменти швидкого оперативного якісного розроблення/модифікації та підтримки ПЗ для проміжного компонентного/модульного рівня ІС.

КЛС має підтримувати мінімум 6 інтерфейсів взаємодії з конкретним типом учасника проекту в залежності від прав та функціональних можливостей:

- з обмеженим доступом для постійних/потенційних відвідувачів Website (Рис. 2.11) з можливістю оперативно знайти необхідну інформацію;
- з обмеженим персоналізованим доступом для користувачів (Рис. 2.12);

- з доступом без обмежень для адміністратора КЛС/Website (Рис. 2.13) з можливістю корегувати структуру Website/КЛС, відповідних шаблонів Webpage/контенту, прав доступу учасників, правил розсилки контенту;

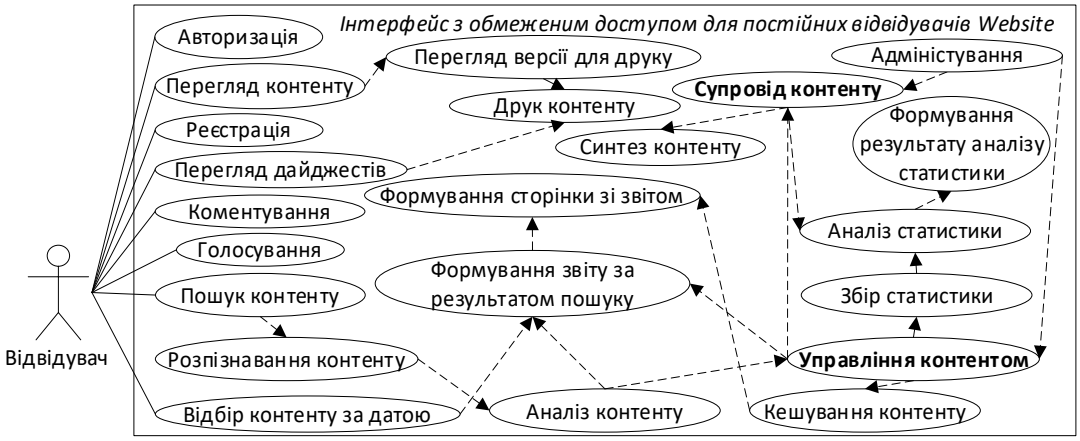


Рис. 2.11. Діаграма use case для обмеженого доступу відвідувачів КЛС

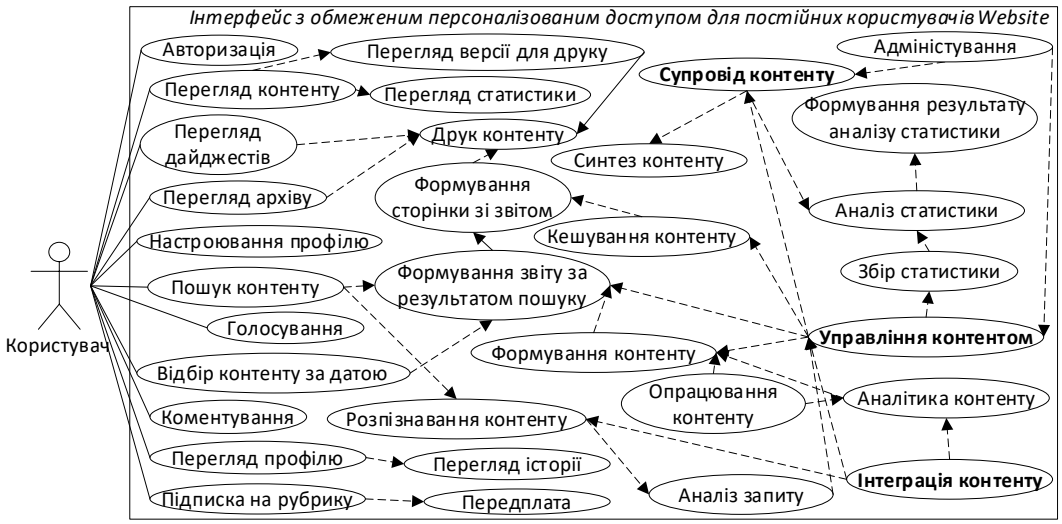


Рис. 2.12. Діаграма use case для обмеженого доступу користувачів Website

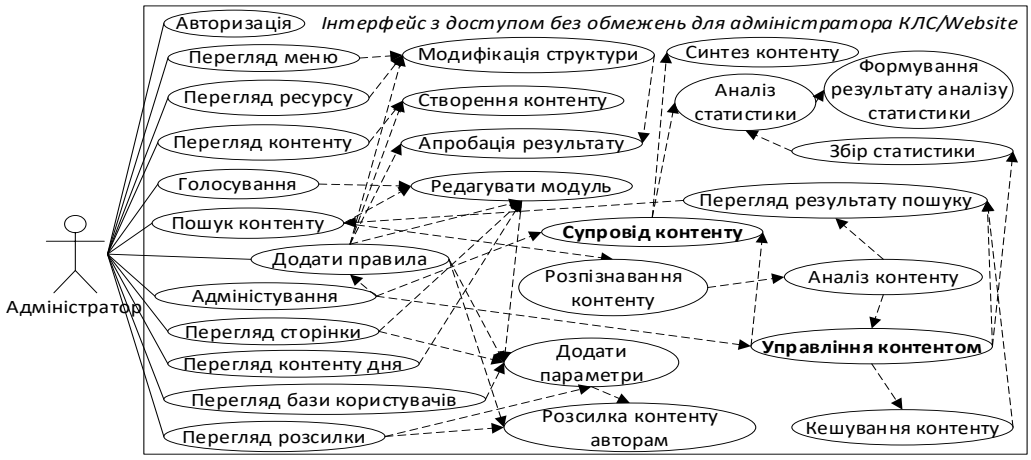


Рис. 2.13. Діаграма use case для вільного доступу адміністратора Website/КЛС

- з вільним доступом до певних модулів КЛС для модератора (Рис. 2.14) з можливістю корегувати параметри/правила/конфігурацію ІПП, фільтрування, аналізу, моніторингу, категоризації тощо контенту;
- з вільним доступом до певних модулів КЛС для аналітика (Рис. 2.15);
- з частковим обмеженим доступом для автора/лінгвіста контенту Website КЛС (Рис. 2.16) [586-587].

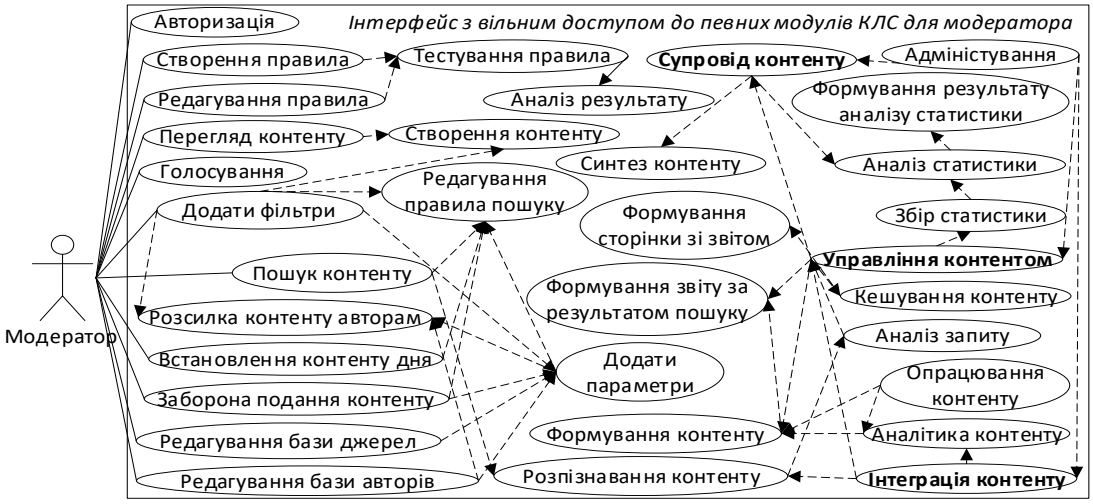


Рис. 2.14. Діаграма use case для доступу модераторів Website КЛС

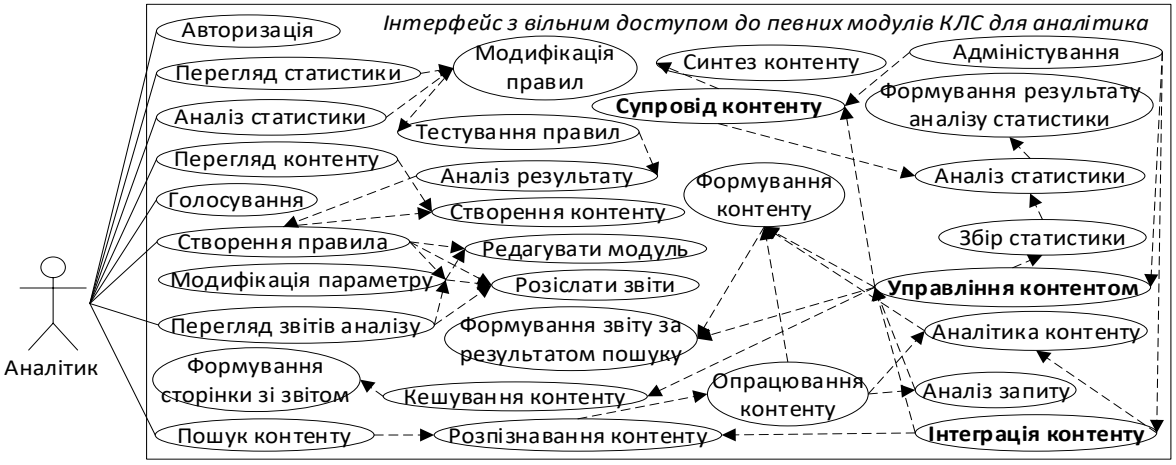


Рис. 2.15. Діаграма use case для аналітиків Website/КЛС

Визначення функціональних вимог для модулів супроводу, управління та інтеграції текстових даних КЛС спонукає розробленню загальної структури відповідних ІС [586-587]. Коректно розроблений Website сприяє полегшенню взаємодії учасників проекту з КЛС та відповідно підтримує можливість зростання функціональності відповідних КЛС розв’язку конкретної NLP-задачі.

Рис. 2.16. Діаграма use case для лінгвістів/авторів контенту Website КЛС

Модуль супроводу контенту формує на основі статистичних даних взаємодії постійних користувачів/відвідувачів Website множину актуальних нерелевантних запитів для подальшого генерування рейтингового за популярністю списку тематичних сюжетів/запитів потенційно релевантного контенту. Цей список застосовують як вхідні дані для модуля інтеграції даних з різних достовірних джерел (інформаційних ресурсів) та для постійних авторів унікального контенту. Автор має можливість ознайомитися з таким списком для створення актуального релевантного для постійної аудиторії/ІС/модулів текстового контенту в КЛС (наприклад, в ЗМІ-системах інформації, рекомендаційних системах, системах аналізу психологічного стану особи, інтерфейсах безмовного доступу тощо) на основі підібраного та інтегрованого текстового контенту з різних достовірних інформаційних джерел як підґрунтя дослідження при генеруванні контенту.

Лінгвіст окрім створення унікального контенту може поновлювати або розробляти нові лінгвістичні е-словники (не лише слів, ключових слів та стійких словосполучень, але морфем, флексій, виключень, основ тощо), але тематичні та інші спеціальні, а також підбирати корпуси текстів для навчання КЛС.

Модератор розробляє різні правила опрацювання текстового контенту на основі досліджень лінгвіста, потреб автора, статистичних даних аналітика щодо

популярності результатів ІІП тематичного контенту (особливо при його малих обсягах або відсутності за частотою відмов від переходів з пошукових систем). Також модератор реалізує правила фільтрування контенту при інтеграції з різних джерел, внутрішньому ІІП контенту за запитами користувачів, анотуванні та реферуванні контенту, ідентифікації дублів в БД/СД, кешуванні інформаційних блоків як етап управління контенту та аналізі персоналізованих профілів/історій дій користувачів та визначенні тематичних сюжетів як етап супроводу контенту. При необхідності у співпраці з лінгвістом модератор формує правила синтезу та розпізнавання мовлення, текстової аналітики та генерації текстів, а також опрацювання/формування відповідного текстового масиву даних [586-587].

Аналітик розробляє різні правила збору/збереження/аналізу статистичних даних функціонування КЛС та дій/подій діяльності постійної цільової аудиторії у визначенні часові проміжки певної періодичності. Також аналітик генерує правила статистичного аналізу динаміки/частоти реалізації етапів TCLC КЛС для подальшого ідентифікації тематичної/змістовної зацікавленості постійної (за діями користувачів Website) або потенційної (за діями унікальних відвідувачів) цільової аудиторії. Своєчасна оперативна реакція на зміни зацікавленості цільової аудиторії сприяє модифікувати напрями інтеграції контенту для підтримки зростання числа прямих/ІІПС/ресурсів відвідувань з досягнутою конверсією, повторних/унікальних/регіональних/тематичних відвідувань КЛС, що в свою чергу приводить до збільшення обсягів цільової аудиторії Website. Також модифікуються правила збирання/збереження/аналізу статистичних даних рейтингів/рубрик контенту/авторів, функціонування Website, періодики активності користувачів/відвідувачів Website відповідно до об’єктів КЛС.

2.4.2. Основні додаткові вимоги мережних, програмних та технічних інструментів програмної реалізації типової КЛС

Формування функціональних вимог для модуля інтелектуального аналізу текстових потоків контенту в КЛС відповідно конкретизує додаткові вимоги мережного, програмного та технічного середовища реалізації типової КЛС,

зокрема, для супроводу/управління/інтеграції контенту Website/КЛС/Webpage (Таблиця 2.4) . Модуль супроводу контенту є допоміжним інструментом для адміністраторів та аналітиків Website/КЛС. Модуль управління контентом – для користувачів, відвідувачів, адміністраторів та модераторів Website/КЛС. Модуль інтеграції контенту – для авторів, лінгвістів та модераторів Website/КЛС .

Таблиця 2.4

Інструменти інтелектуального аналізу текстових потоків контенту [586-587]

Інструменти	Опис
HTTPS, FTP, HTTP, RMI-IIOP, GIOP, IIOP	Протоколи зв'язку між Webserver та користувачем.
SOAP, REST/ Atom	Протокол/правила доступа/взаємодії об'єктів
SSL, TLS	Сертифікати безпечної звязку домену/адресата
CGI, Python, R, PHP, Apache, API	Інтеграція Webserver із джерелами контенту.
HTML,CSS,WML,HDML,XML,XHTML,JavaScript	Підтримка гіпертекстових посилань.
GifCam,Flash,JavaScript,CSS,audio/video format, VRML	Підтримка мультимедійних ефектів.
IMAP, SMTP, POP3, UDP, LMTP, XML-RPC, CMIP	Підтримка інтерактивної взаємодії/зв'язку.
Python, PHP, R, JavaScript	Реалізація процесів NLP-задачі.
Joomla, WordPress, Drupal, LiteDiary, SiMan CMS,	Системи керування контентом.
Django, Tornado, Pyramid, Flask, TurboGears	Webframework на Python
Zend, FuelPHP, CakePHP, Phalcon, Yii, CodeIgniter, Symfony, Laravel	Webframework на PHP
ECM, CMIS, WSDL	WebService керування контентом
EDGE, UMTS, GPRS, WAP, VPN	Підтримка мобільного доступу/обчислень.
CORBA, UML, DCOM, COM, ORB, SWIG	Створення розподілених об'єктів.
СУБД MySQL, filesystem, ОС, Oracle	Збереження та опрацювання даних.

Вибір NLP-фахівців між CMS та Webframework для розроблення проекту КЛС залежить від результатів аналізу їх переваг/недоліків. Головна перевага Webframework у наявності в них великого спектра інструментів для повноцінного розроблення/підтримки будь-якого Web-застосунку. Не треба шукати/створювати окремі бібліотеки для кожної окремої задачі та вирішувати питання сумісності. Webframework – це як лего-конструктор (Таблиця 2.5).

Модель Text Mining контенту безпосередньо пов'язана з процесом ML – пошуком моделі з колекцією функцій, алгоритму і гіперпараметрів, які знаходять кращі результати на навчальних даних для оцінювання наперед невідомих даних. Процес складається зі створення навчального набору (корпусу), аналізу методів вилучення функцій та попереднього опрацювання – перетворення тексту у числові дані для подальшого розуміння процесами ML на основі класифікації і кластеризації тексту. Оскільки для Text Mining контенту застосовують ML,

необхідна мова програмування з великою множиною вбудованих/додаткових наукових та обчислювальних бібліотек як Python (Таблиця 2.6) [505-511].

Таблиця 2.5

Порівняння CMS та Webframework [505-511, 586-587]

Характеристика	CMS	Webframework
Простота супроводу проекту КЛС.	+/-	+
Наявність закладених в ПЗ множини бізнес-процесів	+/-	+
Можлива та відносно проста реалізація бізнес-процесів, не закладених в ПЗ.	+/-	+
КЛС-проекти легко масштабуються та модернізуються	+	+
Рішення працюють значно швидко.	-	+
Рішення витримують велике навантаження	-	+
Підтримка високого рівня безпеки.	+	+
Терміни розробки типового функціоналу короткі.	+	-
Наявність більше за базових компонентів бізнес-логіки рівня програми	+/-	-
Необхідність реалізувати багато функцій індивідуально для конкретної КЛС.	+/-	+
Для розробки не потрібне розуміння бізнес-процесів, які необхідно реалізувати.	+	-
Вбудована підтримка багатьох бізнес-процесів, наприклад, опрацювання замовлень	+	-
Для адміністрування/модернізації КЛС не потрібні спеціалізовані звання/навички	+	+/-

Таблиця 2.6

Інструменти Python для реалізації Text Mining контенту [505-511, 586-587]

Бібліотека	Характеристика	Особливості
Scikit-Learn	Розширення бібліотеки SciPy (Scientific Python) для підтримки інтерфейсу програми (API) при узагальненому ML.	Заснована на Cython з підтримкою високопродуктивних бібліотек C (Boost, LibSVM, LAPACK тощо), розширення ScikitLearn, поєднує високу продуктивність з простотою застосування методів аналізу малих/середніх множин даних. Відкритий вихідний код, комерційно доступне розширення забезпечує єдиний інтерфейс для багатьох моделей класифікації, регресії, кластеризації, розміризації та перехресної перевірки/налаштування гіперпараметрів.
Yellowbrick	Множина візуальних діагностичних засобів для аналізу/інтерпретації результатів ML, додаток Scikit-Learn API.	Надає прості та інтуїтивно зрозумілі візуальні інструменти вибору функцій, моделювання та налаштування гіперпараметрів, управління процесом вибору моделей найбільш ефективного опису текстових даних.
NetworkX	Комплексний пакет аналізу графів для допомоги створення, упорядковування, аналізу та маніпулювання складними мережевими структурами.	Не бібліотека ML або Text Mining контенту, але використання структур даних графа дозволяє кодувати складні зв'язки, які алгоритми графів здатні аналізувати і знаходити семантичні особливості, а тому є важливим інструментом для аналізу тексту.
spaCy	Інструмент реалізації якісного NLP-процесу на основі сучасних складних алгоритмів через простий і зручний API.	Підтримка попереднього опрацювання тексту в межах підготовки до глибокого навчання. Застосовують для створення ІС вилучення інформації або аналізу природної мови на великих обсягах тексту.
Gensim	Надійний, ефективний і простий інструмент семантичного моделювання тексту без викладача.	Створена для пошуку подібності у текстах, підтримує тематичне моделювання для методів приховано-семантичного аналізу та має інші бібліотеки ML (наприклад, word2vec) [200].
NLTK	Пакет NLP-інструментів (Natural Language Tool-Kit).	Містить корпус, лексичні ресурси, граматики, NLP-алгоритми та попередньо навчені моделі для реалізації швидкого опрацювання текстових даних з різних природних мов.
pandas	Аналіз даних.	Аналіз числових даних.
TextBlob	Розширення NLTK	Вилучення іменних словосполучень, PoS-маркування, токенізація, сентимент-аналіз, класифікація та CA

Опрацювання природної мови є перспективним AI-напрямом штучного інтелекту для розуміння та інтерпретації людської мови комп’ютерами. Застосування NLP-методів, ML та найкращих інструментів для інтерпретації текстових даних дозволяє КЛС своєчасно та оперативно провести аналіз та зробити відповідні висновки/прогноз або обрати оптимальне рішення у відповідь на відповідну множину вхідних даних. NLP-методи включають токенізацію, нормалізацію тексту та очищення даних. У стандартному форматі застосовують різні ML-методи для найкращої інтерпретації та розуміння даних. Наприклад, це включає застосування релевантних методів моделювання для класифікації е-листів типу спам/неспам або оцінювання настроїв твіту в Twitter. Також застосовують новітні, складніші методи для моделювання тематики, видобування ключових слів або генерації тексту на основі глибинного навчання.

Технологія розроблення КЛС – це підтримка повної/часткової автоматизації бізнес-процесів (в тому числі опрацювання природної мови) для розв’язку конкретної NLP-задачі. В КЛС на основі підтримки бізнес-процесів завдання, підпроцеси, інформація, повідомлення, документи, контент, тощо передаються для реалізації відповідних дій/подій від одного типу актора (учасника) до наступного згідно колекції закладених процедурних/асоціативних правил розширених NLP-моделей з більш багатими наборами функцій аналізу тексту. Контекст контенту подають/реалізують як NLP-функції та організовують їх візуальну інтерпретацію для аналітиків/модераторів для контролю процесу відбору моделі. Зазвичай аналізують складні зв'язки, витягнуті з тексту, на основі методів аналізу графів. КЛС інтерпретує, реалізує і управляє потоком робіт (бізнес-процесом) на основі ПЗ у вигляді модулів, які аналізують та впроваджують інтерпретацію процесу, взаємодіють з об’єктами/суб’єктами потоку робіт і звертаються до відповідних модулів/інструментів за необхідності.

КЛС автоматизує бізнес-процес розв’язку конкретної NLP-задачі та реалізує правила взаємодії об’єктів/суб’єктів процесу. Ці моменти взаємодії (діалогу) є основними аспектами втрат через невизначеність/неоднозначність інтерпретації вхідних даних (розуміння синтаксичного/семантичного аналізу

тексту та вибір/реалізація відповідного продукційного/асоціативного правила). Розв’язком такої проблеми може виступати масштабування аналізу тексту в багатопроцесорних КЛС за допомогою Spark та реалізація аналізу тексту через глибинне навчання. Результатом реалізації NLP-проекту може бути не лише самостійна КЛС для розв’язку конкретної NLP-задачі, але і програмний вбудований модуль в ІС типу Internet-видавництво, дистанційного навчання, Internet-видання, Internet-журнал, Internet-газета, Internet-магазин продажу контенту як електронні книги, аудіо відео, фото, ПЗ, тощо [505-511, 586-587].

Розроблення множини функціональних вимог для побудови типової КЛС сприяє створення для розробників та NLP-фахівців узагальненої ІТ реалізації відповідних ІС/модулів для суттєвого зменшення обсягу часу/ресурсів на проектування/побудови/впровадження/модернізації/вдосконалення відповідних програмних NLP-модулів. Вимоги до результатів/регламенту функціонування КЛС, шляхів подання/передачі/збереження/модифікації/інтерпретації/знищення текстових/службових даних залежать від реалізації підсистем інтелектуального аналізу текстових потоків контенту як супровід/управління/інтеграції контенту.

Вимоги до сумісності та шляхів обміну/взаємодії текстовими/службовими даними з іншими ІС/модулями/учасниками складаються з умов реалізації та підтримки опрацювання текстових масивів контенту у HTML/XML-форматі.

Супровід регламентних та організаційних вимог до учасників/модулів, їх кваліфікації та складу, регламенту/часу експлуатації ІС, повноважень та прав для взаємодії із ІС тощо надають можливість підтримувати на відповідному рівні функціонування КЛС, оперативно/якісно впровадити/реалізувати КЛС, та своєчасно повномасштабно аналізувати результати апробації діяльності ІС та основних підсистем інтелектуального аналізу текстових потоків контенту.

Ергономічні вимоги до КЛС полягають у комфортності інструментів управління ІС, раціональному компонуванні програмних/інтерфейсних модулів, зручності/оперативності обслуговування/супроводу/підтримки ІС, естетичному дизайні інтерактивного користувацького інтерфейсу. КЛС мають забезпечувати

відповідного рівня захист/безпеку персональних даних та інших компонентів ІС від несанкціонованого доступу, знищення, втрати, пошкодження інформації.

2.5. Основні результати та висновки розділу

Розроблена ІТ опрацювання україномовного текстового контенту на відміну від існуючих підтримує принцип модульності типової архітектури КЛС для розв'язку конкретної задачі ОПМ та аналізу множини параметрів та метрик ефективності функціонування системи відповідно до поведінки цільової аудиторії. Розроблено загальну структуру КЛС для опрацювання текстового контенту українською мовою та концептуальну схему/модель функціонування типової КЛС на основі моделювання взаємодії основних процесів і компонентів системи, що дало змогу вдосконалити ІТ інтелектуального аналізу текстового потоку на основі опрацювання інформаційних ресурсів. Проаналізовано особливості проектування та розроблення комп'ютерних лінгвістичних систем на основі визначення основних етапів як графемний, морфологічний, лексичний, синтаксичний семантичний аналіз/синтез україномовного тексту для розв'язку конкретної NLP-задачі. Зроблена та конкретизована постановка проблеми опрацювання україномовного тексту на основі визначення функціональних особливостей інтелектуального аналізу текстового потоку. Загальний аналіз проблеми аналізу україномовного тексту та визначення основних проблем опрацювання україномовного тексту дало можливість сформулювати основні етапи та вимоги до проекту типової КЛС розв'язку конкретної NLP-задачі. Ідентифікація основних характеристик КЛС та обґрунтування реалізації проекту типової КЛС дало можливість визначити очікувані ефекти від відповідної реалізації проекту. На основі аналізу вхідних/вихідних потоків контенту комп'ютерної лінгвістичної системи визначені та сформульовані функціональні вимоги до проекту типової КЛС, її програмних модулів, мережних, програмних та технічних інструментів програмної реалізації ІС. Основні результати розділу опубліковані у роботах [210-212, 219, 257, 313, 407, 586-587, 803-822, 849-861, 875-878, 883-887, 896-905, 912-917, 929-939, 944-953, 963].

РОЗДІЛ 3
МОДЕЛЮВАННЯ КОМП'ЮТЕРНОЇ ЛІНГВІСТИЧНОЇ СИСТЕМИ ОПРАЦЮВАННЯ УКРАЇНСЬКОЇ МОВИ

3.1. Схематичне моделювання структури КЛС

3.1.1. Концептуальна схема функціонування типової КЛС

КЛС на основі NLP-методів для аналізу текстових/аудіо даних є вже невід'ємною частиною людського повсякденного життя [1-34]. Від імені користувача деякі КЛС переглядають великий обсяг інформації Інтернету та пропонують нові персоналізовані механізми/техніки/інструменти взаємодії з комп'ютером [35-67], наприклад, через спам-фільтри трафіку е-пошти, ІІПС, віртуальних персоналізованих помічників, ІС автоматичного перекладу тощо.

КЛС з підтримкою аналізу природної мови знаходяться на перетині експериментальних досліджень [68-73] і практичного розроблення зазвичай комерційного ПЗ [74-85]. КЛС аналізу мовлення та текстової аналітики безпосередньо взаємодіють з користувачем через підтримку зворотного зв'язку, який суттєво та постійно впливає на функціонування ПЗ та результати аналізу.

Потенціал впровадження аналізу природної мови в КЛС/ІС/модулів на сьогодення постійно експоненційно зростає [68]. Непропорційно великий обсяг NLP-застосунків зазвичай реалізується великими кампаніями із-за складності проектів та необхідністю їх комерціалізації [74-85]. В міру поширення можливостей впровадження у повсякдення КЛС стають менш помітними, маскуючи складність їх реалізації. Паралельно розвиток науки з великих даних та комп'ютерної лінгвістики особливо на основі неангломовних природних корпусів текстів ще не досяг рівня, необхідного для спрощення, оптимізації та стандартизації процесів розроблення відповідного лінгвістичного ПЗ [585-594].

КЛС для розв'язку великого обсягу конкретних NLP-задач тільки починають поширюватися і в перспективі автоматизують більше процесів, які зараз вирішують через додаткові форми та вибір/натискання опцій/кнопок. Для розроблення ІТ реалізації відповідного лінгвістичного ПЗ та забезпечення

високої надійності КЛС необхідно враховувати сучасні перспективні наукові методи ML, аналізу даних, великих даних, та на основі аналізу гіпотез [585-594].

Для підтримки функціонування типової КЛС та роботи основних процесів при розв'язку конкретної NLP-задачі необхідно та достатньо реалізувати основні підсистеми як клієнтську, серверну та технологічну (Рис. 3.1) [585-594, 875-878].

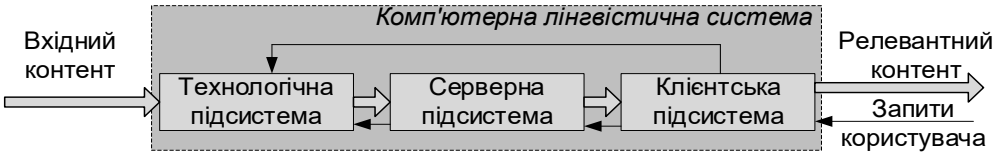


Рис. 3.1. Концептуальна схема функціонування типової КЛС

Основні процеси функціонування типової КЛС на основі інтелектуального аналізу текстового потоку для розв'язку конкретної NLP-задачі [875-878]:

- технологічне опрацювання вхідних потоків контенту:
  - пошук та розпізнавання контенту з відповідних джерел;
  - накопичення в хмарі аналізованого контенту з джерела;
  - збереження в базі/сховищі даних інформації про місце розташування знайденого контенту у відповідному джерелі;
  - попереднє опрацювання розпізнаного контенту в хмарі;
  - аналіз та маркування/класифікація розпізнаного контенту за ступенем релевантності по відношенню до змісту та мети КЛС;
  - інтеграція контенту за умови значення його ступеня релевантності/актуальності більшого за граничну величину;
  - збереження інтегрованого релевантного контенту в БД;
  - формування образу (описових службових даних) інтегрованого релевантного контенту та його збереження в БД;
- управління контентом на основі аналізу та опрацювання тексту через серверну підсистему (Рис. 3.2) на основі даних з клієнтської підсистеми та модуля супроводу контенту [585-594, 875-878]:
  - опрацювання потоків запитів користувачів з клієнтської підсистеми для формування коректного виразу ІІП та подальше кешування популярного контенту через серверну підсистему;

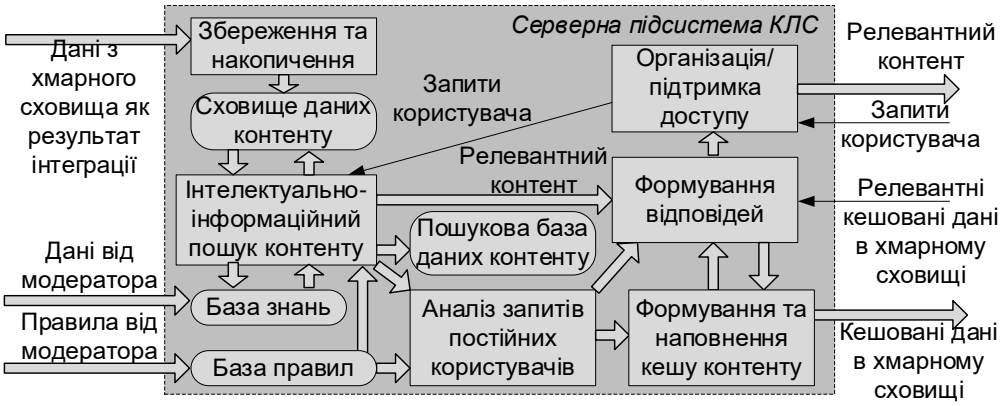


Рис. 3.2. Загальна концептуальна схема функціонування серверної підсистеми

- генерування множини оперативних релевантних актуальних звітів згідно відповідних запитів користувачів КЛС/Website;
- аналіз популярних запитів користувачів в певні проміжки часу для генерування стандартних кешованих в хмарі звітів;
- підтримка ефективного ІІП релевантного контенту згідно запитів користувача/відвідувача КЛС/Website [585-594, 875-878], зокрема, реалізація ІІП лексичного, символьного, атрибутивного, асоціативного, лінгвістичного тощо;
- управління взаємодією контентно-ресурсних елементів КЛС;
- підтримка хмарних обчислень для оперативності доступу до сховищ/баз даних/правил/фільтрів КЛС/Website;

– супровід контенту на основі аналізу та синтезу інформації, в тому числі службового контенту розв’язку конкретної NLP-задачі:

- формування та поповнення сховищ/баз даних/правил/фільтрів КЛС/Website в хмарі/ІС як результат технологічного процесу та процесів управління/інтеграції/супроводу контенту:
  - контентних/реферативних/анотованих/бібліографічних сховищ/баз даних/правил/фільтрів КЛС/Website;
  - семантичних NLP-засобів (правил, словників, бази знань, наприклад, онтології) для ІІП контенту, зокрема, бази лінгвістичних правил (графемного, морфологічного, лексичного, синтаксичного та семантичного);

- бази даних постійних користувачів КЛС та їх профілів;
- бази даних службового контенту функціонування КЛС;

- супровід ІТ взаємодії інформаційно-ресурсних елементів КЛС;
- аналіз статистики/ефективності функціонування КЛС та взаємодії з нею постійної аудиторії в певні проміжки часу;
- генерування сценаріїв прогнозу взаємодії користувачів з КЛС;
- поповнення/модифікація правил аналізу статистичних даних.

Сховище/базу даних КЛС/Website реалізують за трирівневою схемою [585]:

- файлове сховище для накопичення відповідного контенту;
- база знань для ІІП/модифікації/підтримки цього контенту;
- накопичення та відповідне опрацювання службового контенту.

Серверна підсистема КЛС сформована з частини функціональних компонентів модулів управління/супроводу/інтеграції/ контенту, зокрема:

- ІІП контенту на основі лінгвістичного аналізу запитів;
- формування та наповнення кешу часто запитуваних користувачами та відвідувачами інформаційних блоків релевантного популярного контенту для оперативного доступу;
- інтерактивного доступу до відповідних профілів/опцій КЛС;
- аналізу запитів користувачів для накопичення кешу контенту;
- збереження та накопичення інформаційних блоків в хмарі;
- витягування кешованого контенту з хмари за запитом користувача або його знищення із-за настання непопулярності;
- поповнення/модернізація модератором правил та бази знань ІІП/аналізу контенту як запитів постійних користувачів;
- аналізу запитів користувачів для генерування релевантних звітів.

3.1.2. Схематична модель типової КЛС

Анотована база даних КЛС є основою ІІП-модуля Website. Оперативний та якісний ІІП за контекстом актуального контенту забезпечує високу його релевантність для користувача КЛС. Застосовування в ІІП-модулі анотованих <L

сприяє реалізації ефективного ІІП релевантного контенту без інформаційного шуму (Рис. 3.3) [585-594, 875-878]. КЛС має забезпечувати [585-594, 875-878]:

- генерування Webpage за шаблоном та змісту Website;
- збереження та підтримки в актуальному стані кешу/наповнення Webpage/Website згідно потреб цільової аудиторії;
- надання оперативного доступу до Website всіх типів користувачів.

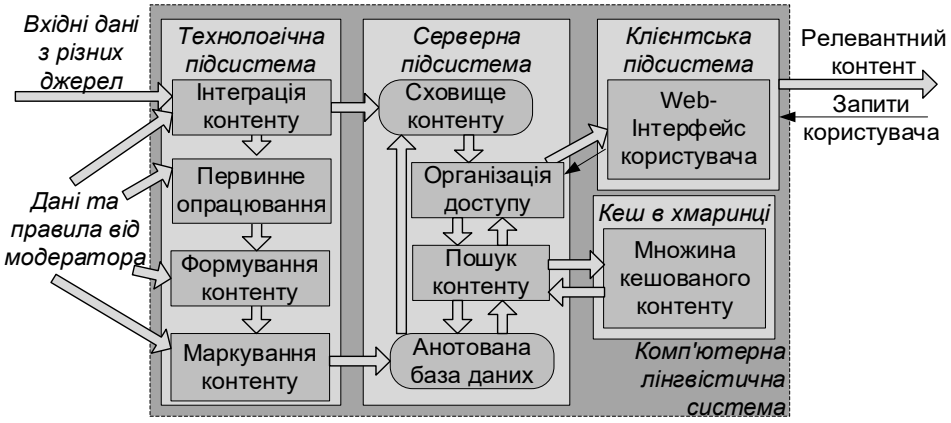


Рис. 3.3. Схематична модель типової КЛС

Mashup-ІС S_{MSS} полягає в формуванні множини інтегрованого контенту з Інтернет-ресурсів згідно потреб цільової аудиторії та конкретних запитів користувачів для зручної навігації по Website/Webpage [585-594, 875-878]:

$$S_{MSS} =< X, W, Q, Y, \varpi, \rho >, \tag{3.1}$$

де X – множина одночасно інтегрованого контенту з Інтернет-ресурсів W, Q – користувацькі запити до Website/Webpage Mashup-ІС S_{MSS}, Y – множина релевантного контенту як результат ІІП за запитом користувача/відвідувача Website/Webpage Mashup-ІС S_{MSS}; ϖ – оператор в інтеграції контенту з Інтернет-ресурсів W та ρ – оператор навігації в базах/сховищах даних/контенту/фільтрів.

Інтеграція множини даних X з різних джерел W, в тому числі з Website, полягає в їх об'єднанні згідно відповідної колекції умов U_{ϖ} в одному Website або Webpage з метою використання різного типу контенту зі збереженням його основних ознак, характеристик подання і можливості подальшого опрацювання:

$$X = \varpi(W, U_{\varpi}). \tag{3.2}$$

Інтеграція має забезпечувати користувача Website/Webpage Mashup-IC S_{MSS} сприймати інтегрований контент як єдиний інформаційний простір з використанням великих СД, в тому числі хмари, та якісний/оперативний ІІП релевантного контенту за запитами згідно колекції умов ІІП U_{ρ} [875-878]:

$$Y = \rho(Q, X, U_{\rho}). \quad (3.3)$$

Зручна навігація в Website/Webpage Mashup-IC S_{MSS} сприяє реалізації можливості підтримки користувача шукати релевантний та актуальний для нього контент по всьому доступному інформаційному просторі ІС з найбільшою повнотою і точністю при найменших витратах на зусилля з його боку [875-878].

КЛС є спеціалізованою ІС, СППР або мультиагентною системою [954-957] для розв'язку конкретної NLP-задачі на основі множини інтегрованого контенту з різних джерел згідно потреб цільової аудиторії та конкретних запитів користувачів для зручної навігації по Website/Webpage з врахування статистики функціонування КЛС, історії дій та персональних профілів користувачів та історії запитів/переходів з ІІПС [585-594, 875-878].

Типову формальну модель КЛС S_{CLS} подамо як кортеж [585-594, 875-878]:

$$S_{CLS} =< X, W, C, K, Y, D, S_{IAC}, M_{LA}, M_{\nu}, M_{\varpi_1}, M_{\varpi_2}, M_{\rho_1}, M_{\rho_2}, M_{\rho_3}, M_{\rho_4}, M_{\upsilon},$$
$$\nu, \varpi_1, \varpi_2, \rho_1, \rho_2, \rho_3, \rho_4, \upsilon >, \quad (3.4)$$

де X – вхідні дані в КЛС з різних джерел інформації W; Y – вихідний релевантний контент з КЛС як результат ІІП згідно запитів користувачів/відвідувачів; M_{LA} – модуль лінгвістичного аналізу контенту як складової ІАТПК-підсистеми S_{IAC}; M_{ν} – модуль генерування/модифікації правил функціонування всіх модулів від модератора КЛС (наприклад, правил оновлення кешу, інтеграції контенту з різних джерел інформації, лінгвістичного ІІП тощо); M_{ϖ_1} – модуль наповнення неструктурованої БД інтегрованим контентом X; M_{ϖ_2} – модуль наповнення структурованої БД на основі опрацьованого інтегрованого контенту C; M_{ρ_1} – модуль генерування результатів згідно запитів відвідувачів; M_{ρ_2} – модуль генерування результатів згідно запитів користувачів; M_{ρ_3} – модуль опрацювання

кешу для формування звітів на популярні запити від користувачів КЛС; M_{ρ_4} – модуль наповнення/модифікації кешу; M_{υ} – модуль генерування статистичних результатів функціонування КЛС/модулів та діяльності користувачів D; ν – оператор генерування/модифікації правил функціонування всіх модулів від модератора КЛС; ϖ_1 – оператор наповнення неструктурованої БД інтегрованим контентом X; ϖ_2 – оператор наповнення структурованої БД на основі опрацьованого інтегрованого контенту C; ρ_1 – оператор генерування результатів згідно запитів відвідувачів; ρ_2 – оператор генерування результатів згідно запитів користувачів КЛС; ρ_3 – оператор опрацювання кешу для формування звітів Y на популярні запити від користувачів КЛС; ρ_4 – оператор наповнення/модифікації кешу КЛС даними K; υ – оператор генерування статистичних результатів функціонування КЛС/модулів та діяльності користувачів КЛС [875-878].

На Рис. 3.4 подана загальна структура запропонованої типової КЛС розв'язку конкретної NLP-задачі на основі функціональних можливостей та взаємодії з хмарами [875-878].

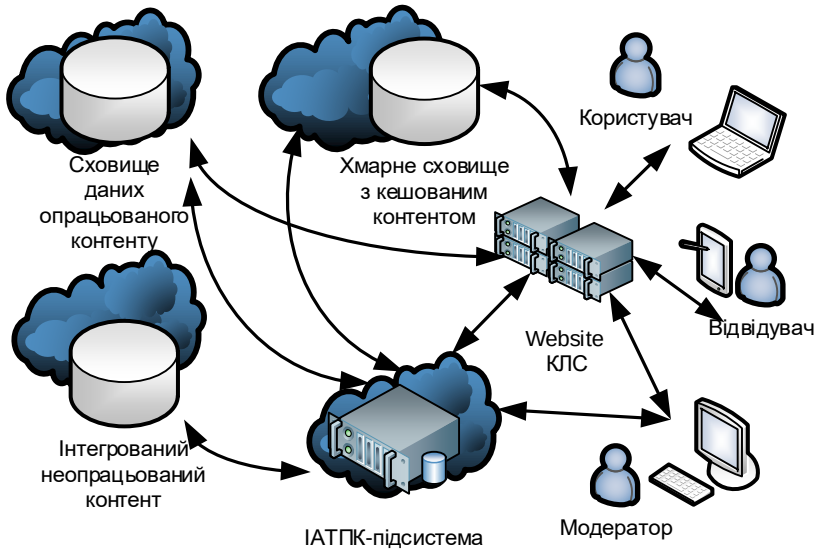


Рис. 3.4. Структура процесу функціонування типової КЛС

Колекція інтегрованого неопрацьованого контенту X міститься в базі даних на основі Non SQL [585-594, 875-878]. Колекція інтегрованого опрацьованого контенту C міститься в сховищі/базі даних на основі SQL. З C формується

наповнення K для хмари на основі статистики D популярних запитів Y від користувачів за певний проміжок часу [585-594, 875-878]. Колекція D є специфічною БД/СД кешованого актуального популярного контенту C для оптимізації функціонування КЛС на основі вбудованих/ модифікованих/ додаткових сервісів в хмарі [585-594, 875-878]. Ці сервіси є результатом роботи модератора, які оновлює правила кешування в КСЛ і/або Website, оновлення даних в SQL базі [585-594, 875-878], ІІП/ІАТПК/ керування/ інтегрування/ супроводу інтегрованого та службового контенту, інтеграції неопрацьованого контенту в Non SQL базі та збору/накопичення статистики функціонування КСЛ/Website [585-594, 875-878].

Основою ІАТПК-підсистеми є основні NLP-процеси КЛС [875-878].

3.2. Формальне моделювання основних NLP-процесів КЛС

3.2.1. Формальна модель комп'ютерної лінгвістичної системи для опрацювання україномовного текстового контенту

Природні мови визначаються не правилами, а контекстом *застосування*, яке реконструюють для комп'ютерного опрацювання [68-73, 506-511, 958-983]. Часто ідентифікуємо значення вживаних слів в поєднанні з іншими співрозмовниками [68]. Словосполучення *золота рибка* означає як морську істоту, так людину з короткою пам'яттю, або істота/людина, яка виконує бажання, тому співрозмовники мають погодитися зі спільним розумінням контексту [506-511]. Відповідно мовлення/мова обмежується менталітетом, регіоном, суспільством і рівнем освіти. Передача змісту/сенсу найпростіше для співрозмовників з подібним життєвим досвідом, освітою, місцем проживання тощо. Тому автоматизувати розуміння промови для розв'язку конкретної NLP-задачі через відповідну КЛС досить складний кропіткий процес, особливо для синтетичних мов, зокрема для української мови [585-594].

Загальну формальну модель КЛС подано колекцією [585-594]:

$$S_{LA} =< X, Y, C, D, R, \alpha, \beta, \gamma, \delta, \lambda, o, \iota, \varsigma, \mu >, \qquad (3.5)$$

де X – вхідний текстовий масив даних; Y – кортеж вихідного опрацьованого тексту згідно призначення КЛС; C – множина проміжного контенту, який опрацьовується на відповідному рівні в КЛС; D – допоміжні словники; R – множина правил опрацювання контенту; α – оператор ФА або ГА тексту; β – оператор МА тексту; γ – оператор ЛА контенту; δ – оператор СА контенту; λ – оператор семантичного аналізу; o – оператор онтологічного аналізу контенту; ι – оператор референційного аналізу контенту; ς – оператор структурного аналізу контенту; μ – оператор ПА контенту.

В порівнянні з формальними мовами (предметні/речові/об'єктні) природні мови є універсальнішими, але менш формалізованими [68-73, 506-511, 958-983]. Використовуємо часто одне слово для опису декількох значень (наприклад, *краб* – морська істота, блюдо, туманність в сузір'ї Тельця та кокарда на кашкетів моряків) в залежності від змісту діалогу (наприклад, опис емоцій дайвінгу, вечері, прочитаної книги, відвідування музею, перегляду історичного фільму тощо лише для слова *краб*) [506-511]. Для збереження множини значень для кожного слова мова має бути надмірною (перевищення кількості інформації для передавання/зберігання повідомлення над її ентропією). Тобто не можливо наперед визначити точне значення змісту для кожної асоціації (кожна лінгвістична змінна є неоднозначною за замовчуванням). Лексична і структурна неоднозначність є великим досягненням природньої мови, наприклад, для генерування нових ідей, проявів творчості [68-73, 506-511, 958-983].

Незалежно від NLP-задачі процес опрацювання україномовних текстів в довільній КЛС подано як послідовність обов'язкових операторів для проведення змістовного структурного аналізу вхідного текстового контенту:

інформаційне джерело $\rightarrow$ *вхідний текст* $\rightarrow$ *графемний аналіз* $\alpha \rightarrow$ *морфологічний аналіз* $\beta \rightarrow$ *лексичний аналіз* $\gamma \rightarrow$ *синтаксичний аналіз* $\delta \rightarrow$ *семантичний аналіз* $\lambda \rightarrow$ *структурований текстовий контент*

Додатковими операторами є такі аналізи як прагматичний μ (видобування знань), онтологічний o та референційний ι (формування міжфразових єдностей). Їх застосування залежить від складності та мети розв'язку NLP-задачі [586].

Основний процес лінгвістичного аналізу текстового контенту подано:

$$Y = \mu \circ o \circ \varsigma \circ \iota \circ \lambda \circ \delta \circ \gamma \circ \beta \circ \alpha, \quad (3.6)$$

$$Y = \mu(C_\mu, D_\mu, R_\mu, o(C_o, D_o, R_o, \varsigma(C_\varsigma, D_\varsigma, R_\varsigma, \iota(C_\iota, D_\iota, R_\iota, , \lambda(C_\lambda, D_\lambda, R_\lambda, \delta(C_\delta, D_\delta, R_\delta,$$
$$\gamma(C_\gamma, D_\gamma, R_\gamma, \beta(C_\beta, D_\beta, R_\beta, \alpha(C_\alpha, D_\alpha, R_\alpha, X))))))))), \quad (3.7)$$

де множини текстового контенту $C = \{C_\mu, C_o, C_\varsigma, C_\iota, C_\lambda, C_\delta, C_\gamma, C_\beta, C_\alpha\}$, лінгвістичних словників $D = \{D_\mu, D_o, D_\varsigma, D_\iota, D_\lambda, D_\delta, D_\gamma, D_\beta, D_\alpha, \}$ та множини продукційних/асоціативних правил $R = \{R_\mu, R_\iota, R_o, R_\varsigma, R_\lambda, R_\delta, R_\gamma, R_\beta, R_\alpha\}$.

Основний лінгвістичний процес опрацювання текстової україномовної інформації для розв'язку конкретної NLP-задачі складається з дев'яти етапів:

Етап 1. Графемний аналіз α текстової україномовної інформації X [586]:

$$C_\alpha = \alpha(X, D_\alpha, R_\alpha), \quad (3.8)$$

$$C_\alpha = \alpha_7 \circ \alpha_6 \circ \alpha_5 \circ \alpha_4 \circ \alpha_3 \circ \alpha_2 \circ \alpha_1, \quad (3.9)$$

де X – вхідний текстовий масив даних; α – оператор ГА; C_α – графемна структура вхідного тексту; D_α – графемні словники та бібліотеки; R_α – правила графемного аналізу; α_1 – OCR-оператор [534-535, 716, 862]; α_2 – оператор графемного розбору вхідного тексту X на розділи (інформаційні блоки), абзаци та речення; α_3 – оператор графемного розбору лінгвістичних ланцюжків на окремі слова; α_4 – оператор формування множини нерозпізнаних ланцюжків; α_5 – оператор ідентифікації та маркування нерозпізнаних ланцюжків як числа, дати, незмінних зворотів, скорочень, власних та географічних назв тощо; α_6 – оператор маркування нетекстових ланцюжків як спецсимволи, формули, рисунки, таблиці тощо; α_7 – оператор генерування маркованої лінійної послідовності слів C_α із службовими знаками та зв'язками.

ГА заміняється ФА у разі розпізнавання змісту людського мовлення [586].

Етап 2. Морфологічний аналіз β текстового контенту C_β полягає в ідентифікації, аналізі та визначенні форми і структури слів, зокрема:

$$C_\beta = \beta(C_\alpha, D_\beta, R_\beta), \quad (3.10)$$

$$C_\beta = \beta_3 \circ \beta_2 \circ \beta_1 \text{ або } C_\beta = \beta_3 \circ \beta_4 \circ \beta_1, \quad (3.11)$$

де β_1 – оператор морфологічної сегментації (англ. Morphological Segmentation) графемно розпізнаного ланцюжка символів (слів/лексем); β_2 – оператор лематизації (англ. Lemmatization) лексем; β_3 – POST-оператор (розмічування частин мови) для сегментованих слів; β_4 – оператор стемінг (англ. Stemming) слів [68-73, 506-511, , 585-594, 958-983].

Класичний загальний алгоритм морфологічного аналізу [585-594].

Крок 1. Морфологічна сегментування вхідного ланцюжка символів (заміна ГА для коротких англомовних повідомлень, і доповнення до ГА для великих корпусів англомовних текстів, а для україномовних всіх видів текстів – окремий крок для маркування слів на дві множини як зразу ідентифікованих (наприклад, прийменники), чи неможливо ідентифікувати (іменник не в називному відмінку).

Крок 2. Лематизація (приведення до нормальної форми на основі аналізу словників) або стемінг – визначення основ (словоформ із відсіканням закінчень).

Крок 3. Ідентифікація граматичної категорії кожного слова та колекції їх відповідних властивостей по відношенню вживання в конкретному місці тексту. (наприклад, колекція для іменника: рід, відмінок, особа тощо).

Крок 4. Формування лінійної послідовності морфологічних структур.

Етап 3. Лексичний аналіз γ текстового контенту C_γ у проміжному етапі аналізу послідовності лексем для генерування дерева розбору на рівні СА [586]:

$$C_\gamma = \gamma(C_\beta, D_\gamma, R_\gamma), \tag{3.12}$$

$$C'_\gamma = \gamma_2 \circ \gamma_1,\ C'_\gamma = \gamma_5 \circ \gamma_4 \circ \gamma_3 \text{ або } C'_\gamma = \gamma_5 \circ \gamma_4. \tag{3.13}$$

де γ_1 – оператор сегментації мовлення (англ. Speech Segmentation) для ідентифікації/уточнення слів/словосполучень/лексем після МА або у разі не коректної інтерпретації при ФА (зазвичай виконується паралельно з ФА та МА в циклічному процесі) [586]; γ_2 – оператор розпізнавання мовлення (англ. Speech Recognition, SR) [85, 644] або мовлення-у-текст (англ. Speech-To-Text, STT) [359, 405] в залежності від змісту NLP-задачі; γ_3 – оператор оптичного розпізнавання символів (англ. Optical Character Recognition, OCR) як друга частина після ГА та МА для уточнення некоректних моментів розпізнавання з врахуванням

розпізнаних сусідніх лексем; γ_4 – оператор токенізації/сегментації слів (англ. Tokenization, Word Segmentation) як підготовка даних для побудови дерева розбору при СА; γ_5

Етап 4. Синтаксичний аналіз δ текстового контенту C_δ полягає в побудові дерева розбору залежностей слів в послідовності лексем на основі їх категорій:

$$C_\delta = \delta(C_\gamma, D_\delta, R_\delta),\ C_\delta = \delta_3 \circ \delta_2 \circ \delta_1, \quad (3.14)$$

де δ_1 – оператор реалізації індукції граматики (англ. Grammar induction) [416-430]; δ_2 – оператор ідентифікації/ліквідації неоднозначності меж або порушення речення (англ. Sentence Breaking, Sentence Boundary Disambiguation) [19-23, 160, 958-983]; δ_3 – оператор синтаксичного парсингу (англ. Parsing) фраз/речень для побудови дерева СА [112-114, 140-147, 295, 471-479, 958-983].

Етап 5. Семантичний аналіз λ текстового контенту C_λ полягає

$$C_\lambda = \lambda(C_\delta, D_\lambda, R_\lambda),\ C_\lambda = \lambda_2 \circ \lambda_1, \quad (3.15)$$

де λ_1 – оператор ідентифікації лексичної семантики з генеруванням колекції значень кожної лексеми тексту; λ_2 – оператор ідентифікації реляційної семантики взаємозалежностей змісту лексем тексту [19-23, 287-297, 958-983].

Класичний загальний алгоритм семантичного аналізу.

Крок 1. Лексеми порівнюють зі змістовними значеннями словника.

Крок 2. Формування множин ймовірнісних для кожного фрагменту тексту/речення/фрази альтернативним сем відповідно для лексем.

Крок 3. Попереднє взаємозвя’зування змісту лексем у єдину структуру.

Крок 4. Генерування упорядкованої колекції логічних записів суперпозицій з семантичних класів лексем і базисних лексичних функцій.

Крок 5. Знаходження/маркування неточності, протирічь, некоректності та неоднозначності змісту отриманого результату на основі лексичного словника.

Семантичний аналіз на сьогодення використовують не у більшості КЛС,

Етап 6. Референційний аналіз ι формування міжфразових єдностей C_ι.

$$C_\iota = \iota(C_\lambda, D_\iota, R_\iota). \quad (3.16)$$

Референційний аналіз часто є частиною семантичного аналізу [212]. Для слов'янських мов при аналізі великих корпусів текстів найкраще виносити як окремий етап (наприклад для аналізу переписки соціальної групи/спільноти в соціальних мережах або інших діалогів для ідентифікації логічних змістовних зав'язків між дописами різних учасників із-за суб'єктивізму мовлення кожного.

Класичний загальний алгоритм референційного аналізу [212].

Крок 1. Контекстний аналіз маркованих фрагментів текстового контенту C_λ, наприклад, аналіз займенника/сполучника *що* в залежності від контексту для відокремлення цента єдності або локальних референцій типу *його*, *який*, *цей*.

Крок 2. Актуальне членування речення маркованих фрагментів текстового контенту C_λ для ідентифікації тематичних структур на основі тем/рем.

Крок 3. Ідентифікація регулярної повторюваності контексту/теми/реми.

Крок 4. Виокремлення дубльованої номінації лексичних одиниць тексту.

Крок 5. Ідентифікація синонімізації лексичних одиниць тексту.

Крок 6. Виокремлення імплікацій на основі ситуативних зв'язках.

Крок 7. Ідентифікація тотожності референції (наприклад, співставлення лексичних одиниць тексту з об'єктом/суб'єктом/явищем діалогу/зображення).

Етап 7. Структурний аналіз ς текстового контенту C_ς на основі ступеня збігу лексичних термінологічних одиниць єдності фрагментів тексту.

$$C_\varsigma = \varsigma(C_\iota, D_\varsigma, R_\varsigma) \text{ або } C_\varsigma = \varsigma(C_\lambda, D_\varsigma, R_\varsigma). \quad (3.17)$$

Аналогічно до референційного аналізу часто є частиною СЕМ для коротких текстів/повідомлень, або взагалі не використовують [212]. Для великих корпусів текстів як додатковий етап ліквідації маркованої неточності при СЕМ.

Класичний загальний алгоритм структурного аналізу [212].

Крок 1. Формування/поповнення базової множини риторичних відношень міжфразових єдностей на основі результатів референційного аналізу або/і СЕМ.

Крок 2. Генерування нелінійної мережі/графу міжфразових єдностей.

Етап 8. Онтологічний аналіз о текстового контенту C_o на основі результатів СЕМ [13-15, 113, 136, 139, 143, 146-147, 163, 294, 297, 474, 477, 479, 490, 491, 579, 803-822] та референційного/структурного аналізів при потребі:

$$C_{\text{о}} = \text{о}(C_{\varsigma}, D_{\text{о}}, R_{\text{о}}),\ C_{\text{о}} = \text{о}(C_{\iota}, D_{\text{о}}, R_{\text{о}}) \text{ або } C_{\text{о}} = \text{о}(C_{\lambda}, D_{\text{о}}, R_{\text{о}}). \quad (3.18)$$

Етап 9. Прагматичний аналіз μ текстового контенту C_{μ} застосовують для визначення структури тексту з врахуванням контексту речень при формуванні абзаців, розділів та діалогів. ПА є суттєвим доповненням СЕМ, референційного та структурного аналізів, якщо вони не сприяли ліквідації маркованої неточності. В деяких випадках достатньо ПА застосовувати зразу ж після СЕМ. Також є незамінним етапом підготовки даних для видобування знань з корпусів тексту.

$$Y = \mu(C_{\mu}, D_{\mu}, R_{\mu}, C_{\lambda}, [C_{\text{о}}, C_{\varsigma}, C_{\iota}],),\ Y = \mu_2 \circ \mu_1, \quad (3.19)$$

де μ_1 – оператор ідентифікації семантики поза окремими реченнями/фразами; μ_2 – оператор опрацювання текстів через вищого рівня NLP-додатки, наприклад, для імітування розумної поведінки та очевидного розуміння природної мови.

3.2.2. Моделі графемного та фонологічного аналізу тексту українською мовою

В залежності від кокетної NLP-задачі застосовують графемний (аналітика тексту) або фонологічний (розпізнавання мовлення) аналіз текстового контенту. ФА полягає у дослідженні структури, організації та інтерпретації звуків мовлення X конкретної природної мови (додаток А, таблиця А.2) на основі фонемних, фонетичних та просодичних правил R_{α} та словників D_{α} аналогів [534-535, 716, 862]. ГА є рекурсивним розбором тексту X з врахуванням лінгвістичних ознак графем різних мов (в тому числі і не природних, наприклад, математичних, програмних, псевдомов тощо) на основі правил розпізнавання рядків певної мови R_{α} та словників D_{α} еталонних моделей графем, зокрема:

$$C'_{\alpha} = \alpha(C_{\alpha}, D_{\alpha}, R_{\alpha}, X),\ C'_{\alpha} \supseteq C_{\alpha}. \quad (3.20)$$

ГА може бути OCR-частиною – кодування/розпізнавання тексту на зображенні в ланцюжок символів для е-подання [534-535, 716, 862].

В залежності від задач є такі методи для формування правил R_{α} ФА[716]:

- Загальна фонологія α'_1 (правила організації фонем в різних мовах).
- Описова фонологія α'_2 (ідентифікація фонеми мови або діалекту).
- Історична фонологія α'_3 (зміни у фонемах, складі мови за період).

- Сегментна фонологія α'_4 (аналіз фонем, складів, фонетичних слів, синтагм та фраз).
- Надсегментна фонологія α'_5 (аналіз інтонації, тону, наголосу, ритму, темпу та пауз).
- Фонетичний аналіз α'_6 (аналіз звукового складу мови).
- Фонематичний аналіз α'_7 найменшої одиниці фонологічного рівня.

ГА є елементарним парсингом тексту (Рис. 3.5) з врахуванням особливостей графем різних мов та використанням спеціальних символів, об'єктів та позначок [211-212, 534-535, 716, 862]. Графемою є атомарна змістовна лінгвістична (графемна) одиниця писемного тексту (знак, символ, спеціальний символ, об'єкт як рисунок тощо). Метою ГА є формування моделі графемної структури вхідного тексту та генерування графемних правил (регулярних виразів) ідентифікації/класифікації графемних одиниць в послідовності символьних рядків/графем та зав'язків між ними. Метою першого рівня ГА – ідентифікації графем – є маркування змістовно самостійних послідовностей символів, лексем у цих послідовностях та визначення основної мови вхідного текстового контенту/фрагментів на основі апріорних еталонів графем (Рис. 3.5) [211-212, 534-535, 716, 862]. Кортеж еталонних моделей графем найкраще описати формальною граматикою (скороченні позначення критеріїв подані в таблиці А.3) [211-212]. Паралельно парсингу/ідентифікації графем їх класифікують/маркують згідно закладених правил (Таблиця 3.1) [211-212, 534-535, 716, 862].

Таблиця 3.1

Правила ідентифікації графем у вигляді рядків[211-212, 534-535, 716, 862]

№	Правило	Розшифровка	Пояснення	Позиція 1	Позиція N	Всі позиції
1	EmpStr	empty string	Пустий рядок	–	–	Space
2	FulStr	full string	Повний рядок	Symbol	Symbol	–
3	IncRgt	incomplete right	Неповний праворуч	Symbol	Space	–
4	IncLgt	incomplete left	Неповний ліворуч	Space	Symbol	–
5	SmtInc	symmetric incomplete	Симетрично неповний	Space	Space	–

Розглянемо класичну граматику Хомські $Grammar$ з алфавітом $Alphabet$ та термами $Terms$ [211-212]:

$$Grammar =< Alphabet, Terms, Symbol, PrRules >, \quad (3.21)$$

$$Alphabet =< Gr, Terms >, \quad (3.22)$$

Terms :=<А, а, Б, б, В, в, Г, г, Ґ, ґ, Д, д, Е, е, Є, є, Ж, ж, З, з, И, и, Й, й, І, і, Ї, ї, К, к, Л, л, М, м, Н, н, О, о, П, п, Р, р, С, с, Т, т, У, у, Ф, ф, Х, х, Ц, ц, Ч, ч, Ш, ш, Щ, щ, ь, Ю, ю, Я, я, Э, э, ы, ъ, ', A, a, B, b, C, c, D, d, E, e, F, f, G, g, H, h, I, i, J, j, K, k, L, l, M, m, N, n, O, o, P, p, Q, q, R, r, S, s, T, t, U, u, W, w, V, v, X, x, Y, y, Z, z, Ä, ä, Ö, ö, Ü, ü, Ą, ą, Ć, ć, Ę, ę, Ł, ł, Ń, ń, Ó, ó, Ś, ś, Ź, ź, Ż, ż, ß>.

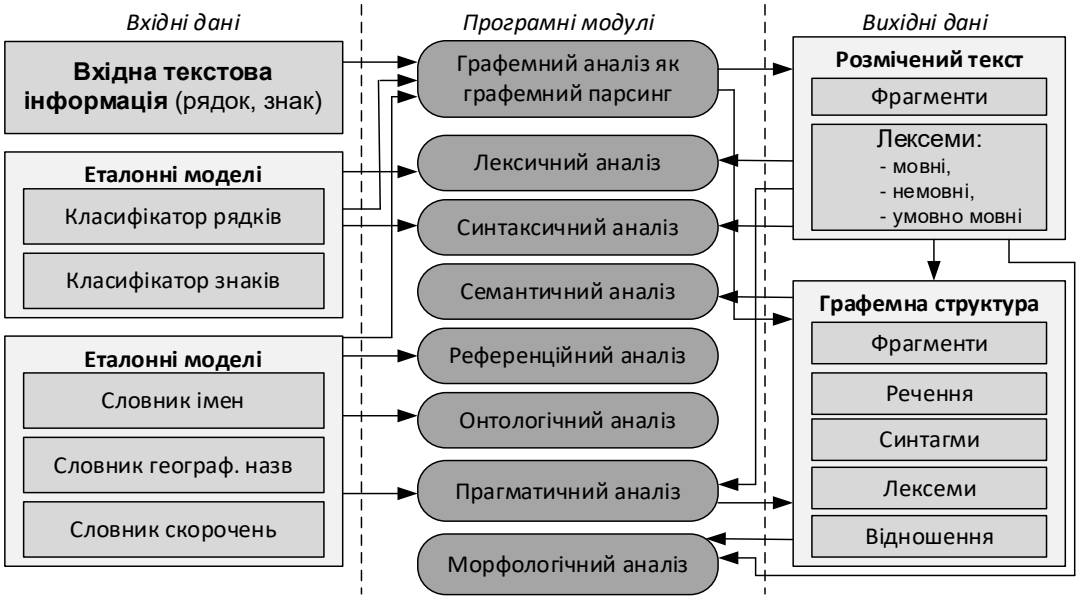


Рис. 3.5. Загальна схема процесу графемного аналізу вхідного тексту

В табл. А.4 додатку А поданий список правил класифікації/маркування графем для еталонних моделей (табл. А.3) згідно продукційних правил [211]:

PrRules := < Symbol→Λ, Symbol→Symbol *Gr*, *Gr*→Λ, *Gr*→*Gr'*, *Gr*→*Gr'Gr*, *Gr*→*Gr Sp*, *Gr*→*Gr Sb*, *Sb*→*Ssg*, *Sb*→*Ssb*, *Sb*→*Dgt*, *Sb*→*Ltr*, *Sp*→*_*, *Ltr*→', *Ltr*→*Vwl*, *Ltr*→*Cnl*, *Ltr*→*Sml*, *Ltr*→*Cpl*, *Ltr*→*Rus*, *Ltr*→*Ukr*, *Ltr*→*Pol*, *Ltr*→*Ger*, *Ltr*→*Eng*, *Ltr*→*Cyr*, *Ltr*→*Lat*, *Ssb*→*Msb*, *Ssb*→*Bsb*, *Ssb*→*Osb*, *Cpl*→*Rcp*, *Cpl*→*Ucp*, *Cpl*→*Pcp*, *Cpl*→*Gcp*, *Cpl*→*Ecp*, *Cpl*→*Ccp*, *Cpl*→*Lcp*, *Sml*→*Rsm*, *Sml*→*Usm*, *Sml*→*Psm*, *Sml*→*Gsm*, *Sml*→*Esm*, *Sml*→*Csm*, *Sml*→*Lsm*, *Lat*→*Lsm*, *Lat*→*Lcp*, *Cyr*→*Csm*, *Cyr*→*Ccp*, *Eng*→*Esm*, *Eng*→*Ecp*, *Ger*→*Gsm*, *Ger*→*Gcp*, *Pol*→*Psm*, *Pol*→*Pcp*, *Ukr*→*Usm*, *Ukr*→*Ucp*, *Rus*→*Rsm*, *Rus*→*Rcp*, *Lcp*→X, *Lcp*→V, *Lcp*→Q, *Lcp*→*Lcv*, *Lcp*→*Lcc*, *Lsm*→x, *Lsm*→v, *Lsm*→q, *Lsm*→*Lsv*, *Lsm*→*Lsc*, *Ccp*→Й, *Ccp*→Ь, *Ccp*→*Csv*, *Ccp*→*Ccc*, *Csm*→й, *Csm*→ь, *Csm*→*Csv*, *Csm*→*Csc*, *Ecp*→X, *Ecp*→V, *Ecp*→Q, *Ecp*→*Lcv*, *Ecp*→*Lcc*, *Esm*→x, *Esm*→v, *Esm*→q, *Esm*→*Lsv*, *Esm*→*Lsc*, *Gcp*→X, *Gcp*→V, *Gcp*→Q, *Gcp*→Ü, *Gcp*→Ö, *Gcp*→Ä, *Gcp*→*Lcv*, *Gcp*→*Lcc*, *Gsm*→x, *Gsm*→v, *Gsm*→q, *Gsm*→ß, *Gsm*→ü, *Gsm*→ö, *Gsm*→ä, *Gsm*→*Lsv*, *Gsm*→*Lsc*, *Pcp*→Ż, *Pcp*→Ź, *Pcp*→Ś, *Pcp*→Ó, *Pcp*→Ń, *Pcp*→Ł, *Pcp*→Ę, *Pcp*→Ć, *Pcp*→Ą, *Pcp*→*Lcv*, *Pcp*→*Lcc*, *Ps*→ż, *Psm*→ź, *Psm*→ś, *Psm*→ó, *Psm*→ń, *Psm*→ł, *Psm*→ę, *Psm*→ć, *Psm*→ą, *Psm*→*Lsv*, *Psm*→*Lsc*, *Ucp*→Ґ, *Ucp*→Ї, *Ucp*→І, *Ucp*→Є, *Ucp*→*Ccv*, *Ucp*→*Ccc*, *Usm*→ґ, *Usm*→ї, *Usm*→і, *Usm*→є, *Usm*→*Csv*, *Usm*→*Csc*, *Rcp*→Ъ, *Rcp*→Э, *Rcp*→Ы, *Rcp*→*Ccv*, *Rcp*→*Ccc*, *Rsm*→ъ, *Rsm*→э, *Rsm*→ы, *Rsm*→*Csv*, *Rsm*→*Csc*, *Lcc*→Z, *Lcc*→W, *Lcc*→T, *Lcc*→S, *Lcc*→R, *Lcc*→P, *Lcc*→N, *Lcc*→M, *Lcc*→L, *Lcc*→K, *Lcc*→J, *Lcc*→H, *Lcc*→G, *Lcc*→F, *Lcc*→D, *Lcc*→C, *Lcc*→B, *Lcv*→Y, *Lcv*→U, *Lcv*→O, *Lcv*→I, *Lcv*→E, *Lcv*→A, *Lsc*→z, *Lsc*→x, *Lsc*→w, *Lsc*→t, *Lsc*→s, *Lsc*→r, *Lsc*→q, *Lsc*→p, *Lsc*→n, *Lsc*→m, *Lsc*→l, *Lsc*→k, *Lsc*→j, *Lsc*→h, *Lsc*→g, *Ls*→f, *Lsc*→d, *Lsc*→c, *Lsc*→b, *Lsv*→y, *Lsv*→v, *Lsv*→u, *Lsv*→o, *Lsv*→i, *Lsv*→e, *Lsv*→a, *Ccc*→Щ, *Ccc*→Ш, *Ccc*→Ч, *Ccc*→Ц, *Ccc*→Х, *Ccc*→Ф, *Ccc*→Т, *Ccc*→С, *Ccc*→Р, *Ccc*→П, *Ccc*→Н, *Ccc*→М, *Ccc*→Л, *Ccc*→К, *Ccc*→З, *Ccc*→Ж, *Ccc*→Д, *Ccc*→Г, *Ccc*→В, *Ccc*→Б, *Csv*→Я, *Csv*→Ю, *Csv*→У, *Csv*→О, *Csv*→И, *Csv*→Е, *Csv*→А, *Csc*→щ, *Csc*→ш, *Csc*→ч, *Csc*→ц, *Csc*→х, *Csc*→ф, *Csc*→т, *Csc*→с, *Csc*→р, *Csc*→п, *Csc*→н, *Csc*→м, *Csc*→л, *Csc*→к, *Csc*→з, *Csc*→ж, *Csc*→д, *Csc*→г, *Csc*→в, *Csc*→б, *Csv*→я, *Csv*→ю, *Csv*→у, *Csv*→о, *Csv*→и, *Csv*→е, *Csv*→а >.

Продукційні правила $PrRules$ застосовують для ідентифікації, класифікації та маркування змістовних графемних одиниць аналізу вхідного тексту контенту X (слова, скорочення, стійкі словосполучення як фразеологізми та метафори,

межі речень та цитат/сарказмів за пунктуацією, смайликів, географічні та власних назв, абревіатур, слова з апострофами тощо) на етапі парсингу з врахуванням мови фрагментів текстів. Вимоги ідентифікації графемної одиниці в послідовності символів для подальшого морфологічного аналізу слів [211]:

1) символьна послідовність легко ідентифікується та класифікується;

2) завелика послідовність для ідентифікації значення за словником;

3) замала послідовність для ідентифікації багатьох значень.

4) кількість графемних одиниць завелика для розбиття вибірки.

3.2.3. Морфологічний аналіз української мови

МА полягає в ідентифікації, аналізі та визначенні форми і структури слів у природомовному тексті з використанням методів як морфологічна сегментація β_1 (англ. Morphological Segmentation), лематизація β_2 (англ. Lemmatization), POST β_3 (розмічування частин мови) та стемінг β_4 (англ. Stemming, Таблиця А.5) [19-23, 256-259, 534-535], зокрема:

$$C'_\beta = \beta(C_\beta, D_\beta, R_\beta), \tag{3.23}$$

де $C'_\beta \subseteq C_\beta$, $C'_\beta = \beta_3 \circ \beta_2 \circ \beta_1$ (достатньо для англомовних коротких текстів визначеної тематики) або $C'_\beta = \beta_3 \circ \beta_4 \circ \beta_1$ (для більшості випадків повідомлень різної тематики). Найкращий спосіб для слов'янських мов [169, 273, 534-541]:

$$C'_\beta = \beta_3 \circ \beta_2 \circ \beta_4 \circ \beta_1. \tag{3.24}$$

Лематизація β_2 – це трансформування словоформи в лему (нормальна словникова форма) [19-23, 404]. Зазвичай при трансформації використовують словник для подання слів у фактичній формі [19-23]. В іншому випадку – видалення лише флексій та повернення до леми [19-23, 404, 256-259, 534-535].

Морфологічна сегментація β_1 – це поділ слів на окремі морфеми для ідентифікації їх класу (Таблиця А.6-А.8) [19-23, 404, 256-259, 534-535, 716, 882]. Складність прямо пропорційна до складності морфології (структури слів) конкретної природної мови (Таблиця А.9-А.10) [534-535, 716, 862]. Англійська мова має суттєво просту морфологію, особливо флективну морфологію, і, отже,

часто повністю ігнорують цю задачу та відповідно моделюють всі можливі форми слова (наприклад, для to run [rʌn] – run, runs, ran, running, для to work [wərk] –work, works, worked, working) як окремі слова [256-259, 534-535]. У таких мовах, як турецька або індійська мова, такий підхід неможливий, оскільки кожен словниковий запис має тисячі можливих форм слів [534-535]. Слов'янські мови досить складні та мають багато закінчень для одного слова в залежності від відмінку [19-23, 256-259, 534-535]. Наприклад, дієслово бігати (англ. to run) має суттєво великий словниковий запас в слов'янських мовах, тому класифікуємо однокореневі лексеми за частинами мови [19-23, 256-259, 534-535]. Тільки для першої версії перекладу залежно від структури речення, може бути 36 варіантів вибору відповідної форми дієслова українською мовою та 60 варіантів для російської мови (Таблиця 3.2) [19-23, 256-259, 534-535].

Таблиця 3.2

Форми дієслова бігати для різних мов в залежності від контексту [534-535]

№	Англійська	Українська	Російська	№	Англійська	Українська	Російська
1	run	бігати	бегать	31	running	бігають	бегают
2		біг	бег	32		бігаючи	бегая
3		біжите	бежите	33		біжить	бежит
4		біжать	бегут	34		біжучи	бегучи
5		біжи	беги	35		бігаючи	бегающий
6		біжіть	бегите	36			бегающая
7	I run	бігаю	бегаю	37			бегающее
8		біжу	бегу	38			бегающие
9	let's run	біжимо	бежим	39			бегающего
10		бігаємо, бігаєм	бегаем	40			бегающей
11		біжімо	побежали	41			бегающих
12	you run	бігаєш	бегаешь	42			бегающему
13		бігаєте	бегаете	43			бегающим
14		бігай	бегай	44			бегающую
15		біжиш	бежишь	45			бегающею
16		бігайте	бегайте	46			бегающими
17	runs	бігає	бегает	47			бегающем
18	ran	бігав	бегал	48		бігши	бегавший
19		бігала	бегала	49			бегавшая
20		бігало	бегало	50			бегавшее
21		бігали	бегали	51			бегавшие
22		бігла	бежала	52			бегавшего
23		бігло	бежало	53			бегавшей
24		бігли	бежали	54			бегавших
25	I will run	бігтиму	побегу	55			бегавшему
26	you will run	бігтимеш	побежишь	56			бегавшим
27	you will run	бігтимете	побежите	57			бегавшую
28	will run	бігтиме	побежит	58			бегавшею
29	we will run	бігтимемо	побежим	59			бегавшими
30	will run away	бігтимуть	побегут	60			бегавшем

Тоді для всіх із 13 форм з табл. 1.1 слова run є по 36 варіантів аналогів українською без врахування контексту конкретного речення – це понад 450 результатів. Окрім загальних основних форм для дієслова *бігти* є понад двох дюжин, менш вживаних слів [534-535], наприклад,

вбігати, вибігати, добігати, забігати, набігати, оббігати, відбігати, перебігати, побігати, підбігати, пробігати, збігати, вбігти, вибігти, добігти, забігти, набігти, відбігти, перебігти, прибігти, підбігти, пробігти, забігати, набігатися, пробігтися, розбігтися, збігтися, убігати, тощо.

Кожне з цих слів має близько 36 варіантів форм залежно від структури речення та контексту (понад 1000 варіантів форм слів). Крім того, слово *run* при перекладі українською мовою також може набувати значення іменника, прикметника, прислівника, прикметника або складного слова [534-535]. До того ж іменник та прикметник мають власну форму відмінювання (7 випадків українською мовою з різними флексіями та відповідними чергуваннями літер в залежності від правил морфології мови), наприклад [534-535]:

- Іменник в різних родах [534-535], наприклад:

біг, бігання, біганина, бігун, бігунка, біженець, біженка, біженство, вибіг, вибігання, забіг, забігайлівка, забігання, набіг, набігання, перебігання, перебіжчик, перебіжчиця, побігайчик, побігеньки, пробіг, пробіжка, перебігання, розбіг, розбіжка, збігання, тощо.

- Прикметник [534-535], наприклад:

побіжний, біговій, біженський, набіганий, збіганий, вибіганий, забіганий, пробіганий, перебіганий, розбіганий, тощо.

- Прислівники [534-535], наприклад:

побіжно, бігом, перебіжкою, набігом, набігу, забігом, вибіганням, забігом, набіганням, перебіганням, перебіжчиком, пробігом, розбігом, збігом, тощо.

- Дієприслівники, утворені майже від кожного дієслова [882], наприклад:

вбігати – вбігаючи, вибігати – вибігаючи, розбігтися – розбігаючись, тощо.

- Складні слова різних частин мови[534-535], наприклад:

автопробіг, велопробіг, мотопробіг.

І це лише для одного слова run. Цю проблему частково вирішує метод розмічування частин мови β_3, тобто розбір на частини мови або граматичне позначення слова в корпусі (тексті) з врахування суміжних та споріднених слів у реченні [534-535]. Багато слів, особливо загальновживані, можуть бути кількома частинами мови. Наприклад, “забіг” може бути іменником (мій забіг на

марафоні) або дієсловом (я забіг у дім) тощо. Для більшості таких випадків ідентифікація можлива з застосуванням алгоритму Мартіна Портера [534-535] – стемінг β_4 – трансформування слова до її основи через відсікання флексій, префіксів та суфіксів із застосуванням відповідного алгоритму та правил без словника на відмінну від лематизації β_1 (Таблиця А.5) [256-259, 534-535]:

- ІІП за таблицею β_4^1;
- Відсічення закінчень/суфіксів за правилами/деревами закінчень β_4^2;
- Застосування правил лематизації β_4^3;
- Стохастичні алгоритми β_4^4;
- Гібридний підхід з комбінації вищеперерахованих β_4^5;
- Відсічення префіксів β_4^6;
- ІІП відповідності β_4^7.

Вибір конкретного алгоритму залежить від призначення КЛС. Зазвичай для слов'янських мов використовують комбінацію цих алгоритмів. Наприклад, для опрацювання/генерування найскладнішого для української мови слова як дієприкметник та засновування його в необхідній формі згідно змісту речення на основі 5 класів морфем (Таблиця 3.3 – Таблиця 3.4) [404, 882], наприклад:

розпил'+а+н+ий, посіj+а+н+ий, запрограм+ова+н+ий, роздрук+ова+н+ий, змарн+і+л+ий, запізн+і+л+ий:

$$[\textit{префікс}] + \{\textit{корінь} + [\textit{інтерфікс}]\} + [\textit{постсуфікс}] + [\textit{суфікс}] + [\textit{закінчення}]. \quad (3.25)$$

Таблиця 3.3

Класи морфем для дієприкметників в українській мові [404, 882]

Клас	Назва	Приклад
I	Основа	*змарн-, роздрук-, загоj-, заспокоj-, розпил'-* і т. д.;
II	Тематичний елемент	*-и(і,ї)-/-а(я)-/-ол(р)о-*;
III	Постфікс формування доконаного і недоконаного виду	*-ува-(-ювa-)/-овува-/-ну-*, наприклад, *атакувати, воєнізувати, гарантувати, інтенсифікувати, наслідувати, організовувати, організувати, телеграфувати, телефонувати, яровизувати, засохнути, промокнути*;
IV	Суфікс	*-л-, -уч-/-юч-, -ач-/-яч-, -н-, -ен-/-єн-, -т-, -ова-* тощо;
V	Закінчення	*-а, -і, -е, -у/-ю, -ий, -о* і т. д.

Таблиця 3.4

Основні лінгвістичні ознаки множин морфем основ дієприкметників [404, 882]

Клас	Назва	Позначення	Приклад
I	перехідність/ неперехідність	t /$\bar{t}$/$t-\bar{t}$	*малює* типу $(t-\bar{t})$
	вигляд основи	$d/\bar{d}/d-\bar{d}$	для $(d-\bar{d})$ дієслово відносно омонімічне (*автоматизувати, досліджувати*, а також *веліти, вінчати, женити*)
	дієвідміна	I/II	Таблиця А.6
II	необхідність/ можливість тематичного елементу (ТЕ)	$a/i/\tilde{a}/\tilde{\imath}/o$ $/atem$	a – необхідний ТЕ *-а-/-я-*(*розпил+я+н+ий, чит +а+н+ий, пис+а+н+ий, леж+а+чий*); i – необхідний ТЕ *-и(і,ї)-* (*змарн+і+л+ий*); $\tilde{a}$ – ТЕ *-а-/-я-* можливий (*оспів+а+н+ий*, але *оспів+ува+н+ий*); $\tilde{\imath}$ – ТЕ *-и(і,ї)-* можливий, але не необхідний (*запізн+ювa+н+ий, запізн+і+л+ий, запізн+ ен+ий, загоїти – загоювати, змусити – змушувати, запізнитися – запізнюватися, узгодити – узгоджувати, вирішити – вирішувати*); o – можливість утворення паралельних форм дієприкметників для основ дієслів інфінітива на *-ор(л)о-* (*колоти – колотий* і *колений*, *пороти – поротий* і *порений, молоти – молотий* і *мелений*); $atem$ – неможливість ТЕ (*вести – ведений*);
III- I IV	можливість приєднання суфікса до основи	у/ў/н/$\emptyset$	у – приєднання до основи *-ува-/-ювa-* або *-овува-* (*застос+овува+н+ий, будувати*); ў – можливість додати до основи *-ува-/-ювa-* або *-овува-* (*заго+ювa+н+ий* або *загої+ти, застосовувати, досліджувати, розпилювати, зачитувати, спізнюватися*); н – можливість утворення паралельних форм дієприкметників для основ із *-ну-* (*прип+ну+т+ий, усуну(ти) – усунутий* і *усунений*; *кину(ти) – кинутий* і *кинений*; *замкну(ти) – замкнутий* і *замкнений*; *верну(ти) – вернутий* і *вернений*; *стисну(ти) – стиснутий* і *стиснений*; *зігну(ти) – зігнутий* і *зігнений*); $\emptyset$ – неможливість суфікса (*запрягти, пекти, опасти*);
V	можливість або необхідність приєднання −ся	ся /$\overline{\text{ся}}$/ся − $\overline{\text{ся}}$	«ся» – необхідність приєднання −ся (*розчервонітися, зажуритися, усміхнутися, намерзатися, сміятися, втомитися*), «$\overline{\text{ся}}$» – неможливість приєднання −ся (*стогнати*), «ся − $\overline{\text{ся}}$» – можливість з −ся і без $\overline{ся}$ (*купати – купатися*).

Для суфіксів дієприкметників аналізують вид (доконаний/недоконаний = $d/\bar{d}$) [404, 882], час (теперішній/минулий = $pres/past$), стан (активний/пасивний = act/pas) та дієвідміну (I/II/I-II), де (I-II) означає, що даний суфікс може приєднуватися до основ як I, так і II дієвідміни (*опалий, зажурений, намерзлий, розчервонілий, змарнілий, мерзлий, промоклий, засохлий, втомлений, усміхнений*). Для закінчень дієприкметників – форму повну/скорочену = $f/\bar{f}$ (Таблиця 3.5).

Таблиця 3.5

Множина прикладів морфем всіх класів із кортежами лінгвістичних ознак [882]

Клас	Приклад	
I	*втруч*-($\bar{t}$, $\bar{d}$, *I*, *a*, $\emptyset$, *ся*)	*розфарб*-(t, d, *I*, $atem$, у, *ся* − $\overline{ся}$)
	кох-(t, $\bar{d}$, *I*, *a*, $\emptyset$, *ся* − $\overline{ся}$)	*вес*-($t-\bar{t}$, $\bar{d}$, *I*, $atem$, $\emptyset$, *ся* − $\overline{ся}$)
	поділ-(t, d, II, $\tilde{\imath}$, $\emptyset$, *ся* − $\overline{ся}$)	*буд*-($t-\bar{t}$, $d-\bar{d}$, *I*, $atem$, у, *ся* − $\overline{ся}$)
	втрач-(t, $d-\bar{d}$, II, $\tilde{\imath}$, $\emptyset$, $\overline{ся}$)	*побуд*-(t, d, *I*, $atem$, у, *ся* − $\overline{ся}$)
	смій-($\bar{t}$, $\bar{d}$, *I*, $\tilde{a}$, $\emptyset$, *ся*)	*привес*-(t, d, *I*, $atem$, $\emptyset$, *ся* − $\overline{ся}$)
	спит-(t, $\bar{d}$, *I*, *a*, $\emptyset$, *ся* − $\overline{ся}$)	*дослідж*-(t, $d-\bar{d}$, *I*, $\tilde{\imath}$, ў, *ся* − $\overline{ся}$)
	стогн-($\bar{t}$, $\bar{d}$, *I*, $\tilde{a}$, $\emptyset$, $\overline{ся}$)	*автоматиз*-($t-\bar{t}$, $d-\bar{d}$, *I*, $atem$, у, *ся* − $\overline{ся}$)
	усміх-($\bar{t}$, d, *I*, $atem$, н, *ся*)	*фарб*-(t, $\bar{d}$, *I*, $atem$, у, *ся* − $\overline{ся}$)
	запізн-($\bar{t}$, d, *I*, $\tilde{\imath}$, ў, *ся*)	*люб*-(t, $\bar{d}$, II, $\tilde{\imath}$, $\emptyset$, *ся* − $\overline{ся}$)
	мол-(t, d, *I*, *o*, $\emptyset$, *ся* − $\overline{ся}$)	*мал′*-($t-\bar{t}$, $d-\bar{d}$, *I*, $atem$, у, *ся* − $\overline{ся}$)
	змарн-($\bar{t}$, d, *I*, *i*, $\emptyset$, $\overline{ся}$)	*нес*-(t, $\bar{d}$, *I*, $atem$, $\emptyset$, *ся* − $\overline{ся}$)

Клас	Приклад
II	*-и(і)- -а(я)- -у(ю)ва- -овува- -ну-*
III	*-ува-(-юва-)/-овува-/-ну-*
IV	*-л-* (I-II,$act,past,d$) *-т-* (I-II,$pas,pres/past,d/\bar{d}$) *-уч-/-юч-* ($I,act,pres,\bar{d}$) *-е(є)н-* (I-II,$pas,pres/past,d/\bar{d}$) *-ач-/-яч-* (II,$act,pres,\bar{d}$) *-н-* (I-II,$pas,pres/past,d/\bar{d}$) *-ува-(-юва-)/-овува-/-ова-* (I-II,$pas,pres/past,d/\bar{d}$)
V	f: *-ий -ого -ою -ої -им -ими -ому -их* $\bar{f}$: *-а -е -і -у -о*

Тепер наведемо правила формування українських дієприкметників на основі перерахованих морфем будують форми (Таблиця А.6-8) [404, 716, 882]. Особлива складність серед цих правил полягає в ідентифікації/генеруванні пасивних дієприкметників минулого часу від дієслів недоконаного виду без суфіксів [404, 882]. Для одних випадків це можливо (*писаний*, *фарбований*), а для інших – ні (*гублений*). Але існує досить незрозумілих випадків (наприклад, *люблений*, *ведений*, *будований*), коли на утворення таких дієприкметників впливає контекст та семантика, що не можливо описати морфологічними правилами граматики [367-379, 416-430]. Але в більшості випадків можливо згенерувати морфологічні продукційні правила ідентифікації/генерування різних форм українських дієприкметників на основі формальної породжувальної граматики Хомські [404, 882]:

$$G = (V, T, S, P),\ N = V \backslash T, \tag{3.26}$$

де V – алфавіт, T – множина термінальних елементів, $S(S \in V)$ – початковий символ, P – множина продукцій (продукційних правил) типу $X \to Y$, кожна з яких має містити у лівій частині хоча б 1 нетермінальний елемент з N (Таблиця 3.6).

Таблиця 3.6

Нетермінальні символи формальної породжувальної граматики G_0 [404, 882]

№	Символ	Визначення
1	D_K	дієприкметник;
2	$D_K(x, y)$	лексема заданого часу і стану (x, y описано в правилі I);
3	$O'(a_1, a_2, a_3, a_4, a_5, a_6)$	основа дієслова, враховуючи ТЕ/суфікс при необхідності: a_1 – перехідність $(t/\bar{t}/t-\bar{t})$; a_2 – вид $(d/\bar{d}/d-\bar{d})$; a_3 – дієвідміна (I/II); a_4– тематичність $(a/i/\tilde{a}/\tilde{\imath}/o/atem)$; a_5 – суфікс (у/ў/н/$\emptyset$), a_6 – закінчення $-$ся (ся /$\overline{ся}$/*ся* $-$ $\overline{ся}$);
4	$O(a_1, a_2, a_3, a_4, a_5)$	основа (без закінчення $-$ся (ся /$\overline{ся}$/*ся* $-$ $\overline{ся}$));
5	$C(x, y, a_3)$	суфікс з ознаками (x – стан, y – час, a_3 – дієвідмінювання);
6	Ф(и, r, n)	флексія з ознаками: и – форма (повна/скорочена = $f/\bar{f}$), r – категорія роду (чоловічий/жіночий/середній = $m/w/k$), n – число (однина/множина = $s/\bar{s}$);

Продукційні правила ідентифікації/генерування різних форм українських дієприкметників на основі формальної породжувальної граматики Хомські [367-379, 404, 416-430, 882]:

I. Правила підстановки для формування граматичних значень породжуваного дієприкметника в текстовому контенті українською мовою [882].

$$D_K \rightarrow D_K(x, y), \tag{3.27}$$

де $x = (act/pas)$; $y = (pres/past)$, наприклад,

$$D_K \rightarrow D_K(pas, pres),\ D_K \rightarrow D_K(act, pres),\ \ldots \tag{3.28}$$

II. Правила підстановки для генерування граматичних значень відповідними морфемами в текстовому контенті українською мовою [404, 882].

II.1: $D_K(act, pres) \rightarrow O'\left(t, \bar{d}, a_3\right)C(act, pres, a_3)\Phi,$ (3.29)

II.2: $D_K(act, past) \rightarrow O'(\bar{t}, d, a_3)C(act, past, a_3)\Phi,$ (3.30)

II.3: $D_K(pas, pres) \rightarrow O'\left(t, d - \bar{d}, a_3\right)C(pas, pres, a_3)\Phi,$ (3.31)

II.4: $D_K(pas, past) \rightarrow O'\left(t, d - \bar{d}, a_3\right)C(pas, past, a_3)\Phi,$ (3.32)

де O, C, Φ – позначення різних морфем без опису/ідентифікації (Таблиця 3.6).

При формуванні опису відповідних морфем для скорочення опускаються маркування ознак, які в конкретному правилі II.i набувають різних значень, наприклад, $C(act, past)$ – скорочення для 2 виразів $C(act, past, a_3)$, тому правило II.1 фактично складається з багатьох варіантів; $O\left(\bar{d}, a_3\right)$ – скорочення для $O\left(a_1, \bar{d}, a_3, a_4, a_5\right)$, де (a_1, a_3, a_4, a_5) приймають різні допустимі значення.

III. Правила підстановки для розкладання дієслівної основи (відокремлення основи слова і ТЕ/суфікса при їх наявності) в тексті українською мовою [882].

III.1: $O'\left(\overline{atem}\right) \rightarrow O\left(\overline{atem}\right)T,$ (3.33)

де T – ТЕ; $\overline{atem}$ –значення ознаки a_4, відмінне від $atem$, тобто $(a/i/\tilde{a}/\tilde{\imath}/o)$.

III.2: $O'\left(\overline{d}, \overline{\emptyset}\right)C(x, y) \rightarrow O\left(\overline{d}, \overline{\emptyset}\right)C_d C(x, y, I),$ (3.34)

де C_d – суфікс дієслова; $\overline{\emptyset}$ – будь-яке значення ознаки, відмінне від $\emptyset$; x та y повинні задовольняти наступній умові: при $x = pas$ необхідно, щоб $y = pres$.

III.3: $O'(atem) \rightarrow O(atem),$ (3.35)

IV. Правила підстановки для ідентифікації/генеруванні ТЕ відповідної морфеми дієприкметника в текстовому контенті українською мовою [404, 882].

$$\textbf{IV.1:}\ (\tilde{a})T\alpha \to O(\tilde{a})\zeta, \tag{3.36}$$

$$\textbf{IV.2:}\ O(\tilde{\imath})T\alpha \to O(\tilde{\imath})\zeta, \tag{3.37}$$

$$\textbf{IV.3:}\ O(a)T \to O(a)a +, \tag{3.38}$$

$$\textbf{IV.4:}\ O(i)T \to O(i)i +, \tag{3.39}$$

$$\textbf{IV.5:}\ O(o)T \to O(o)o +, \tag{3.40}$$

$$\textbf{IV.6:}\ O\left(\overline{d}, II, a\right)TC(act, pres) \to O\left(\overline{d}, II, a\right)a + C(act, pres), \tag{3.41}$$

$$\textbf{IV.7:}\ O\left(d - \overline{d}, I, a\right)TC(pas, pres) \to O\left(d - \overline{d}, I, a\right)a + C(pas, pres), \tag{3.42}$$

$$\textbf{IV.8:}\ O\left(d - \overline{d}, I, i\right)TC(pas, pres) \to O\left(d - \overline{d}, I, i\right) + C(pas, pres), \tag{3.43}$$

$$\textbf{IV.9:}\ (\tilde{a}, II)T\beta \to O(\tilde{a}, II)a + \xi, \tag{3.44}$$

$$\textbf{IV.10:}\ O(\tilde{\imath}, I)T\beta \to O(\tilde{\imath}, I) + \xi, \tag{3.45}$$

де ζ та ξ – скорочення: ζ – довільна голосна, ξ – довільна приголосна; $+$ – межа між морфемами та з'являється після тих, якими не можуть закінчуватися слова.

V. Правила підстановки для ідентифікації/генеруванні при утворенні дієслів відповідною морфемою в текстовому контенті українською мовою [404, 882].

$$\textbf{V.1:}\ O(\text{І}, y)C_d \to O(\text{І}, y)\text{ува} +, \tag{3.46}$$

$$\textbf{V.2:}\ O(\text{І}, y)C_d \to O(\text{І}, y)\text{ова} +, \tag{3.47}$$

$$\textbf{V.3:}\ O(\tilde{y})C_d \to O(\tilde{y}), \tag{3.48}$$

$$\textbf{V.4:}\ O\left(\overline{t}, d, \text{н}\right)C_d \to O\left(\overline{t}, d, \text{н}\right) + C(pas, past), \tag{3.49}$$

$$\textbf{V.5:}\ O(t, d, \text{н})C_dC(pas, pres) \to O(t, d, \text{н})\text{ну} + C(pas, pres), \tag{3.50}$$

VI. Правила підстановки для ідентифікації/генеруванні суфікса дієприкметника відповідною морфемою в тексті українською мовою [404, 882].

$$\textbf{VI.1:}\ C(act, past, I - II) \to \text{л} +, \tag{3.51}$$

$$\textbf{VI.2:}\ O(atem)C(act, pres, I) \to \text{уч} +, \tag{3.52}$$

$$\textbf{VI.3:}\ O\left(\overline{atem}\right)YC(act, pres, I) \to \text{юч} +, \tag{3.53}$$

$$\textbf{VI.4:}\ O(atem)C(act, pres, II) \to \text{ач} +, \tag{3.54}$$

$$\textbf{VI.5:}\ O\left(\overline{atem}\right)YC(act, pres, II) \to \text{яч} +, \tag{3.55}$$

$$\textbf{VI.6:}\ C(pas, pres/past, I - II) \to \text{н} +, \tag{3.56}$$

VI.7: $C(pas, pres/past, I-II) \rightarrow$ т +, (3.57)

VI.8: $O(atem)C(pas, pres/past, I-II) \rightarrow$ ен +, (3.58)

VI.9: $O(\overline{atem})YC(pas, pres/past, I-II) \rightarrow$ єн +, (3.59)

VI.10: $O(atem)C(pas, pres/past, I-II) \rightarrow$ ува +, (3.60)

VI.11: $O(\overline{atem})YC(pas, pres/past, I-II) \rightarrow$ ювa +, (3.61)

VI.12: $C(pas, pres/past, I-II) \rightarrow$ овува +, (3.62)

VI.13: $O(atem)C(pas, pres/past, I-II) \rightarrow$ ова +, (3.63)

VI.14: $O(\overline{atem})YC(pas, pres/past, I-II) \rightarrow$ йова +, (3.64)

VI.15: $O(\overline{atem})X'C(pas, pres/past, I-II) \rightarrow X'$ьова +, (3.65)

де Y – будь-який суфікс/ТЕ; X' – м'яка приголосна, X – довільної приголосна.

VII. Правила підстановки для вибору форми дієприкметника $(f/\bar{f})$ і реалізації флексії відповідною морфемою в тексті українською мовою [404, 882].

VII.1: $\Phi \rightarrow \Phi(f)$, (3.66)

VII.2: $\Phi(f, s) \rightarrow$ ого, им, ому, (3.67)

VII.3: $\Phi(f, m) \rightarrow$ ий, (3.68)

VII.4: $\Phi(f, w) \rightarrow$ ою, ої, (3.69)

VII.5: $\Phi(f, \bar{s}) \rightarrow$ им, ими, их, (3.70)

VII.6: $\Phi \rightarrow \Phi(\overline{f})$, (3.71)

VII.7: $\Phi(\overline{f}, w) \rightarrow$ а, y, (3.72)

VII.8: $\Phi(\overline{f}, k) \rightarrow$ е, (3.73)

VII.9: $\Phi(\overline{f}, \bar{s}) \rightarrow i$, (3.74)

VII.10: $C(pas)\Phi(\overline{f}) \rightarrow$ о, (3.75)

VIII. Правила підстановки для ідентифікації/генеруванні основи на основі словника відповідною морфемою в тексті українською мовою [404, 882].

VIII.1: $O(t-\bar{t},\ d-\overline{d},\ I, atem,$ у$) \rightarrow$ автоматиз+, буд+, мал′+, …, (3.76)

VIII.2: $O(t-\bar{t},\ \overline{d},\ I, atem,\ \emptyset) \rightarrow$ вес+, …, (3.77)

VIII.3: $O(t,\ d-\overline{d},\ II, \tilde{\imath},\ \emptyset) \rightarrow$ втрач+, …, (3.78)

VIII.4: $O(\bar{t},\ \overline{d},\ I, a,\ \emptyset) \rightarrow$ втруч+, …, (3.79)

VIII.5: $O(t,\ d-\overline{d},\ I, \tilde{\imath},\ \tilde{y}) \rightarrow$ дослідж+, …, (3.80)

VIII.6: $O(\bar{t},\ d,\ I, \tilde{\imath},\ \tilde{y}) \rightarrow$ запізн+, …, (3.81)

VIII.7: $O(t,\ \overline{d},\ I, a,\ \emptyset) \rightarrow$ кох+, …, (3.82)

VIII.8: $O(t,\ \overline{d},\ II, \tilde{\imath},\ \emptyset) \rightarrow$ люб+, …, (3.83)

VIII.9: $O(t,\ \overline{d},\ I, atem,\ \emptyset) \rightarrow$ нес+, …, (3.84)

VIII.10: $O(t,\ d,\ I, atem,$ у$) \rightarrow$ побуд+, розфарб+, …, (3.85)

VIII.11: $O(t,\ d,\ II, \tilde{\imath},\ \emptyset) \rightarrow$ поділ+, …, (3.86)

VIII.12: $O(t,\ d,\ I, atem,\ \emptyset) \rightarrow$ привес+, …, (3.87)

VIII.13: $O(\bar{t},\ \overline{d},\ I, \tilde{a},\ \emptyset) \rightarrow$ сміj+, стогн+, …, (3.88)

VIII.14: $O(t,\ \overline{d},\ I, a,\ \emptyset) \rightarrow$ спит+, …, (3.89)

VIII.15: $O(\bar{t},\ d,\ I, atem,$ н$) \rightarrow$ усміх+, …, (3.90)

VIII.16: $O(t,\ \overline{d},\ I, atem,\ y) \rightarrow$ фарб+, …, (3.91)

VIII.17: $O(t,\ d,\ I,$ о$,\ \emptyset) \rightarrow$ мол+, …, (3.92)

VIII.18: $O(\bar{t},\ d,\ I,$ і$,\ \emptyset) \rightarrow$ змарн+, …, (3.93)

……..

IX. Основні морфонологічні правила підстановки формування або ідентифікації дієприкметника в текстовому контенті українською мовою [882].

IX.1: $\alpha_1 + \rightarrow \alpha_1 + j\alpha_2$, (3.94)

де α_1 та α_2 – довільні голосні.

IX.2: $. j + u \rightarrow i$, (3.95)

де j – позначення звуку [j] (йот).

IX.3: $oZ + C(pas, pres) + \Phi \rightarrow aZ + C(act, pres) + \Phi$, (3.96)

де Z – довільна послідовність не довша за 3 символи (чергування *о/а* в основі типу *скочити/скакати*, *ломити/ламати*, *кроїти/краяти*, *клонити/кланятися*, *котити/катати*, *схопити/хапати*, *гонити/ганяти*, *допомогти/допомагати*); група приголосних за тим символом, що чергується *-о-* (тобто що відокремлює його від ТЕ -а-/-я- перед *-у*(*ю*)*ва-*), не може містити більше 3 літер [404, 882].

IX.4.1: *с′* + *W* → *ш* + *W*, (3.97)

IX.4.2: *в′* + *W* → *вл′* + *W*, (3.98)

IX.4.3: *б′* + *W* → *бл′* + *W*, (3.99)

IX.4.4: *д′* + *W* → *дж′* + *W*, (3.100)

IX.4.5: *т′*+ *W* → *ч* + *W*, (3.101)

……..

де *W* = *-е*(*є*)*н-*, *-у*(*ю*)*ва-*, *-ова-*, *-овува-* [404, 882].

IX.5.1: *д* + *W* → *д′* + *W*, (3.102)

IX.5.2: *с* + *W* → *с′* + *W*, (3.103)

……..

де перед суфіксом *W* тверді кінцеві приголосні атематичних основ пом'якшуються: *принес* + *ти* – *принес′* + *ен*+ *ий* [404, 882].

IX.6: *нн* + Ф → *н* + о. (3.104)

X. Графічно-орфографічні правила підстановки для формування або ідентифікації дієприкметника в текстовому контенті українською мовою [882].

X.1.1: $j + a \rightarrow я$, $ja \rightarrow я$, (3.105)

X.1.2: $j + y \rightarrow ю$, $jy \rightarrow ю$, (3.106)

X.1.3: $j + e \rightarrow є$, $je \rightarrow є$, (3.107)

……..

X.2.1: $X' + a \rightarrow X + я$, (3.108)

X.2.2: $X' + y \rightarrow X + ю$, (3.109)

X.2.3: $X' + u \rightarrow X + i$, (3.110)

X.2.4: $X' + i \rightarrow X +$, (3.111)

X.2.5: $X' + e \rightarrow X + є$, (3.112)

XI. Правила стирання показника межі між морфемами дієприкметника в текстовому контенті українською мовою [404, 882].

$$A + B \rightarrow AB, \quad (3.113)$$

де A і B – будь-які морфеми, що до $A + B$ непридатне жодне з правил груп IX-X.

Таке обмеження на A та B перешкоджає несвоєчасному знищенню межі між морфемами до застосування відповідних морфонологічних правил. Якщо яке-небудь морфонологічне правило може бути застосоване, то воно має бути обов'язково застосоване для перешкоджання утворення нісенітниць, наприклад, **котаючий* від *котити* або**качений* від *катати*. Основні властивості [404, 882]:

1. Транзитивність виводимості. Якщо є послідовність $A_0, A_1, \ldots, A_n$, у якій кожний i-тий ланцюжок безпосередньо *виводиться* з i-1 згідно гіпотетичного силогізму $(p \to q, q \to r \vdash p \to r)$, то A_n виводиться з A_0, а послідовність $A_0, A_1, \ldots, A_n$ є *виведенням* A_n з A_0 [882].

2. Безпосередня виводимість. Якщо є 2 послідовності C та D:

$$C = W_1 F W_2,\ D = W_1 H W_2, \tag{3.114}$$

де W_1 і/або W_2 можуть бути порожніми і в граматиці G є правило $F \to H$, то D безпосередньо виводиться з C [404, 882], наприклад, з послідовності за правилом VI.3 $O(t,\ \bar{d},\ I,\ a,\ \emptyset,$ ся $- \overline{\text{ся}})a$ + $C(I,act,pres,\bar{d})$ + Ф безпосередньо виводиться послідовність $O(t,\ \bar{d},\ I,\ a,\ \emptyset,$ *ся* $- \overline{ся})a$ + *юч* + Ф.

XII. Правила маркування лексеми як дієприкметник з множиною ознак у відповідному реченні/фрази (рід, час, відмінок тощо) в українському тексті [882].

$$A'_{x,y,z} \to word_{x,y,z}. \tag{3.115}$$

Приклад повного виведення (генерування/ідентифікації дієприкметника D_K) в текстовому контенті українською мовою [404, 882]:

(I) $D_K(pas, pres)$

(II.3) $O'(t, d, I, atem, y,$ *ся* $- \overline{ся})C(pas, pres, I)Ф$

(III.3) $O(t, d, I, atem, y)C(pas, pres, I)Ф$

(VI.13) $O(t, d, I, atem, y)ова + Ф$

(VII.1) $O(t, d, I, atem, y)ова + Ф(f)$

(VII.3) $O(t, d, I, atem, y)ова + ий$

(VIII.10) $розфарб + ова + ий$

(XI) *розфарбований*

(XII) $A'_{x,y,z} \to розфарбований_{x,y,z}$

Приклад тупикового виведення дієприкметника D_K в українському тексті:

(I) $D_K(act, past)$

(II.2) $O'\left(\overline{t}, d, II, atem, \emptyset, ся - \overline{ся}\right)C(act, past, II)Ф$

(Ш.3) $O\left(\overline{t}, d, II, atem, \emptyset\right)C(act, past, II)Ф$

(VI.1) $O\left(\overline{t}, d, II, atem, \emptyset\right)л + Ф$

(VII.1) $O\left(\overline{t}, d, II, atem, \emptyset\right)л + Ф(f)$

(VII.3) $O\left(\overline{t}, d, II, atem, \emptyset\right)л + ий$

(XI) $O\left(\overline{t}, d, II, atem, \emptyset\right)лий$

(XII) $A'_{x,y,z} \rightarrow лий_{x,y,z}$

Цей ланцюг для української мови неможна далі генерувати [882].

3.2.4. Лексичний аналіз української мови

ЛА – попереднє опрацювання тексту чи мовлення [407, 567], трансформування ланцюга символів в послідовність токенів (кортежів символів за відповідними шаблонами) [402-407, 982]:

$$C'_\gamma = \gamma(C_\gamma, D_\gamma, R_\gamma, C_\beta), \tag{3.116}$$

де $C'_\gamma \subseteq C_\gamma$, $C'_\gamma = \gamma_2 \circ \gamma_1$ (достатньо для трансформування звукового ряду мовлення в друкований текст) або $C'_\gamma = \gamma_5 \circ \gamma_4 \circ \gamma_3$ (для трансформування відсканованого тексту в мовлення), але для більшості випадків достатньо:

$$C'_\gamma = \gamma_5 \circ \gamma_4. \tag{3.117}$$

I. Сегментація мовлення γ_1 (англ. Speech Segmentation) – поділ звукового потоку людського мовлення на окремі слова [405, 982]. У розмові або промові людей майже не ідентифікуються паузи між послідовними словами, тому сегментація є необхідною задачею розпізнавання мови. У більшості розмовних мов звуки як послідовні літери поєднуються один з одним у процесі коартикуляції [405, 982], тому перетворення аналогового сигналу в дискретні символи є досить складним NLP-процесом в КЛС системах з технічної реалізації.

II. Розпізнавання мовлення γ_2 (англ. Speech Recognition, SR) [85, 644] або мовлення-у-текст (англ. Speech-To-Text, STT) – трансформування мовленнєвого сигналу в е-текст [359, 405, 982]. Різні люди вимовляють слова в кожній мові з різними наголосами, швидкістю та інтонаціями. КЛС має розпізнавати широкий спектр вхідних неструктурованих даних як ідентичні до одного еквівалента [405, 982].

III. Оптичне розпізнавання символів γ_3 (англ. Optical Character Recognition, OCR) [534-535, 716, 862] – переведення відсканованого рукописного або друкованого тексту після ГА та МА в послідовність кодів для е-подання з виправленням простих помилок на основі статистики та теорії ймовірності використання послідовності N-грам/слів/закінчень замість незрозумілих символів [535].

IV. Токенізація або сегментація слів γ_4 (англ. Tokenization, Word Segmentation) [110] – це розмежування та категоризація секцій ланцюга вхідних символів для МА [405, 534-535, 982]. Для англійської або української мов це досить тривіальна ситуація, оскільки слова зазвичай розмежовані пробілами (проблеми лише при наявності стилістичних та граматичних помилок в тексті). Однак деякі письмові мови як китайська, корейська, японська та тайська не позначають так межі слів. Тоді токенізація є важливою задачею на основі словникового запасу та морфології слів. Іноді метод застосовують для формування мішка слів (англ. Bag-of-words model, BOW) при Data Mining [13-15, 19-23, 68].

V. Текст-у-мовлення γ_5 (англ. Text-To-Speech, TTS) [110] – трансформування рукописного, машинописного або друкованого тексту на мовленнєвий сигнал для усного подання, наприклад, для людей з вадами зору [405, 982].

3.2.5. Синтаксичний аналіз української мови

СА є основою семантичного аналізу:

$$C'_\lambda = \delta(C_\delta, D_\delta, R_\delta), \tag{3.118}$$

де $C'_\lambda \subseteq C_\lambda$, $C'_\lambda = \delta_3 \circ \delta_2 \circ \delta_1$:

I. Індукція граматики δ_1 (Grammar induction) – генерування формальної граматики для опису синтаксису мови [19-23, 160, 416-430, 958-983].

II. Неоднозначність меж або порушення речення δ_2 (англ. Sentence Breaking, Sentence Boundary Disambiguation) – аналіз присутності/ відсутності відповідних розділових знаків та тексту між ними (крапка не лише позначає закінчення речення, але й скорочення) [19-23, 160, 958-983].

III. Парсинг δ_3 (англ. Parsing) – генерування з вхідної послідовності символів дерева розбору (граматичний аналіз) речення для аналізу граматичної структури згідно із заданою формальною граматикою (Таблиця 3.7) [112-114, 140-147, 157-170, 295, 471-479, 958-983]. Для типового речення існують сотні-тисячі аналізів, більшість з яких э абсолютно безглуздими для носія мови. Існує два основних типи парсингу: залежностей δ_3^1 (англ. Dependency Parsing) та складових δ_3^2 (англ. Constituency Parsing), або у вигляді деякого поєднання цих способів δ_3^3 [112-114, 140-147, 295, 416-430, 471-479, 958-983]. Парсинг залежностей фокусується на взаємозв'язку між словами в реченні (первинні об'єкти та предикати), а парсинг складових – побудові дерева СА з застосуванням імовірнісної контекстно-вільної (стохастичної) граматики (англ. Probabilistic Context-Free Grammar, PCFG) [112-114, 140-147, 295, 416-430, 471-479, 534-535, 716, 862, 958-983].

Наприклад, при синтаксичному розборі/генеруванні речень/фраз вибір флексії відмінка конкретного слова в українській мові прямо залежить від типу основи та частини мови, зокрема, для іменникових груп з врахуванням контексту (Таблиця 3.7, Таблиця А.9 – Таблиця А.10) [534-535, 716, 862, 958-983]:

$$R \to E_1 X \to E_1 C_1 C_2,\ R \to E_2 X \to E_2 C_3 C_4, \tag{3.119}$$

Множина лінгвістичних одиниць X одного типу поряд з множиною лінгвістичних одиниць E_1 іншого типу трансформуються іншим способом $C_1 C_2$, ніж поряд з множиною лінгвістичних одиниць E_2 третього типу – $C_3 C_4$.

Без врахування контексту треба вводити більш дробову класифікацію:

$$R \to X_1 \to C_1 C_2,\ R \to X_2 \to C_3 C_4, \tag{3.120}$$

Таблиця 3.7

Правила формулювання україномовних фраз [212, 219, 709-710]

№	Назва правила	Правило
I.	Вибір структури R	$R \to \#\tilde{S}_{x,y,н,w}\ \tilde{V}_{y,тепер,w}\#.$
II.	Іменна група	1) $\tilde{V}_{x,y,z,з} \to \tilde{S}_{x,y,z,з}\ \tilde{S}_{x',y',р,w}$; 2) $\tilde{S}_{x,y,z,з} \to A_{x,y,z}\ \tilde{S}_{x,y,z,з}$; 3) $K_1\tilde{S}_{x,y,z,w}\ K_2 \to K_1 S^{займ}_{x,y,z,w} K_2$, $K_1 \neq A_{x,y,z}$, $K_2 \neq \tilde{S}_{z'}$; 4) $\tilde{S}_{x,y,z,з} \to S_{x,y,z}$.
III.	Дієслівна група	1) $\tilde{V}_{y,тепер,w} \to V_{y,тепер,w}\ \tilde{S}_{x',y',зн,w'}\ \tilde{S}_{x'',y'',ор,w''}$; 2) $\tilde{V}_{y,тепер,w} \to V_{y,тепер,w}\ \tilde{S}_{x',y',ор,w'}\ \tilde{S}_{x'',y'',зн,w''}$; 3) $\tilde{V}_{y,тепер,w} \to V_{y,тепер,w}\ \tilde{S}_{x',y',зн,w'}$; 4) $\tilde{V}_{y,тепер,w} \to V_{y,тепер,w}\ \tilde{S}_{x',y',ор,w'}$.
IV.	Підстановка слів	1) $S_{ч,y,z} \to син_{y,z},\ldots$; 2) $S_{ж,y,z} \to посмішка_{y,z},\ldots$; 3) $S_{сер,y,z} \to щастя_{y,z},\ldots$ 4) $S^{займ}_{х,од,z,1} \to я_z$; 5) $S^{займ}_{x,од,z,2} \to ти_z$; 6) $A_{x,y,z} \to веселий_{x,y,z}, безмежний_{x,y,z}, мій_{x,y,z}, твій_{x,y,z},\ldots$; 7) $V_{y,тепер,w} \to наповнити_{y,тепер,w},\ldots$

Кожний запис є множиною правил, наприклад, II.1 формує 648 правил [212, 219, 709-710]: $\tilde{S}_{ч,од,н,з} \to \tilde{S}_{ч,од,н,з}\ \tilde{S}_{ч,од,р,1}$; $\tilde{S}_{ч,од,р,з} \to \tilde{S}_{ч,од,р,з}\ \tilde{S}_{ч,од,р,1}$; …; $\tilde{S}_{сер,мн,м,з} \to \tilde{S}_{сер,мн,м,з}\ \tilde{S}_{сер,мн,род,з}$ (Таблиця 3.8). Для генерування дерева речення українською застосовують відповідні правила узгодження флексій (Рис. 3.6-Рис. 3.7) [212, 219, 709-710, 958-983].

Таблиця 3.8

Позначення граматичних категорій іменної/дієслівної групи та складових [212]

Тип	Опис
Імення група	
Іменна група/$\tilde{N}$	прикметник/*А*, іменник/*N*, займенник/$N^{займ}$;
Число/*ЧЛ*	однина/*од*, множина/*мн*;
Рід/*РД*	чоловічий/*ч*, жіночий/*ж*, середній/*с*;
Відмінок/*ВД*	називний/*н*, родовий/*р*, давальний/*д*, знахідний/*з*, орудний/*о*, місцевий/*м*, кличний/*к*;
Особа/ОС	1-ша/*1*, 2-га/*2*, 3-тя/*3*.
Дієслівна група	
Дієслівна група/$\tilde{R}$	дієслово/*R*, в межах іменної групи прикметник/*А*, іменник/*N*;
Число/*ЧЛ*	однина/*од*, множина/*мн*;
Рід/*РД*	чоловічий/*ч*, жіночий/*ж*, середній/*с*;
Особа/*ОС*	1-ша/*1*, 2-га/*2*, 3-тя/*3*;
Час/*ЧС*	теперішній/*тп*, минулий/*мн*, майбутній/*мб*.

Кожен крок є розгортанням символу послідовності (наприклад, $\tilde{R}_{од,тп,3} - R_{од,тп,3}\,\check{S}_{ч,од,з,1}\,\check{S}_{с,од,о,3}$) або заміною (так, $\check{S}_{ч,од,з,1} - S^{займ}_{ч,од,з,1}$). Для такого розгортання треба формувати детальніші типи лінгвістичних одиниць для врахування місцезнаходження в контексті речення [212, 219, 709-710, 958-983], наприклад:

$$Word_{мн,род} \to O^i F_{мн,род},\ Word_{мн,род} \to O^i F^i_{мн,род}, \quad (3.121)$$

де $Word$ – словоформа реченні, 0^i – основа слова типу i $(i = 1,2,3,\dots)$, $F_{мн,род}$ – флексія родового відмінку множини в українські мові, наприклад:

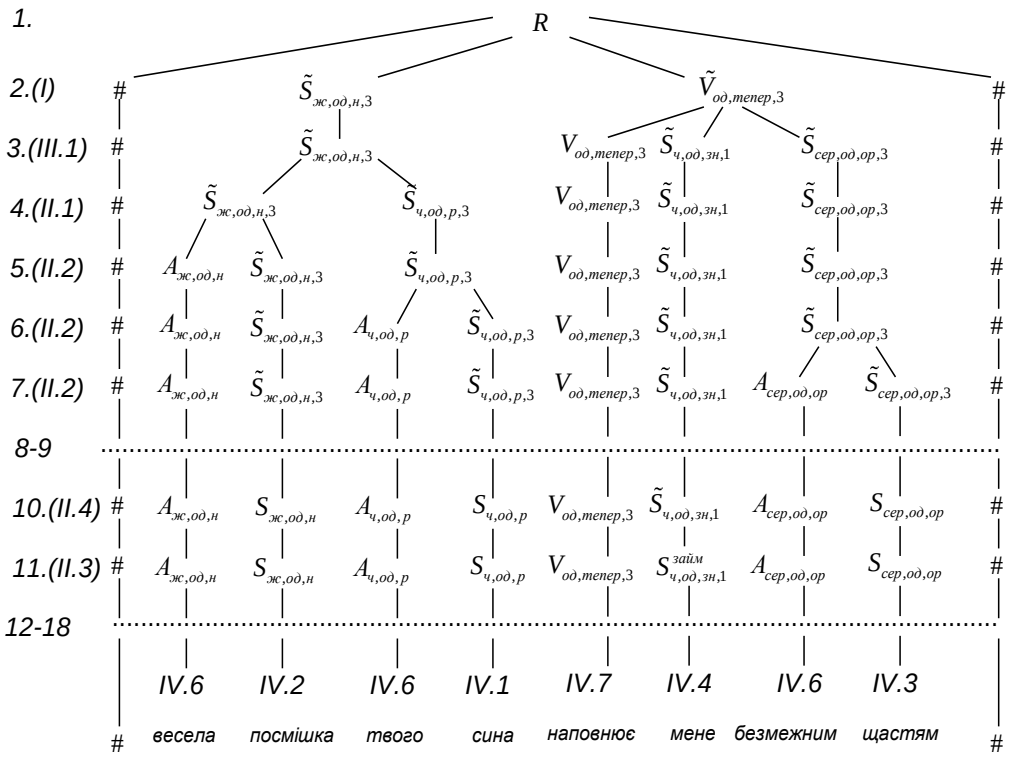


Рис. 3.6. Приклад граматики із фразовою структурою

1. S
2. (I) $\#\tilde{S}_{ж,од,н,3}\ \tilde{R}_{од,тп,3}\#$
3. (III.1) $\#\tilde{S}_{ж,од,н,3}\ R_{од,тп,3}\tilde{S}_{ч,од,з,1}\tilde{S}_{с,од,о,3}\#$
4. (II.1) $\#\tilde{S}_{ж,од,н,3}\ \tilde{S}_{ч,од,р,3}R_{од,тп,3}\tilde{S}_{ч,од,з,1}\tilde{S}_{с,од,о,3}\#$
5. (II.2) $\#A_{ж,од,н}\tilde{S}_{ж,од,н,3}\ \tilde{S}_{ч,од,р,3}R_{од,тп,3}\tilde{S}_{ч,од,з,1}\tilde{S}_{с,од,о,3}\#$
6. (II.2) $\#A_{ж,од,н}\tilde{S}_{ж,од,н,3}\ A_{ч,од,р}\tilde{S}_{ч,од,р,3}R_{од,тп,3}\tilde{S}_{ч,од,з,1}\tilde{S}_{с,од,о,3}\#$
7. (II.2) $\#A_{ж,од,н}\tilde{S}_{ж,од,н,3}\ A_{ч,од,р}\tilde{S}_{ч,од,р,3}R_{од,тп,3}\tilde{S}_{ч,од,з,1}A_{с,од,о}\tilde{S}_{с,од,о,3}\#$
8. (II.4) $\#A_{ж,од,н}S_{ж,од,н}\ A_{ч,од,р}\tilde{S}_{ч,од,р,3}R_{од,тп,3}\tilde{S}_{ч,од,з,1}A_{с,од,о}\tilde{S}_{с,од,о,3}\#$
9. (II.) $\#A_{ж,од,н}S_{ж,од,н}\ A_{ч,од,р}S_{ч,од,р}R_{од,тп,3}\tilde{S}_{ч,од,з,1}A_{с,од,о}\tilde{S}_{с,од,о,3}\#$
10. (II.4) $\#A_{ж,од,н}S_{ж,од,н}\ A_{ч,од,р}S_{ч,од,р}R_{од,тп,3}\tilde{S}_{ч,од,з,1}A_{с,од,о}S_{с,од,о}\#$
11. (II.3) $\#A_{ж,од,н}S_{ж,од,н}\ A_{ч,од,р}S_{ч,од,р}R_{од,тп,3}S^{займ}_{ч,од,з,1}A_{с,од,о}S_{с,од,о}\#$
12. (IV.6) $\#весела\ S_{ж,од,н}\ A_{ч,од,р}S_{ч,од,р}R_{од,тп,3}S^{займ}_{ч,од,з,1}A_{с,од,о}S_{с,од,о}\#$
13. (IV.2) $\#весела\ посмішка\ A_{ч,од,р}S_{ч,од,р}R_{од,тп,3}S^{займ}_{ч,од,з,1}A_{с,од,о}S_{с,од,о}\#$
14. (IV.6) $\#весела\ посмішка\ твого\ S_{ч,од,р}R_{од,тп,3}S^{займ}_{ч,од,з,1}A_{с,од,о}S_{с,од,о}\#$
15. (IV.1) $\#весела\ посмішка\ твого\ сина\ R_{од,тп,3}S^{займ}_{ч,од,з,1}A_{с,од,о}S_{с,од,о}\#$
16. (IV.7) $\#весела\ посмішка\ твого\ сина\ наповнює\ S^{займ}_{ч,од,з,1}A_{с,од,о}S_{с,од,о}\#$
17. (IV.6) $\#весела\ посмішка\ твого\ сина\ наповнює\ мене\ A_{с,од,о}S_{с,од,о}\#$
18. (IV.6) $\#весела\ посмішка\ твого\ сина\ наповнює\ мене\ безмежним\ S_{с,од,о}\#$
19. (IV.3) $\#весела\ посмішка\ твого\ сина\ наповнює\ мене\ безмежним\ щастям\#$

Рис. 3.7. Приклад виведення речення заданої структурної схеми

$$O^1F_{\text{мн,род}} \to O^i\text{ів}(\text{друз} - \text{ів}),\ F^1_{мн,род} \to ів, \qquad (3.122)$$

$$O^2F_{\text{мн,род}} \to O^i\text{ок}(\text{ігра}\text{ш} - \text{ок}),\ F^2_{мн,род} \to ок, \quad (3.123)$$

$$O^3F_{\text{мн,род}} \to O^i\text{ей}(\text{діт} - \text{ей}),\ F^3_{мн,род} \to ей, \quad (3.124)$$

$$O^4F_{\text{мн,род}} \to O^i\text{их}(\text{знайом} - \text{их}),\ F^4_{мн,род} \to их, \quad (3.125)$$

$$O^5F_{\text{мн,род}} \to O^i(\text{машин} -),\ F^5_{мн,род} \to \Lambda. \quad (3.126)$$

……

Вибір відмінку прямого доповнення в словоформі або реченні залежить в українські мові від присутності/відсутності заперечення [212, 219, 709-710]:

$$\tilde{V} \to V^i\tilde{S}_d,\ \tilde{V} \to V^i\tilde{S}^1_d, \quad (3.127)$$

$$\tilde{V} \to \neg V^i\tilde{S}_d,\ \tilde{V} \to \neg V^i\tilde{S}^2_d. \quad (3.128)$$

де $\tilde{V}$ – група дієслова в реченні, V^i – перехідне дієслово в групі дієслова, $\tilde{S}_d$ – пряме доповнення, $\tilde{S}$ – група іменника, $\neg$ – заперечення [212, 219, 709-710], зокрема для речень *школяр пише есе* та *школяр не пише есе* відповідні виведення:

$$X V\tilde{S}_d \underset{X\neq\neg x}{\to} XV^i\tilde{S}_з,\ \tilde{S}^1_d \to \tilde{S}_з \text{ або відповідно } X\neg V\tilde{S}_d \underset{X\neq\neg x}{\to} X\neg V^i\tilde{S}_p,\ \tilde{S}^2_d \to \tilde{S}_p.$$

Використання орудного суб’єкта $\tilde{S}^{sb}$ при віддієслівному іменнику залежить від наявності об’єкту $\tilde{S}^{ob}$ (*аналіз змісту системою*) [212, 219, 709-710, 958-983]:

$$\tilde{S} \to \tilde{S}'\tilde{S}^{ob}\tilde{S}^{sb},\ \tilde{S} \to \tilde{S}'\tilde{S}^{ob}\tilde{S}^{sb^1}, \quad (3.129)$$

$$\tilde{S} \to \tilde{S}'\tilde{S}^{sb},\ \tilde{S} \to \tilde{S}'\tilde{S}^{sb^2}, \quad (3.130)$$

$$\tilde{S}^{ob}\tilde{S}^{sb} \to \tilde{S}^{ob}\tilde{S}_o,\ \tilde{S}^{sb^1} \to \tilde{S}'_o. \quad (3.131)$$

Від семантики контексту при коректній ідентифікації та виправленні граматичної та стилістичної помилки повністю відмовитися неможливо, необхідно враховувати не лише один символом в лівій частині правил (3.30)-(3.40), яке забезпечує перестановку символів [212, 219, 709-710, 958-983].

Семантичні ознаки досліджують за допомогою множини лексичних та лінгвістичних ресурсів у вигляді словників та бібліотек D, інструментами управління словниками T_D, семантичного маркування ролей λ^3_2 та процесу вкладання слів λ_3 (англ. word embeddings) [184, 646]. Вкладання слів полягає у відображенні слів, словосполучень або фраз із словника D у вектори дійсних чисел E_D для зручності опрацювання, наприклад, на основі Word2Vec [200].

$$S_f = \lambda_2^3(\lambda_3(D, T_D), E_D), \tag{3.132}$$

Врахування семантики значно спрощує граматичне дерево (Рис. 3.8) [212, 219, 709-710, 958-983]. Якщо при розгортанні, замінювані або переписуванні символи (*предки*) з'єднаємо безпосередньо з кінцевими результатами (*нащадками*), отримаємо дерево складових, або синтаксичну структуру. Задані правила (Таблиця 3.7) здатні породжувати і інші не обов'язково змістовні фрази, так як правила II.1 і II.2 є циклічними. Поряд з послідовністю *весела посмішка* можна отримати *весела весела посмішка*, тощо. Кількість фраз в природній мові має бути скінчена [212, 219, 709-710, 958-983]. При правильній побудові дерева розбору речення (Рис. 3.6-Рис. 3.8) та подальшому скороченні (Рис. 3.9) можна провести узгодження відмінків [212, 219, 709-710, 958-983].

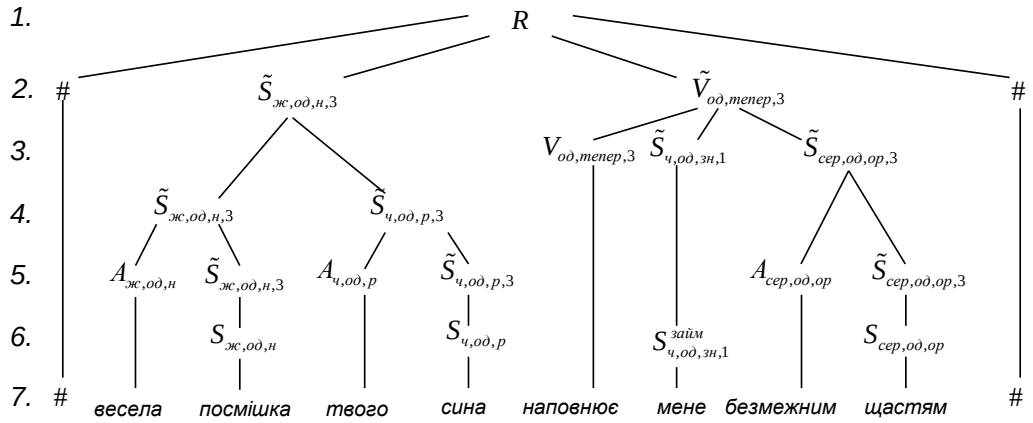


Рис. 3.8. Приклад дерева складових для контекстно-залежної граматики

Узгодження відмінків між лінгвістичними одиницями речення впливає на подальший семантичний аналіз тексту. Наприклад, в українській мові можливо генерувати послідовності лінгвістичних одиниць типу $x_1x_2x_3qx_1'x_2'x_3'$, $x_2x_2x_3qx_2'x_1'x_2'x_3'$, $x_1x_3x_2x_1qx_1'x_3'x_2'x_1'$ тощо (або як XqX') [212, 219, 709-710, 958-983]: ^{a}Саша, ^{b}Софія, ^{c}Катя, ^{d}Данило, … – $^{a'}$спортсмен, $^{b'}$співачка, $^{c'}$художниця,^{d}поет, … *відповідно*. Зокрема, $x \rightarrow (abcd\ldots)$ – послідовність власних імен; $x' \rightarrow (a'b'c'd'\ldots)$ – послідовність професій, узгоджені з власними іменами в роді; q – знак пунктуації.

$$\left.\begin{array}{l}1.\ R \to RY_i x_i^{'},\\ 2.\ x_i^{'}Y_j \to Y_j x_i^{'},\\ 3.\ RY_i \to x_i R,\\ 4.\ R \to q.\end{array}\right\} i,j = 1,2,3, \tag{3.133}$$

де x_i, x_i', q – основні лінгвістичні одиниці; R, Y_i – допоміжні лінгвістичні одиниці; R – початковий символ як індикатор типу генерування ланцюга.

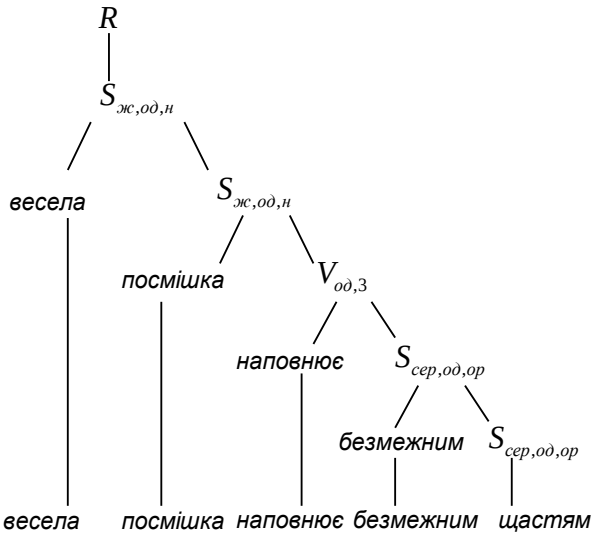


Рис. 3.9. Приклад регулярної граматики (тип 3)

Для ефективнішого дослідження помилок використовують аналіз мета-даних лінгвістичних ознак та характеристик оригіналу тексту [212, 219, 709-710], зокрема, жанр контенту, наявність/відсутність діалекту, суржика, сленгу, термінології та ймовірність написання носієм мови або результату перекладу.

3.2.6. Семантичний аналіз української мови

Семантичний аналіз на сьогодення використовують не у більшості КЛС, але з поступовим впровадженням ШІ в повсякденність пересічної людини ця задача має бути вирішена та спрощена (Рис. 3.10) [567]. Чим складніша граматика мови, тим складніше провести СЕМ [19-23, 219, 289, 535]:

$$C'_{\mu} = \lambda(C_{\lambda}, D_{\lambda}, R_{\lambda}), \tag{3.134}$$

де $C'_{\mu} \subseteq C_{\mu}$, $C'_{\mu} = \lambda_2 \circ \lambda_1$.

Лінгвістична семантика λ_1 (окремих слів у контексті) полягає у:

$$C''_{\mu} = \lambda_1^6 \circ \lambda_1^5 \circ \lambda_1^4 \circ \lambda_1^3 \circ \lambda_1^2 \circ \lambda_1^1,\ C''_{\mu} \subseteq C'_{\mu}. \tag{3.135}$$

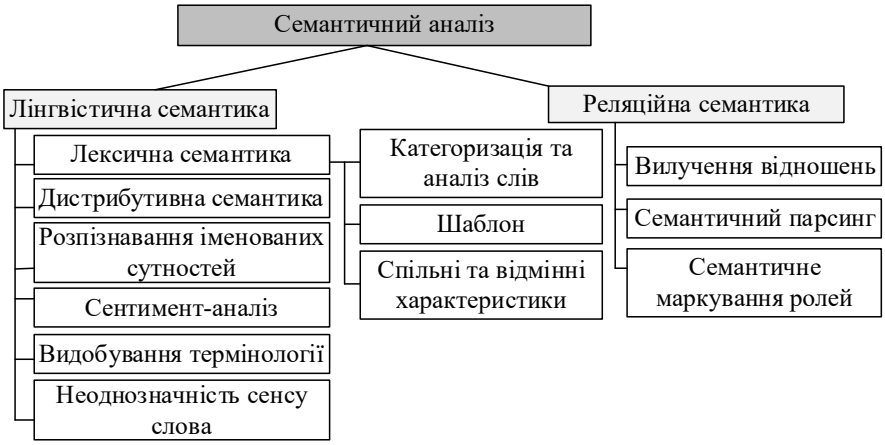


Рис. 3.10. Класифікація основних методів семантичного аналізу тексту

I. Лексична семантика λ_1^1 (англ. Lexical Semantics) – аналіз значень окремих лексичних елементів слів/лексем/морфем [19-23, 219, 153, 165, 184, 535], на відмінну від семантики речень, для конструювання семантичної мережі на основі [265, 289, 353, 433-441, 489, 558, 725-802, 823-848, 906-926]:
   1) Аналізу та категоризації слів (Таблиця А.9) [534-535, 716, 862];
   2) Генерування множин відмінних/спільних характеристик у лексикосемантичних схемах різних мов (Таблиця А.10) [534-535];
   3) Формування шаблону – відношення сенсу лексичних/ фразеологічних одиниць (зміст та лексика) на основі синтаксису до змісту конкретного речення, аналізу паронімів, омонімів службових і знаменних слів, гіпонімії, гіперонімії, антонімії, синонімії [567].

II. Дистрибутивна семантика λ_1^2 (англ. Distributional Semantics) – розрахунок ступені семантичного наближення між лінгвістичними одиницями на основі їх розподілу (дистрибуції) у великих масивах лінгвістичних даних (текстових корпусів) [19-23, 219, 153, 165, 184-188, 534-535, 567, 740-747].

III. Неоднозначність сенсу слова λ_1^3 (англ. Word-Sense Disambiguation, WSD) – визначення конкретного сенсу слова із множини ймовірносних в конкретному реченні для покращення результатів роботи КЛС [633], наприклад, при аналізі дискурсу, підвищення релевантності ІІПС, визначення когерентності (цілісності) тексту (англ. Coherence), анафори

вірша (англ. Anaphora Resolution) та підсумковим виведенням (англ. Inference) [19-23, 219, 153, 165-167, 184, 534-539, 567, 633].

IV. Розпізнавання іменованих сутностей λ_1^4 (англ. Named-Entity Recognition, NER) [523] або ідентифікація об'єктної сутності, фрагментація об'єктної сутності, видобуток об'єктної сутності – видобування з неструктурованого тексту інформації щодо наявності певних іменованих сутностей відповідних категорій як дати, відсотки, грошові значення, кількості, час, географічні місця, імена організацій або людей, тощо [19-23, 219, 153, 165, 184, 534-535, 567]. Наявність великої літери на початку слова не вирішує цю проблему – початок речення або віршованого рядка також починається з великої літери. В німецькій мові всі іменники починаються з великої літери. Крім того іменовані сутності часто охоплюють кілька слів, лише деякі з них пишуться з великої літери. У французькій, українській та іспанській мовах не використовують великі літери у прикметникових іменах. В німецькі мові всі іменники пишуться з великої літери. Китайська, корейська, японська чи арабська взагалі не мають великих літер.

V. Видобування термінології λ_1^5 (англ. Terminology Mining, Terminology Extraction, Term Recognition, Glossary Extraction, Term Extraction) – це автоматичне вилучення відповідних термінів із відповідного корпусу [114, 160, 402, 815, 959]. Одним із перших кроків до моделювання домену знань є збір словникового запасу термінів доменів як лінгвістичних ознак змісту тексту [567]. Підходи до видобування термінів застосовують лінгвістичні процесори (позначення частини мови, фрагментація частин текстів) для вилучення термінологічних кандидатів [567], тобто синтаксично правдоподібних термінологічних іменних фраз або ключових слів, наприклад для рубрикації [19-23, 219, 153, 165, 184, 534-535, 567].

Аналіз тональності тексту, мультимодальний аналіз настроїв або сентимент-аналіз λ_1^6 (англ. Opinion Mining, Sentiment Analysis) – це виявлення суб'єктивної емоційно-забарвленої множини контенту [285, 635]

(позитивного, негативного або нейтрального) в текстовому масиві даних конкретного автора по відношенню до відповідного об'єкта, суб'єкта, події або явища тематичної ПО [19-23, 219, 153, 165, 184, 567]. Це корисно для виявлення тенденцій громадської думки для онлайн маркетингу, в соціальних мережах, формуванні політичної пропаганди тощо [106, 177, 266, 508, 523-524, 537, 543, 624, 631].

Реляційна семантика λ_2 (семантика окремих речень) полягає у:

$$C'''_{\mu} = \lambda_2^3 \circ \lambda_2^2 \circ \lambda_2^1, C'''_{\mu} \subseteq C'_{\mu}. \quad (3.136)$$

I. Вилучення відношень λ_2^1 (англ. Relationship Extraction) – ідентифікація взаємозв'язків іменних сутностей (наприклад, родинних зав'язків, колег, ворогів тощо) [19-23, 219, 153, 165, 184, 534-535, 567].

II. Семантичний парсинг λ_2^2 (англ. Semantic Parsing) – подання семантики частини тексту (зазвичай речення) у вигляді логічного формалізму (DRT парсинг, Discourse Representation Theory) або графу (AMR парсингу, Abstract Meaning Representation) [19-23, 219, 153, 165, 184, 567], наприклад:

$$\exists a, b, e: f(a, \text{хотіти\_01}) \wedge f(e, \text{гуляти\_01}) \wedge f(b, \text{дитина}) \wedge$$
$$\wedge g_0(a, b) \wedge g_1(a, e) \wedge g_0(e, b). \quad (3.137)$$

AMR формат: (a / хотіти-01 : arg0 (b / дитина) : arg1 (e / гуляти-01 : arg0 b)).

III. Семантичне маркування ролей λ_2^3 (англ. Semantic Role Labelling, Implicit Semantic Role Labelling Below) – призначення словам або словосполученням міток семантичної ролі у реченні, наприклад, роль мети, агента, або результату [19-23, 219, 153, 165, 184, 567] за алгоритмом:

$$C''''_{\mu} = \lambda_2^{35} \circ \lambda_2^{34} \circ \lambda_2^{33} \circ \lambda_2^{32} \circ \lambda_2^{31}, C''''_{\mu} \subseteq C'''_{\mu}. \quad (3.138)$$

1) Виокремлення речення або фрагмента тексту λ_2^{31}.
2) Визначення в реченні семантичних предикатів λ_2^{32} (дієслівні та іменникові фрейми).
3) Роз'єднання визначених семантичних предикатів λ_2^{33}.
4) Ідентифікація елементів визначеного фрейму λ_2^{34}.

5) Класифікація ідентифікованих елементів фрейму λ_2^{35}– призначення семантичної ролі в реченні.

3.2.7. Прагматичний аналіз української мови

ПА застосовують для визначення структури тексту з врахуванням контексту речень при формуванні абзаців, розділів та діалогів [542-546]. Основна задача – ідентифікація контексту таких лінгвістичних одиниць як речення та формування семантичного взаємозв'язку цих лінгвістичних елементів [547-551]. ПА є обов'язковим компонентом в СШІ для інтерпретації людського діалогу, мовлення та відповідного його аналізу (Рис. 3.11) [542-551].

$$Y = \mu(C_\mu, D_\mu, R_\mu), Y = \mu_2 \circ \mu_1. \quad (3.139)$$

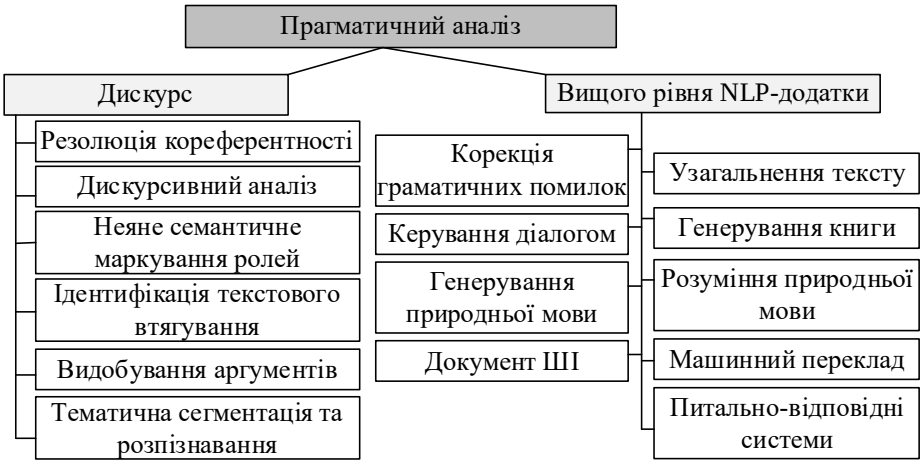


Рис. 3.11. Класифікація основних методів прагматичного аналізу тексту

Дискурс μ_1 (семантика поза окремими реченнями) полягає у:

$$Y' = \mu_1^6 \circ \mu_1^5 \circ \mu_1^4 \circ \mu_1^3 \circ \mu_1^2 \circ \mu_1^1, Y' \subseteq Y. \quad (3.140)$$

I. Резолюція кореферентності μ_1^1 (англ. Coreference Resolution) – ідентифікація в наступному тексті наступних слів-згадок або виразів про об'єкти, суб'єкті, явища та події, які ідентифіковані в попередньому масиві тексту [542-551]. Резолюція анофори є відповідним прикладом цієї задачі (співставлення займенників в подальшому тексті із іменниками або іменами з попереднього фрагменту тексту) [542-551]. Іншою задачею є ідентифікація мостових відношень – виразів, які мають відношення до певних об'єктів, суб'єктів, явищ або подій [542-550], наприклад, в тексті

«вона передала сердечні вітання від онучки Софії бабусі» вираз «сердечні вітання» є виразним посиланням та мостовим відношенням (належать конкретному суб'єкту та призначені конкретному іншому суб'єкту відношень у відповідному тексті) [542-550].

II. Дискурсивний аналіз μ_1^2 (англ. Discourse Analysis) полягає у етапах:

1) Дискурсивний парсинг μ_1^{21} (англ. Discourse Parsing) – ідентифікація дискурсної структури зв'язаного тексту, тобто характеристики дискурсних взаємозв'язків між реченнями (наприклад, тлумачення, удосконалення, контраст) [542-550].
2) Розпізнавання та класифікація мовленнєвих актів μ_1^{22} (англ. Recognizing and Classifying the Speech Acts) у частині фрагменту тексту (наприклад, запитання типу так-ні, питання змісту, твердження, висловлювання тощо) [542-550].

III. Неявне семантичне маркування ролей μ_1^3 (англ. Implicit Semantic Role Labelling) складається з наступних етапів [542-550]:

1) Семантичне маркування ролей S_R фрагменту тексту.
2) Ідентифікація множини семантичних ролей S_N, які явно не реалізовані в поточному реченні відповідного тексту.
3) Класифікація на відповідні аргументи $S_N = S_1 \cup S_2$, які:
    i. явно реалізовані в інших фрагментах тексту S_1,
    ii. не вказані та не реалізовані у всьому тексті S_2.
4) Семантичне маркування ролей першої множини S_1 у локальному аналізованому фрагменті тексті.

IV. Ідентифікація текстового втягування μ_1^4 (англ. Recognizing Textual Entailment) – порівняння двох фрагментів тексту *A* і *B* для формування відповідного висновку щодо відповідності дійсності *A* [542-550] та:

1) логічного істинного виведення з нього змісту *B* фрагменту,
2) логічного заперечення з нього змісту *B* фрагменту тексту,
3) можливості *B* фрагменту тексту бути істинними чи хибними.

V. Тематична сегментація та розпізнавання μ_1^5 (англ. Topic Segmentation and Recognition) – поділ відповідного фрагменту тексті на сегменти різної тематики та ідентифікація тематики цих сегментів [542-550].

VI. Видобування аргументів μ_1^6 (англ. Argument Mining) – автоматичне вилучення та ідентифікація аргументованих структур із тексту, які включають передумову, висновки, схему аргументації та взаємозв'язок між основним та допоміжним аргументом або основним та контраргументом у дискурсі [542-550]. Використовують у багатьох різних жанрах, наприклад, для якісної оцінки змісту соціальних мереж для політиків та відповідних дослідників. Також використовують для аналізу наукових статей, оглядів товарів, онлайн публікацій ЗМІ, юридичних документів, Internet-дебатів та діалогічних доменів.

Опрацювання природної мови через вищого рівня NLP-додатки μ_2 імітують розумну поведінку та очевидне розуміння природної мови та на сьогодення в загальному поділені на такі класи [542-550]:

I. Узагальнення тексту μ_2^1 (англ. Automatic Summarization, Text Summarization) – генерування читабельного дайджесту/анотації [542-550] як підсумку у вигляді фрагменту тексту із загального аналізованого тексту (наприклад, наукової статті, публікації у газеті або журналі).

II. Корекція граматичних помилок μ_2^2 (англ. Grammatical Error Correction) – ідентифікація та виправлення граматичних/стилістичних помилок на всіх рівнях лінгвістичного аналізу (фонологія/орфографія, морфологія, лексика, синтаксис, семантика, прагматика) [542-550].

III. Машинний переклад μ_2^3 (англ. Machine Translation) – автоматичний переклад тексту з однієї людської мови на іншу з використанням всіх рівнів лінгвістичного аналізу, особливо граматики, семантики та фактів про реальний світ тощо на основі розв’язку AI-повної (англ. AI-complete, AI-hard) задачі реферативного перекладу контенту [542-550].

IV. Керування діалогом μ_2^4 (англ. Dialogue Management, Dialogue System, Conversational Agent, CA) – організація спілкування програми з людиною, відмінне від чат-ботів, з використанням тексту, промови, графіки, тактильності, жестів тощо для двостороннього зв'язку [542-550].

V. Питально-відповідні системи μ_2^5 (англ. Question Answering) – визначення правильної відповіді на основі аналізу людського типового запитання [542-550] (наприклад, "Яка столиця України?") та відкритого/складного запитання (наприклад, "У чому сенс буття?").

VI. Генерування природньої мови μ_2^6 (англ. Natural Language Generation, NLG) – перетворення контенту з баз/сховищ даних або семантичних намірів на читабельну конкретну людську мову через алгоритм [542-550]:

1) Визначення вмісту μ_2^{61} (яку інформацію згадувати в тексті).
2) Структурування документа μ_2^{62} (шаблон передачі контенту).
3) Агрегація (об’єднання подібних речень для поліпшення читабельності та природності текстового контенту).
4) Лексичний вибір μ_2^{63} (додавання слів до понять).
5) Генерація виразів-посилань μ_2^{64} (створення виразів, що ідентифікують об’єкти та регіони). Це завдання також включає прийняття рішень щодо займенників та інших видів анафори.
6) Реалізація μ_2^{65} (генерування тексту з врахуванням правил синтаксису, морфології та орфографії, наприклад, використання дієслів у необхідних часових відмінках).
7) При необхідності використання методів машинного навчання μ_2^{66} (найчастіше LSTM) на великому наборі вхідних даних та відповідних (написаних людиною) вихідних текстів, наприклад для генерування текстових підписів до зображення.

VII. Розуміння природньої мови μ_2^7 (англ. Natural Language Understanding, NLU) – перетворення тексту на логічні структури для керування NLP-програмами, тобто ідентифікація семантики з множини можливих у формі

організованих нотацій понять природної мови [542-550]. Введення та створення мовленнєвої мета-моделі та онтології є ефективним та емпіричним. Для побудови формалізації семантики використовують явну формалізацію на противагу з неявними припущеннями, наприклад, про замкнений світ (англ. Closed-World Assumption, CWA – помилкове будь-яке твердження, про який не відомо його вірність) проти відкритого (англ. Open-World Assumption OWA – істинність твердження не залежить від знання спостерігача про його вірність), або суб'єктивне ТАК/НІ проти об'єктивного ІСТИНЕ/ХИБНЕ [542-550].

VIII. Генерування книги μ_2^8 (англ. Book Generation) – створення повноцінних книг на основі NLG-технології μ_2^6, продукційних/асоціативних правил, нейронної мережі, фактичних знаннях та узагальненні тексту μ_2^1 [542-550].

IX. Документ ШІ μ_2^9 (англ. Document AI) – платформа навчання агента витягувати конкретні необхідні дані з різних типів документів [542-550]. Призначена для користувачів без досвіду ШІ, ML та NLP швидко отримувати доступ до необхідного контенту, прихованого в текстах, наприклад, для адвокатів, бізнес-аналітиків та бухгалтерів.

3.3. Приклади моделювання процесів розв'язку типових NLP-задач

3.3.1. Формальна модель КЛС ідентифікації вірусних заголовків новин

Модель КЛС ідентифікації вірусних заголовків новин подано як:

$$S_{vh} = \lambda_1^6 \circ \psi_{NN} \circ \kappa_1 \circ \eta_n \circ \beta_3 \circ \lambda_1^4 \circ \gamma_4. \qquad (3.141)$$

де γ_4 – токенізація; λ_1^4 – розпізнавання іменованих сутностей; β_3 – розмічування частин мови; η_n – Ngrams (послідовності елементів та їх частоти); κ_1 – кластеризація; ψ_{NN} – ML на основі нейронних мереж (Neural Networks, NN); λ_1^6 – застосування SentiWordNet (лексично-семантичний тезаурус для аналізу тональності тексту) [51].

3.3.2. Виправлення граматичних та стилістичних помилок

Ще актуальною NLP-технологією є виправлення помилок, основними задачами якої є ідентифікація помилки, виправлення помилки та навчання користувача [51-53]. Корекція граматичних помилок μ_2^2 є одним із підпроцесів виправлення різних типів помилок. Виправлення помилок подано:

$$S_{ec} = \mu_2^2(X) \tag{3.142}$$

де S_{ec} – результат ідентифікації та виправлення граматичних/стилістичних помилок на всіх рівнях лінгвістичного аналізу (фонологія/орфографія, морфологія, лексика, синтаксис, семантика, прагматика); μ_2^2 – основний процес виправлення помилок в текстовому масиві даних X.

Детальний процес виправлення помилок подано як:

$$S_{ec} = \mu_{24}^2 \circ \mu_{23}^2 \circ \mu_{22}^2 \circ \mu_{21}^2, \tag{3.143}$$

де $X' = \mu_{21}^2(X, R_\mu, R_e, D_l, D_g, D_\beta, \beta_3, \delta_3, \mu_1^1)$ – перевірка на відповідність шаблону на основі складних багатошарових NLP-правил R_μ на основі регулярних виразів R_e, лексичних D_l та граматичних D_g словників, POS-тегів D_β та розмічування частин мови β_3, дерев розбору через парсинг δ_3 та резолюції кореферентності μ_1^1;
$X'' = \mu_{22}^2(X', D_t, D_\beta, D_d, \eta_n, \mu_{22}^{21}, \mu_{22}^{22})$ – статистичні методи уточнення виправлень (η_n – Ngrams аналіз на основі множини токенів D_t, POS-тегів D_β, та історії залежностей аналогів D_d; μ_{22}^{21} – ідентифікація помилки через статистику використання з аналогічними словами/ помилками в тексті; μ_{22}^{22} – виправлення на ймовірніший варіант із можливих аналогів);
$X''' = \mu_{23}^2(X'', D_c, R_\eta, D_a, \eta_n, \mu_{23}^{21}, \mu_{23}^{22}, \mu_{23}^{23}, \mu_{23}^{24}, \mu_{23}^{25})$ – машинне навчання на основі множини класифікаторів D_c для конкретної мови, Ngrams аналізу η_n (наприклад, біграми з відповідним аналізом лівого та правого контексту з подальшим заміною кращого варіанту з POS Ngrams моделі), правил машинного навчання R_η як вибору з кількох правильних варіантів або ідентифікації правильного але рідкісного застосування, процесів виявлення μ_{23}^{21} та виправлення μ_{23}^{22} помилки на

основі ML, анотування даних D_a для навчання μ_{23}^{23}, вибору ознак для навчання μ_{23}^{24} та навчання класифікатора μ_{23}^{25} використовуючи метод випадковий ліс, логістична регресія або інший (іноді узгодження шаблонів і прості статистичні дані не можуть узагальнити варіанти рішень, наприклад, ідентифікація та вибір в англійській мові прийменника або артикля, а в українській мові дієприкметника; деривaційна словотвірна морфологія; run-on sentences, тобто самостійні або підрядні речення не поєднані сполучником або пунктуацією);

$S_{ec} = \mu_{24}^{2}(X''', D_c, R_\eta, D_a, \eta_n, \mu_{24}^{21}, \mu_{24}^{22})$ – нейронний машинний переклад на основі процесів Noisy channel translation μ_{24}^{21} (перевірка правопису, відповідей на запитання, розпізнавання мови та машинного перекладу на основі знаходження передбачуваного слова з даним словом, де символи якимось чином зашифровані) та Round-trip translation μ_{24}^{22} (двосторонній переклад з вихідної мови на цільову для оцінки якості/точності результату).

N-gram модель призначає ймовірності реченням і послідовностям слів на основі підрахунку N-грамів наприклад за припущенням Маркова [51-53]:

$$p(x_j^n) \approx \prod_{i=1}^{n} p(x_i | x_{i-1}), \tag{3.144}$$

де x_j^n – j-тий ланцюг або речення з n слів; x_i – поточне слово в j-тому ланцюгу або реченні; x_{i-1} – попереднє слово в j-тому ланцюгу або реченні. Для ідентифікації помилки треба визначити граматичні або стилістичні ознаки за допомогою правил розмічування частин мови β_3, парсингу: залежностей δ_3^1 та складових δ_3^2 або у вигляді деякого поєднання цих способів δ_3^3 [51-53].

3.4. Основні результати та висновки розділу

Розроблена загальна структура КЛС опрацювання текстового контенту українською мовою та концептуальна схема/модель функціонування типової КЛС на основі моделювання взаємодії основних процесів та компонентів ІС.

Здійснено моделювання основних NLP-процесів КЛС за рахунок взаємодії основних процесів/компонентів ІС та адаптованих до української мови методів лінгвістичного опрацювання текстового контенту на основі графемного,

морфологічного, лексичного, синтаксичного, семантичного, структурного, онтологічного та прагматичного аналізу дозволила вдосконалити ІТ інтелектуального аналізу текстового потоку для розв’язку конкретної задачі ОПМ. Це забезпечило адаптацію процесів ОПМ для аналізу україномовного текстового контенту. Розроблена та описана формальна модель комп’ютерної лінгвістичної системи для опрацювання україномовного текстового контенту, що дало змогу визначити основні структурні елементи та оператори опрацювання природної мови на кожному рівні аналізу тексту як графемного/фонологічного, морфологічного, синтаксичного, семантичного, референційного, структурного, онтологічного та прагматичного. У зв’язку зі складністю морфології української мови детальна увага приділена саме опису моделі морфологічного аналізу текстового контенту.

Наведені приклади моделювання процесів розв’язку типових NLP-задач як КЛС ідентифікації вірусних заголовків новин та виправлення граматичних та стилістичних помилок.

Основні результати розділу опубліковані у роботах [19-23, 110, 112-114, 136, 139-147, 160, 163, 211-212, 256-259, 287-297, 402-407, 474, 471-477, 479, 490, 491, 534-535, 579, 803-822, 875-878, 882, 954-957, 958-983].

РОЗДІЛ 4
АРХІТЕКТУРА КОМП'ЮТЕРНОЇ ЛІНГВІСТИЧНОЇ СИСТЕМИ ОПРАЦЮВАННЯ КОНТЕНТУ УКРАЇНСЬКОЮ МОВОЮ

4.1. Загальна архітектура комп'ютерних лінгвістичних систем

4.1.1. Основні процеси комп'ютерних лінгвістичних систем

Розглянемо архітектурні шаблони проектування КЛС на основі супроводу життєвого циклу ML-моделі для моніторингу/управління конвеєром (інформаційним потоком) контенту (Рис. 4.1).

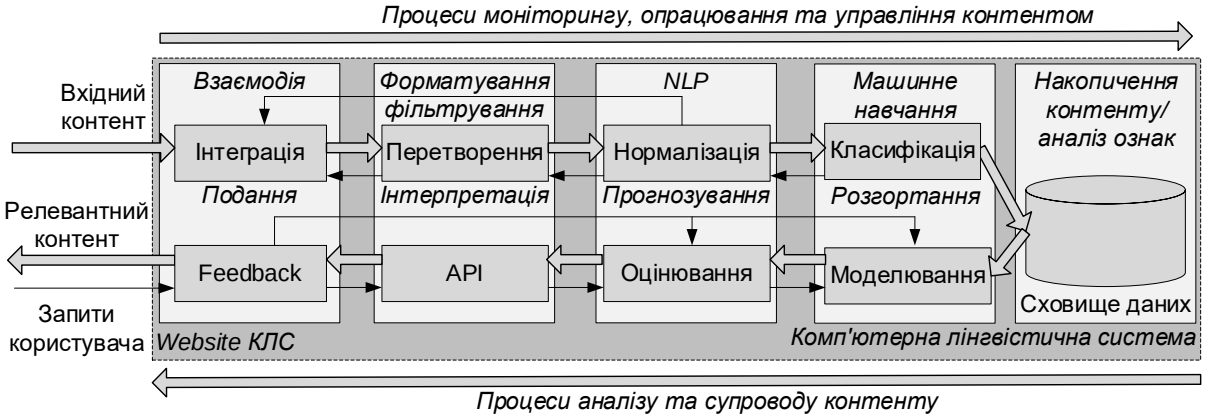


Рис. 4.1. Схема моніторингу/управління конвеєром контенту КЛС

Стандартний конвеєр опрацювання контенту реалізує ітераційний процес, що складається з етапів створення і розгортання ML-процесу [506-511, 1009]. Процес моніторингу/управління конвеєром контенту має ще складатися з додаткових етапів (Рис. 4.1) для покращення якості/оперативності/ефективності розв'язку NLP-задач [958-983]. На етапі побудови неопрацьований інтегрований контент фільтрується від шуму/дублів та форматується в придатну форму для подальшого опрацювання/управління, проведення над ним експериментів, передачі ML-моделям класифікації/кластеризації/прогнозування/оцінювання тощо [984-1008]. На етапі аналізу та супроводу контенту відбувається розгортання змісту для визначення найкращої ML-моделі проведення оцінок/прогнозів, які безпосередньо впливають на постійного користувача та цільову аудиторію.

4.1.2. Основі складові компоненти комп’ютерних лінгвістичних систем

На основі Feedback (зворотного зв’язку) та вихідних даних моделі цільова аудиторія взаємодіє з КЛС, що сприяє адаптації обраної моделі навчання. П’ять стадій відповідних процесів визначають основні архітектурні принципи побудови типових КЛС. Для процесів моніторингу, опрацювання та управління контентом це – взаємодія, форматування/фільтрування, NLP, машинне навчання [506-511, 1009] та накопичення даних в СД. Для процесів аналізу та супроводу контенту відповідно це – аналіз ознак, розгортання, прогнозування, інтерпретація та подання контенту/результату. На стадії взаємодії необхідні набір правил інтеграції контенту з множини достовірних джерел в певні часові проміжки. Також паралельно необхідний набір правил перевірки даних введених від користувача КЛС як попередній етап для стадії форматування/фільтрування згідно колекції наперед закладених модератором правил та контенту з СД. Наступна стадія NLP є підготовним проміжним етапом для машинного навчання та накопичення даних. Стадія машинного навчання може приймати різні форми від SQL-запитів до різних програмних модулів. Процес супроводу більш простіший для реалізації, ніж етап управління при умові, що останній коректно реалізований, особливо при NLP-аналізі, при якому створюють додаткові лексичні ресурси та артефакти (словники, перекладачі, регулярні вирази тощо), від яких напряму залежить ефективність функціонування КЛС (Рис. 4.2).

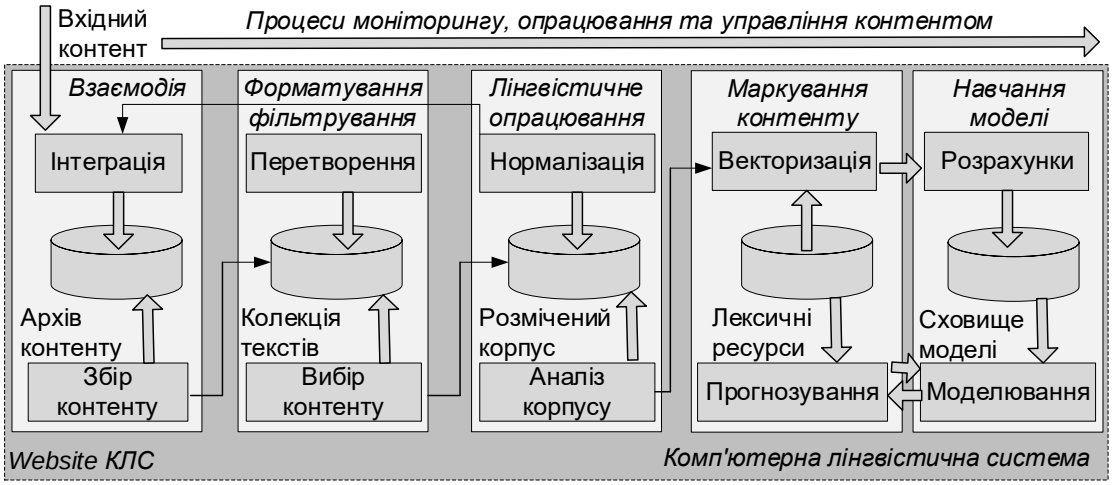


Рис. 4.2. Схема опрацювання конвеєра контенту КЛС

Процес переходу від неопрацьованого тексту до розгорнутої моделі машинного навчання складається з послідовності додаткових перетворень контенту. По-перше, вхідний текстовий контент перетворюється в вхідний корпус як колекція текстів, накопичується і зберігається в СД. Вхідний контент далі групують, фільтрують, форматують, лінгвістично опрацьовують, маркують, нормалізують та перетворюють у вектори для подальшого опрацювання. При остаточному перетворенні модель/моделі (Рис. 4.3) тренують на векторному корпусі, створюють узагальнене подання вихідного контенту для подальшого застосування при розв'язку конкретної NLP-задачі.

4.1.3. Загальна архітектура комп'ютерних лінгвістичних систем на основі машинного навчання

Архітектура КЛС на основі ML з прискореним або навіть автоматичним генеруванням моделі має підтримувати та оптимізувати перетворення контенту з легкістю тестування і налаштовування. Процес генерування оптимальної ML-моделі є складним циклічним алгоритмом, основними етапами якого є формування колекції ознак, вибору моделі та корегування гіперпараметрів. Після кожної ітерації результати оцінюються для визначення оптимальної колекції ознак, моделей і параметрів розв'язку конкретної NLP-задачі при відповідних вхідних даних (рис. 1.5) [506-511, 1009-1010].

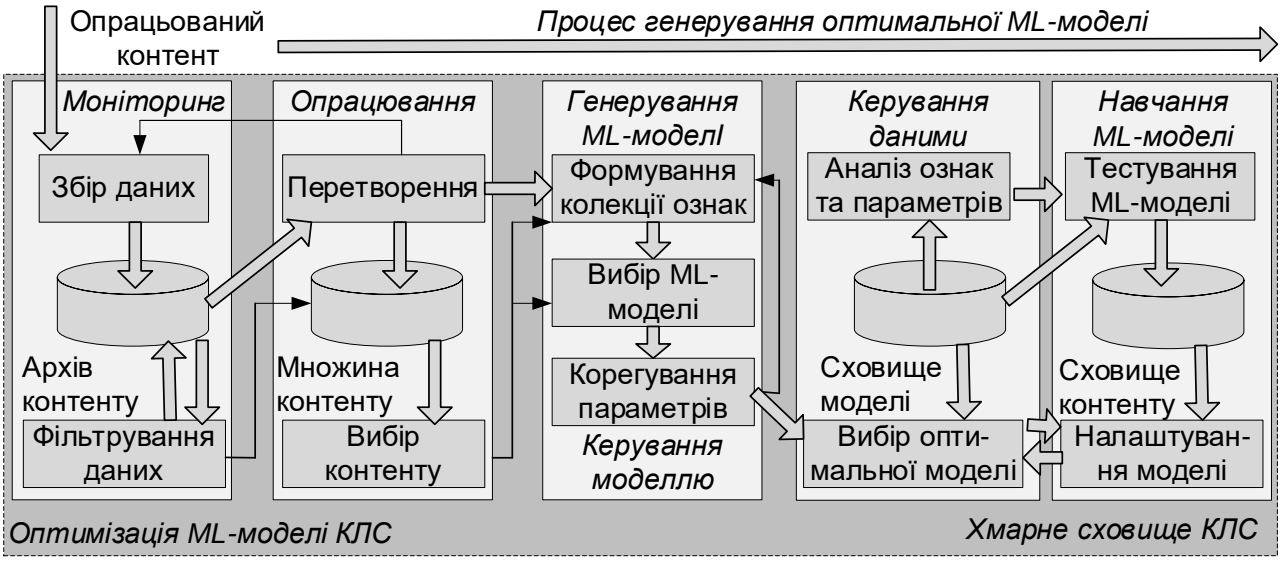


Рис. 4.3. Процес формування та оптимізації моделі машинного навчання

Згідно з [506-511, 1009-1011] є 3 основні напрями статистичного ML: клас моделей, форма моделі та навчена модель. Клас моделей визначає взаємозв'язок між змінними та сформованою метою (наприклад, лінійна модель, рекурентна нейронна мережа тощо). Форма моделі є конкретний компонент моделі: колекція ознак, алгоритм або колекція гіперпараметрів. Навчена модель є форма моделі, яка навчена на певному наборі даних та адаптована для прогнозування. КЛС складаються з багатьох навчених моделей, побудованих під час їх вибору, що створює та оцінює форми моделі.

Будь-який текст природньою мовою на початках як вхідний контент в КЛС є колекцією не випадкових неструктурованих даних. Але зазвичай текст сформований на основі певних лінгвістичних правил для можливості розуміння цих даних. Мета модуля інтеграції перетворити цю колекцію не випадкових неструктурованих даних на структуровані/напівструктуровані на основі полів (записів) або розмітки для зручної інтерпретації модулями КЛС. ML-методи (наприклад, навчання за вчителем) дозволяють навчати (і перекваліфіковувати) статистичні моделі в міру зміни мови при NLP-процесах. Генеруючи ML-моделі на контекстно-залежних корпусах, КЛС можуть застосовувати вузькі семантичні значення для підвищення точності без необхідності додаткової інтерпретації.

Формально ML-модель української мови має вхідну неповну фразу доповнити відсутніми словами/словосполученнями, найбільш ймовірнісними для завершення змісту висловлювання згідно попереднього тексту (аналіз контексту для подальшого вгадування/прогнозування сенсу). Зазвичай грамотно та коректно побудований текст є передбачуваним на основі його зв'язності. Обчислення ентропії (ступеня невизначеності/неочікуваності) розподілу ймовірності моделі української мови вимірює ступінь передбачуваності тексту. Так незакінчені фрази *Київ - столиця...* та *сонце сходить на...* мають низьку ентропію і статистичні мовленнєві моделі з великою ймовірністю вгадають продовження *України* і *сході* відповідно. А вирази з високою ентропією як *ми йдемо в гості до...* та *я зустрів сьогодні...* пропонують багато варіантів продовження (батьків, друзів, сусідів, колег є однаково ймовірними без аналізу

попереднього контексту). Мовленнєві моделі здатні робити виведення або ідентифікувати зв'язки між лексемами. Формально модель використовує контекст для ідентифікації вузького простору прийняття рішень з множини невеликого числа варіантів. Застосування статистичних ML-методів (з вчителем та без нього) дозволяє генерувати мовленнєві моделі витягання сенсу з текстів для підтримки його передбачуваності. Спочатку ідентифікують характерні ознаки контенту для передбачення мети. Текстові дані надають безліч можливостей для вилучення поверхневих ознак на основі парсингу та розбиттям речень/висловлювань/фраз (наприклад, мішок слів), а також на основі витягнутих морфологічних/синтаксичних/семантичних характеристик. Особливу увагу приділяють лінгвістичним/контекстним/структурним ознакам.

1. Прикладом аналізу лінгвістичної ознаки може слугувати ідентифікація переважаючої статі в фрагменті тексту новин (ролі гендеру) в різних контекстах [1012] для виявлення гендерних упереджень щодо тематики публікацій. В гендерному аналізі тексту слова в жіночому і чоловічому роді використовують для формування частотної оцінки гендерних ознак, тобто

$$Sing_{GS} =< X_{Sentence}, W_{Male}, W_{Female}, W_{Unknown}, W_{Both}, f_{gendetize}, >, \quad (4.1)$$

де $X_{Sentence}$ – аналізоване речення/висловлювання; W_{Male} – множина слів з ознакою чоловіка; W_{Female} – множина слів з ознакою жінка; $W_{Unknown}$ – множина слів з невідомою гендерною ознакою; W_{Both} – множина слів з ознакою чоловіка та жінки; $f_{gendetize}$ – оператор ідентифікації гендерного класу речення.

$$Sing_{GS} = f_{gendetize}(X_{Sentence}, W_{Male}, W_{Female}, W_{Unknown}, W_{Both}), \quad (4.2)$$

$$Sing_{GS} = \begin{cases} N_{Male} > 0, N_{Female} = 0 \ \rightarrow \ male \\ N_{Male} = 0, N_{Female} > 0 \ \rightarrow \ female \\ N_{Male} > 0, N_{Female} > 0 \ \rightarrow \ both \\ unknown \end{cases} \quad (4.3)$$

де N_{Male} – кількість слів з ознакою чоловіка в аналізованому реченні $X_{Sentence}$; N_{Female} – кількість слів з ознакою жінка в аналізованому реченні $X_{Sentence}$.

Також треба визначити частоту слів, ознак роду і речень у всій публікації:

$$Sing_{TS} =< X_{Text}, S_{NG}, N_{Sentence}, W_{NG}, f_{countgender} >, \quad (4.4)$$

$$Sing_{TS} = f_{countgender}(N_{Sentence}, S_{NG}, W_{NG}, f_{gendetize}(X_{Text}, X_{Sentence})), \quad (4.5)$$

де X_{Text} – аналізований текст публікації; S_{NG} – множина кількостей маркованих за гендерною ознакою речень $X_{Sentence}$ аналізованого тексту X_{Text}; $N_{Sentence}$ – кількість речень аналізованого тексту X_{Text}; W_{NG} – множина кількості слів кожної гендерної ознаки для кожного маркованого речення $X_{Sentence}$; $f_{countgender}$ – оператор ідентифікації та класифікації/маркування всіх речень аналізованого тексту X_{Text} за гендерною ознакою на основі $f_{gendetize}$.

$$Sing_{TS} = \begin{matrix}^{N_{Sentence}} \\ \Big[\\ _{1}\end{matrix} \begin{matrix} S_{NG}[Sing_{GS}] \mathrel{+}= 1 \\ W_{NG}[Sing_{GS}] \mathrel{+}= len(X_{Sentence}) \end{matrix} \quad (4.6)$$

Для гендерної ідентифікації необхідно пропарсити вихідний текст публікацій з подальшим маркуванням речень та слів на основі бібліотеку NLTK:

$$Sing_{TP} =< X_{Text}, S_{Sentence}, N_{Sentence}, W_{Word}, N_{word}, f_{parsegender}, f_{pcent} >, \quad (4.7)$$

$$Sing_{TP}=f_{pcent}(f_{parsegender}(N_{Sentence}, W_{Word}, N_{word}, f_{countgender}(S_{Sentence})), \quad (4.8)$$

$$Sing_{TS} = \begin{matrix}^{N_{Gender}} \\ \Big[\\ _{k=1}\end{matrix} \begin{matrix} pcent_k = \left(\frac{W_{NG_k}}{total}\right) * 100 \\ N_{Sentence_k} = S_{NG}\left[Sing_{GS_k}\right] \\ print(pcent_k, Sing_{GS_k}, N_{Sentence_k}) \end{matrix} \quad (4.9)$$

$$total = \sum_{i=1}^{N_{Gender}} W_{NG_i}, \quad (4.10)$$

$$S_{Sentence} = \bigcup_{i}^{N_S} X_{Sentence_i}, \quad (4.11)$$

$$X_{Sentence_i} = \bigcup_{i}^{N_{WS}} W_{Sentence_i}, W_{Word} = \bigcup_{i}^{total} W_{Word_i}, \quad (4.12)$$

де N_{word} – кількість слів аналізованого тексту X_{Text}; N_{Gender} – кількість класифікацій за гендерною ознакою (в конкретному випадку – 4); W_{NG_k} – кількість слів в реченнях певної гендерної ознаки; S_{NG} – множина кількостей речень в аналізованому тексті певної гендерної ознаки; $pcent_k$ – відсоток належності тексту публікації до певної гендерної ознаки; $Sing_{GS_k}$ – конкретна гендерна ознака; $N_{Sentence_k}$ – кількість речень в аналізованому тексті конкретної гендерної ознаки; $S_{Sentence}$ – множина ідентифікованих парсингом речень в аналізованому тексті X_{Text}; W_{Word} – колекція множин ідентифікованих

парсингом слів в кожному реченні аналізованого тексту X_{Text}; W_{Word} – множина всіх слів тексту X_{Text}; $total$ – кількість всіх слів в аналізованому тексті X_{Text}.

Такий детермінований механізм демонструє, як зміст/частота використання слів/словосполучень (особливо стереотипних) впливають на передбачуваність контенту згідно попереднього контексту (гендерна ознака вбудована безпосередньо в українську мову – кожний іменник має рід). Але мовленнєві ознаки не завжди мають визначальне значення, наприклад, множина і час застосовують для аналізу мови/процесів/дій/подій в часі.

2. Прикладом аналізу контекстної ознаки може слугувати аналіз настроїв або сентимент-аналіз тексту (емоційне забарвлення при обговоренні конкретної тематики відповідною групою людей). Зазвичай застосовують при комплексному аналізі зворотного зв'язку від користувачів, наприклад, е-комерції, полярності повідомлень або реакцій на події/явища, наприклад, в соціальних мережах або при політичних/економічних дискусіях/форумах тощо. При поверхневому сентимент-аналізі застосовують зазвичай механізм гендерної класифікації (позитивно/негативно/нейтрально забарвлене слово). Наприклад, для позитивного – *чудовий*, *прекрасний*, *правдивий*, негативного – *лінивий*, *поганий*, *дратівливий*, та нейтрального – *білий*, *сонячний*, *космічний*. Але настрій не є особливістю мови та залежить від сенсу слів/словосполучень відповідно до навколишнього контексту тексту, наприклад, для слова *кумедний* є декілька інтерпретацій передачі настрою, зокрема, позитивна – *смішний клоун*, негативна – *кумедний одяг*, та нейтральна – *кумедний кіт* або *кумедна іграшка*. Слово *гострий* зі слова *перець* або *ніж* – позитивне значення при покупці, але зі слова *біль* та *ніж* в кримінальній справі – негативне значення. Також заперечення перетворює сенс позитивного тексту з позитивним словами в негативне і навпаки, наприклад, *ми дуже багато очікували від відпочинку на морі сонячними гарними днями, але обіцяна курортна база відпочинку все спаскудила* (одне негативне слово *спаскудила* перекреслило всі попередні позитивні) або *дощ, прохолода та вітер не стали перепонами гарно відпочити в чудовій компанії*.

Лише завдяки машинному навчанню в таких випадках можна отримати передбачуваність тексту та виявити емоційну забарвленість згідно контексту. Апріорі детермінований/структурний підхід втрачає гнучкість контексту і сенсу, тому більшість мовленнєвих моделей враховують розташування слів в контексті, використовуючи ML-методи для прогнозування.

Основний метод розроблення простих мовленнєвих моделей є мішок слів як частота сумісної появи слів у вузькому, обмеженому контексті (Рис. 4.4).

1) *інтелектуальна інформаційна система* → ***інтелект інформ систем***
2) *інтелектуальний інформаційний пошук* → ***інтелект інформ пошук***
3) *опрацювання інформаційних ресурсів* → ***опрацюв інформ ресурс***
4) *система електронної комерції* → ***систем електр комерц***
5) *комп'ютерна лінгвістична система* → ***комп'ютер лінгвіст систем***
6) *аналіз природної мови* → ***аналіз природ мов***
7) *опрацювання природної мови* → ***опрацюв природ мов***
8) *опрацювання текстового контенту* → ***опрацюв текст контент***
9) *аналіз текстового контенту* → ***аналіз текст контент***
10) *пошук текстового контенту* → ***пошук текст контент***
11) *лінгвістичний аналіз контенту* → ***лінгвіст аналіз контент***
12) *лінгвістичний аналіз тексту* → ***лінгвіст аналіз текст***

	аналіз	***електр***	***інтелект***	***інформ***	***комерц***	***комп' ютер***	***контент***	***лінгвіст***	***мов***	***опрацюв***	***пошук***	***природ***	***ресурс***	***систем***	***текст***
аналіз	0														
електр	0	0													
інтелект	0	0	0												
інформ	0	0	2	0											
комерц	0	1	0	0	0										
комп'ютер	0	0	0	0	0	0									
контент	2	0	0	0	0	0	0								
лінгвіст	2	0	0	0	0	1	1	0							
мов	1	0	0	0	0	0	0	0	0						
опрацюв	0	0	0	1	0	0	1	0	1	0					
пошук	0	0	1	1	0	0	1	0	0	0	0				
природ	1	0	0	0	0	0	0	0	2	1	0	0			
ресурс	0	0	0	1	0	0	0	0	0	1	0	0	0		
систем	0	1	1	1	1	1	0	1	0	0	0	0	0	0	
текст	2	0	0	0	0	0	3	1	0	1	1	0	0	0	0

Рис. 4.4. Матриця частот сумісної появи слів

Таке оцінювання сприяє визначити ймовірнісне сусідство та з невеликих фрагментів тексту визначити їх значення. Далі, використовуючи методи статистичного виведення, можна передбачити порядок слів. Це досить просто для англомовних текстів, де слова не відмінюються. Для україномовних тестів краще застосовувати не мішок слів, а мішок основ слів. Наприклад, для 12 словосполучень як 3-грам (36 слів) без врахування відмінювання отримаємо

матрицю розміром 20×20, а з врахуванням відмінювання, роду та особи (аналіз лише основ слів) – 15×15. Причому для української мови місце розташування основ в 3-грамі зазвичай не важливе та має часто однозначну за змістом ймовірність сумісності появи, наприклад, *інформаційний ресурс* (*інформ ресурс*) та *ресурс інформації* (*інформ ресурс*). Модель мішок слів/основ також розширюють аналізом спільної появи стійких словосполучень та фрагментів виразів, що мають велике значення для ідентифікації сенсу тексту. Вирази *зелений край скатертини* (межа) та зелений край батьківщини (місцевість) у вигляді 3-грами несуть змістовно різне навантаження. Тобто лише для слова *край* є декілька тлумачень (межа об'єкту, шматок, закінчення дії/стану, особлива місцевість, місце проживання, адміністративно-територіальна одиниця). Статистичний аналіз n-грам дозволяє виокремити закономірності контексту.

Мовленнєві моделі на основі аналізу n-грам контексту вимагають здатності досліджувати відношення тексту до деякої цільової змінної. Застосування аналізу лінгвістичних і контекстних ознак сприяє формуванню загальної передбачуваності тексту. Але для їх ідентифікації та подальшого використання потрібна здатність парсити/визначати лінгвістичні одиниці мови.

3. Прикладом аналізу структурної ознаки може слугувати конструювання онтології для реалізації ІІП. Поряд лінгвістичними та контекстними ознаками тоді необхідно ідентифікувати та опрацьовувати одиниці мови високого рівня для визначення словника операцій до корпусу тексту. Різні одиниці мови опрацьовують на різних рівнях, та коректна реалізація NLP-методів на основі ML має важливе значення для оперативної та коректної ідентифікації лінгвістичного контексту (структури взаємозв'язку сем). На основі типового шаблону висловлювання (твердження або проста фраза) у вигляді *суб'єкт* → *дієслово* → *об'єкт* → *визначення об'єкту* (*підмет* →*присудок* →*додаток*) конструюють онтології, які визначають конкретні відношення між сутностями. Вони ж дозволяють вирішити проблему відсутності обов'язкового порядку слів в українському реченні для ідентифікації його семантики. Доцільно застосовувати

для задач де постійно треба опрацьовувати великі обсяги текстових даних та присутня довгострокова ресурсна підтримка проекту.

Семантичний аналіз полягає не лише в ідентифікації змісту тексту, але і в генеруванні структур даних, до яких можна застосувати логічні міркування. Текстові семантичні (тематично-змістовні) подання (англ. Thematic Meaning Representations, TMR) застосовують для кодування речень у вигляді предикатних структур на основі логіки першого ступеня або лямбда-числення (λ-числення). Мережеві/графові структури застосовують для кодування взаємодій предикатів відповідних ознак тексту. Потім реалізують обхід для аналізу центральності термінів або суб’єктів та причин відношень між елементами. Аналіз графів зазвичай не є повним СЕМ, але допомагає сформувати частину важливих логічних рішень або висновків. Семантика, синтаксис і морфологія дозволяють додавати дані до простих текстових рядків з лінгвістичним значенням та генерувати новий змістовний текстовий контент.

В даний час природна мова є однією з найбільш часто використовуваних форм контенту. Його аналіз дозволяє збільшити корисність додатків даних і зробити їх невід'ємною частиною повсякденного життя. Для масштабованого аналізу та машинного навчання тексту в першу чергу потрібні сучасні знання та корпуси тексту відповідної ПО. Наприклад, якщо у сфері фінансів КЛС має ідентифікувати фінансові терміни, абревіатури акцій та назви компаній. Тому документи у корпусі ПО повинні містити ці сутності. Тобто розроблення будь-якої КЛС починається з отримання текстових даних відповідного типу і формування корпусу із структурними та контекстними ознаками ПО.

4.2. Метод графемного аналізу української мови

4.2.1. Основні регулярні вирази графемного аналізу

Для ГА текстових рядків найкраще застосовувати регулярні вирази (англ. Regular Expression, RE) як алгебраїчні позначення для ознак множини символьних ланцюжків. Зазвичай застосовують в розробленні/підтримці кожного типу комп’ютерних мов (програмування, комунікаційних протоколах,

розмітки даних, специфікацій та проектування), функціонування текстових редакторів, та ПЗ опрацювання тексту, особливо при ІІП за шаблоном або в колекціях текстових корпусів ПО. Ідентифікація/пошук фрагмента/рядка за шаблоном в послідовності символьних ланцюжків реалізують для знаходження всіх збігів або першого. В шаблонах застосовують спецсимволи **[**, **]**, ^, \, **-**, **?**, *****, **+**, **.**, **$**, **|**, **(**, **)**, **_**, **{**, **}** тощо, в тому числі /, але останній не є RE, але його межами. Найпростіший RE є кортежем простих символів (Таблиця 4.1) для розпізнавання першого або всіх подібних до шаблону входжень послідовностей знаків.

Таблиця 4.1

Регулярні вирази ГА текстів українською мовою для розпізнавання всіх знаків

№	RE	Розпізнавання	Приклад та результат
1	/контент/	точної послідовності символів підрядка з врахуванням регістру	Структурна схема лінгвістичного аналізу текстового ***контенту***
2	/к/	конкретного символу з врахуванням регістру	Контент-аналіз застосовують для аналізу пото***к***ів ***к***онтенту
3	/-/	конкретного спецсимволу	Контент-аналіз застосовують
4	/[кК]онтент/	точної послідовності символів без врахування регістру 1-ого знаку	***Контент***-аналіз застосовують для аналізу потоків ***контенту***
5	/[онві]/	або **о**, або н, або **в**, або **і**	К***он***те***н***т-а***н***ал***і***з застосо***в***ують
6	/[0123456789]/	Будь-яка цифра в послідовності рядка	RE чутливі до регістру– правила ***1***, ***2*** та ***4*** дають різні результати
7	/[0123]/	або **0**, або **1**, або **2**, або **3**	RE чутливі до регістру– правила ***1***, ***2*** та 4 дають різні результати
8	/[0-9]/	Будь-яка цифра в послідовності рядка	RE чутливі до регістру– правила ***1***, ***2*** та ***4*** дають різні результати
9	/[а-я]/	Будь-яка літера українського алфавіту нижнього регістру	К***онтент-аналіз застосовують***
10	/[А-Я]/	Будь-яка літера українського алфавіту верхнього регістру	***К***онтент-аналіз застосовують
11	/[А-Яа-я]/	Будь-яка літера українського алфавіту без врахування регістру	***Контент-аналіз застосовують***
12	/[A-Z]/	Будь-яка літера англійського алфавіту верхнього регістру	***RE*** чутливі до регістру– правила 1, 2 та 4 дають різні результати
13	/[^А-Я]/	Будь-який символ, окрім літера англійського алфавіту верхнього регістру	К***онтент-аналіз застосовують для аналізу потоків контенту***
14	/[^Кк]/	Будь-який символ, окрім літер **К** та **к**	К***онтент-аналіз застосовують для аналізу пото***к***ів ***к***онтенту***
15	/[^\.]/	Будь-який символ, окрім знаку крапки **.**	***Контент-аналіз застосовують***
16	/[к^]/	або **к**, або ^	аналіз пото***к***ів ***к***онтенту
17	/х^у/	Шаблон рядка **х^у**	функція ***х^у***
18	/^[А-Я]/	Будь-яка літера українського алфавіту верхнього регістру на початку рядка	***К***онтент-аналіз застосовують для аналізу потоків контенту в КЛС
19	/^а/	Літера **а** на початку рядка	Контент-аналіз застосовують
20	/контенту?/	Наявність/відсутність необов'язкового символу **у** в підрядку	Структурна схема лінгвістичного аналізу текстового ***контенту***
21	/зв'?зок/	Для пошуку необов'язково враховувати апостроф **'**, і його часто упускають	Структурна ознака описує зв***'***язок між лінгвістичними лексемами.
22	/лін.вітиска/	Позначення будь-якого симолу	лін***г***вістика або лін***г***вістика

№	RE	Розпізнавання	Приклад та результат
23	/б.гу/	Позначення будь-якого симолу	Зараз змагання з ***бігу***, тому я ***біжу***. Я ***бігу***н, тому ***біжу*** естафету
24	/ї*/	Будь-який рядок без **ї** або довільне кількість **ї**	***МА лексеми провадять на основі її особистої множини ознак***
25	/її*/	Будь-який рядок з однією або більше **ї**	МА лексеми провадять на основі ***її*** особисто***ї*** множини ознак
26	/[нжтлдчз]*/	або без або довільна кількість або **н** або **ж** або **т** або **л** або **д** або **ч** або **з**	***Віддалено ллється на ланах нашого життя беззмінне збіжжя знання як обличчя особистого досвіду***!
27	/[нжтлдчз]/	або **н** або **ж** або **т** або **л** або **д** або **ч** або **з**	Ві***дд***а***л***е***н***о ***лл***є***т***ься ***н***а ***л***а***н***ах ***н***ашого ***ж***и***тт***я бе***зз***мі***нн***е збі***ж******ж***я з***н***а***нн***я як об***л***и***чч***я особис***т***ого ***д***осві***д***у!
28	/[0-9]*/	або без або довільна кількість одного елементу з діапазону 0-9	***RE чутливі до регістру– правила 1, 2 та 4 дають різні результати***
29	/[0-9][0-9]*/	Одна цифра з діапазону 0-9 обов’язкова, інша ні, але якщо є – довільна кількість однієї з 0-9	RE чутливі до регістру– правила ***1***, ***2*** та ***4*** дають різні результати
30	/[0-9]+/	довільна кількість різних цифр з 0-9	Спецсимвол знаку питання ? для RE-правил ***20***-***21***
31	/[нжтлдчз]+/	один або **н** або **ж** або **т** або **л** або **д** або **ч** або **з** або декілька, або довільна їх комбінація	Ві***дд***а***л***е***н***о ***лл***є***т***ься ***н***а ***л***а***н***ах ***н***ашого ***ж***и***тт***я бе***зз***мі***нн***е збі***ж******ж***я з***н***а***нн***я як об***л***и***чч***я особис***т***ого ***д***осві***д***у!
32	/[нжтлдчз]{2}/	точно два або **н** або **ж** або **т** або **л** або **д** або **ч** або **з**	Ві***дд***алено ***лл***ється на ланах нашого жи***тт***я бе***зз***мі***нн***е збі***ж******ж***я зна***нн***я як обли***чч***я особистого досвіду!
33	/аналіз.*аналіз/	Ідентифікація рядка з використанням подвійного слова **аналіз**	Контент-***аналіз*** застосовують для ***аналіз***у потоків контенту в КЛС
34	/^В/	**В** на початку рядка	***В*** наш час в Інтернет все є.
35	/^Контент-аналіз$/	розпізнання конкретної фрази	***Контент-аналіз***␣␣
36	/␣$/	позначення пробілу в кінці рядка	Контент-аналіз␣ застосовують␣
37	/^Контент-аналіз\. $/	розпізнання конкретної фрази з крапкою та пробілу в кінці рядка	***Контент-аналіз.***␣␣
38	/^/[А-Я]\. $/	розпізнавання всіх можливих речень	***В наш час в Інтернет все є.***␣␣
39	/\bаналіз\b/	розпізнавання конкретної набору символів (слова) з врахуванням меж	Контент-***аналіз*** застосовують для аналізу потоків контенту
40	/\b19\b/	розпізнавання слова як числа	Йому виповнилось ***19*** в 2019.
41	/\b3\b/	розпізнавання слова в межах	Ціна -***3***$ за 13 одиниць.
42	/\b5\b/	розпізнавання слова в межах	Ціна -5€ за ***5*** одиниць.
43	/ML|МН/	розпізнавання скорочень ML або МН	Реалізація КЛС на основі ***МЛ***
44	/контент(у|ний)/	розпізнавання слова з різними флексіями	***Контентний*** аналіз застосовують до великих потоків ***контенту***
45	/№␣[0-9]+␣*/	1 цифра з будь-яким числом пробілів	В␣колонці␣ ***№ 3***␣␣␣␣␣
46	/(№␣[1-9]+␣*)*/	розпізнавання довільного числа послідовності **№** та будь-якої цифри	В ␣ колонках ␣ ***№ 1***␣␣ та ␣ ***№ 3***␣␣, але не в №␣13␣

RE чутливі до регістру– правила 1, 2 та 4 дають різні результати. Використання спецсимолів **[** та **]** вирішує проблему чутливості RE до регістру. Рядок символів в середині **[]** реалізує диз’юнкцію значень при співпадінні. RE-правило 6 розпізнає будь-яку цифру в послідовності символів рядка.

Спецсимвол тире **-** в середині **[]** для RE-правил 8-12 дозволяє не перелічувати всі символи, а вказує будь-який символ у відповідному діапазоні.

Наприклад, Шаблон /[3-6]/ вказує на будь-який з символів 3, 4, 5 або 6, а /[в-ж]/ вказує на один із символів в, г, д, або ж при графемному аналізу вхідного тесту.

Спецсимвол карет або циркумфлекс **^** всередині **[]** для RE-правил 13-18 несе різне змістовне навантаження в залежності від місця розташування. Якщо на початку зразу ж після **[** означає, всі знаки після нього відхиляються в аналізованому рядку символів (RE 13-15). Карет ^ має 3 призначення: для позначення початку рядка (не всередині **[]** – RE 18-19); для позначення заперечення всередині **[]** (RE 13-15); просто для позначення карет ^(RE 16-17).

Спецсимвол знаку питання **?** для RE-правил 20-21 дозволяє позначати в шуканому рядку необов'язкові символи. Це корисно у випадках, коли можуть бути як присутні/відсутні символи в певній послідовності, що не вирішують **[]**. В **[]** – можна позначити відсутність конкретного символу з діапазону можливих, але не описують відсутність взагалі будь-якого символу, невідмінну від **?**.

Спецсимвол знаку крапки **.** для RE-правил 22-23 дозволяє позначити місцерозташування будь-якого символу в послідовності аналізованого рядку.

Якщо спецсимвол **?** є відсутність або присутність одного символу, то подвоєння символу можемо подати через спецсимвол ***** (RE 26-29), який означає відсутність конкретного символу або RE перед ***** в RE або довільна його кількість в послідовно розміщених в розпізнаному рядку, тобто результатом може бути і рядок без цього символу. Тому для знаходження хоча б одного символу із можливої послідовності однакових двох – приклад RE 29, а для двох різних – 30.

Спецсимвол знаку **+** для RE-правил 30-31 дозволяє позначити один або декілька випадків, що безпосередньо передують символу/RE. Для точного позначення кількості (наприклад рівно 2 рази) застосовують **{}** (RE 32).

Спецсимвол знаку крапки **.** часто застосовують разом з спецсимволом ***** для позначення будь-якого рядка символів (RE 33).

Анкор (англ. Anchor – якір, прив'язка) – це спеціальний символ (наприклад, знаки катет ^ або долара **$**) конкретизації місцерозташування RE в символьному рядку. В деякий випадках карет ^ позначає початок рядка (RE 34). Знак долара **$** розпізнає закінчення рядка (RE 35-36). Зворотний слеш \ дозволяє розпізнати

спецсимволи в символьному рядку вхідного тесту (RE 37-38). Анкори **\b** та **\B** ідентифікують присутність та відсутність меж слова відповідно (RE 39-42). Слово – будь-який кортеж з цифр, підкреслень або літер (без спецсимволів).

Для організації вибору альтернативи між, наприклад, синонімами застосовують операцію диз'юнкції на основі спецсимволу **|** (RE 43-46). Поєднання спецсимовлів | всередині **()** дозволяє організувати розпізнавання диз'юнкції лише для певного шаблону з врахуванням різних флексій/префіксів (RE 44). Спецсимволи **()** застосовують для організації лічильників типу ***** (RE 46). Різниця в тому, що ***** застосовують для одного символу, а не цілої послідовності.

Для складних диз'юнктивних операторів RE при групуванні з різних спецсимолів застосовують поняття пріоритетності (Таблиця 4.2): **()** → *****, **+**, **?**, **{}** → **рядок**, **^**, **$** → **|** від найвищого до найнижчого (розмежування через символ →) (). Жадібні шаблони RE типу /[а-я]*/ розпізнають нуль або більше літер та на мають збігів, розширюючи ідентифікацію, щоб покрити стільки рядків, скільки можуть. Нежадібні RE на основі *? та +? знаходять найменший можливий текст.

RE типу /␣*/ застосовують для позначення відсутності або присутності деякої кількості пробілів, оскільки навколо завжди можуть бути додаткові пробіли. Існують псевдоніми для загальних діапазонів, які можуть використовуватися переважно для збереження типу графем (Таблиця 4.3).

Таблиця 4.2

Регулярні вирази для розпізнавання ключових слів, стоп-слів та маркерів

№	RE	Розпізнавання
1	/але/ /аналіз/	простий (але неправильний) шаблон – враховує інші можливі варіанти входження послідовності символів у вхідний рядок
2	/[аА]ле/ /[аА]наліз/	без врахування регістру першої літери, але нажаль враховує інші випадки, наприклад, м**але**ча або к**аналіз**ація
3	/\b[аА]ле\b/ /\b[аА]наліз\b/	з врахуванням меж слова (без літер, підкреслення та цифр з обох боків) – для але добре, але слова аналізу вже ігнорує
4	/[^а-яА-Я][аА]наліз[а-я]/	перед **аналіз** немає жодної літери без врахування регістру, а після довільна літера малого регістру українського алфавіту
5	/\b[аА]наліз[а-я]*/	перед **аналіз** немає жодної літери, підкреслення або цифри, а після довільна літера малого регістру українського алфавіту або жодної
6	/(^|\b[аА]ле\b/ /(^|\b[аА]наліз([а-я]*|$)/	до пункту 5 додається можливість зустріти слова аналіз на початку або в кінці рядка, коли жодного символу не існує в цих позиціях
7	/[0-9]+ (\$|грн\.|EU)/	цілого значення ціни в грн., або валюті США/Євросоюзу
8	/[0-9]+\,[0-9][0-9] грн\./	дійсного значення ціни в грн.
9	/(^|\W)[0-9]+(\,[0-9][0-9])? (\$|грн\.|EU)?\b/	дійсного значення ціни у валюті України/США/Євросоюзу на рівні слова в реченні/висловлюванні/фразі

№	RE	Розпізнавання
10	/(^|\W)[0-9]{0,5}(\,[0-9][0-9])? (\$|грн\.|EU)?\b/	дійсного значення ціни у валюті України/США/Євросоюзу на рівні слова з врахуванням обмеження кількості цифр перед комою
11	/\b[6-9]+␣*(UAH|₴|грн\.|[Гг]грив(ня|ні|ень))\b/	рядків із значенням ціни > 5 у валюті України з врахуванням різних варіантів позначень та скорочень
12	/\b[0-9]+(\,[0-9]+)?␣*(UAH|₴|грн\.?)\b/	рядків із дійсним значенням ціни у валюті України з врахуванням наявності/відсутності різних варіантів позначень та скорочень

Коректно побудовані RE дозволяють уникнути помилок припущення (зайве розпізнає) та заперечення (випадково пропустили). Зниження загального коефіцієнта помилки для ГА передбачає дві антагоністичні умови для генерування колекції RE як збільшення відкликання (мінімізація помилкових ігнорувань) та підвищення точності (мінімізація помилкових розпізнавань).

Таблиця 4.3

Основні псевдоніми RE для загальних діапазонів ГА

№	Діапазон	RE	Розпізнавання	Приклад
1	[␣\n\t\f\r]	\s	будь-які знаки пробілів та табуляцій	аналіз␣контенту
2	[^\s]	\S	жодного знаку пробілів та табуляцій	***а***наліз␣контенту
3	[0-9]	\d	будь-якої цифри з діапазону	***1***4␣лютого␣2005
4	[^0-9]	\D	жодної цифри з діапазону	14␣лютого␣2005
5	[а-яА-Я0-9_]	\w	будь-яка літера, цифра та підкреслення	***к***онтент-аналіз
6	[^\w]	\W	жодна літера, цифра та підкреслення	контент-аналіз
7	\b[0-9]*\b	*	жодна або декілька попереднього RE	***вже 22 рік***
8	\b[0-9]+\b	+	одна або декілька попереднього RE	вже ***2022*** рік
9	\b[0-9]?\b	?	точно відсутнє або один раз присутнє	22 ***рік 2 століття***
10	\b[0-9]{2}\b	{n}	певна кількість повторень	***22*** рік 2 тисячоліття
11	\b[0-9]{1,2}\b	{n,m}	в діапазоні певна кількість повторень	***22*** рік ***2*** тисячоліття
12	\b[0-9]{2,}\b	{n,}	принаймні певна кількість повторень	***22*** рік 2 тисячоліття
13	\b[0-9]{,2}\b	{,m}	до певної кількості повторень	22 рік ***2*** тисячоліття
14	[0-9]{1,}*[0-9]{1,}	*	спеціальне позначення знаку *	значення ***5*93***
15	1[0-9]{1}\.0[0-9]{1}	\.	спеціальне позначення знаку крапки	дата ***14.02***
16	[а-я]\?	\?	спеціальне позначення знаку питання	контент-аналі***з?***
17	[а-я]\n[а-я]	\n	спеціальне позначення знаку нового рядка	контент-аналі***з*** ***к***онтент-моніторинг
18	[б-я])\t[а-я])	\t	спеціальне позначення знаку табуляції	а) ***б) с)***
19	s/текст/контент/	s/x/y/	заміна/уточнення слова іншим	текст → контент
20	s/([0-9]+)/<\1>/	s/R/R'/	заміна/уточнення виразу шаблоном	27 → <27>
21	/x(.*)y\1z/	/(.*)\1/	повторення двічі певного рядку/виразу	x***A***y***A***z
22	/x(.*)y(.*)z\1w\2u/	/()()\1\2/	дублі двох виразів у певних місцях	/x***A***y***B***z***A***w***B***u/
23	/(?:x|y)(z|)text x\1/	/(?:|)(|) \1/	групування, без фіксації шаблону	**x w** text x **w**
24	/(?![яЯ]) [А-Яа-я]+/	/(?!x)y/	будь-який рядок, який не починається з я	***к***онтент-аналіз

RE /{9}/ є розпізнаванням точно 9 випадків попереднього символу/виразу, RE /а.{3}я/ – послідовності *а…я*, RE /{3,12}/ – від 3 до 12 попереднього символу/виразу, RE /(5,)/ – принаймні 5 випадків попереднього символу/виразу, а RE /(,13)/ – до 13 випадків попереднього символу/виразу. Спецсимвол s перед RE дозволяє замінити вираз на шаблон. Спецсимвол \k вказує місцерозташування символу/фрази/виразу як дубль першого елемента в групі

захоплення, тобто шаблону в (), де k – номер дужок або груп захоплення. Таким чином, спецсимволи **()** мають подвійну функцію в RE: для групування умов та визначення порядку застосування операторів. Для групування, без фіксації отриманого шаблону у реєстрі застосовують RE виду (?: шаблон) як групу, що не захоплює вираз. При застосуванні RE визначають ранг використання у черзі. RE типу (?: шаблон) є позитивним твердженням (RE 23).

Оператор (?=шаблон) є позитивним при ідентифікації шаблону з нульовою шириною, тобто покажчик відповідності не просунувся. Оператор (?!шаблон) є позитивним, якщо шаблон не збігається, є нульовою шириною і курсор не просувається. Негативні твердження зазвичай використовують при аналізі складної моделі контенту, коли треба вилучити особливий випадок (RE 24).

4.2.2. Основні етапи графемного аналізу україномовних текстів

Графемний аналіз є попереднім опрацюванням та перетворенням тексту у певний маркований та стиснутий формат для наступних NLP-процесів (Рис. 4.5):

вилучення контенту → виокремлення абзаців → виокремлення речень в межах абзацу → виокремлення лексем в межах речення → маркування лексем тегами для МА як розмічування частинами мови.

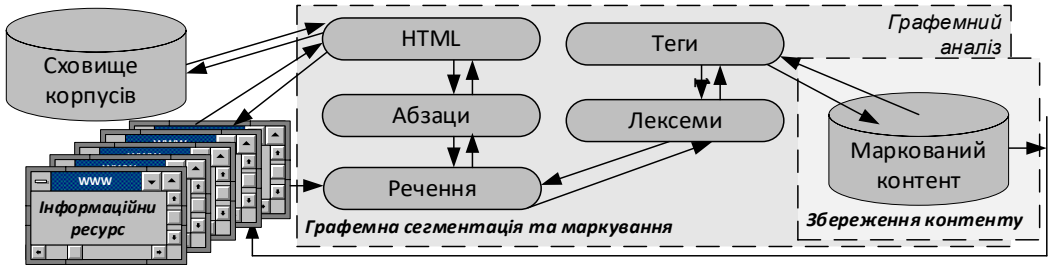


Рис. 4.5. Розбиття контенту, графемна сегментація та маркування

На перших етапах інтеграції контенту з різних джерел необхідно реалізовувати процеси фільтрації, доступу та обчислення розмірів тексту на основі застосування стандартного API попереднього графемного опрацювання розбиття документів через виконання наступної послідовності NLTK-методів:

1) $f_{raw}()$ – організація доступу до попередньо неопрацьованого тексту;
2) $f_{html}()$ – ліквідація нетекстового змісту, сценаріїв та тегів стилів;
3) $f_{patas}()$ – ідентифікація окремих абзаців з тексту контенту;

4) $f_{sents}()$ – ідентифікація окремих речень з тексту контенту;

5) $f_{tokens}()$ – ідентифікація окремих лексем із тексту контенту;

6) $f_{mark}()$ – графемне маркування ідентифікованих лексем на основі RE;

$$T_{marked} = f_{mark}(f_{tokens}(f_{sents}(f_{patas}(f_{html}(f_{raw}(X_{content})))))), \quad (4.13)$$

та при необхідності додаткових методів, наприклад, додавання тегів або розбір речень, перетворення анотованого тексту в структури даних як дерева, або виділення окремих елементів XML. Для ідентифікації та вилучення основного контенту з інформаційного ресурсу з наперед невизначеною структурою та високою варіативністю документів з різних джерел застосовують $f_{html}()$ на основі бібліотеки Python readability-lxml, яка видаляє всі аномальні артефакти, залишаючи тільки текст. При опрацюванні HTML-тексту $f_{html}()$ використовує колекцію формальних RE для ідентифікації та видалення навігаційних меню, оголошень, тегів сценаріїв і CSS, а потім створює нове дерево об'єктної моделі контенту, витягує текст з вихідного дерева та вбудовує в новостворене дерево.

4.2.3. Особливості графемного аналізу україномовних текстів

Задачі з векторизації, вилучення функцій та ML значною мірою залежать від здатності КЛС ефективно розбити текстовий контент на складові компоненти, зберігаючи початкову структуру. Точність і чутливість ML-моделей залежать від ефективності ідентифікації зав'язків лексем з відповідним контекстом в тексті.

Абзаци містять повні ідеї контексту та є структурною одиницею контенту. На основі NLTK реалізують оператор $f_{patas}()$ як генератор абзаців, який визначається як блоки тексту, розділені двома символами подачі рядків. Оператор $f_{patas}()$ сканує всі файли і передає кожен HTML-текст RE-конструктору, вказуючи, що аналіз HTML-розмітки повинен виконуватися через lxml HTMLparser. Отриманий об'єкт зберігає деревовидну структуру, якою можна переміщатися за допомогою оригінальних HTML-тегів і елементів.

Якщо абзаци є структурними одиницями контенту, то речення є семантичними одиницями. Як абзац, що виражає одну ідею, речення містить повну думку, яку автор сформулював і виразив множиною слів. Графемна

сегментація – це поділ тексту на речення для подальшого опрацювання шляхом позначення слів частинами мови при МА. Оператор $f_{sents}()$, викликаючи $f_{patas}()$ і повертаючи ітератор (генератор), сортує всі речення з усіх абзаців.

Оператор $f_{sents}()$ обходить всі абзаци, обрані оператором $f_{patas}()$, і використовує оператор $f_{words}()$ для виконання фактичної графемної сегментації. Внутрішньо оператор $f_{tokens}()$ використовує $f_{mark}()$, модель, попередньо навчену RE-правилами розпізнавання/ідентифікації для різних видів лексем, розділових знаків, скорочень, географічних назв, абревіатур та інших знаків, які служать ознаками початку/закінчення речення або табуляції. Розділові знаки не завжди мають однозначне тлумачення, наприклад, або є ознакою закінчення речення, але вони також присутні в датах, абревіатурах, скороченнях, багатокрапці тощо. Визначення меж речення не завжди є легкою задачею. Пунктуація має вирішальне значення для ідентифікації меж слів (комами, пробілами, двокрапками) та для виявлення деяких аспектів сенсу (знаки питання, знак оклику, лапки). Для деяких завдань, таких як тегування частин мови, аналіз чи синтез мовлення, іноді необхідно відноситися до пунктуаційних знаків так, ніби вони є окремими словами. При аналіз мовлення знаки пунктуації заміняють паузи, наголоси та зміна динаміки інтонації.

Лексемізація – це процес отримання лексем (синтаксичнозакодованих ланцюжків символів) і для його реалізації застосовують оператор $f_{words}()$ на основі RE, який виділяє через $f_{mark}()$ маркери за пробілами та розділовими знаками та повертає список алфавітних та не алфавітних символів. Як і розмежування речень, розпізнавання лексеми не завжди є легкою задачею: наявність розділових знаків в лексемі, розділові знаки як незалежні лексеми, лексеми з дефісами та без, лексеми як скорочені форми слів (одне або декілька слів). Для цих випадків вибираються різні інструменти вибору маркерів.

Будь-яке висловлювання – це мовленнєвий корелят речення. Наявність лексем типу дисфлюенції (втрата швидкості мовлення, наприклад довша пауза при роздумах) несе не так семантичне навантаження як емоційне. Вигуки типу *ммм*, *ох*, *ах* тощо є наповнювачами чи заповненими паузами та також є

емоційним забарвленням, але не семантичним. Незакінчене слово з подальшим повторенням та його закінченням або просто з повторенням є фрагментом, який не несе семантичне навантаження, а лише емоційне. Тому при проведенні ФА в залежності від мети розв'язку конкретної задачі через КЛС важливо приймати до уваги (відповідно маркувати) або ігнорувати деякі види пунктуації (трикрапки, знаки оклику тощо), дисфлюенції, дублі-фрагменти, вигуки тощо. Якщо КЛС є лише транскрипцією промови, тоді треба ігнорувати такі фонеми, щоб позбутися втрати швидкості мовлення. Але вони дозволяють визначити психологічний стан мовця та його емоційний стан, ідентифікувати особливість авторського мовлення спікера при зміні тембру голосу, актуальні при прогнозуванні майбутнього слова, оскільки вони сигналізують, що спікер перезапускає висловлювання/ідею, і тому для розпізнавання промови розглядаються звичайні лексеми як фонеми.

Маркування лексеми як леми (сукупність лексичних форм, що мають однакову основу, ту саму основну частину мови та той самий зміст слова) або як словоформи (повна інфлексована або похідна форма слова) є суттєвою відмінністю для проведення наступного етапу МА як лемматизацію або стемінг, тобто ідентифікації основ слів. Для багатьох NLP-задач на англійській мові достатньо маркувати відповідні лексеми як словоформи, аде для української мови – ні, треба ще ідентифікувати основи слів (наприклад на основі аналізу флексії за деревом закінчень в додатку В).

Є два способи визначення слів з ігноруванням пунктуації – розпізнавання лексем як типів (кількість різних слів $|V|$ у множині слів корпусу, тобто потужність алфавіту корпусу, де елемент алфавіту/словника – унікальне слово) та токенів (загальна кількість N слів аналізованого корпусу), тобто $|V| \leq N$. Найбільший корпус Google N-grams містить 13 мільйонів типів серед тих, що відображаються ≥ 40, тому справжня кількість є набагато більшою.

Співвідношення між кількістю типів $|V|$ і кількість токенів N називається **законом Хердана** (Herdan, 1960) або **законом Хіпса** (Heaps, 1978):

$$|V| = kN^x, \tag{4.14}$$

де k та x – позитивні константи при $0 < x < 1$.Значення x залежить від розміру корпусу та жанру, для великих корпусів x коливається в межах [0,67;0,75], коли розмір словника для тексту зростає значно швидше, ніж квадратний корінь довжини його слів. Іншим показником кількості слів у мові є кількість лем, а не типів слів (наприклад, Oxford English Dictionary – понад 615 000 записів).

4.3. Метод морфологічного аналізу української мови

4.3.1. Особливості морфологічного аналізу україномовних текстів

Морфологія ідентифікує форму речей, а в текстовому аналізі – форму окремих слів/лексем. Лексемами є як слова, так і знаки пунктуації, дозволяють чіткіше провести наступний СА. Структура слів допомагає визначити множину, рід, час, особу, відмінювання тощо. МА є складною задачею, так як в більшості мов є багато винятків з правил і особливих випадків. Основним завданням МА є ідентифікація частин слів для віднесення їх до певних класів (тег) частин мови. Наприклад, іноді важливо зрозуміти, чи є іменник в однині або множині, або є власною назвою. Також треба часто знати, чи дієслово має невизначену форму, минулий час, або це дієприкметник. Отримані частини мови потім застосовують для генерування більших структур (фрагменти/фрази), або цілі дерева слів, які потім використовують для побудови структур даних семантичного міркування.

Після ГА маємо доступ до лексем в реченнях в абзацах інтегрованих текстів контенту, що дає можливість застосувати МА для маркування слів з колекції лексем частинами мови (наприклад, дієслова, іменники, прийменники, прикметники), які вказують на роль слова в контексті речення. В українській мові одне і те ж слово зазвичай в залежності від флексій може приймати різні ролі. Маркування частинами мови на основі МА-правил полягає в додаванні відповідного тегу до кожного слова із колекції лексем, який містить інформацію про визначення слова та його роль у поточному контексті. МА-правила застосовують для розроблення модулів/підсистем ідентифікації ключових слів, рубрикації тексту (Рис. 4.6), машинного перекладу, виправлення помилок, а також для психологічного аналізу людини, семантичного аналізу тощо. При

ідентифікації слів для подальшої класифікації атрибут rub_id описує рубрику, до якої належить конкретне ключове слово (Таблиця 4.4) [535].

Таблиця 4.4

Приклади українських та англійських слів/flag для ідентифікації ключових слів

№	Українське	Англійське	№	Українське	Англійське
1	курсорний/V	cursoriness/17,13	39	буферизувати/ABGH	buffer/18,9,13,17,10,23
2		cursorily	40	відформатувати/AB	format/1,20,17
3		cursor/9,13,17,10	41	кодувати/ABGH	code/17,2,23,10,12,18,9
4		cursory/16	42	кешувати/ABGH	cache/9,17,18,10,13
5	кирилічний/V	Cyrillic	43	кука/ab	hook/10,23,9,18,13,17
6	кілобітовий/V	kilobit/17	44	клавіатурний/V	keyboard/18,9,13,23,10,17
7	кілобіт/efg		45	клавіатура/ab	
8	кілобайтовий/V	kilobyte/17	46	кодосумісний/V	code/17,2,23,10,12,18,9
9	кілобайт/efg		47		code compatible
10	кодек/efg	coder/2,13	48		compatible/17,5
11	кодер/efg		49		compatibleness/13
12	консольний/V	consoled/7	50		compatibility/5,13,17
13		consoler/13	51		compatibly/5
14	консоль/ij	console/23,8,10	52	кодогенератор/efg	code/17,2,23,10,12,18,9
15	Кобол/e	COBOL	53		generators/1
16		Cobol/13	54		generator/17,13
17	кілобод/efg	kilobaud/13	55	конфігуратор/efg	configuration/1,17,13
18	копілефт/e	Copyleft/19,18,17	56		configure/1,10,17,9,8
19	хакер/efg	hacker/13	57	криптозахищений/V	crypto-protected/7,21
20	хеш/e	hash/1,10,17,9	58	криптографічний/V	cryptographic
21	таймер/efg	timer/13	59		cryptographically
22	стек/efgo	stack/13	60		cryptography/13,17
23	спам/e	spam/13	61	копірайт/e	copyright/13,17,18,9,10,23
24	смайл/ef	smile/10,13,9,17,18	62	комутований/V	switch/10,8,23,13,18,17,9,12
25	сайт/ef	site/9,17,12,13	63	конкатенація/ab	concatenate/22,17,9,10
26	рестарт/ef	restart/8	64	комбосписок/ab	combo/13,17
27	рекурсія/ab	recursion/13	65		box/9,18,17,12,23,10,13
28	процесор/efg	processor/13,17	66		list/12,13,18,9,15,10,23,22,17
29	проксі	proxy/17,13	67	крос-компілятор/efg	cross/13
30	принтер/efg	printer/1,13	68		compilable/7
31	подкаст/e	podcast/13	69		compilation/17,1,13
32	плотер/efg	plotter/13,9,17,10	70		compile/1,17,9,2,10
33	піксель/efg	pixel/17,13	71		compiler/2,17
34	опція/ab	option/10,9,13,17	72		compiler's
35	оффлайн/e	offline/13	73	крос-асемблер/efg	cross-assembler/3,13,17
36	онлайн/e	online/13	74	фрейм/efg	frame/17,18,9,12,10,13,23
37	модем/efg	modem/17,13	75	файл/ef	file/6,9,18,17,10,13,23
38	сплайн/efg	spline/13,17,9	76	сигнатура/ab	signature/13,17

4.3.2. Порівняння словників та основних правил для морфологічного аналізу україномовних та англомовних текстів

Прапор (flag) атрибута визначає властивості цього ключового слова (частини мови, до якої воно належить). У тематичних словниках кожне слово має свою властивість, наприклад, a b c d o – різні типи іменників, A – дієслова, V – прикметники (Рис. 4.7) [535]. Для порівняння складності у тематичних словниках

(всього 23 правила) кожне англомовне слово має також властивість, наприклад, числа 1-23 – це номери правил типу PFX (префікси, 1-7 правила) і SFX (суфікси і закінчення, 8-23 правила) та опишіть деякі іменники для англійських слів (Рис. 4.8). Наприклад, правила типу PFX описують модифікацію деяких іменників для англійських слів із префіксами: re-(rule PFX 1), de- (правило PFX 2), dis- (правило PFX 3), con- (правило PFX 4), in- (правило PFX 5), pro- (правило PFX 6) і un- (правило PFX 7) [535].

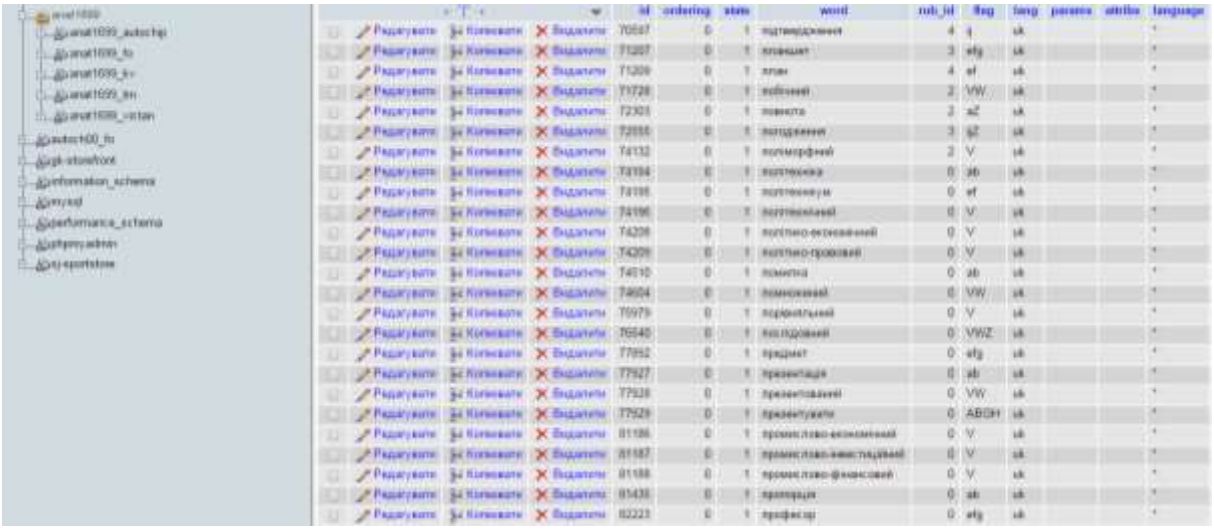


Рис. 4.6. Відношення ключових слів в базі даних КЛС рубрикації тексту

```
# Групи a b c d o
#
# -- Перша відміна: іменники жіночого та чоловічого та середнього роду
#
# -- Друга відміна: іменники чоловічого роду із закінченням на -ар -ир
#                   наголошені (Мінська група на -ар -ир)
#
# -- Друга відміна: іменники чоловічого роду з чергуванням -і -о
#
# -- Числівники -ять, -сят, -сто
#
#

SFX a Y 235

#
# ОДНИНА (множина перенесена в гр. b)
#
# Спочатку перша відміна
#
# тверда група а Називному відмінку однини з закінченням на -а
# однина
SFX a   a н [^жчшщ]а  # хата  хати  (Р.)
SFX a   a і [^гґкх]а  # хата  хаті  (Д.М.)
SFX a   a у а    # хата  хату  (З.)
SFX a   a ою [^жчшщ]а  # хата  хатою  (О.)
```

Рис. 4.7. Словники класифікації іменників для українських слів

Правила типу SFX описують як деякі модифікації іменників для англійських слів із суфіксами або закінченнями (Рис. 4.8) [1013]:

- *-able* [^aeiou], *-able* ee, *-able* [^aeiou]e (правило SFX),
- *-d* e, *-ied* [^aeiou]y, -ed [^ey], *-ed* [aeiou]y (правило SFX 9),

- *-ing* e, *-ing* [^e] (правило SFX 10) та - *ieth* y, *-th* [^y] (правило SFX 11),
- *-ment* (правило SFX 14) та *-ion* e, *-ication* y, *-en* [^ey] (правило SFX 15),
- *-ings* e, *-ings* [^e] (правило SFX 12) та -'*s* (правило SFX 13),
- *-iness* [^aeiou]y, *-ness* [aeiou]y, *-ness* [^y] (правило SFX 16),
- *-ies* [^aeiou]y, *-s* [aeiou]y, *-es* [sxzh], *-s* [^sxzhy] (правило SFX 17),
- *-r* e, *-ier* [^aeiou]y, *-er* [aeiou]y, *-er* [^ey] (правило SFX 18),
- *-st* e, *-iest* [^aeiou]y, *-est* [aeiou]y, *est* [^ey] (правило SFX 19),
- *-ive* e, *-ive* [^e] (правило SFX 20) та *-ly* (правило SFX 21),
- *-ions* e, *-ications* y, *-ens* [^ey] (правило SFX 22),
- *-rs* e, *-iers* [^aeiou]y, *-ers* [aeiou]y, *-ers* [^ey] (правило SFX 23). Літери e та y біля суфіксів є маркерами рішень.

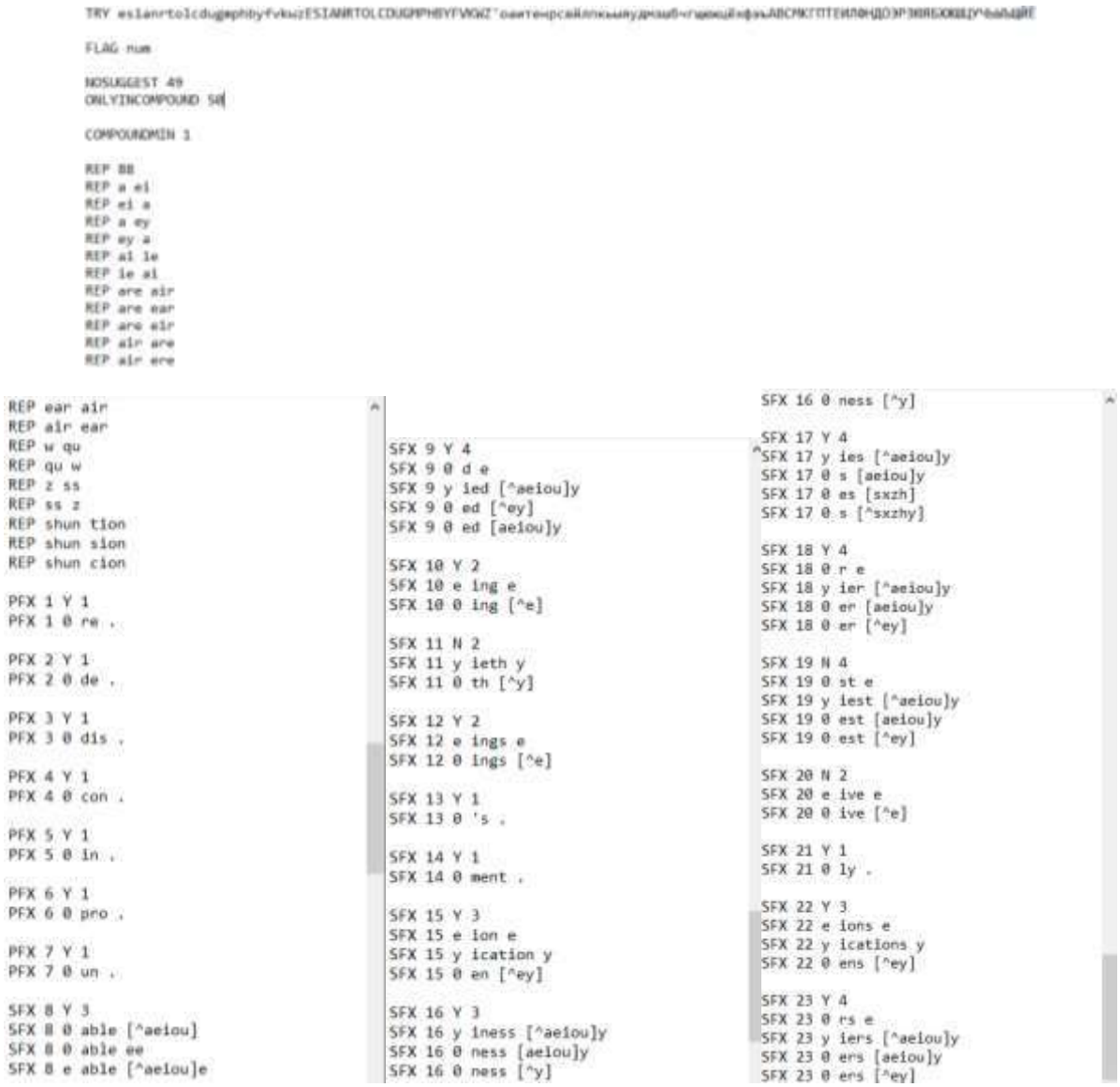


Рис. 4.8. Словники класифікації іменників для англійських слів

4.3.3. Основні правила ідентифікації іменників при аналізі україномовного текстового контенту

Файл афіксів (частин слів, що приєднують до кореня і привносять граматичне або словотворче значення, елементи словотворення, наприклад, префікс, суфікс, постфікс, флексія) має тип файлу *.aff і може містити додаткові атрибути – правила зведення до основи слова (Рис. 4.9) [535].

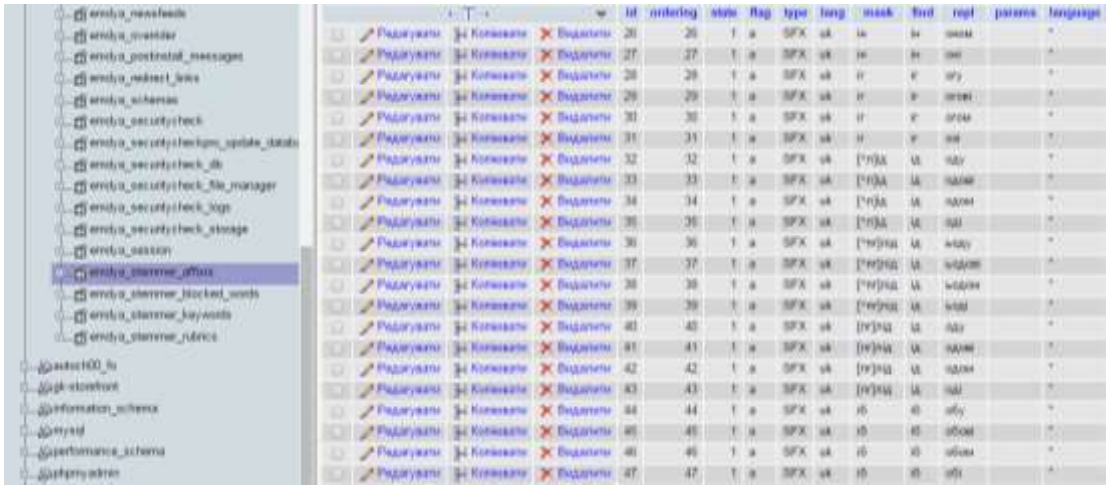


Рис. 4.9. Правила зведення до основи слова типу іменник

Позначення SET застосовують зазвичай для ідентифікації послідовності частин афіксів і довідників. REP формує таблицю підстановки для виправлення кількох символів для слів. TRY ідентифікує послідовності для заміни. SFX і PFX ідентифікують типи суфіксів і префіксів, які марковані афіксами для слів [1013].

Прапор атрибута *flag* визначає тип слова, маска атрибута *mask* показує правило ідентифікації закінчення, значення атрибуту *find* – закінчення слова в називному відмінку, значення атрибуту *repl* – закінчення слова в неназивному відмінку. У квадратних дужках наводяться винятки з правил.

Наприклад, перший рядок (ordering 26) описує конкретний приклад розпізнавання іменників групи а з чергуванням *-і -о* та флексією *-ін* називного відмінку в орудному відмінку (флексія *-оном*), а наступний запис (ordering 27) такі ж іменники, але в місцевому відмінку (флексія *-оні*), але не розпізнає інші правила тієї є групи або інших груп в давальному відмінку – флексії *-онові* та *-ону* (Рис. 4.10). Третій запис (ordering 28) вже розпізнає іменники з чергуванням *-і -о* із флексіями *-іг* називного відмінку в давальному відмінку – флексія *-огу*,

але не розпізнає інші правила тієї ж групи а (це роблять правила 29-31 відповідно): -*огові* (Д.М.), -*огом* (О.), -*озі* (М.). Дев'ятій запис (ordering 34) вже розпізнає іменники із флексіями на -[^*л*]*ід* називного відмінку не після -*л* в орудному з флексією -*одом*, але не розпізнає інші правила тієї ж групи а (відповідно правила 32-33 та 35): -[^*л*]*оду* (Д.Р.), -[^*л*]*одові* (Д.), -[^л]*оді* (М.).

```
# тверда група в Називному відмінку однини з закінченням на -а
# Множина
SFX b   а       0       [^клн]а         # хата  хат                     (Р.)
#    [^ВКЛНРШЧ] А       >       -А,-            # хата > хат                    (Р.)
#    [ШЧ] А             >       -А,ЕЙ           # миша > мишей                  (Р.)
SFX b   а       0       [^ст]ла         # щогла  щогл                   (Р.)
SFX b   ла      ел      [ст]ла          # мітла  мітел                  (Р.)
#    [^К] В А           >       -А,-            # глава > глав                  (Р.)
#    [^Р] К В А         >       -А,-            # буква > букв                  (Р.)
#    Р К В А            >       -ВА,ОВ          # церква > церков               (Р.)
SFX b   а       0       [^сжзвм]на              # батьківщина  батьківщин       (Р.)
SFX b   на      ен      [сжвм]на                # сосна  сосен                  (Р.)
SFX b   на      н       изна            # тризна  тризн                 (Р.)
SFX b   на      ен      озна            # борозна  борозен                      (Р.)
SFX b   а       0       [аеєоуиіїяю]ка          # автоматика  автоматик (Р.)
SFX b   ка      ок      [^аеєоуиіїяю]ка         # відпустка  відпусток          (Р.)
# МР 2002.01.25
#    Ш К А              >       -ШКА,ЩОК        # дошка > дощок         (Р.)
# сестра - в /о

# м'яка група в Називному відмінку однини з закінченням на -я
# множина
SFX b   я       ь       [^іеуиаон'ьлтдр]я               # Вася  Вась    (Р.)
# МР 2001.09.02 родовий відмінок множини на -ря
SFX b   я       0       [^ео]ря         # буря  бур             (Р.)
# +++ ДК - зміни надані від vesna@ 11.06.02
SFX b   я       ь       еря             # вечеря  вечерь        (Р.)
SFX b   оря     ір      оря             # зоря  зір     (Р.)
# --- ДК - зміни надані від vesna@
#
# МР богиня кухня бойня вишня; 2001.09.02 -ЬНЯ,ЕНЬ
SFX b   я       ь       [іеиаоюуїяє]ня          # богиня  богинь        (Р.)
SFX b   ня      онь     [кх]ня          # кухня  кухонь         (Р.)
SFX b   йня     єнь     йня             # бойня  боєнь         (Р.)
SFX b   ьня     ень     ьня             # вітальня  віталень    (Р.)
SFX b   ня      ень     [^іеиаоюуїяєнкхйь]ня            # вишня  вишень         (Р.)
# МР 2001.08.26 -лля -ття -ддя  2001.09.02 -[вмб]ля 2001.09.14 -стя
# МР 2001.09.16  [^С] [С] -сля  (тесля - тесль) ?
SFX b   ля      ель     [^лоеуаієюяїиск]ля              # будівля  будівель     (Р.)
SFX b   я       ь       [лоеуаієюяїиск]ля               # Валя  Валь    (Р.)
SFX b   я       ей      адя             # попадя  попадей       (Р.)
SFX b   я       ів      [^да]дя         # дядя  дядів           (Р.)
SFX b   я       ь       [^тс]тя         # Катя  Кать    (Р.)
SFX b   ля      ей      лля             # рілля  рілей          (Р.)
#    Т Л Я              >       -Я,ЕЙ           # рілля > рілей         (Р.)
SFX b   я       ів      уддя            # суддя  суддів         (Р.)
SFX b   дя      ей      аддя            # баддя  бадей          (Р.)
SFX b   тя      ей      ття             # стаття  статей        (Р.)
SFX b   я       ь       [^о]стя         # причастя  причасть    (Р.)
SFX b   я       ей      остя            # гостя  гостей         (Р.)
# МР                                      Мар'я > Марей ?
SFX b   'я      ей      [^р]'я          # сім'я  сімей          (Р.)
```

Рис. 4.10. Приклад правил морфологічного розбору українських іменників

REP визначає таблицю підстановки для виправлення кількох символів в українських словах [535], наприклад: REP 5; REP сч щ; REP уюч увальн; REP ююч ювальн; REP ємн містк; REP обез зне. Префікс заперечної форми (Рис. 4.11):

- Прикметники із закінченням на *-ий*;
- Прикметники короткої форми на *-ен* змінюються так само як і повної (ясен - ясний...).

Наявність відношення заблокованих модератором слів (Рис. 4.12) [535], зокрема тих, які не можуть бути ключовими, дозволяє зменшити обсяг перевірки при класифікації текстів (Рис. 4.13) [535]. Для ідентифікації ключових слів важливо правильно розпізнати прикметники у будь-якому відмінку, роді та числі (Рис. 4.14) [535].

```
PFX Z Y 1
PFX Z   0       не      .               #  голосний > неголосний


# УВАГА!! не можна використовувати /Y з прапорцем /Z!!
# Сировиною групи /Y є прикметники у порівняльній формі (не 'простий/Y', а 'простіший/Y')

PFX Y Y 1
PFX Y   0       най     .               #  голосніший > найголосніший


# тотожні слова з префіксом "в-" утворені від слів з префіксом "у-"
# можуть створювати надто довгі рядки словоформ, поки що не використовується
PFX X Y 1
PFX X   в       у       в               #  вбити > убити
```

Рис. 4.11. Приклад правил ідентифікації заперечної форми українських слів

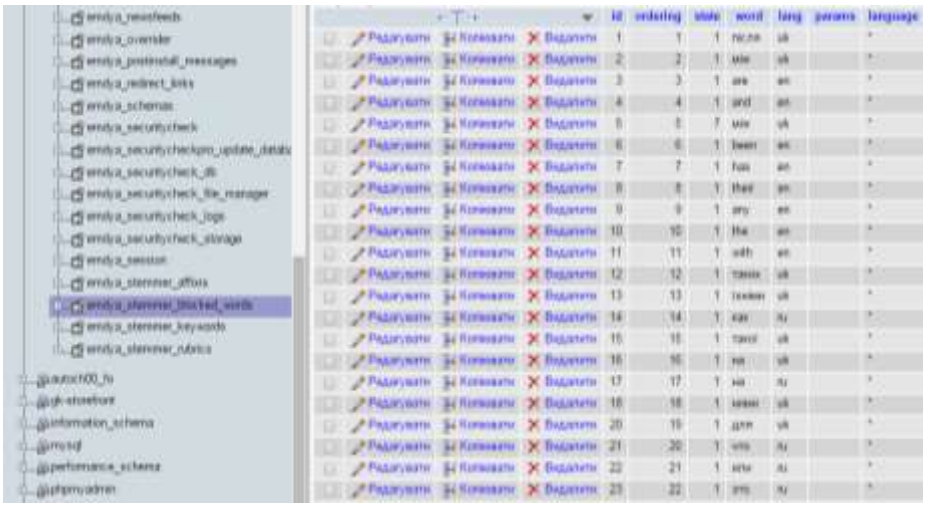


Рис. 4.12. Відношення заблокованих модератором слів

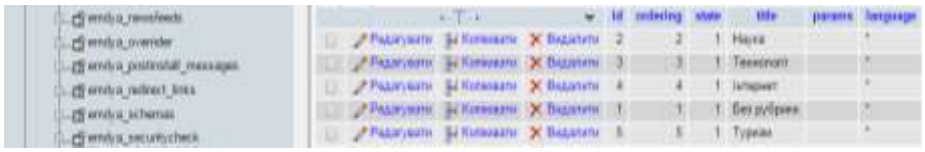


Рис. 4.13. Відношення рубрик

```
# Прикметники із закінченням на -ий
# прикметники короткої форми на -ен змінюються так само як і повної (ясен - ясний...)
#
# Чоловічого роду
SFX V   ий      ого     [^ц]ий          #  відісланий  відісланого      (Р.З.)
SFX V   ий      ому     [^ц]ий          #  відісланий  відісланому      (Д.М.)
SFX V   ий      им      ий              #  відісланий  відісланим       (О. Мн:Д.)
SFX V   ий      ім      ий              #  відісланий  відісланім       (М.)
# Жіночого роду
SFX V   ий      а       [^ц]ий          #  відісланий  відіслана        (Н.)
SFX V   ий      ої      [^ц]ий          #  відісланий  відісланої       (Р.)
SFX V   ий      ій      ий              #  відісланий  відісланій       (Д.)
SFX V   ий      у       [^ц]ий          #  відісланий  відіслану        (З.)
SFX V   ий      ою      [^ц]ий          #  відісланий  відісланою       (О.)
# Середнього роду
SFX V   ий      е       ий              #  відісланий  відіслане        (Н.)
# Множина
SFX V   ий      і       ий              #  відісланий  відіслані        (Н.)
SFX V   ий      их      ий              #  відісланий  відісланих       (Р.)
SFX V   ий      ими     ий              #  відісланий  відісланими      (О.)
#
# Прикметники на -лиций
#
# Чоловічого роду
SFX V   ий      ього    [^у]ций         #  білолиций  білолицього       (Р.З.)
SFX V   ий      ьому    [^у]ций         #  білолиций  білолицьому       (Д.М.)
# Жіночого роду
SFX V   ий      я       [^у]ций         #  білолиций  білолиця  (Н.)
SFX V   ий      ьої     [^у]ций         #  білолиций  білолицьої        (Р.)
SFX V   ий      ю       [^у]ций         #  білолиций  білолицю  (З.)
SFX V   ий      ьою     [^у]ций         #  білолиций  білолицьою        (О.)
#
# Чоловічого роду
SFX V   ий      ого     уций            #  куций  куцого        (Р.З.)
SFX V   ий      ому     уций            #  куций  куцому        (Д.М.)
# Жіночого роду
SFX V   ий      а       уций            #  куций  куца  (Н.)
SFX V   ий      ої      уций            #  куций  куцої         (Р.)
SFX V   ий      у       уций            #  куций  куцу  (З.)
SFX V   ий      ою      уций            #  куций  куцою         (О.)
#
# Прикметники із закінченням на -ій/-їй
#
# Чоловічого роду
SFX V   ій      ього    ій              #  синій  синього       (Р.)
```

Рис. 4.14. Приклад правил морфологічного розбору українських прикметників

Опишемо кожний маркований клас множини правил МА іменників:

- Клас I для іменників маркованих прапорцем *flag* як *a*, *b*, *c*, *d* або *o*:
  - 1 відміна: іменники жіночого та чоловічого та середнього роду;
  - 2 відміна: іменники чоловічого роду із закінченням на *-ар*, *-ир*, наголошені (мішана група на *-ар*, *-ир*);
  - 2 відміна: іменники чоловічого роду з чергуванням *-і*, *-о*;
  - Числівники *-ять*, *-сят*, *-сто*;
- Клас II для іменників маркованих прапорцем *flag* як *e*, *f*, *g* або *h*:

- Іменники другої відміни чоловічого роду з нульовим закінченням;
- Іменники другої відміни чоловічого роду з закінченням на *-о*;

```
SFX A Y 402

# зворотня форма - в групі /B
# складний майбутній час - /AG та складний зворотній час зворотньої форми - /BH
#
# МН (окрім Я, Ти, Він) для всіх закінчень, крім "йти" (йшов, йшла, йшло), та -сти
SFX A   ти      ла      [^с]ти          #  абонувати  абонувала         (Вона)
SFX A   ти      ло      [^с]ти          #  абонувати  абонувало         (Воно)
SFX A   ти      ли      [^с]ти          #  абонувати  абонували (Ми, Ви, Вони)
SFX A   ти      в       [аеиіїоуя]ти            #  абонувати  абонував  (Я, Ти, Він)

# Т -вати (Т недоконаної форми, МБ доконаної)
SFX A   вати    ю       [ауюя]вати              #  абонувати  абоную            (Я)
SFX A   вати    єш      [ауюя]вати              #  абонувати  абонуєш           (Ти)
SFX A   вати    є       [ауюя]вати              #  абонувати  абонує            (Він)
SFX A   вати    ємо     [ауюя]вати              #  абонувати  абонуємо  (Ми)
SFX A   вати    єте     [ауюя]вати              #  абонувати  абонуєте          (Ви)
SFX A   вати    ють     [ауюя]вати              #  абонувати  абонують          (Вони)
# МН -ати
# НФ -вати
SFX A   вати    й       [ую]вати                #  абонувати  абонуй            (Ти)
SFX A   вати    ймо     [ую]вати                #  абонувати  абонуймо  (Ми)
SFX A   вати    йте     [ую]вати                #  абонувати  абонуйте  (Ви)
SFX A   ти      й       [ая]вати                #  ставати  ставай              (Ти)
SFX A   ти      ймо     [ая]вати                #  ставати  ставаймо    (Ми)
SFX A   ти      йте     [ая]вати                #  ставати  ставайте    (Ви)
# Т -рвати, -звати (Т недоконаної форми, МБ доконаної)
SFX A   ати     у       [рз]вати                #  рвати  рву                   (Я)
SFX A   ати     еш      [рз]вати                #  рвати  рвеш          (Ти)
SFX A   ати     е       [рз]вати                #  рвати  рве                   (Він)
SFX A   ати     емо     [рз]вати                #  рвати  рвемо         (Ми)
SFX A   ати     ете     [рз]вати                #  рвати  рвете         (Ви)
SFX A   ати     уть     [рз]вати                #  рвати  рвуть         (Вони)
# НФ -вати
SFX A   ати     и       [рз]вати                #  рвати  рви           (Ти)
SFX A   ати     імо     [рз]вати                #  рвати  рвімо         (Ми)
SFX A   ати     іть     [рз]вати                #  рвати  рвіть         (Ви)
#
# Дієслова з закінченням -зати з чергуванням з/ж (без чергування - гр. /I)
# Т -зати (Т недоконаної форми, МБ доконаної)
SFX A   зати    жу      зати            #  казати  кажу         (Я)
SFX A   зати    жеш     зати            #  казати  кажеш        (Ти)
SFX A   зати    же      зати            #  казати  каже         (Він)
SFX A   зати    жемо    зати            #  казати  кажемо       (Ми)
SFX A   зати    жете    зати            #  казати  кажете       (Ви)
SFX A   зати    жуть    зати            #  казати  кажуть       (Вони)
# МН -зати
# НФ -зати
# варіанти закінчень: -жи (д.ф. зв'яжи), -ж (різати, зарізати - ріж), і в гр. /I -зай (вирізати - вирізай)
SFX A   зати    ж       ізати           #  різати  ріж          (Ти)
SFX A   зати    ж       мазати          #  мазати  маж          (Ти)
SFX A   зати    жи      казати          #  казати  кажи                 (Ти)
SFX A   зати    жи      [еия]зати               #  лизати  лижи                 (Ти)
```

Рис. 4.15. Приклад правил морфологічного розбору українських дієслів

- Клас III для іменників маркованих прапорцем *flag* як *i*, *j* або *k*:
  - Третьої відміни без чергування;
  - Другої відміни середнього роду з закінченням на *-о*, *-а*, *-я*;
  - Другої відміни на приголосну без зак. *-і* в місцевому;
- Клас IV для іменників маркованих прапорцем *flag* як *l*, *m* або *n*:
  - Третьої відміни з чергуванням;
  - Четвертої відміни середнього роду з закінченням на -а, *-я*;

- Група V для іменників маркованих прапорцем *flag* як р: патроніми чоловічого (ч.) та жіночого (ж.) роду однини (о.) та множини (м.) від чоловічих імен.

Для подальшого СА доречним є правильно розпізнати дієслова (Рис. 4.15).

Більш детально опишемо кожний маркований клас множини правил розпізнавання іменників з вказанням їх загальної кількості N (Таблиця 4.5). Всього для МА україномовних іменників використовують біля 1300 правил опрацювання суфіксів та закінчень з врахування чергувань літер (Таблиця А.11).

Таблиця 4.5

Основні МА-правила для маркування іменників при розмічуванні частини мови

Клас	*flag*	N	Особливості МА-правил
I	*a*	248	Для однини: • 1 відміна: іменники жіночого та чоловічого та середнього роду. • 2 відміна: чол. роду на *-ар*, *-ир*, наголошені (мішана група на *-ар*, *-ир*). • 2 відміна: іменники чоловічого роду з чергуванням *-і* та *-о*. • числівники *-ять*, *-сят*, *-сто*.
I	*b*	384	Для множини: • 1 відміна: іменники жіночого та чоловічого та середнього роду. • 2 відміна: чол. роду на *-ар*, *-ир*, наголошені (мішана група на *-ар*, *-ир*). • 2 відміна: іменники чоловічого роду з чергуванням *-і* та *-о*. • іменники в множині із закінченням на *-и*.
I	*c*	54	2 відміна, в род. відмінку однини із закінченням *-а*/*-я*, а саме означення: • істот та осіб: *студента*, *моря*, *Любомира*; • предметів, які піддаються лічбі: *зошита*, *ножа*, *олівця*; • власні населених пунктів: *Ужгорода*, *Тернополя*; • водних об'єктів з наголошеною флексією: *Дніпра*; • виміри: *квадрата*, *міліметра* (але *віку*, *року*); • визначень: *відмінка*; • архітектури: *парника*, *коридора*, *гаража*.
I	*d*	44	• кличний відмінок; • перша відміна (для закінчень [*ая*]); • 2 відміна (закінчення [*рнгдблвк*] з чергуванням *о-і* та випаданням *е*, *о*).
I	*o*	53	Для множини: • 1 відміна: жін./чол./сер. роду з чергуванням *о/і* та появою *о*(*е*) в род. відм.; • 2 відміна: середнього роду на *-о* з чергуванням *о/і* в род. відмінку множини.
II	*e*	19	Для однини: • тверда група іменників з закінченням на *-о*; • тверда група іменники з нульовим закінченням; • тверда група іменників з нульовим закінченням на шиплячий; • мішана група з нульовим закінченням на шиплячі; • м'яка група з закінченням на *-й*, *-ій* або *-ь*.
II	*f*	25	Для множини: • тверда група іменники з нульовим закінченням; • тверда група іменників з нульовим закінченням на шиплячий; • мішана група з нульовим закінченням на шиплячі; • тверда група іменників з закінченням на *-о*; • м'яка група з закінченням на *-й*, *-ій* або *-ь*; • група іменників з закінченням на *-ття*, *-ттів*, викл. з гр. /*і*; • співпадає з род. відм. одн.;

Клас	*flag*	N	Особливості МА-правил
			• іменники з закінченням на *-ок* і випаданням *о* перенесені в гр. а.
II	*g*	3	родовий відмінок другої відміни на *-а*.
II	*h*	5	• друга відміна (потрібні для закінчень на приголосні крім *ЙЖЧШЩ*); • іменники другої відміни чоловічого роду з нульовим закінченням; • співпадає з давальним.
III	*i*	47	Для однини: • іменники третьої відміни жіночого роду з нульовим закінченням; • із закінченням на шиплячий, крім *-ь*; • іменники другої відміни середнього роду з закінченням на -о, *-а* або *-я*; • м'яка група на *-е*, крім шиплячих; • мішана група на шиплячий перед *-е*; • відприкметникові іменники.
III	*j*	66	Для множини: • іменники третьої відміни жіночого роду з нульовим закінченням; • із закінченням на шиплячий, крім *-ь*; • іменники другої відміни середнього роду з закінченням на *-о, -а* або *-я*; • м'яка група на *-е*, крім шиплячих або мішана група; • мішана група на шиплячий перед *-е*.
III	*k*	8	• кличний; • іменники третьої відміни жіночого роду з нульовим закінченням; • однина жіночого роду прикметникових іменників.
IV	*l*	40	Для однини: • іменники третьої відміни з чергуванням; • іменники четвертої відміни середнього роду з закінченням на *-а* або *-я*; • 2 відміна чоловічого роду на *-о*[*дв*]*ець* з випаданням *е* і чергуванням *о-і*; • мішана група 2 відміни на *-яр*; • 2 відміна чол. роду на *-ар*/*-ир*, наголошені (мішана група на - *ар*/*-ир*).
IV	*m*	66	Для множини: • іменники третьої відміни з чергуванням; • іменники четвертої відміни середнього роду з закінченням на *-а*/*-я*; • 2 відміна чоловічого роду на *-о*[*дв*]*ець* з випаданням *е* і чергуванням *о-і*; • мішана група 2 відміни на *-яр*; • 2 відміна чол. роду на *-ар*/*-ир*, наголошені (мішана група на *-ар*/*-ир*).
IV	*n*	9	• з чергуванням *і е* або з чергуванням *і о*; • із закінченням на шиплячий крім *-ь*; • іменники третьої відміни жіночого роду з нульовим закінченням; • мішана група 2 відміни на -яр м'яка група на *-ар*/*-ир*.
IV	*q*	2	• мішана група 2 відміни на *-яр*; • м'яка група на *-ар*/*-ир* (наголошені закінчення при відмінюванні).
V	*p*	222	патроніми чоловічого/ жіночого роду однини та множини від чоловічих імен.

4.3.4. Правила розпізнавання україномовних прикметників та дієслів

Генерувати термінальні ланцюжки англійською мовою досить складно (але МА-правил набагато менше, ні в українській), оскільки наявність артиклів і зв'язок груп іменників один з одним відповідним прийменником робить дерево довшим і ширшим. Генерування термінальних ланцюжків в українській мові ускладняється відмінками та родовими відмінностями зворотів терміну, що вживається в контексті. Для ідентифікації ключових слів не достатньо розпізнати іменники (біля 1300 RE-правил), треба ще ідентифікувати прикметники – всього

99 RE-правил для українських текстів (Таблиця 4.6-Таблиця 4.7, Таблиця А.12) [535]. А для коректного СА та СЕМ, в тому числі побудови онтології необхідно розпізнати дієслова на основі понад 800 RE-правил (Таблиця Д.1 додатку Д).

Таблиця 4.6

Основні МА-правила для маркування прикметників як частини мови [535]

flag	N	Особливості МА-правил розпізнавання прикметників
V	83	• однини із закінченням на *-ий*; • однини короткої форми на *-ен* змінюються так само як і повної (*ясен - ясний*...); • із закінченням на *-лиций*; • із закінченням на *-ій*/*-їй*; • множини із закінченням на *-ій*/*-їй*; • присвійні від іменників 1-ої відміни - назв людей на *-ин*; • присвійні від іменників 2-ої відміни на *-ів* (тверда група); • присвійні від іменників 2-ої відміни на *-їв*.
U	13	• м'яка група присвійних із закінченням на *-ів -> -ев*; • множини із закінченням на *-ів*.
W	3	формування прислівника з прикметника, середній рід порівняльної форми прикметників відповідає відповідному прислівнику у порівняльній формі (*міцніший - міцніше*).

Таблиця 4.7

Основні RE типу SFX українських прикметників на основі goroh.pp.ua [269-276]

№	Flag	Рід	F1	F2	RE	Число	Ознака	Приклад 1	Приклад 2	Відм.	№
1	V	ч	ий	ого	[^ц]ий	одн	на -ий	текстовий	текстового	Р.З.	1
2				ому					текстовому	Д.М.	2
3				им	ий				текстовим	О.Мн:Д.	3
4				ім					текстовім	М.	4
5		ж		а	[^ц]ий				текстова	Н.	5
6				ої					текстової	Р.	6
7				ій	ий				текстовій	Д.	7
8				у	[^ц]ий				текстову	З.	8
9				ою					текстовою	О.	9
10		с		е	ий				текстове	Н.	10
11		-		і		мн			текстові		11
12				их					текстових	Р.	12
13				ими					текстовими	О.	13
14		ч		ього	[^у]ций	одн	на -лиций	білолиций	білолицього	Р.З.	14
15				ьому					білолицьому	Д.М.	15
16		ж		я					білолиця	Н.	16
17				ьої					білолицької	Р.	17
18				ю					білолицю	З.	18
19				ьою					білолицькою	О.	19
20		ч		ого	уций			куций	куцого	Р.З.	20
21				ому					куцому	Д.М.	21
22		ж		а					куца	Н.	22
23				ої					куцої	Р.	23
24				у					куцу	З.	24
25				ою					куцою	О.	25
26		ч	ій	ього	ій		на -ій/-їй	крайній	крайнього	Р.	26
27				ьому					крайньому	Д.	27
28				ім					крайнім	О.Мн.:Д.	28
29		ж		я					крайня	Н.	29
30				ьої					крайньої	Р.	30
31				ю					крайню	Д.	31
32				ьою					крайньою	О.	32
33		с		є					крайнє	Н.	33
34		-	й	-	[іі]й	мн			крайні	Н.	34

№	Flag	Рід	F1	F2	RE	Число	Ознака	Приклад 1	Приклад 2	Відм.	№
35				х					крайніх	Р.	35
36				ми					крайніми	О.	36
37		ч	їй	його	їй	одн		безкраїй	безкрайого	Р.З.	37
38				йому					безкрайому	Д.	38
39				їм					безкраїм	О.М.Мн.:Д.	39
40		ж		я					безкрая	Н.	40
41				йої					безкрайої	Р.	41
42				ю					безкраю	З.	42
43				йою					безкрайою	О.	43
44		с		є					безкрає	Н.	44
45		ч	-	ого	[їи]н		присвійні від іменників 1-ої відміни - назв людей на -ин	мамин	маминого	Р.	45
46				ому					маминому	Д.	46
47				им					маминим	О. Мн:Д.	47
48				ім					маминім	М.	48
49		ж		а					мамина	Н.	49
50				ої					маминої	Р.	50
51				ій					маминій	Д.М.	51
52				у					мамину	З.	52
53				ою					маминою	О.	53
54		с		е					мамине	Н.	54
55				і		мн			мамині	Н.	55
56				их					маминих	Р.	56
57				ими					маминими	О.	57
58		ч	ів	ового	ів	одн	присвійні від іменників 2-ої відміни на -ів, тверда група	татів	татового	Р.	58
59				овому					татовому	Д.	59
60				овим					татовим	О. Мн:Д.	60
61				овім					татовім	М.	61
62		ж		ова					татова	Н.	62
63				ової					татової	Р.	63
64				овій					татовій	Д.М.	64
65				ову					татову	З.	65
66				овою					татовою	О.	66
67		с		ове					татове	Н.	67
68		-		ові		мн			татові	Н.	68
69				ових					татових	Р.	69
70				овими					татовими	О.	70
71		ч	їв	євого	їв	одн	присвійні від іменників 2-ої відміни на -їв, тверда група	Вереміїв	Веремієвого	Р.	71
72				євому					Веремієвому	Д.	72
73				євим					Веремієвим	О. Мн:Д.	73
74				євім					Веремієвім	М.	74
75		ж		єва					Веремієва	Н.	75
76				євої					Веремієвої	Р.	76
77				євій					Веремієвій	Д.М.	77
78				єву					Веремієву	З.	78
79				євою					Веремієвою	О.	79
80		с		єве					Веремієве	Н.	80
81		-		єві		мн			Веремієві	Н.	81
82				євих					Веремієвих	Р.	82
83				євими					Веремієвими	О.	83
1	U	ч	ів	евого	ів	одн	м'яка група присвійні на -ів, -ев	вчителів	вчителевого	Р.	84
2				евому					вчителевому	Д.	85
3				евим					вчителевим	О. Мн:Д.	86
4				евім					вчителевім	М.	87
5		ж		ева					вчителева	Н.	88
6				евої					вчителевої	Р.	89
7				евій					вчителевій	Д.М.	90
8				еву					вчителеву	З.	91
9				евою					вчителевою	О.	92
10		с		еве					вчителеве	Н.	93
11		-		еві		мн			вчителеві	Н.	94
12				евих					вчителевих	Р.	95
13				евими					вчителевими	О.	96
1	W		ий	о	[^жчшщ]ий	-	прислівник	надісланий	надіслано	-	97
2			ій	ьо	ій			синій	синьо		98
3			їй	йо	їй			безкраїй	безкрайо		99

4.3.5. Модифікований алгоритм стеммера Портера

КЛС маркує слова вхідного тексту частинами мови (уточнює після ГА теговані/марковані лексеми як слова) на основі RE-правил та аналізу флексій (згідно древа закінчень з додатку В) як іменників однини відповідного роду та відмінку, іменників множини відповідного відмінку, прикметників, прислівників, дієслів, особових займенників тощо (кожна з колекцією ознак).

МА-модуль повертає колекцію списків абзаців, кожний з яких є списками речень, а ті є списками лексем, в тому числі слів, позначених частинами мови.

Періодичний проміжний аналіз вхідного/інтегрованого текстового контенту дозволяє оцінити, як змінюється тематичний корпус з плином часу.

У процесі аналізу підрахуємо кількість абзаців, речень і слів, а також збережемо кожну унікальну лексему в додатковому проміжному словнику. Якщо лексеми/слова не існувало до цього часу в словнику лексем/основ слів, маркуємо як нову та зберігаємо в проміжному словнику для аналізу модератором. Підрахуємо кількість контенту та категорій в корпусі вхідного текстового контенту і формуємо словник зі статистичним зведенням корпусу, що містить: загальну кількість інтегрованого контенту і категорій; загальна кількість абзаців, речень і слів; кількість унікальних лексем; лексичне різноманіття як відношення кількості унікальних лексем до їх загальної кількості; середня кількість абзаців у контенті; середня кількість речень на абзац; загальний час опрацювання.

Оскільки корпус зростає в міру того, як нові дані збираються, попередньо переробляється і стискаються, МА-метод дозволить обчислити ці ознаки і проаналізувати їх динаміку змін. Це є важливим інструментом контент-моніторингу для ідентифікації можливих проблем в КЛС, наприклад, в ML-моделі значна зміна лексичного різноманіття та кількість абзаців на контент впливає на якість моделі. Тобто МА-метод та ГА-методи окрім ідентифікації лексем та прямого маркування слів частинами мови застосовують для збору додаткової інформації при визначенні величини змін корпусу з метою своєчасного початку подальшої векторизації та реструктуризації ML-моделі.

Основним етапом МА-методу є ідентифікація основ слів (стемінг) без врахування флексій (суфіксів та закінчень) та в деяких випадках – префіксів. За змістом флексій ідентифікують частину мови слова (Рис. 4.16).

```
var $ADJECTIVE =// Прикметник
'/(ими|ій|ий|а|е|ова|ове|ів|є|їй|єє|еє|я|ім|ем|им|ім|их|іх|ою|йми|іми|у|ю|ого|ому|ої)$/';
// Дієприкметник
 var $PARTICIPLE = '/(ий|ого|ому|им|ім|а|ій|у|ою|ій|і|их|йми|их)$/';
//Дієслово
 var $VERB = '/(сь|ся|ив|ать|ять|у|ю|ав|али|учи|ячи|вши|ши|е|ме|ати|яти|є)$/';
 var $NOUN =// Іменник
'/(а|ев|ов|е|ями|ами|еи|и|ей|ой|ий|й|иям|ям|ием|ем|ам|ом|о|у|ах|иях|ях|ы|ь|ию|ью|ю|
ия|ья|я|і|ові|ї|ею|єю|ою|є|еві|ем|єм|ів|їв|\'ю)$/';
```

Рис. 4.16. Приклад ідентифікації форм флексій відповідно до частини мови

Для наступного СА цього недостатньо (маркувати слово лише частиною мови), треба ще визначити наприклад для іменника/прикметника рід/відмінок тощо. Класичний алгоритм стеммера Портера працює послідовним відсіканням закінчень і суфіксів. Для англомовних текстів це не є проблемою, так як там флексій дуже мало. Для українських слів треба застосовувати модифікований (розширений) алгоритм стеммера Портера з перевіркою як додаткових флексій в залежності від частини мови (згідно дерева закінчень), так і отримані основи слів із словником основ для ідентифікації наявного слова (Рис. 4.17).

Алгоритм 4.1. Модифікований алгоритм стеммера Портера

Етап 1. Ідентифікувати наступну лексему як слово w_i ($w_s = w_i$).

Етап 2. Перевірити з словником стоп-слів $D_{w_{sw}}$ чи w_s є службовим словом. Якщо так, то $i = i + 1$ та перейти до етапу 1, інакше – до етапу 3.

Етап 3. Перейти до кінця слова w_s. Розпізнати флексію f_1^i в w_s із всіх можливих (Рис. 4.16 – обирається найдовша, наприклад, у w_s=*текстова* обираємо закінчення f_1^i=*ова*, а не $f_1^i \neq a$) з RE типу слів як $R_{adjectival}$, R_{noun}або R_{verb} та при наявності видалення флексії f_1^i (Рис. 4.18).

Етап 4. Збереження флексії f_1^i у тегу слова w_i.

Етап 5. Маркувати w_i. як тип $m_{adjectival}^{w_i}$, $m_{noun}^{w_i}$або $m_{verb}^{w_i}$ відповідно.

Етап 6. Знаходження видаленої флексії f_1^i в дереві флексій $T_{flection}$ (додаток В – обирається найдовша). Перевірка вмісту піддререва $T_{flection}^{f_1}$ з наявним

закінченням слова f_2^i $(f = f_2^i + f_1^i)$. Якщо w_s. закінчується на f_2^i та має відповідник в $T_{flection}^{f_1}$, то зберігаємо в $f_i = f$ та видаляємо в w_s.

Етап 7. Отриману основу w_s початкового слова w_i перевіряємо із змістом словника основ D_{w_s} слів української мови. При відсутності відповідника зберігаємо $< w_i, w_s >$ в додатковому тимчасовому проміжному словнику $D_{<w_i,w_s>}$ для модератора та перехід до етапу 1, інакше перехід до етапу 4.

Етап 8. Аналіз флексії та наявності/відсутності чергування літер в основі/флексіях слів $< w_i, w_s >$ і аналогу основи слова в D_{w_s} згідно відповідного RE-правила МА для ідентифікації додаткових ознак аналізованого слова w_i.

Етап 9. Дописування ідентифікованих лінгвістичних ознак розпізнаної частини мови до тегу слова w_i типу $m_{adjectival}^{w_i}$, $m_{noun}^{w_i}$або $m_{verb}^{w_i}$ відповідно. Збереження результатів у відповідний словник D_{w_i} аналізованого тексту.

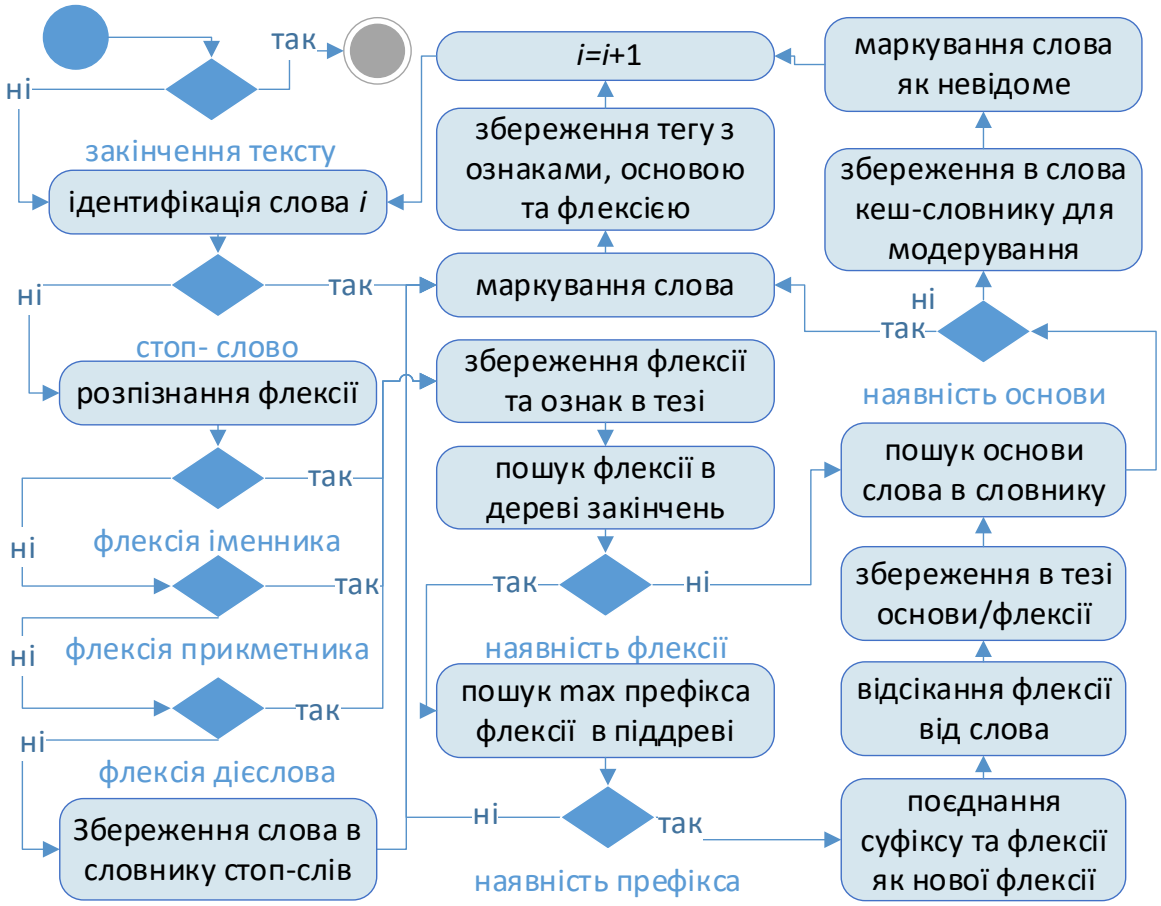


Рис. 4.17. Модифікований алгоритм стемінгу

```
var $VOWELS = '/(а|е|і|і|о|у|и|е|ю|я)/' # українські голосні літери
var RV # частина слова після першої VOWELS. RV=0, якщо VOWELS=0 в W_i відсутні.
# Всі перевірки f проводяться над RV. Літери перед RV не беруть участь взагалі.
# Так, при перевірці PARTICIPLE наступні а/я також повинні бути всередині RV.
var R1 # частина W_i після 1-го рядка VOWELS+NOVOWELS.
var R2 # частина R1 після 1-го рядка VOWELS+NOVOWELS.
# Наприклад, W_i=інформаційний: RV = нформаційний, R1 = формаційний, R2 = маційний.
Class 1. ADVERB # дієприслівник
  Group 1: '/[а|я]?(в|вши|вшися)/'
  Group 2: '/(ив|ивши|ившися)/'
Class 2. ADJECTIVE # прикметник
  '/(а|е|і|и|ими|іми|ій|ий|ім|ім|им|ього|ого|ьому|ому|іх|их|ую|юю|ая|яя|ою|ею)/'
Class 3. PARTICIPLE # дієприкметник
  Group 1: '/[а|я]?(вш|ювa|ува|уч|ючl)/'
  Group 2: '/(нн|н|ячи|ачи|ова|ову|ем)/'
Class 4. REFLEXIVE = '/(ся|сь)/' # рефлексивна флексія
Class 5. VERB # дієслово
  Group 1: '/[а|я]?(ла|е|ете|йте|ли|люю|й|в|ем|емо|ний|ло|ть|но|ють|ні|ть|еш)/'
  Group 2: '/(ила|ела|ена|йте|ите|єте|юй|уй|ій|ай|ало|ив|или|имо|ений|ило|іло|
                ено|ють|ать|ені|ять|іть|ить|иш|ую|ю)/' #
  Class 6. NOUN = '/(а|ев|ов|і|тя|е|ами|іями|ями|ей|ею|ями|ям|ії|и|ою|ій|ой|ий|
                й|им|им|ім|ам|ом|о|у|ах|ях|ую|ю|ія|я)/' # іменник
  Class 7. SUPERLATIVE = '/(ш|іш)/' # ступінь порівняння - найдовший, миліший
  Class 8. DERIVATIONAL = '/[ість]?/' # словотворча флексія - милість, щедрість
  Class 9. ADJECTIVAL # (ADJECTIVE|PARTICIPLE+ADJECTIVE): падаюча = пада+юч+а.
```

Рис. 4.18. Класи лінгвістичних ознак флексій морфологічного аналізу

Зростання обсягів RE-правил МА збільшує у геометричній прогресії навантаження на КЛС лише із-за розпізнавання флексій та основ словоформ. Для англомовних текстів складність менша із-за декількох параметрів, наприклад для іменників 2 відмінка – 2 флексії в множині (s|es). Для німецької мови складність збільшується – 4 відмінка (але флексії майже не змінюються, змінюються лише артиклі), разом пишуться словосполучення з ≥ 2 слів тощо. В українській мові – 7 відмінків іменника, та кожному з яких змінюються флексії в залежності від роду та множини/однини та деякі слова мають в деяких відмінках різні закінчення (наприклад для втручання в місцевому відмінку два варіанти – втручанню, втручанні), крім того часто присутнє чергування літер.

Тому для українські слів простий класичний алгоритм стемінгу Портера не підходить (приведення слова до основи-кореня із відсіканням флексій). Краще поєднати такий алгоритм з пошуком/перевіркою отриманих проміжних результатів з деревом флексій (щоб не перебирати всі можливі флексії) та з вмістом тематичних словників основ з множиною RE-правил ідентифікації ознак (класифікація за частинами мови). Лише для рубрикації тексту на основі ідентифікації слів достатньо провести МА лише для деяких іменникових груп

(прикметники з іменниками та іменники з іменниками) без аналізу слів інших частин мови (розпізнавання за деревом флексій – не прикметник та не іменник – ігнорувати, крім того ключові мають знаходитися поряд і лише між іменниками можуть інколи бути 1 прийменник. Достатньо ідентифікувати в тексті саме основи іменників/прикметників/абревіатур та аналізувати їх ймовірність скупчення в різних частинах контенту відносно загального обсягу.

Класичний алгоритм стемінгу – Стеммер Портера – не використовує словників основ слів, а лише застосовує послідовно множину RE-правил відсікання флексій відповідно до особливості конкретної мови. Алгоритм працює з окремими словами без аналізу та врахування контексту. Не враховують лінгвістичні ознаки як особливості словотворення (префікс, суфікс тощо) та частини мови (іменник, дієслово тощо). Основу складають такі прийоми до слів:

- відсікання флексії від аналізованого слова (для українських слів можлива реалізація з перевіркою отриманих основ та флексій з аналогами в БД).
- слово має незмінну флексію (для більшості українських слів неможлива умова, але можна ідентифікувати частки, сполучники, прийменники, деякі іншомовного походження іменники, абревіатури тощо).
- змінює флексію при відмінюванні через випадання/чергування літер.
- зміна флексії слова та словотворення відповідає конкретному RE-правилу, наприклад при утворенні слів з деяких дієслівних груп:

(ов)*ува(ти|нню|нням|нні|ння|ли|ло|ла|вшись|вши|в|вся|всь|лися|лись|тися|тись).

- зміна флексії слова як виключення з RE-правил.
- закінчення слова збігається з конвертним RE-правилом ідентифікації флексії, але саме слово немає флексії: *вітер*, але *відер*.
- більшість коротких слів є незмінними (достатньо словника стоп-слова).

Такі прийоми значно ускладнюють алгоритм стемінгу українських слів. Тому спочатку аналізують розповсюджені флексії, наприклад, для 1 літери ц (34), щ (110), ф (214), б (281), п (341), ж (353), з (581), г (636), л (754), с (914), ч (959), д (1038), н (2531), р (2709) або 1-4 літер (Таблиця 2.2). Флексій ≥ 5

(наприклад, max(йтесь)=6837, max(ванням)=4656 – додаток В) значно менше серед ключових слів, тому для швидкості/оперативності розв'язку в деяких NLP-задачах КЛС їх ігнорують, але для СА/СЕМ це не припустимо.

4.3.6. Особливості застосування морфологічного аналізу

Багато NLP-задач не вимагають повної реалізації всіх NLP-процесів від графемного до прагматичного аналізів. Наприклад для ідентифікації ключових слів достатньо привести графемний та морфологічний аналіз (алгоритм 4.2). А ось перед майже будь-яким NLP-процесом текст має бути нормалізований.

Алгоритм 4.2. Скорочене наївне опрацювання текстового контенту

Етап 1. Груба токенізація (або графемний аналіз) спецсимолів вхідного тексту.

Крок 1.1. Зчитування тексту та видалення повторних пробілів підряд та тегів при їх наявності (якщо текст інтегрований з Web-ресурсу), послідовно маркуючи службові символи початку/кінця абзацу/заголовку/тексту тощо.

Крок 1.2. Графемний парсинг та сегментування між службовими символами або тегами вхідного тексту X, послідовно маркуючи кожну послідовність неалфавітних символів як лексеми та розпізнаючи алфавітні послідовності між пробілами та іншими спецсимволами (наприклад, цифри та пунктуацію) згідно RE-правил як слова-лексеми для формування списку S ідентифікованих алфавітних токенів як слова w_i.

Крок 1.3. Відсортувати список $S \rightarrow S_A$. ідентифікованих токенів w_i за алфавітом, підраховуючи входження однакових ланцюжків та формуючи алфавітно-частотний словник D_a, запис якого є у вигляді *кількості появи – слово*.

Крок 1.4. Переведення всіх літер верхнього реєстру в нижній та перерахунок входжень слів-токенів у алфавітно-частотний словник $D_A \rightarrow D_a$.

Крок 1.5. Відсортувати та зберегти словник $D_a \rightarrow D_N$. ідентифікованих слів w_i за зменшенням частоти появи (в германських мовах в топі будуть артиклі, займенники, прикметники та сполучники, а в слов'янських мовах більшість слів з одною основою та різними флексіями будуть займати різні рядки списку, що значно спотворює картину реального розподілу слів в тексті).

Етап 2. Сегментування/токенізація слів аналізованого текстового контенту.

Крок 2.1. Сегментація слів на основі словників, метрик типу ймовірність помилки в слові, моделі статистичної послідовності, попередньо навченої від сегментованих корпусів текстів (між пробілами, пунктуацією тощо).

Крок 2.2. Токанізація на основі RE-правил маркованих лексем типу послідовності неалфавітних символів як токенів (дати, ціни, URL-адреси, хештеги, адреси е-пошти тощо), пунктуації (як кінця речення, чи межі підрядного речення), мішаних лексем алфавітно-неалфавітних символів (скорочень, складних дефісних слів, із апострофом тощо), рядків з символами верхнього реєстру (як початку речення, географічних назв, власних назв, абревіатур) та їх нормалізація при необхідності (наприклад, *к.т.н.* $\rightarrow$ *ктн* як окреме слова-токен або *ML* як *машинне навчання*).

Крок 2.3. Аналіз токенів з символами верхнього реєстру (крім при наявності великих лише перших літер) для маркування на основі RE-правил скінченних автоматів або як абревіатур або передачі емоції.

Крок 2.4. Маркування неідентифікованих лексем D_x та неоднозначності (наприклад, апостроф як частина слова тощо).

Етап 3. Лематизація множини розпізнаних та маркованих алфавітних токенів тексту як лем, ідентифікованих як слова аналізованого тексту.

Крок 3.1. Нормалізація токенів на основі ідентифікації афіксів з дерева закінчень (додаток В) як стенокардія маркованих токенів-слів (приведення слова до початкової форми на основі RE-правил МА з Таблиця 4.1-Таблиця 4.7 та Таблиця А.11-А.12 додатку А для ідентифікації коренів та афіксів через алгоритм 4.1 модифікованого стеммера Портера), тобто визначення чи аналізовані токени мають однаковий корінь та різняться лише флексією з послідовним ідентифікацією частини мови аналізованих слів з подальшим маркуванням їх як леми з всіма супутніми лінгвістичними ознаками.

Крок 3.2. Перегрупування та перерахунок частот слів в алфавітно-частотному словнику $D_N \rightarrow D_l$ з врахування нормалізованих слів-лем на кроці 3.1.

Етап 4. Додатковий аналіз неідентифікованих лексем $D_x \neq \varnothing$ шляхом ітераційного поєднання частих пар символів/рядків в межах слів-лексем (наприклад, чи лексеми між пробілами або іншими знаками пунктуації *контент-аналіз*, *Web-сайт*, *контент-моніторинг* або *Web-resource* є одним словом, чи двома) через кодування бітової пари, або BPE [73, 1014] на основі стиснення тексту [73, 1015] для подальшого можливої ідентифікації слів, їх маркування та нормалізації.

Крок 4.1. Формування набору символів рівних колекції властивостей з $D_x \neq \varnothing$. Кожне слово подаємо як послідовність символів плюс спеціальний символ кінцевого слова або спецсимвол, наприклад тире – в межах лексеми (наприклад, *контент-*, *Web-*, *контент-* або *Web-*). Позначаємо $i = 0$.

Крок 4.2. Обчислення кількості n_l кожної пари символів/рядків $\forall(s_k^x, s_j^x)$ як появи основ слів у вхідному тексті при $\forall\{s_k^x \in D_x,\ s_j^x \in D_l\}$ або $\forall\{s_k^x \in D_l,\ s_j^x \in D_x\}$, які знаходяться поряд та розмежовані спецсимволом тире (складні слова) крапка (дата), кома (дійсне число) і/або пробілом, або їх комбінацією, але не знаками пунктуації, цифрами та іншими спецсимволами.

Крок 4.3. Формування алфавітно-частотного словника D'_x на основі $\forall(s_k^x, s_j^x)$. Визначення кількості входжень унікальних лексем в $D'_x \to h = |D'_x|$.

Крок 4.4. Знаходження $n_l = max$ найчастішої пари $a_i = (s_k^x, s_j^x)$ в D'_x, де $\exists(s_k^x, s_j^x) \in D'_x$, $\exists\{s_k^x \in D_x, s_j^x \in D_l\}$ або $\exists\{s_k^x \in D_l, s_j^x \in D_x\}$.

Крок 4.5. Заміна a_i новим поєднанням/злиттям символом/рядком $b_i = s_k^x\ s_j^x$.

Крок 4.6. Вилучення з D'_x значення $s_k^x\ s_j^x$ та з D_x значень s_k^x або s_j^x відповідно.

Крок 4.7. Обчислення кількості появи у вхідному тексті b_i, появи s_k^x та s_j^x при $\exists s_k^x \in D_l$ і/або $\exists s_j^x \in D_l$ відповідно, коли вони окремо вживані (не поряд).

Крок 4.8. Включення в D_l значення b_i, та частоти його входження. Перезапис значень частот в D_l для s_k^x та s_j^x при $s_k^x \in D_l$ і/або $s_j^x \in D_l$ відповідно.

Крок 4.9. Позначаємо $i = i + 1$. Якщо $h > 0$, $D_x \neq \varnothing$ для $\forall n_l > 1$ та в D'_x є хоч 1 марковане значення b_i, то перехід до кроку 4.4, інакше ($h = 0$ та $D_x = \varnothing$ або

$D_x \neq \varnothing$ при $\forall n_l = 1$ або $\forall s_k^x$ не знайдено жодної неунікальної пари $\forall s_j^x$ для утворення $a_i = (s_k^x, s_j^x)$) – до етапу 5.

Етап 5. Сегментування речень в аналізованому контенті (пункти 4.1.4 -4.1.5).

4.4. Метод лексичного аналізу української мови

4.4.1. Особливості методу лексичного аналізу україномовних текстів

Процес лексичного аналізу україномовного тексту C'_γ полягає в парсингу, сегментації та токенізації кожного речення окремо, яким властивий не строгий порядок слів, але паралельно сталий порядок розташування окремих лінгвістичних одиниць. В повному простому україномовному реченні з прямим порядком слів структурна схема є умовно фіксованою. Основними лексичними категоріями відповідного речення є іменна та дієслівна групи. Граматика типу 0 за класифікацією Н. Хомські не доречна для таких речень із-за складності реалізації. При контекстно-залежній граматиці застосовують конкретні обмеження, зокрема, на структуру україномовного речення з деякою множиною варіацій. На основі синтаксичних правил генерування україномовних речень з частковим порядком слів (наприклад, немає строгого порядку для підмета та присудка в реченні, але прикметник зазвичай є перед іменником або іншим прикметником, якщо це не поетичний уривок, також лексичні одиниці *іменникової групи* розміщуються навколо підмета тощо), виведемо лексичну схему для іменної групи $\tilde{S}$ на основі регулярних виразів:

$$\tilde{S} = ([A]\{0, n\}[S]\{1, m\}|[P]), \tag{4.15}$$

де $A = a_1 a_2 a_3 \dots a_{N-1} a_N$ – послідовність прикметників, а запис $[A]\{0, n\}$ – вибір від 0 до n прикметників з $a_1 a_2 a_3 \dots a_{N-1} a_N$, причому $n<N$; $S = s_1 s_2 s_3 \dots s_{M-1} s_M$ – послідовність іменників, а запис $[S]\{1, m\}$ – вибір від 1 до m іменників з $s_1 s_2 s_3 \dots s_{M-1} s_M$, причому $m<M$; $P = p_1 p_2 p_3 \dots p_{K-1} p_K$ – послідовність займенників, а запис $[P]$ – вибір 1 займенника з $p_1 p_2 p_3 \dots p_{K-1} p_K$; запис $(x|y)$ – вибір або x, або y; значення a_i та s_j узгоджуються за родом, числом та відмінком.

Відповідно для *дієслівної групи* лексична схема на основі RE-виразів:

$$\tilde{V} = ([V]\{1,n\}[\widetilde{S'}]\{0,m\}|[\widetilde{S'}]\{0,m\}[V]\{1,n\}), \quad (4.16)$$

де $V = v_1 v_2 v_3 \dots v_{N-1} v_N$ – послідовність дієслів, а запис $[V]\{1,n\}$ – вибір від 1 до n дієслів з $v_1 v_2 v_3 \dots v_{N-1} v_N$, причому $n<N$; $\widetilde{S'} = \tilde{S}_1 \tilde{S}_2 \tilde{S}_3 \dots \tilde{S}_{M-1} \tilde{S}_M$ – послідовність іменникових груп, а запис $[\widetilde{S'}]\{0,m\}$ – вибір від 0 до m іменникових груп з $\tilde{S}_1 \tilde{S}_2 \tilde{S}_3 \dots \tilde{S}_{M-1} \tilde{S}_M$, причому $m<M$; запис $(x|y)$ – вибір або x, або y; узгодження між v_i та $\tilde{S}_j$ провадиться за особою, родом та числом.

Лексична схема *українського речення* на основі RE-виразів:

$$R = ([\widetilde{S'}]\{0,1\}[\widetilde{V'}]\{0,1\}|[\widetilde{V'}]\{0,1\}[\widetilde{S'}]\{0,1\}), \quad (4.17)$$

де $\widetilde{V'} = \tilde{V}_1 \tilde{V}_2 \tilde{V}_3 \dots \tilde{V}_{N-1} \tilde{V}_N$ – послідовність дієслівних груп, а запис $[\widetilde{V'}]\{0,1\}$ – вибір від 0 до 1 дієслівних груп з $\tilde{V}_1 \tilde{V}_2 \tilde{V}_3 \dots \tilde{V}_{N-1} \tilde{V}_N$ з наявністю присудка; $\widetilde{S'} = \tilde{S}_1 \tilde{S}_2 \tilde{S}_3 \dots \tilde{S}_{M-1} \tilde{S}_M$ – послідовність іменникових груп, а запис $[\widetilde{S'}]\{0,1\}$ – вибір від 0 до 1 іменникових груп з $\tilde{S}_1 \tilde{S}_2 \tilde{S}_3 \dots \tilde{S}_{M-1} \tilde{S}_M$ з наявністю підмета; запис $(x|y)$ – вибір x або y; узгодження між $\tilde{V}_i$ та $\tilde{S}_j$ провадиться за особою, родом та числом.

Основними лексичними ознаками дієслівної групи є час, число, особа.

Для порівняння лексична схема іменникової групи на основі RE-виразу для англомовного речення:

$$\tilde{S} = (article[A]\{0,n\}[S]/of[A]\{0,n\}[S]/\{0,m\}|[P]). \quad (4.18)$$

Лексична схема дієслівної англомовної групи на основі RE-виразу:

$$\tilde{V} = [V][\widetilde{S'}]\{0,m\}. \quad (4.19)$$

Лексична схема для англомовного речення основі RE-виразу:

$$R = [\widetilde{S'}][\widetilde{V'}]. \quad (4.20)$$

Узгодження відмінків між лексичним одиницями україномовного речення впливає на подальший синтаксичний та семантичний аналіз контенту:

$$\left.\begin{array}{l} 1.\ R \to RY_i x_i^{'}, \\ 2.\ x_i^{'} Y_j \to Y_j x_i^{'}, \\ 3.\ RY_i \to x_i R, \\ 4.\ R \to q. \end{array}\right\} i,j = 1,2,3, \quad (4.21)$$

де x_i, x_i', q – основні лексичні одиниці; R, Y_i – допоміжні лексичні одиниці; R – початковий символ як індикатор типу генерування ланцюга речення.

Етапи лексичного формування ланцюга токенів $x_2 x_1 x_1 x_3 q x_2^{'} x_1^{'} x_1^{'} x_3^{'}$:

1. R
2. (1) RY_3x_3'
3. (1) $RY_1x_1'Y_3x_3'$
4. (1) $RY_1x_1'Y_1x_1'Y_3x_3'$
5. (1) $RY_2x_2'Y_1x_1'Y_1x_1'Y_3x_3'$
6. (2)$RY_2Y_1x_2'x_1'Y_1x_1'Y_3x_3'$
7. – 11. ..... (2; 5 разів)
12. (3) $x_2RY_1Y_1Y_3x_2'x_1'x_1'x_3'$
13. – 15. ..... (3; 3 рази)
16. (4) $x_2x_1x_1x_3qx_2'x_1'x_1'x_3'$

4.4.2. Приклади застосування методу лексичного аналізу україномовних текстів

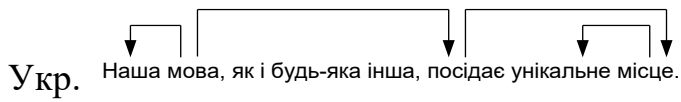
Приклад лексичного генерування типу $\{xqx'\}$: $\overset{a}{\text{Саша}}$, $\overset{b}{\text{Софія}}$, $\overset{c}{\text{Катя}}$, $\overset{d}{\text{Данило}}$, … – $\overset{a'}{\text{спортсмен}}$, $\overset{b'}{\text{співачка}}$, $\overset{c'}{\text{художниця}}$,$\overset{d}{\text{поет}}$, … *відповідно*, де x $(abcd\ldots)$ – послідовність власних імен, x' $(a'b'c'd'\ldots)$ – послідовність професій, узгоджені з власними іменами; q – тире. Будь-яке дієслово має здатність виступати доповненням: *моя дитина вподобала книгочитання*. Цей процес теоретично може бути повторений необмежене число разів: *він книгочитанняцікаводумає про книгочитанняцікавість*, тобто

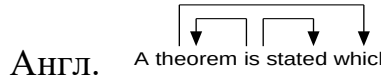
$$\textit{Він}\ \overbrace{\textit{книго}}^{a}\overbrace{\textit{читання}}^{b}\overbrace{\textit{цікавість}}^{c}-\textit{думає про}-\overbrace{\textit{книго}}^{a'}\overbrace{\textit{читання}}^{b'}\overbrace{\textit{цікавість}}^{c'}.$$

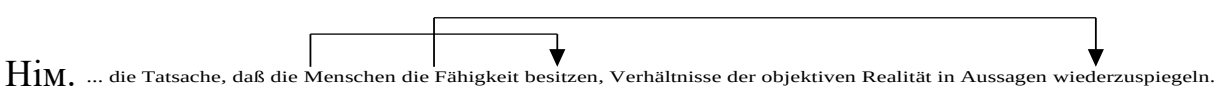
Мова, яка полягає з ланцюжків вигляду $abcd\ldots d'c'b'a'$ (складених із символів a_1, a_2, a_3, a_1', a_2', a_3'), породжується граматикою із 6 правил:

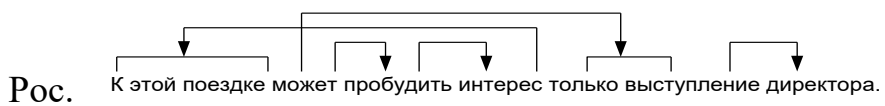
$$\left.\begin{matrix} I \to a_i I a_i^{'} \\ I \to a_i a^{'} \end{matrix}\right\} i = 1,2,3. \qquad (4.22)$$

Такі граматики не забезпечують, наприклад, природного опису для так званих непроектних конструкцій як з розривами, перетином ( , ) або обрамленням ( , ) напрямів синтаксичної залежності (Рис. 4.19).









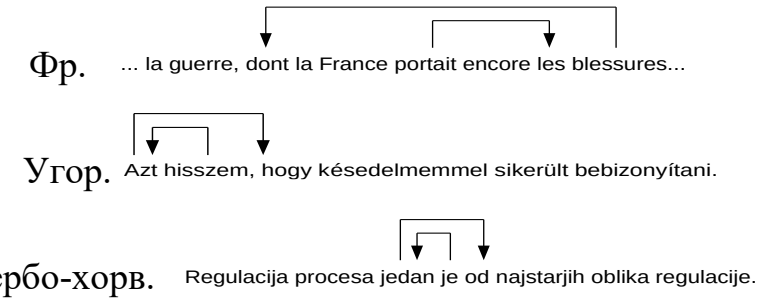


Рис. 4.19. Приклади природного опису для так званих непроектних конструкцій

Для опису таких конструкції речень застосовують:

1. Праве підпорядкування: назва курсу, лист бумаги, une regle stricte, give him,

2. Ліве підпорядкування: основний курс, белый лист, cette regle, good advice.

3. Послідовне підпорядкування (Рис. 4.20):

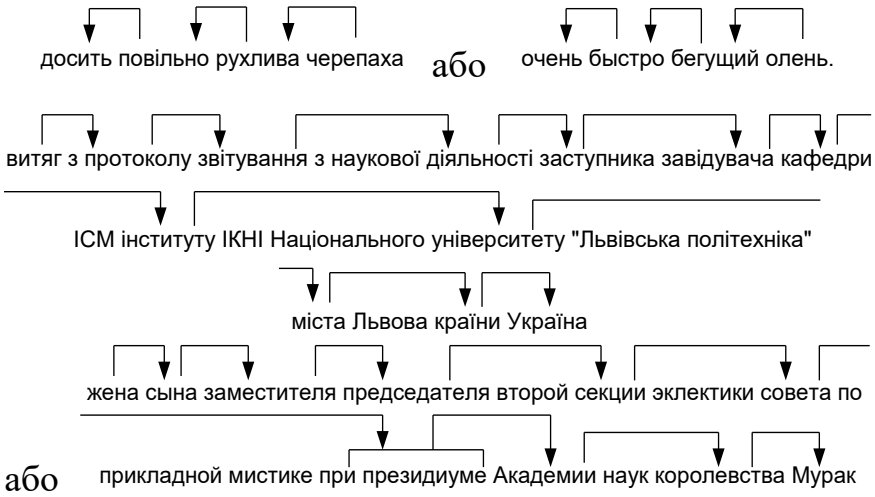


Рис. 4.20. Приклади природного опису послідовного підпорядкування

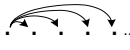
Лише при коректній ідентифікації та розпізнавання непроектних конструкцій можна провести граматично-синтаксичний аналіз україномовних речень для побудови відповідно дерев залежностей складових цих речень.

4.5. Метод синтаксичного аналізу української мови

4.5.1. Особливості синтаксичного аналізу україномовних текстів

Синтаксис – це множина реляційних правил формування речень/фраз, яка зазвичай визначається граматикою. Речення – це лінгвістичні одиниці мови для генерування сенсу та кодування інформації. Метою СА є демонстрація значущих зв’язків між словами на основі поділу речення на частини, або між лексемами в

деревоподібній структурі C'_{λ}. Синтаксис є необхідною основою для міркувань про систему понять або семантики, адже він є важливим інструментом для визначення ступеня впливу слів один на одного при генеруванні фраз. Наприклад, СА ідентифікує прийменникову фразу *в потяг* та іменну фразу *чемодан в потяг* як складові дієслівної фрази *заніс чемодан в потяг*.

Для будь-якого термінального ланцюжка, що виводиться (Рис. 4.21-Рис. 4.22), наявне таке виведення в кожному реченні займає $\leq k$ останніх позицій справа. Необхідне виконання множини вимог, які призводять до послідовного виведення типу  або вкладеного  :

Приклад 4.1. $P = \{\tilde{S}_{x,y,z} \to S_{x,y,z}\tilde{S}_{x',y',p}, \tilde{S}_{x,y,z} \to \tilde{A}_{x,y,z}\tilde{S}_{x,y,z}, \tilde{S}_{x,y,z} \to S_{x,y,z},$
$\tilde{A}_{x,y,z} \to \{дуже, досить, точно, просто, суттєво, \dots\}A_{x,y,z}, \tilde{A}_{x,y,z} \to A_{x,y,z},$
$S_{ж,y,z} \to система_{y,z}, \dots, A_{x,y,z} \to інформаційний_{x,y,z}, простий_{x,y,z}, \dots,$
$S_{ч,y,z} \to запит_{y,z}, користувач_{y,z}, ресурс_{y,z}, бізнес_{y,z}, \dots\}$

Приклад 4.2. $P = \{\tilde{S}_{x,y,z} \to S_{x,y,z}\tilde{S}_{x',y',p}, \tilde{S}_{x,y,z} \to \tilde{A}_{x,y,z}S_{x,y,z}, \tilde{S}_{x,y,z} \to S_{x,y,z},$
$\tilde{A}_{x,y,z} \to \{дуже, досить, точно, просто, суттєво, \dots\}A_{x,y,z}, \tilde{A}_{x,y,z} \to A_{x,y,z},$
$S_{ж,y,z} \to школа_{y,z}, \dots, S_{ч,y,z} \to сміх_{y,z}, школяр_{y,z}, Львів_{y,z}, \dots,$
$S_{с,y,z} \to місто_{y,z}, \dots, A_{x,y,z} \to веселий_{x,y,z}, запальний_{x,y,z}, дитячий_{x,y,z} \dots\}$

Іншим виведенням є використання більшого обсягу пам'яті, наприклад почати виведення з $\tilde{S}_{ч,од,н} \to дуже\ A_{\text{ч,од,н}}\ A_{\text{ч,од,н}}\ S_{\text{ч,од,н}}\ S_{\text{ч,од,р}}\ S_{\text{ж,од,р}}\ S_{\text{с,од,р}}\ S_{ч,од,р}$.

$\tilde{S}_{ч,од,н}$
$\tilde{A}_{ч,од,н}\tilde{S}_{ч,од,н}$
досить $A_{ч,од,н}\tilde{S}_{ч,од,н}$
досить простий $A_{ч,од,н}\tilde{S}_{ч,од,н}$
досить простий інформаційний $\tilde{S}_{ч,од,н}\tilde{S}_{ч,од,р}$
досить простий інформаційний $S_{ч,од,н}\tilde{S}_{ч,од,р}$
досить простий інформаційний запит $\tilde{S}_{ч,од,р}$
досить простий інформаційний запит $\tilde{S}_{ч,од,р}\tilde{S}_{ч,од,р}$
досить простий інформаційний запит $S_{ч,од,р}\tilde{S}_{ч,од,р}$
досить простий інформаційний запит користувача $\tilde{S}_{ч,од,р}$
досить простий інформаційний запит користувача $\tilde{S}_{ч,од,р}\tilde{S}_{ж,од,р}$
досить простий інформаційний запит користувача $S_{ч,од,р}\tilde{S}_{ж,од,р}$
досить простий інформаційний запит користувача ресурсу $\tilde{S}_{ж,од,р}$
досить простий інформаційний запит користувача ресурсу $\tilde{S}_{ж,од,р}\tilde{S}_{ч,од,р}$
досить простий інформаційний запит користувача ресурсу $S_{ж,од,р}\tilde{S}_{ч,од,р}$
досить простий інформаційний запит користувача ресурсу системи $\tilde{S}_{ч,од,р}$
досить простий інформаційний запит користувача ресурсу системи $S_{ч,од,р}$
досить простий інформаційний запит користувача ресурсу системи бізнесу

Рис. 4.21. Процес виведення україномовного ланцюжка для прикладу 4.1

Рис. 4.22. Процес виведення україномовного ланцюжка для прикладу 4.2

Бувають випадки в текстовому контенті, коли необмежену глибину виведення має не лише праве, але і ліве послідовне підпорядкування, наприклад, за рахунок підрядних речень з службовим словом *який, що, коли* тощо (тваринка, яку врятувала Софія). Рис. 4.23 ілюструє фразу з глибиною 22 і є абсолютно правильною з граматичної точки зору (як і її український варіант). Більш того, ніщо не заважає продовжити фразу вліво *на волю в обійми зеленої пахучої трави*.

Kivánom, hogy valamint az agyag[23] ɯlelx karjai[22] közül kibontakozni[21] akary[20] kocsikerïik[19] rettentx nyikorgбsбtyl[18] megriadt[17] juhбszkutya[16] bundбjбba[15] kapaszkody[14] kullancs[13] kidülledt fïilszemïibxl[12] alбcseppent[11] kɯnnyeseppben[10] visszatьkrɯzxdx[9] holdvilág fïinyïitxl[8] illuminбlt[7] rablylovagvбr[6] felvonyhidjбbyl[5] kiбlly[4] vasszegek[3] kohïziys erejïnek[2] hatбsa[1] évszézadokra összetartja annak materiáját, aképpen tartsa össze ezt a társaságot az igaz szeretet.

Я хочу, аби справжнє кохання скріпило цю компанію так, як на століття скріплює матеріал мосту дія[1] єднальної сили[2] цвяхів[3], що торчать[4] з підіймального мосту[5] розбійницького феодального замку[6], освяченного[7] місячним світлом[8], що відображається[9] в краплині[10], яка витікає[11] з витріщеного ока[12] клеща[13], що вчепилася[14] в шерсть[15] вівчарки[16], наполоханої[17] жахливим скрипом[18] возових колес[19], що прагнуть[20] вирватися[21] з обіймів[22] грязюки[23].

Рис. 4.23. Приклад з новели Г. Фехера – жартівливий тост з [369, 695]

Українська мова дозволяє генерувати фрази з необмеженою кількістю послідовно підрядних зліва направо конструкцій типу $Y_1 Y_2 \ldots Y_i \ldots$ (необмежене

праве підпорядкування), і при цьому в кожній з конструкцій X_i можливо необмежене ліве підпорядкування – послідовність ланцюжків $\dots Y_{ij} \dots Y_{i3} Y_{i2} Y_{i1}$; проте усередині послідовності Y_{ij} подальше необмежене розгортання неможливе [1016]. Згідно з правилами української мови Y_i трактують як прості речення, що є кожне наступне додатковим визначником до попереднього, а Y_{ij} – як препозитивні дієприкметникові звороти [404, 882, 1016].

4.5.2. Алгоритм синтаксичного аналізу україномовних текстів

Граматика $G' = \langle D', D_1', I', R' \rangle$ має основний словник $D' = N_1, N_2, \dots, N_n$ символів і правил вигляду $R' = \{Y \to ZN_i, X \to N_i\}$, де $Y \in D_1'$ і $Z \in D_1'$ [369]. Кожному з N_i відповідає деяка регулярна граматика $G_i' = \langle D, D_1^i, N_i, R_i \rangle$, де D – основний словник $\forall G_i'$, D_1^i – допоміжний словник при $D_1^i \cap D' = N_i$ та $D_1^i \cap D_1' = N_i$; N_i – початковий символ; правила схеми вигляду $R_i = \{C \to eE, C \to c\}$ (заголовними латинськими символами є нетермінальні, а рядковими – термінальні). Нетермінальні словники граматик G_i' попарно не перетинаються. Об'єднанням:

$$G = G' \cup G_1' \cup G_2' \cup \dots \cup G_n', \tag{4.23}$$

де головний словник D у всіх G_i', а допоміжний додатковий словник та схема:

$$D_1 = D' \cup D_1' \cup D_1^1 \cup D_1^2 \cup \dots \cup D_1^n, \tag{4.24}$$

$$R = R' \cup R_1 \cup R_2 \cup \dots \cup R_n. \tag{4.25}$$

Граматика G є спеціальною та еквівалентна автоматній, наприклад:

$$R' = \begin{cases} I \to BN_1 \\ B \to CN_1 \\ C \to BN_2 \\ C \to EN_3 \\ E \to EN_4 \\ E \to N_2 \end{cases}, R_1 = \begin{cases} N_1 \to bP_1 \\ P_1 \to aQ_1 \\ Q_1 \to aQ_1 \\ Q_1 \to c \end{cases}, R_2 = \{N_2 \to d, R_3 = \begin{cases} N_3 \to aP_3 \\ N_3 \to bQ_3 \\ N_3 \to cW_3 \\ P_3 \to a \\ Q_3 \to b \\ W_3 \to dW_3 \\ W_3 \to eW_3 \\ W_3 \to d \end{cases}, R_4 = \begin{cases} N_4 \to cP_4 \\ P_4 \to b \end{cases} \tag{4.26}$$

Алгоритм 4.3. Алгоритм синтаксичного аналізу речення.

Етап 1. Згенерована послідовність без обмежень породжується направо за рахунок N_i як синтаксичної групи або речення на основі правил R'.

Етап 2. Будь-яке з N_i на основі R_i необмежено розгортається у вигляді дерева (Рис. 4.24) справа наліво – в ланцюжок термінальних символів як слова.

Проаналізувати синтаксичну структуру речення – це ідентифікувати порядок слів залежно від синтаксичної структури та взаємозв'язків, що визначається обов'язково згідно аналізу по відношенню до сусідів та чимось похідним/вторинним. Доцільним є видозмінити граматику, щоб обидві частини предикати (Рис. 4.24) були деревами синтаксичних відношень [369]. Лінії з індексами описують синтаксичні зв'язки різних типів; символи $A, B, C, \ldots$ – синтаксичні категорії.

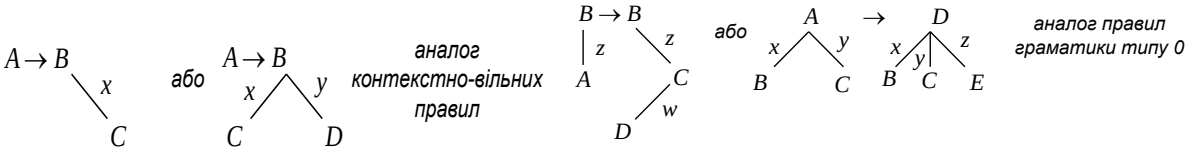


Рис. 4.24. Правила побудови дерева

У результаті отримують синтаксичні структури (а не фрази) мови як частину граматики, що породжує. Іншу частину цієї граматики складає розрахунок в українській мові – з обов'язковим обліком логічного виведення лінійних послідовностей слів, вирішуючи проблему розривних складових.

4.6. Метод семантичного аналізу української мови

Семантичний аналіз полягає не лише в ідентифікації змісту тексту, але і в генеруванні структур даних, до яких можна застосувати логічні міркування. Текстові семантичні (тематично-змістовні) подання (англ. Thematic Meaning Representations, TMR) застосовують для кодування речень у вигляді предикатних структур на основі логіки першого ступеня або лямбда-числення (λ-числення).

Мережеві/графові структури застосовують для кодування взаємодій предикатів відповідних ознак тексту. Потім реалізують обхід для аналізу центральності термінів або суб'єктів та причин відношень між елементами.

Аналіз графів в тому числі онтології O зазвичай не є повним СЕМ, але допомагає сформувати частину важливих логічних рішень/висновків на основі таксономії понять X [163]:

$$O: \text{ПСУМ} \rightarrow Concepts. \quad (4.27)$$

Результатом СЕМ на основі онтологічної моделі правил синтаксису української мови O є зважені орієнтовані графи семантики тексту:

$$O = < Concepts, Relationships, Functions >, \quad (4.28)$$

де $Relationships$ – кортеж відношень між концептами ПО української мови; $Concepts$ – кортеж концептів ПО опису правил української мови; $Functions$ – кортеж функцій інтерпретації концептів/правил української мови [163].

Таксономія концептів задає синтаксис мови як кореневий концепт онтології:

$$Concepts_{\mu}: < R_{Snt} > \rightarrow C'_{\mu}. \quad (4.29)$$

Оптимальне визначення кортежу відношень між цими концептами та кортежу правил української мови, формалізованих дескриптивною логікою DL, дозволить ефективно опрацьовувати українські тексти [163]:

$$Concepts =< R_{Mrp}, R_{Pnc}, R_{Str}, R_{Snt}, R_{Smn} >, \quad (4.30)$$

де кортежі концептів морфології R_{Mrp}, пунктуації R_{Pnc}, структури R_{Str}, синтаксису R_{Snt} (Рис. 4.25) та семантики R_{Smn} [163].

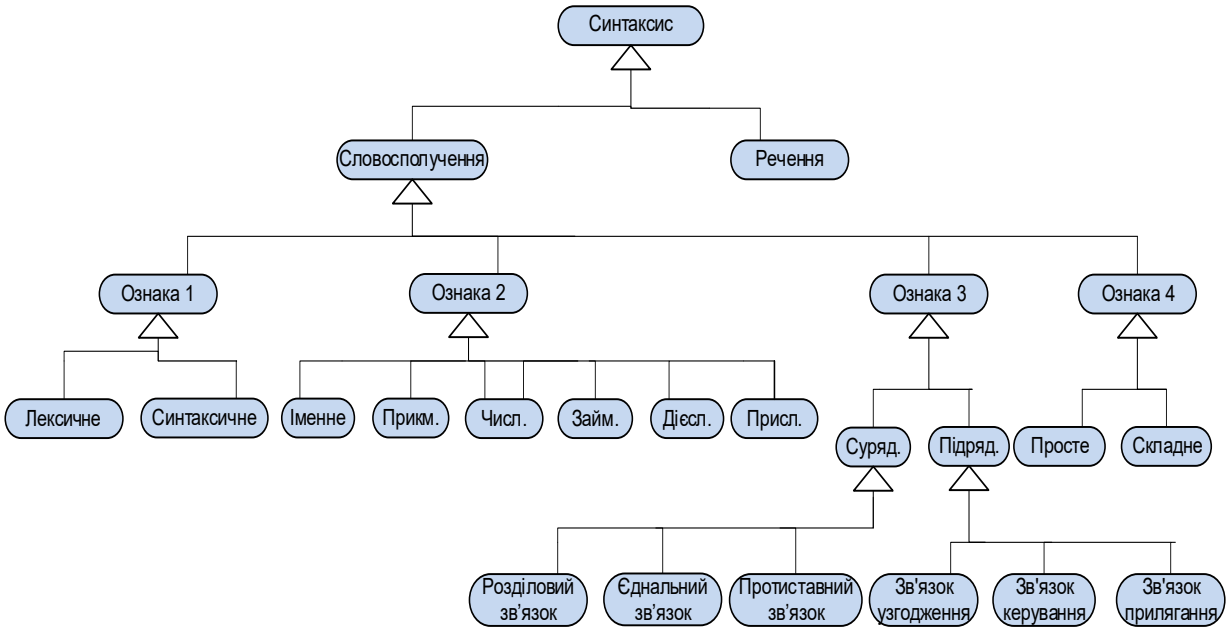


Рис. 4.25. Діаграма класів для ієрархії виду *Синтаксис Словосполучення*

При СЕМ для ідентифікації множини семів відповідного тексту та їх взаємозв'язку спочатку на основі результатів СА будують семантичний граф відношень лінгвістичних одиниць з врахування частин мови слів:

$$C'_{\mu} = \lambda(C_{\lambda}, D_{\lambda}, R_{\lambda}, Concepts_{\mu}), \quad (4.31)$$

$$Concepts_{\mu} =< C_{WrdCmb}, C_{SntCmb} >, \quad (4.32)$$

де C_{WrdCmb} – кортеж концептів утворення словосполучень; C_{SntCmb} – кортеж концептів генерування речень в українській мові (Рис. 4.26) [163].

Кортеж C_{WrdCmb} згідно правил синтаксису української мови (Рис. 4.26):

$$C_{WrdCmb} =< Sgn_1^{Wrd}, Sgn_2^{Wrd}, Sgn_3^{Wrd}, Sgn_4^{Wrd} >, \quad (4.33)$$

де Sgn_i^{Wrd} – кортеж властивостей генерування словосполучень [163].

Кортеж Sgn_1^{Wrd} згідно правил синтаксису української мови (Рис. 4.26):

$$Sgn_1^{Wrd} =< Sgn_{Lxc}^{I}, Sgn_{Snt}^{I} >, \quad (4.34)$$

де Sgn_{Lxc}^{I} – кортеж лексичних ознак генерування словосполучень; Sgn_{Snt}^{I} – кортеж синтаксичних ознак генерування словосполучень [163];

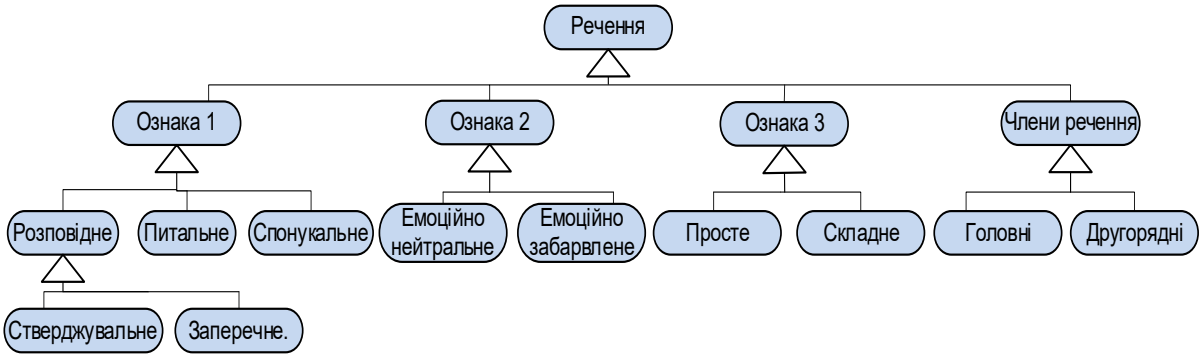


Рис. 4.26. Діаграма класів для ієрархії виду *Речення*

$$Sgn_2^{Wrd} =< Sgn_{Nou}^{II}, Sgn_{Adc}^{II}, Sgn_{Nmr}^{II}, Sgn_{Prn}^{II}, Sgn_{Vrb}^{II}, Sgn_{Adv}^{II} >, \quad (4.35)$$

де Sgn_{Nou}^{II} – кортеж іменних властивостей; Sgn_{Adc}^{II} – кортеж прикметникових властивостей; Sgn_{Nmr}^{II} – кортеж числівних властивостей; Sgn_{Prn}^{II} – кортеж займенникових властивостей; Sgn_{Vrb}^{II} – кортеж дієслівних властивостей; Sgn_{Adv}^{II} – кортеж прислівних властивостей [163];

$$Sgn_3^{Wrd} =< Sgn_{Crd}^{III}, Sgn_{Inf}^{III} >, \quad (4.36)$$

де Sgn_{Crd}^{III} – кортеж сурядних та Sgn_{Inf}^{III} – кортеж підрядних властивостей [163];

$$Sgn_4^{Wrd} =< Sgn_{SmWd}^{IV}, Sgn_{CmWd}^{IV} >, \quad (4.37)$$

де Sgn_{SmWd}^{IV} – кортеж простих та Sgn_{CmWd}^{IV} – кортеж складних властивостей.

Кортеж Sgn_{Crd}^{III} описує складові властивості речення зв’язку:

$$Sgn_{Crd}^{III} =< Sgn_{AdCm}^{Crd}, Sgn_{CnCm}^{Crd}, Sgn_{DvCm}^{Crd} >, (4.38)$$

де Sgn_{AdCm}^{Crd} – кортеж властивостей розділового, Sgn_{CnCm}^{Crd} – кортеж властивостей єднального та Sgn_{DvCm}^{Crd} – кортеж властивостей протиставного зв'язків [163].

$$Sgn_{Inf}^{III} =< Sgn_{CtCm}^{Inf}, Sgn_{MgCm}^{Inf}, Sgn_{AgCm}^{Inf} >. (4.39)$$

де Sgn_{CtCm}^{Inf} – кортеж властивостей узгодження; Sgn_{MgCm}^{Inf} – кортеж властивостей керування; Sgn_{AgCm}^{Inf} – кортеж властивостей прилягання [163].

Кортеж концептів генерування речень в українській мові (Рис. 4.26):

$$C_{SntCmb} =< Sgn_1^{Snt}, Sgn_2^{Snt}, Sgn_3^{Snt}, Sgn_{SnMb}^{Snt} >,.(4.40)$$

де властивості генерування речень в українській мові згруповані у Sgn_i^{Snt} – кортежі властивостей генерування речень в українській мові; Sgn_{SnMb}^{Snt} – кортеж властивостей ідентифікації членів речення;

$$Sgn_1^{Snt} =< Sgn_{NrSn}^{I}, Sgn_{PrSn}^{I}, Sgn_{InSn}^{I} >, \quad (4.41)$$

де Sgn_{NrSn}^{I} – кортеж властивостей генерування розповідних речень; Sgn_{PrSn}^{I} – кортеж властивостей генерування питальних речень; Sgn_{InSn}^{I} – кортеж властивостей генерування спонукальних речень;

$$Sgn_2^{Snt} =< Sgn_{EmNt}^{II}, Sgn_{EmCl}^{II} >, \quad (4.42)$$

де Sgn_{EmNt}^{II} – кортеж властивостей генерування емоційно-нейтральних речень; Sgn_{EmCl}^{II} – кортеж властивостей генерування емоційно-забарвлених речень;

$$Sgn_3^{Snt} =< Sgn_{SlSt}^{III}, Sgn_{ClSt}^{III} >, \quad (4.43)$$

де кортеж концептів утворення Sgn_{SlSt}^{III} простих та Sgn_{ClSt}^{III} складних речень;

$$Sgn_{SnMb}^{Snt} =< Sgn_{MnStMb}^{SnMb}, Sgn_{SdStMb}^{SnMb} >, \quad (4.44)$$

де Sgn_{MnStMb}^{SnMb} – кортеж властивостей ідентифікації головних членів речення; Sgn_{SdStMb}^{SnMb} – кортеж властивостей ідентифікації другорядних членів речення;

$$Sgn_{NrSn}^{I} =< Sgn_{AfSt}^{NrSn}, Sgn_{NgSt}^{NrSn} >. \quad (4.45)$$

де Sgn_{AfSt}^{NrSn} – кортеж властивостей генерування стверджувальних речень; Sgn_{NgSt}^{NrSn} – кортеж властивостей генерування заперечних речень.

Для генерування простого речення Sgn_{SlSt}^{III} аналізують ознаки (Рис. 4.27):

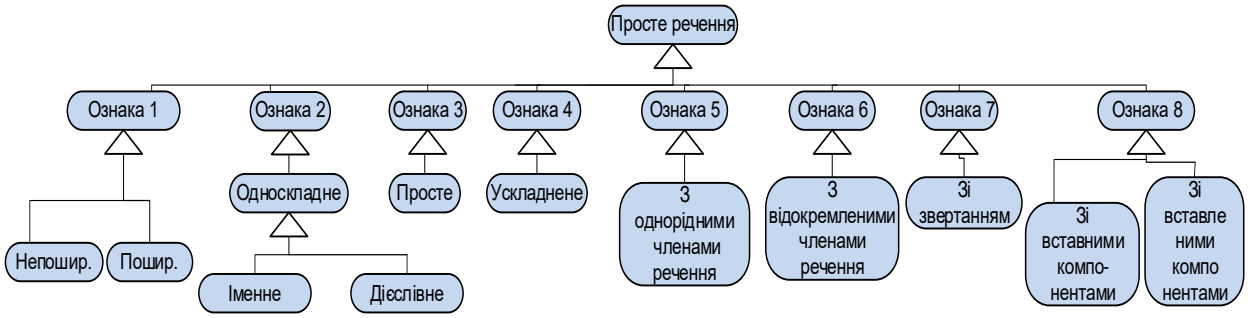


Рис. 4.27. Діаграма класів для ієрархії виду *Просте речення*

$$Sgn_{SlSt}^{III} =< Sgn_1^{SlSt}, Sgn_2^{SlSt}, Sgn_3^{SlSt}, Sgn_4^{SlSt}, Sgn_5^{SlSt}, Sgn_6^{SlSt}, Sgn_7^{SlSt}, Sgn_8^{SlSt} >. \quad (4.46)$$

де Sgn_i^{SlSt} – кортеж властивостей генерування простих речень.

Аналогічно формують кортежі для ідентифікації членів речення Sgn_{SnMb}^{Snt} (Рис. 4.28-Рис. 4.29) та складного речення Sgn_{ClSt}^{III} (Рис. 4.30).

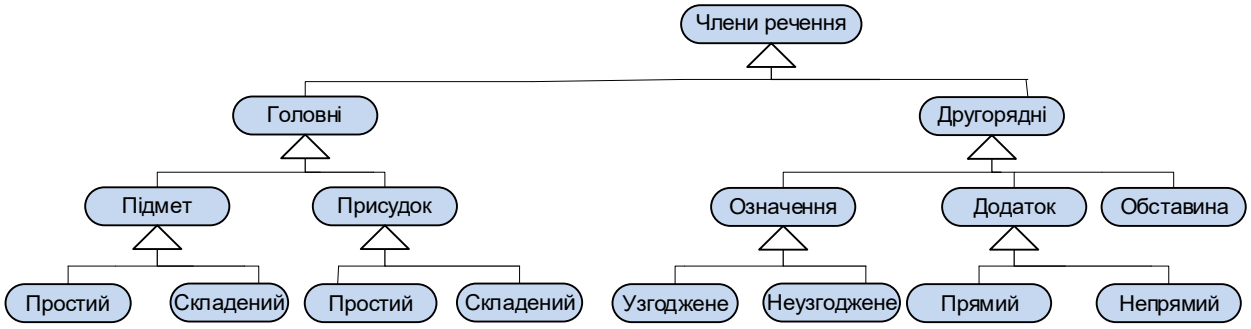


Рис. 4.28. Діаграма класів для ієрархії виду *Члени речення*

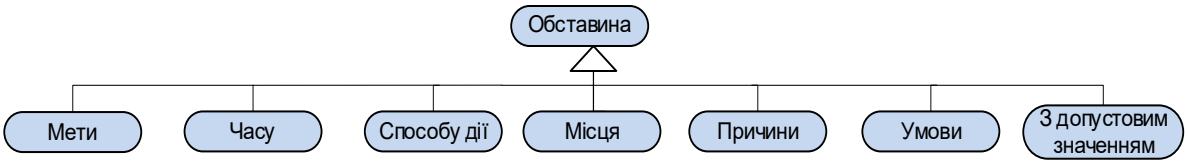


Рис. 4.29. Діаграма класів для ієрархії виду *Обставина*

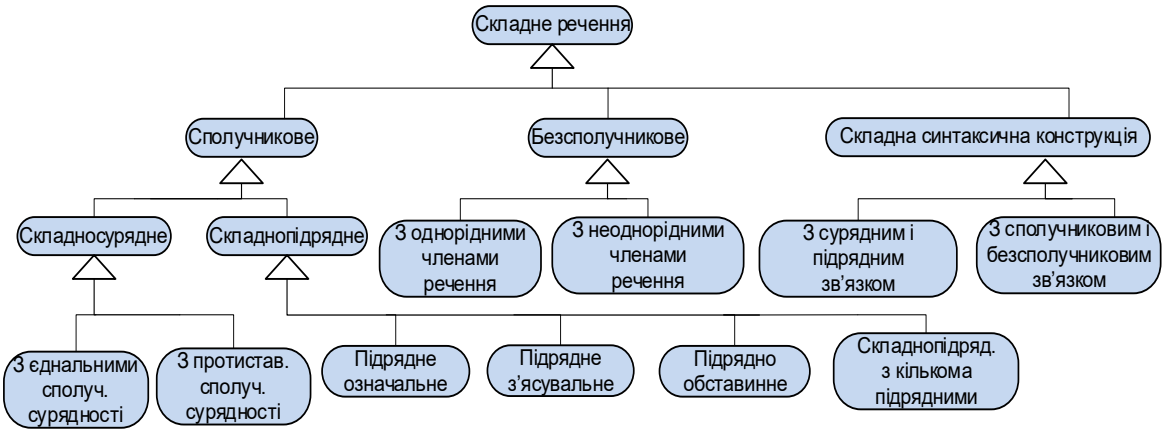


Рис. 4.30. Діаграма класів для ієрархії виду *Складне речення*

На основі інструменту Protégé 3.4.7 реалізовано ієрархії класів та підкласів синтаксичних концептів та правил української мови (Рис. Б.1 додатку Б) [163]. Рис. Б.2 додатку Б відображає ієрархію класів БЗ на прикладі фрази:

Словосполучення з єдиальним сурядним зв'язком має деякий єдиальний сполучник

В SWRL Rules розроблені правила синтаксу на основі SWRL(Рис. Б.3) [163]. На основі Open SPARQL Query panel розроблені Query, наприклад для *Складне_безсполучникове_речення, Речення* та *Словосполучення* (Рис. Б.4) [163].

Процес видобування даних з україномовного тексту на основі онтології синтаксису дозволяє доповнювати концептуальні зважувальні графи контенту.

4.7. Метод прагматичного аналізу української мови

4.7.1. Особливості прагматичного аналізу української мови

Прагматика досліджує залежність значення від контексту текстового контенту автора та з врахуванням його попередніх знань, намірів, мети тощо в порівнянні від семантики, яка аналізує саме значення в залежності від результатів ГА, МА, ЛА та СА в межах конкретного тексту. Прагматика є продовженням СЕМ з врахуванням особливості контексту аналізованого тексту з врахування неоднозначності висловлювань аналізованого тексту на основі аналізу особливостей авторських висловлювань в попередніх аналогічних текстах, спираючись на час, місце, спосіб, мету та інші обставини розмови.

В ПА при розв'язку неоднозначності авторського висловлювання в конкретному аналізованому тексті з врахуванням особливостей авторського мовлення в попередніх подібних промовах найкраще застосовувати моделі прогнозування слів, наприклад, N-граматичні моделі мови (англ. Language Model, LM) [510]. Кожний спікер як особистість з унікальним життєвим досвідом має не лише свій власний словник тематичних слів, але і унікальний почерк використання цих слів та їх послідовності у певному контексті відповідного тематичного спрямування. У висловлюванні «*лінгвістична система опрацьовує* …» наступне слово залежить не лише від контексту, але і від так званого

мовленнєвого почерку автора тексту: *текст*, *контент*, *текстовий контент*, *вхідні дані*, *вхідну інформацію*, *інтегровані дані*, *авторський контент*, *публікації* тощо. Фраза «*включіть свою виконану лабораторну роботу ...*» на відмінну від «*додайте свою виконану лабораторну роботу ...*» має ширше значення та суттєво залежить не лише від контексту але і від мовця (включити може означати як завантажте розроблене ПЗ на комп'ютері або у сенсі додати її як пункт до якогось списку тощо). Учасники діалогу інтуїтивно розуміють зміст на основі свого досвіду спілкування з автором фрази. Для прагматичного аналізу необхідне введення моделей, які визначають імовірність для кожного наступного слова. Вони також призначені для призначення вірогідності мети висловлювання для коректного машинного перекладу, ідентифікації/виправленні граматичних та стилістичних помилок, розпізнавання рукописного тексту або мови.

Кожна мова має особливі статистичні параметри, і аналіз вірогідності поя лише літер та їх сполучень як N-грам відповідної мови дає можливість ідентифікувати саму мову або стиль автора (Рис. 4.31 – з більшою вірогідністю автор еталону написав Уривок 1).

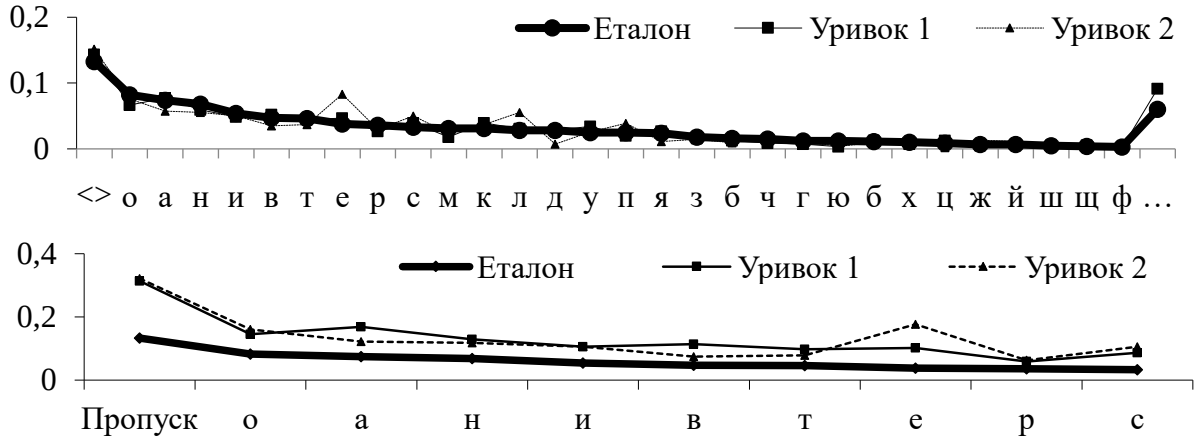


Рис. 4.31. Вірогідність появи літер в еталоні та аналізованих уривках

Для українських текстів статистичними параметрами стилів є вірогідності появ голосних, приголосних, пропуски між словами, а також м'яких і сонорних груп приголосних [962]. Вірогідність також важлива для посилення комунікації [1017]. Фізик Стівен Хокінг використовував прості рухи для вибору слів з меню

для синтезу мовлення. Для таких ІС прогнозування слова доречно застосувати для формування пропозицій щодо списку ймовірних слів для меню [506-511].

4.7.2. Основні правила моделювання мови на основі N-грам

Однією із найрозповсюдженою та найпростішою в реалізації для англомовних текстів є LM – N-грам, яка присвоює ймовірності до речень або послідовностям слів [506-511]. Для україномовних текстів краще застосовувати таку LM до послідовності основ слів без врахування флексій (інакше будуть отримані некоректні результати ПА) для розрахунку $P(b|a)$ – ймовірності появи основи слова b після послідовності основ a. Врахування слів в N-грам LM в україномовних текстах доречно при ідентифікації граматичних помилок.

$$P(систем|комп'ютер\ лінгвіст), \quad (4.47)$$

$$P(системи|комп'ютерні\ лінгвістичні), \quad (4.48)$$

$$P(систему|комп'ютерну\ лінгвістичну). \quad (4.49)$$

Один із найкращих способів обчислити таку ймовірність – провести статистичний аналіз на великих корпусах текстах відповідного автора або відповідного тематичного спрямування з достовірних Інтернет-джерел:

$$P(систем|комп'ютер\ лінгвіст) = \frac{N(комп'ютер\ лінгвіст\ систем)}{N(комп'ютер\ лінгвіст)}, \quad (4.50)$$

$$P(систем|комп'ютер\ лінгвіст) = \frac{P(комп'ютер\ лінгвіст\ систем)}{P(комп'ют\ лінгвіст)}. \quad (4.51)$$

Це дає ймовірнісний результат на певний часовий проміжок, тому що мова є творчою, не однорідною, поновлюється словник, постійно розвивається як загалом, так для конкретного мовця – автора тексту. Для аналізу відповідної випадкової лінгвістичної події $A_i = комп'ют$ знаходять $P(A_i)$ для обчислення вірогідності появи певної послідовності лінгвістичних подій на основі ланцюгового правила або загального правила добутку (chain rule of probability):

$$P(A_1A_2 \dots A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1^2) \dots P(A_n|A_1^{n-1}), \quad (4.52)$$

$$P(A_1A_2 \dots A_n) = \prod_{i=1}^{n} P\left(A_i|A_1^{i-1}\right). \quad (4.53)$$

Для аналізу послідовності N основ слів $x_1x_2 \dots x_n$ або x_1^n ($x_1x_2 \dots x_{n-1} \rightarrow x_1^{n-1}$) при $A_1 = x_1$, $A_2 = x_2$, $A_3 = x_3$, ..., $A_n = x_n$ обчислюють:

$$P(x_1 x_2 \dots x_n) = P(x_1^n) = P(x_1)P(x_2|x_1)P(x_3|x_1^2) \dots P(x_n|x_1^{n-1}), \quad (4.54)$$

$$P(x_1^n) = \prod_{i=1}^{n} P\left(x_i|x_1^{i-1}\right). \quad (4.55)$$

Ланцюгове правило відображає зв'язок між загальною вірогідністю появи конкретної послідовності основ та умовною вірогідністю появи основи слова конкретними попередніми в цій послідовності основами слів. Врахування всієї динаміки появи в тексті всіх основ слів у відношенні до послідовностей інших основ слів є надлишковим/малоефективним процесом із-за мінливості мови/промови в часі. Прогнозування моделі 2-грам полягає в апроксимації динаміки появи лише останніх кілька основ слів в заданій послідовності:

$$P(систем|лінгвіст) = \frac{N(лінгвіст\ систем)}{N(лінгвіст)}, \quad (4.56)$$

$$P(систем|лінгвіст) = \frac{P(лінгвіст\ систем)}{P(лінгвіст)}. \quad (4.57)$$

Для прогнозу умовної ймовірності наступної основи слова застосовуємо Марківське припущення (ймовірність слова залежить лише від попереднього):

$$P(x_n|x_1^{n-1}) \approx P(x_n|x_{n-1}). \quad (4.58)$$

Для прогнозу умовної ймовірності наступної основи слова в N-грамі на основі метрики максимальної (найбільшої) правдоподібності (англ. Maximum Likelihood Estimation, MLE) обчислюємо:

$$P(x_n|x_1^{n-1}) \approx P(x_n|x_{n-k+1}^{n-1}). \quad (4.59)$$

На основі обчислюємо ймовірність повної послідовності основ слів:

$$P(x_1^n) \approx \prod_{i=1}^{n} P(x_i|x_{i-1}). \quad (4.60)$$

Знаходимо MLE-оцінку для параметрів моделі N-грам, статистично проаналізувавши відповідний текстовий корпус, і нормалізувавши частоти появ основ слів та їх послідовностей в межах [0;1]:

$$P(x_n|x_{n-1}) = \frac{N(x_{n-1}x_n)}{\sum_x N(x_{n-1}x)} = \frac{N(x_{n-1}x_n)}{N(x_{n-1})}. \quad (4.61)$$

Наприклад, для трьох речень міні-корпусу (умовно теги   є межами одного речення) обчислимо Марківського припущення 2-грам появи основ слів:

* КЛС опрацьовує текстовий контент* на основі *NLP-процесів *

* Інтеграція текстового контенту є одним із основних процесів КЛС *

* КЛС розв'язує конкретну NLP-задачу для відповідного контенту *

$$P(\text{КЛС}|<p>) = \frac{2}{3};\ P(\text{інтегр}|<p>) = \frac{1}{3};\ P(\text{опрац}|\text{КЛС}) = \frac{1}{3};$$

$$P(</p>|\text{контент}) = \frac{1}{3};\ P(\text{контент}|\text{текст}) = \frac{2}{3};\ P(\text{задач}|\text{NLP}) = \frac{1}{2}.$$

Оцінювання MLE-параметра для моделі N-грам як відносної частоти:

$$P(x_n|x_{n-k+1}^{n-1}) = \frac{N(x_{n-k+1}^{n-1}x_n)}{N(x_{n-k+1}^{n-1})}. \quad (4.62)$$

Алгоритм 4.4. Алгоритм аналізу оцінок MLE-параметра для моделі N-грам.

Етап 1. Вхідний текст пропарсити та розбити на окремі фрази (речення) $R_1R_2 \dots R_m$, маркуючи кожний початок-закінчення відповідним тегом . Ліквідувати всі не алфавітні символи. Великі літери перевести в малі. Видалити службові слова при необхідності (для певних NLP-задач).

Етап 2. Застосувати стемінг Портера для отримання відповідно послідовності основ слів $x_{i1}x_{i2} \dots x_{in_i}$ основ слів $\forall R_i$ з врахуванням нормалізації слів.

Етап 3. Отримати на вхід запити $Q_1Q_2 \dots Q_k$ як послідовності слів шуканих даних. Знайти $\forall Q_j$ для кожного слова $y_{j1}y_{j2} \dots y_{jk_j}$ основу через стемінг.

Наприклад для фрази пошукового запиту Q_j:

Методи та засоби опрацювання інформаційних ресурсів систем електронної контент комерції

y_{j1}	y_{j2}	y_{j3}	y_{j4}	y_{j5}	y_{j6}	y_{j7}	y_{j8}	y_{j9}	y_{j10}
метод	*та*	*засіб*	*опрац*	*інформ*	*ресурс*	*систем*	*електрон*	*контент*	*комерц*
58	190	25	62	122	83	170	89	408	300

Етап 4. Провести статистичний аналіз входження основ слів та послідовностей основ слів запиту у аналізований текст.

Основи слів аналізованого тексту		x_{i1}	x_{i2}	x_{i3}	x_{i4}	x_{i5}	x_{i6}	x_{i7}	x_{i8}	x_{i9}	x_{i10}
		метод	*та*	*засіб*	*опрац*	*інформ*	*ресурс*	*систем*	*електрон*	*контент*	*комерц*
x_{i1}	*метод*	0	**8**	0	**6**	0	0	0	0	**1**	0
x_{i2}	*та*	**2**	0	**5**	**1**	**7**	0	**2**	0	0	**1**
x_{i3}	*засіб*	0	**2**	0	**14**	0	0	0	0	0	0
x_{i4}	*опрац*	0	0	0	0	**46**	0	0	**1**	**3**	**4**
x_{i5}	*інформ*	0	0	0	0	0	**64**	**9**	0	0	0
x_{i6}	*ресурс*	0	**7**	0	0	0	0	0	**1**	0	0
x_{i7}	*систем*	0	**8**	0	**1**	0	0	0	**21**	0	0
x_{i8}	*електрон*	0	0	0	0	0	0	0	0	**72**	**10**
x_{i9}	*контент*	0	**10**	0	0	0	0	0	0	0	**73**
x_{i10}	*комерц*	0	**6**	0	0	0	0	0	0	**176**	0

Етап 5. Знайти вірогідності появи 2-грам в аналізованому тексті. В кожному рядку значення ділимо на y_{ji}, де i номер рядка після нормалізації.

Основи слів аналізованого тексту		x_{i1}	x_{i2}	x_{i3}	x_{i4}	x_{i5}	x_{i6}	x_{i7}	x_{i8}	x_{i9}	x_{i10}	y_{ji}
		метод	*та*	*засіб*	*опрац*	*інформ*	*ресурс*	*систем*	*електрон*	*контент*	*комерц*	
x_{i1}	*метод*	0	**0,18**	0	**0,1**	0	0	0	0	**0,02**	0	58
x_{i2}	*та*	**0,01**	0	**0,03**	**0,005**	**0,035**	0	**0,01**	0	0	**0,005**	190
x_{i3}	*засіб*	0	**0,08**	0	**0,16**	0	0	0	0	0	0	25
x_{i4}	*опрац*	0	0	0	0	**0,74**	0	0	**0,016**	**0,048**	**0,064**	62
x_{i5}	*інформ*	0	0	0	0	0	**0,52**	**0,074**	0	0	0	122
x_{i6}	*ресурс*	0	**0,084**	0	0	0	0	0	**0,012**	0	0	83
x_{i7}	*систем*	0	**0,047**	0	**0,006**	0	0	0	**0,124**	0	0	170
x_{i8}	*електрон*	0	0	0	0	0	0	0	0	**0,81**	**0,112**	89
x_{i9}	*контент*	0	**0,025**	0	0	0	0	0	0	0	**0,179**	408
x_{i10}	*комерц*	0	**0,02**	0	0	0	0	0	0	**0,053**	0	300

Отримані матриці в більшості випадків будуть розрідженими. Фраза та різні варіації (множина/однина та відмінки) *система електронної контент-комерції*: $P(\text{систем електрон контент комерц}) == P(\text{електрон}|\text{систем})P(\text{контент}|\text{електрон})\ P(\text{комерц}|\text{контент}) =$ =0,124×0,81×0,179=0,01797876.

З кожним наступним множенням вірогідність зменшується. Застосовуючи логарифмування ймовірностей (англ. log probabilities) дозволить оперувати не настільки малими значеннями для розрахунку точності.

$$\prod_{i=1}^{n} P_i = e^{\sum_{i=1}^{n} \log P_i}, \tag{4.63}$$

4.8. Основні результати та висновки розділу

Розроблена загальна архітектура комп'ютерних лінгвістичних систем на основі основних процесів опрацювання інформаційних ресурсів як інтеграція, супровід та управління контентом, а також з застосуванням методів інтелектуального та лінгвістичного аналізу текстового потоку з використанням технології машинного навчання. Удосконалено ІТ інтелектуального аналізу текстового потоку на основі опрацювання інформаційних ресурсів, що дало змогу адаптувати загально типову структуру модулів інтеграції, управління та супроводу контенту для розв'язку різних задач ОПМ та підвищити ефективність функціонування КЛС на 6-9%. Це стало можливим завдяки поєднанню адаптованих до української мови методів лінгвістичного аналізу, вдосконаленої ІТ опрацювання інформаційних ресурсів, МН та множини метрик оцінювання

ефективності функціонування КЛС. Основний принцип побудови таких КЛС полягає на модульності, що полегшує їх побудову згідно вимог щодо наявності відповідних процесів для розв'язку конкретної задачі ОПМ. Описано основні NLP-методи на основі регулярних виразів узгодження з шаблонами при графемному та морфологічному аналізах україномовних текстів.

Удосконалено методи ОПМ на основі регулярних виразів узгодження з шаблонами, що дало змогу адаптувати методи токенізації та нормалізації тексту каскадами простих підстановок регулярних виразів та кінцевих автоматів. Визначені основні допустимі операції регулярних виразів як об'єднання та диз'юнкція символів/ланцюжків/виразів, оператори лічби та прецедентності, а також анкори як спецсимволи ідентифікації присутності/відсутності символів в RE. Визначені основні етапи токенізації та нормалізації українського тексту каскадами простих підстановок регулярних виразів та кінцевих автоматів.

Удосконалено метод МА україномовного тексту на основі сегментації та нормування слова, сегментації речення та модифікованого алгоритму стемінгу Портера як ефективного засобу ідентифікації афіксів лем для можливості розмічування аналізованого слова, що дало змогу підвищити точність пошуку ключових слів на 9%.

Реалізовані та описані алгоритми сегментації та нормування слова, сегментації речення та модифікований стеммінг Портера як ефективний спосіб ідентифікації афіксів лем для можливості розмічування аналізованого слова. На відмінну від класичного алгоритму Портера (не має високої точності навіть для англомовних текстів) модифікований є адаптованим саме для української мови та дає точний результат в межах 85-93% випадків в залежності від якості, стилю, жанру тексту та відповідно наповнення словників КЛС. Описано алгоритм мінімальної редакційної відстані рядків українських текстів як мінімальна кількість операцій, необхідних для перетворення одного в інший.

Основні результати розділу опубліковані у роботах [163, 535, 958-983, 984-1008].

РОЗДІЛ 5
ЗАСТОСУВАННЯ МЕТОДІВ ЛІНГВІСТИЧНОГО ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ТЕКСТІВ УКРАЇНСЬКОЮ МОВОЮ

5.1. Ідентифікація ключових слів контенту на основі технології Web Mining

5.1.1. Особливості визначення ключових слів україномовного тексту

Ідентифікація ключових слів текстового контенту $\varsigma\,(C, U, R, D, T) \to C'$ є відображенням вхідного текстового контенту C в новий стан C', який на відмінну від попереднього доповнений множиною ключових слів як основні маркери змісту тексту. Для цього лінгвістично досліджують багаторівневу лінійну (послідовності) та при необхідності ієрархічну/мережеву (взаємозв'язки) структуру тексту як: символи, N-грами, морфологічні ознаки, ваги слів та словосполучень, ознаки речень та взаємопов'язаних єдностей (Рис. 5.1).

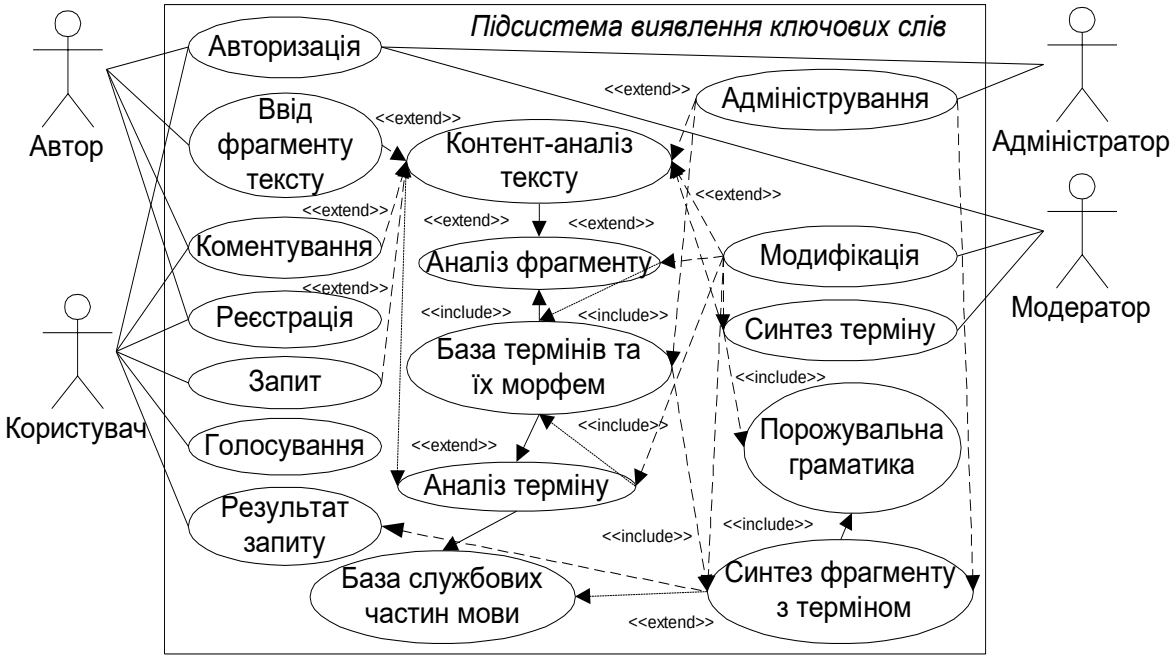


Рис. 5.1. Діаграма варіантів використання ідентифікації ключових слів

Технологія Web Mining ґрунтується на використанні методів інтелектуального аналізу потоку інформаційного контенту для ідентифікації закономірностей в Інтернет або Web-site. Основною технології Web Mining є Text

Mining, який застосовують для вилучення структурованих/неструктурованих даних з Web-page, Web-site, структур посилань тощо.

Алгоритм 5.1. Ідентифікація ключових слів контенту на основі Web Mining

Етап 1. Інтеграція/завантаження текстового контенту для подальшого аналізу.

Етап 2. Графемний аналіз текстового контенту C.

Крок 1. Форматування вхідного текстового контенту, наприклад для українського тексту однакові апострофи.

Крок 2. Видалення службової частини контенту C, наприклад тегів.

Крок 3. Видалення несимвольної частини контенту C, наприклад дат, чисел, фінансової позначень, математичних формул, зображень тощо. Видалення спецсимволів, які не входять в абетку, окрім службових як пробіл, апостроф.

Крок 4. Аналіз абревіатур і скорочень контенту C. Якщо $\leq n$ вживанні в тексті та відсутні в словнику D, тоді крок 5, інакше крок 6.

Крок 5. При необхідності редагувати тематичний словник D, наприклад додати нові скорочення або абревіатури.

Крок 6. Сегментація вхідного масиву тексту C на речення та абзаци з відповідним маркуванням відповідних меж.

Крок 7. Сегментація послідовності символів речень контенту C на лексеми.

Етап 3. Морфологічний аналіз україномовного тексту C.

Крок 1. Виділення основ (словоформ без флексій).

Крок 2. Аналіз отриманої флексії для визначення частини мови

Крок 3. Маркування слова відповідною частиною мови

Крок 4. Словоформи маркуються колекцією морфологічних ознак: відмінок, рід, відмінювання, однина/множина, особа тощо).

Крок 5. Якщо частина мови слова є іменником, маркувати як потенційно можливе ключове слово. Якщо частина мови слова є прикметником, маркувати його та наступне слово (якщо воно іменник) як словосполучення, яке потенційно може бути ключовим словом.

Крок 6. Формування лінійного ланцюжка маркованих структур.

Етап 4. Лексичний аналіз україномовного тексту C.

Крок 1. Пошук основи в словнику основ для подальшої нормалізації з врахуванням частини мови вживання в конкретному місці тексту C.

Крок 2. Нормалізація маркованих морфологічних структур.

Крок 3. Сегментація та аналіз ланцюжка нормалізованих лексем контенту C на токени та типи слів з врахуванням маркованих меж речень.

Крок 4. Формування колекцій токенів (послідовностей символів за відповідними шаблонами) як лексем з подальшою ідентифікацією їх типів з врахуванням взаємозв'язків їх в текстовому контенті C.

Крок 5. Якщо розмірність текстового контенту $\leq N_1$, тоді етап 9, інакше етап 5.

Етап 5. Синтаксичний аналіз текстового контенту C.

Крок 1. Виокремлення лексем $U_1 \in U$ для текстового контенту C.

Крок 2. Ідентифікація послідовності токенів як виразу або речення.

Крок 3. Ідентифікація іменної групи виразу на основі словника основ слів D.

Крок 4. Визначення дієслівної групи речення на основі словника основ слів D.

Крок 5. Формування дерева розбору зліва направо лінгвістичних змінних.

Крок 6. Аналіз іменної групи речення для текстового контенту C.

Крок 7. Аналіз дієслівної групи речення для текстового контенту C.

Крок 8. Дослідження синтаксичних категорій словоформами.

Крок 9. Якщо не кінець контенту C, то перехід до кроку 2, інакше до етапу 9.

Етап 6. Семантичний аналіз україномовного тексту C.

Крок 1. Токени виразів порівнюють з семантичними класами словника D.

Крок 2. Визначення для конкретного речення морфо-семантичних аналогів.

Крок 3. Поєднання токенів у загальну структуру.

Крок 4. Генерування кортежу суперпозицій лексичних функцій і семантичних класів.

Етап 7. Референційний аналіз для визначення міжфразових єдностей тексту C.

Крок 1. Контекстний аналіз контенту C для ідентифікації локальних референцій (який, цей, його) і виділення висловлювання – ядра єдності.

Крок 2. Тематичний аналіз для виокремлення тематичної структури.

Крок 3. Ідентифікація тотожності референцій; синонімізації, дублювання та повторної номінації токенів; імплікації на основі ситуативних зв'язків.

Етап 8. Структурний аналіз текстового контенту C.

Крок 1. Ідентифікація базового кортежу риторичних зв'язків між єдностями.

Крок 2. Побудова нелінійної мережі єдностей.

Етап 9. Ідентифікація множини ключових слів контенту $\varsigma\ (C, U, R, D, T) \to C'$.

Крок 1. Формування алфавітно-частотного словника $Vocab = \upsilon(C, D, R)$.

Крок 2. Ідентифікація термів $(Noun \in U_1) \cap (Noun \in Vocab)$ – іменників, словосполучень іменників, прикметника з іменником або абревіатур.

Крок 3. Формування скороченого списку слів, частоти яких відповідають умовам формування потенційних ключових слів – $Filter \subseteq Vocab$.

Крок 4. Визначення рівня унікальності $\forall Noun\ Unicity(Noun)$, $Noun \in Filter$.

Крок 5. Розрахунок Nmb_{Smb} (кількість знаків без пробілів) для $Noun \in Filter$ при $Unicity \geq 80$.

Крок 6. Розрахунок Us_{Fr} (частоти використання ключових слів). Для термів з $Nmb_{Smb} \leq 2000$ частота $Us_{Fr} \in (6; 8]\%$, з $2000 > Nmb_{Smb} < 3000$ частота $Us_{Fr} \in [4; 6]\%$, з $Nmb_{Smb} \geq 3000$ частота $Us_{Fr} \in [2; 4)\%$.

Крок 7. Розрахунок ймовірності вживання ключових слів Bs_{Fr} (на початку тексту), Is_{Fr} (в середині текстового контенту) та Es_{Fr} (в кінці текстового контенту).

Крок 8. Порівняння значень Bs_{Fr}, Is_{Fr} та Es_{Fr} для визначення пріоритетів ключових слів при умові $Bs_{Fr} \gg Is_{Fr} \gg Es_{Fr}$.

Крок 9. Сортування ключових слів згідно визначених пріоритетів.

Крок 10. Порівняння вмісту $Filter \subseteq Vocab$ зі списком $Thematic \in D$.

Крок 11. Формування нового списку токенів $Resvoc = Filter \cap Thematic$.

Крок 12. Формування колекції ключових слів C' з $KeyWords \in Resvoc$, $KeyWords = \{Noun, Unicity \geq 80, Nmb_{Smb}, Us_{Fr}, Bs_{Fr}, Is_{Fr}, Es_{Fr}\}$.

5.1.2. Метод ідентифікації ключових слів україномовного контенту

Аналіз текстового потоку контенту C для ідентифікації ключових слів зазвичай реалізують на законі Зіпфа і зводять до вибору слів із середньою частотою появи. Це просто реалізувати для англомовних текстів. Для україномовних текстів це не спрацює. Треба адаптувати алгоритми парсеру та стемінгу до української мови на основі тематичних частотних словників основ.

Алгоритм 5.2. Адаптація алгоритмів парсеру/стемінгу українських текстів

Етап 1. На основі парсеру ідентифікують множину слів, що мають частоту появи в певній межі, наприклад, 4-6% при ≤ 2000 кількості знаків без пробілів;

Етап 2. На основі парсеру та стемінгу генерують підмножину часто-вживаних семантично-навантажених слів через вилучення/маркування слів зі словника заблокованих, наприклад, таких як прийменники, сполучники, займенники, дієслова, частки тощо;

Етап 3. Якщо ключовим словом є прикметник (флексія нормалізованого слова **ий**), тоді по тексту знаходять всі основи справа від нього та будується для них частотний словник. Ті словосполучення, що використані більше за відповідне порогове значення (але менше цього прикметника) і є ключовими словами. Величину порогового значення визначає модератор. Поновлюють множину ключових слів.

Етап 4. Якщо ключовим словом є іменник (флексія слова не **ий**), тоді досліджують всі основи та їх флексії з обох боків від нього.

Крок 1. Аналізують всі слова зліва від іменника на присутність флексій **ий** та порівнюються з частотним словником. Ідентифікується множина слів, які вживані найчастіше за порогове значення – це і є нові ключові слова.

Крок 2. Аналізуються всі основи та їх флексії справа – без флексії **ий** та флексій інших частин мови, окрім іменників, порівнюються з частотним словником, за яким визначається множина ключових слів.

Етап 5. Нову підмножину порівнюють із тематичним словником основ україномовних слів для формування множини ключових слів;

Етап 6. При умові відсутності аналога слова додавання його в тематичний словник основ слів через буферний словник (редагує модератор) для накопичення статистики для різного стилістичного текстового контенту;

Експериментальною базою для відповідного дослідження обрано 100 наукових статтей Вісника НУ «Львівська політехніка» серії «Інформаційні системи та мережі» (http://science.lp.edu.ua/sisn), двох номерів 783 (http://science.lp.edu.ua/SISN/SISN-2014) та 805 (http://science.lp.edu.ua/sisn/vol-cur-805-2014-2). Для досягнення мети дослідження розроблено ІС (Рис. 5.2), розміщену на ресурсі Victana (http://victana.lviv.ua/index.php/kliuchovi-slova) з використанням наступних інструментів: CMS Joomla! для е-каркасу ІС, PHP для реалізації алгоритму, MySQL для зберігання даних та словників, HTML для реалізації розмітки Web-pages та CSS для опису стилів Web-pages.

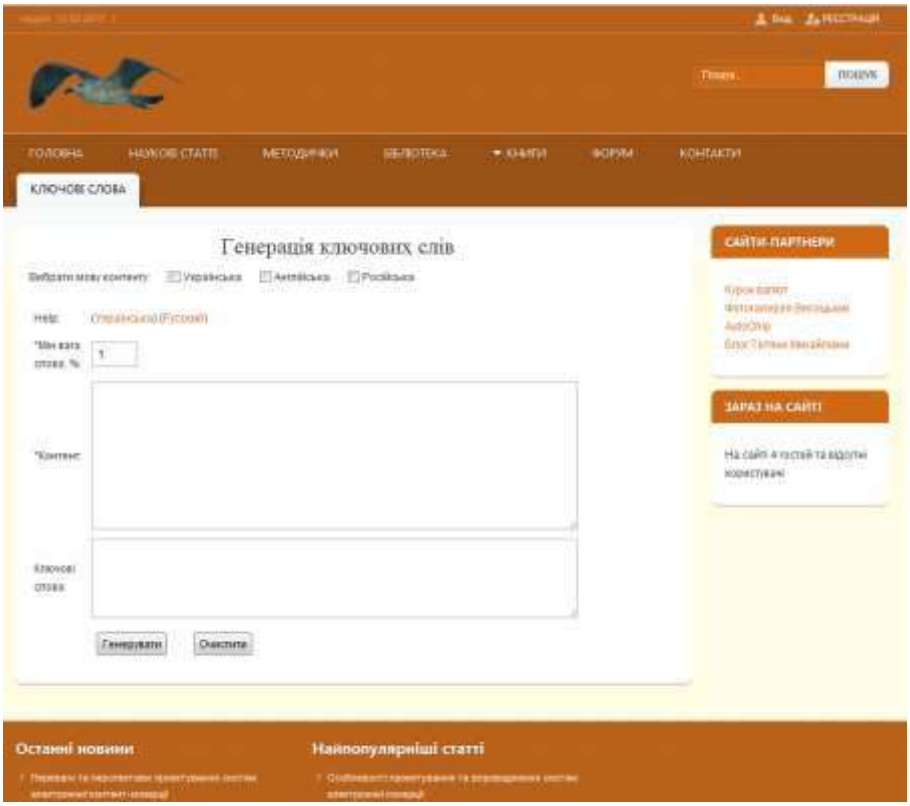


Рис. 5.2. Діалогове вікно ІС ідентифікації ключових слів в текстовому контенті

Розроблена ІС має такі основні компоненти.

1. Діалоговий дружній користувацький Web-інтерфейс на Web-page меню *Ключові слова* з такими розділами (Рис. 5.2):

- *Вибрати мову контенту* – одну/декілька мов аналізованого тексту.

- *Мін. вага слова*, % – відсоток ваги ключового слова до загальної кількості слів тексту, після якого будуть обиратись ключові слова; формат - XX.XX, в межах [00.01 - 99.99]; обов'язкове для заповнення поле.
- *Help* – коротка інструкція українською мовою в окремому Web-page.
- *Контент* – поле для аналізованого текстового контенту.
- *Ключові слова* – поле для виведення ІС множини ключових слів.
- *Генерувати* – запуск процесу ідентифікації ключових слів.
- *Очистити* – очищення поля вводу *Контент*.
- *Повторюваність слів, раз* – кількість повторень ключового слова в тексті.
- *Рекомендовані рубрики* – перелік тематичних рубрик згідно ключових слів.

2. Основні відношення БД: основи слів; заборонені слова; рубрики; правила приведення до основи слова.

3. PHP-функції опрацювання текстового контенту:

- get_keywords() – формування списку ключових слів.
- get_word() – запис правил приведення до основи слова.
- explode_str_on_words() – очищає отриманий контент від заблокованих слів, спецсимволів тощо.
- blocked_words() – формує список заблокованих слів в залежності від обраної мови контексту.
- count_words() – розрахунок частот ключових слів.
- set_keywords() – запис ключових слів до БД при їх відсутності.
- recommend_rubric() – формування списку рекомендованих рубрик.
- function error() – опрацювання помилок, направлення листа адміну ІС.

Дослідження динаміки функціонування модуля визначення колекції ключових слів із 100 науково-технічних статей проведено у два етапи із аналізом:

- змісту тематичного словника та множини заблокованих слів.
- уточнених на основі ML змісту тематичного словника та множини заблокованих слів, так як з кожною наступною перевіркою тексту через

відповідний модуль потенційно генерується додаткова колекція невідомих слів (відсутніх і в списку заблокованих і в тематичному словнику).

На кожному етапі модуль реалізовано перевірку тексту статей у два кроки: аналіз всієї статті (Рис. 5.3, а) та без мета-даних (інформація про авторів, назва, авторські ключові слова та анотації декількома мовами, список літератури тощо) (Рис. 5.3, б) для аналізу похибки точності генерування колекції ключових слів при наявності інформаційного шуму.

а) б)

Рис. 5.3. Результати перевірки (http://victana.lviv.ua/index.php/kliuchovi-slova)

5.1.3. Результати експериментального дослідження ідентифікації ключових слів україномовного контенту

Аналіз статистики здійснено на основі порівняння множин визначених авторами статті ключових слів та визначених модулем за двома різними етапами з різними вагами слів в межах [1,5] (в опції **Мін.вага слова, %*) з повними та скороченими текстами робіт (Таблиця 5.1) при середньому арифметичному значенні авторських ключових слів 4,77, які приблизно складаються з 9-10 слів.

Таблиця 5.1

Статистичні дані обсягів аналізованих текстів науково-технічних публікацій

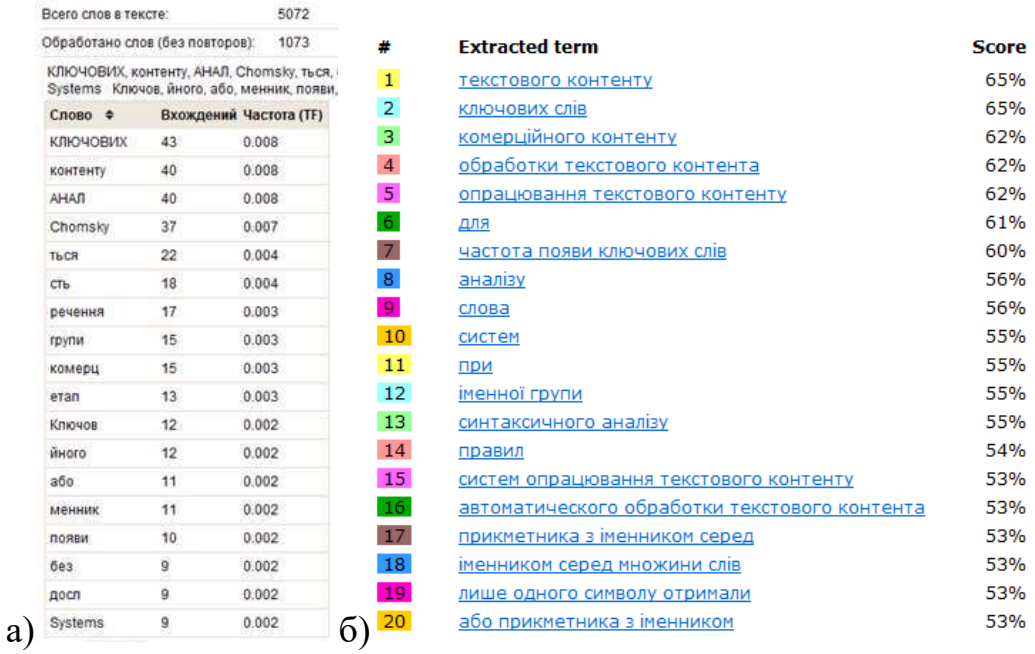
Назва обсягу статті	Крок 1		Крок 2	
	Всього	Середнє арифметичне	Всього	Середнє арифметичне
Сторінок	956	9,56	828	8,28
Абзаців	16497	164,97	15263	152,63
Рядків	42553	425,53	36965	369,65
Слів	345580	3455,8	291247	2912,47
Знаків	2327209	23272,09	1974773	19747,73
Пробілів та знаків	2674889	26748,89	2265917	22659,17

Таблиця 5.2 містить такі позначення: *A* (всього визначених ключових слів при заданій вазі слова), *B* (утворених значущих слів без займенника та дієслів), *C* (співпадіння слів зі авторським списком), *D* (точність співпадіння ідентифікованих ключових слів з авторським), *E* (додаткові визначені ключові слова, але не визначені автором публікації). Відомими ІС ідентифікації ключових слів в межах [100 ÷ 1000] слів є [1039-1043]. Недолік цих ІС – неточність та некоректність опрацювання україномовних текстів при відсутності грамотно побудованих морфологічних словників, словників основ та заблокованих слів. Також основним недоліком більшості таких ІС є обмеженість опрацювання обсягів текстового контенту [100 ÷ 1000] (Рис. 5.4).

Таблиця 5.2

Статистичні дані досліджених змісту текстів науково-технічних публікацій

Назва	Вага слова	Етап 1					Етап 2				
		A	B	C	D	E	A	B	C	D	E
Крок 1	≥ 1	5,46	3,92	2,51	2,08	1,74	7,43	7,03	3,27	3	4,18
	≥ 2	1,08	0,88	0,63	0,59	0,26	2,67	2,64	1,65	1,54	1,12
	≥ 3	0,41	0,38	0,22	0,21	0,16	1,21	1,2	0,85	0,79	0,41
	≥ 4	0,15	0,13	0,09	0,09	0,04	0,46	0,45	0,33	0,31	0,15
	≥ 5	0	0	0	0	0	0	0	0	0	0
Крок 2	≥ 1	6,51	5,02	2,68	2,23	2,37	8,35	7,78	3,25	2,91	4,99
	≥ 2	1,34	1,11	0,74	0,72	0,39	3,12	3,07	1,81	1,67	1,43
	≥ 3	0,51	0,45	0,29	0,27	0,17	1,42	1,4	0,93	0,85	0,54
	≥ 4	0,19	0,17	0,12	0,12	0,05	0,73	0,72	0,45	0,42	0,31
	≥ 5	0,11	0,1	0,06	0,06	0,04	0,33	0,32	0,25	0,23	0,1



а) б)

Рис. 5.4. Результат аналізу статті на а) [1042] та б) [1043]

Найкращою ІС для опрацювання україномовного текстового контенту є [1044] (Рис. 5.5), але вона не ідентифікує множину ключових слів, а лише частоту вживання слів, словосполучень та частин слів. Взагалі не працює з основами слова (слова *ключових* та *ключові* є різними). Розроблений ресурс працює з основами слова та орієнтований на україномовні/англомовні тексти (Рис. 5.6). Для [257] частота вживання ключових слів на Victana: *слово* - 120; *ключовий* - 49; *контент* - 46; *аналіз* - 39; *Chomsky* - 37; *система* - 37. Автори визначили ключові слова: *текст*, *україномовний*, *алгоритм*, *синтаксичний аналіз*, *породжувальні граматики*, *лінгвістичний аналіз*, *контент-моніторінг*, *ключові слова*, *інформаційна лінгвістична система*, *структурна схема речення*. Автори зазвичай більше визначають ключових слів порівняно з закономірностями розподілу частоти слів за Zipf-законом.

Наименование показателя	Значение
Количество символов	35927
Количество символов без пробелов	31118
Количество слов	4354
Количество уникальных слов	1589
Количество значимых слов	2873
Количество стоп-слов	1013
Вода	34.0 %
Количество грамматических ошибок	460
Классическая тошнота документа	8.12
Академическая тошнота документа	4.9 %

Слово	Количество	Частота, %
слів	66	1.52
контент	54	1.24
ключових	45	1.03
chomsky	37	0.85
текст	36	0.83
система	29	0.67
текстовой	24	0.55
граматика	22	0.51
аналізу	21	0.48
крок	21	0.48
речення	18	0.41
chomsky	16	0.37
частота	16	0.37

Семантическое ядро

Фраза/слово	Количество	Частота, %
слів	66	1.52
контент	54	1.24
ключових	45	1.03
ключових слів	42	0.96 / 1.93
chomsky	37	0.85
текст	36	0.83
система	29	0.67
текстового контенту	24	0.55 / 1.10
текстовой	24	0.55
граматика	22	0.51
аналізу	21	0.48
крок	21	0.48
речення	18	0.41

Стоп-слова

Слово	Количество	Частота, %
в	85	1.95
тот	68	1.56
of	60	1.38
п	56	1.29
з	48	1.10
на	45	1.03
слово	40	0.92
the	35	0.80
для	31	0.71
р	29	0.67
і	29	0.67
and	27	0.62
у	26	0.60

Рис. 5.5. Результат аналізу цієї статті на [1044]

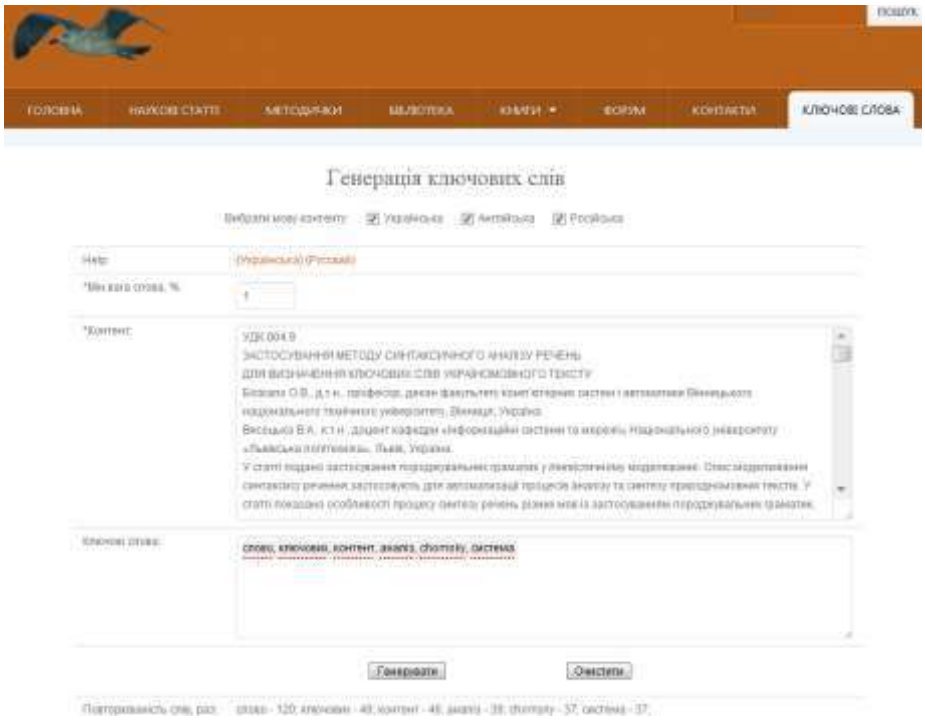


Рис. 5.6. Результат аналізу статті на http://victana.lviv.ua/kliuchovi-slova

Автор статті майже завжди формує за власним розсудом кількість та зміст множини ключових слів в діапазоні від 2 до 10 словосполучень (зазвичай – 3-5). Розроблений модуль визначає різну кількість слів, в залежності від стиля написання відповідного автора, обсягу статті, жанру, тематики та частоти вживання відповідних слів (від 0 до декількох десятків). Збіг множин знайдених ключових слів з авторськими без врахування зайвих слів, визначених авторами (повторюваність > 30 для обсягу тексту понад 4800 слів), складає відповідно для [1044] - 83%; [1043] - 57%; [1042] - 35%; http://victana.lviv.ua/kliuchovi-slova - 90% (Рис. 5.7).

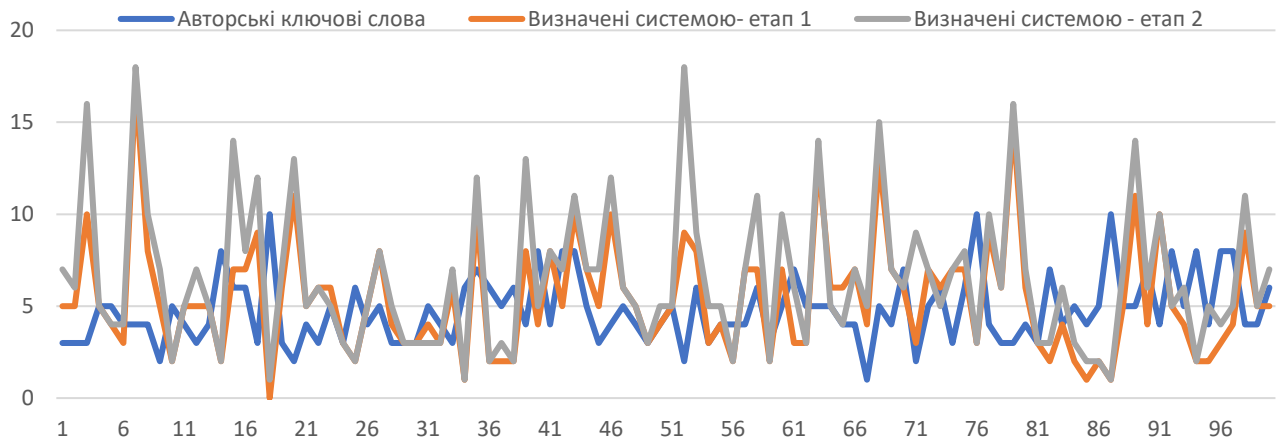


Рис. 5.7. Результати аналізу множини 100 статей

Рис. 5.8 демонструє особливості генерування множини ймовірних ключових слів порівняно з авторською множиною. Автор статті часто визначає більшу кількість слів (A_2) та меншу кількість ключових слів (A_1), ніж реально присутні в тексті. На Рис. 5.8б поданий розподіл щільності тексту в статтях, де кількість 1 – сторінок, 2 – абзаців, 3 – рядків, 4 – слів, 5 – знаків, 6 – пробілів і знаків, 7 – слів на сторінці, 8 – знаків на сторінці, 9 – пробілів та знаків на сторінці.

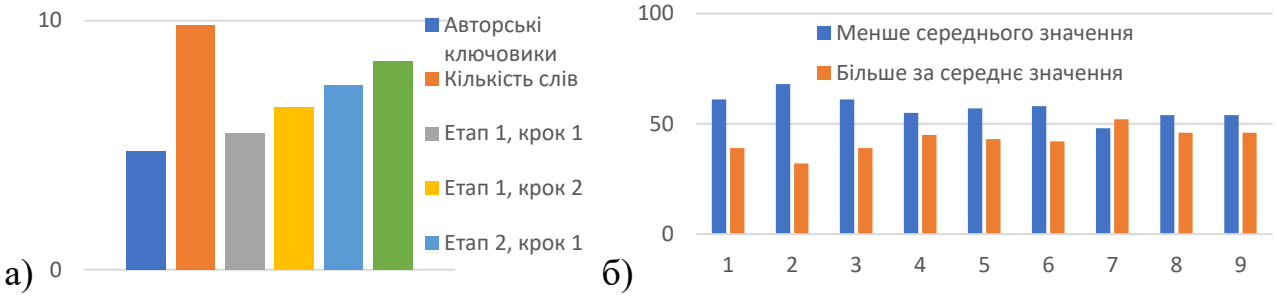


Рис. 5.8. Аналіз перевірки 100 статей (пояснення в Таблиця 5.4)

Таблиця 5.3

Статистичні дані як пояснення до Рис. 5.8

Позначення	Назва стовпця діаграми	Середньоарифметична кількість ключових слів	
		Пояснення	Значення
A_1	Авторські ключовики	визначених автором	4,77
A_2	Кількість слів	містять авторські	9,82
A_3	Етап 1, крок 1	ймовірних ключових слів, знайдених модулем на етапі X та кроку Y (Рис. 5.9-Рис. 5.10)	5,46
A_4	Етап 1, крок 2		6,51
A_5	Етап 2, крок 1		7,43
A_6	Етап 2, крок 2		8,35

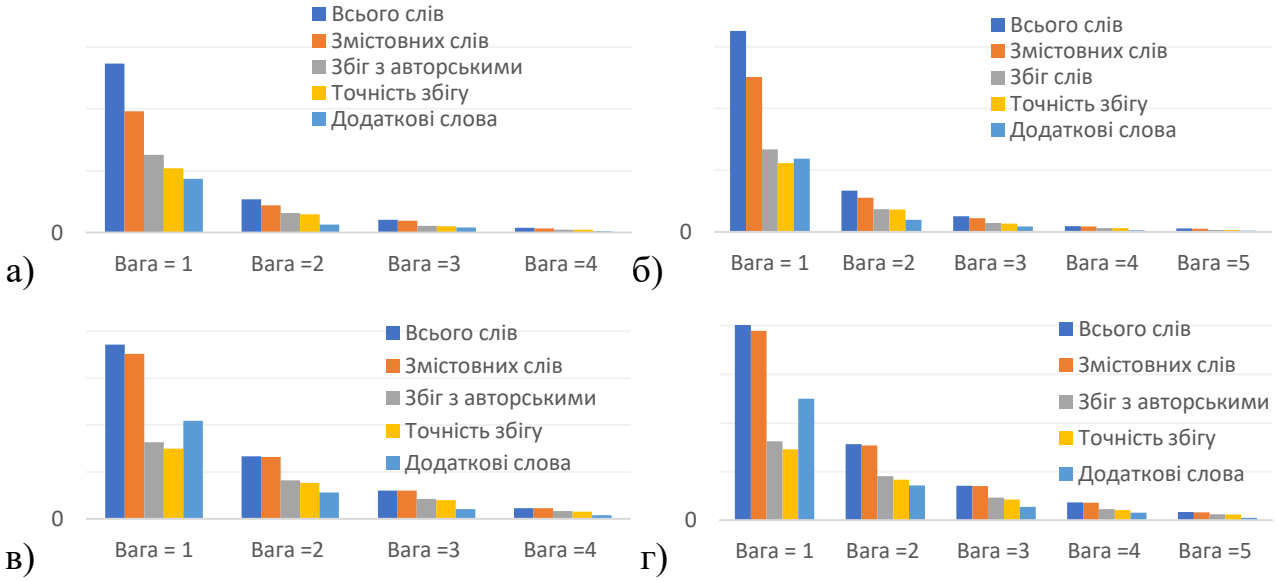


Рис. 5.9. Отримання значущих слів при опрацюванні тексту на: а) етапі 1, крок 1, б) етапі 1, крок 2, в) етапі 2, крок 1 та г) етапі 2, крок 2

Значення A_3 відмінне за значення A_1 на 0,69 (за кількістю, але не за змістом); відповідно A_4 від A_1 на 1,74; A_5 від A_1 на 2,66; A_6 від A_1 на 3,58. Значення A_2 відмінне за значення A_3 на 4,36; відповідно A_2 від A_4 на 3,31; A_2 від A_5 на 2,39; A_2 від A_6 на 1,47. Адаптивна зміна параметрів/правил модуля збільшує колекцію ідентифікованих ключових слів майже вдвічі (наприклад, значення A_1 за A_3 більше в 1,144654; A_6 – в 1,750524; A_5 – в 1,557652; A_4 – в 1,36478).

Загальний приріст значення, отриманий в залежності від модерації словників складає відповідно для A_3 14,46541; A_4 – 36,47799; A_5 – 55,7652; A_6– 75,05241. При порівнянні A_2 більше за $A_3 \div A_6$ маємо ланцюг таких значень як 1,7985; 1,5084; 1,3217; 1,176. Для різних етапів та кроків експерименту опрацювання первинного тексту середній збіг списків виявлених ключових слів з авторськими змінюється у проміжку 52,6-68,5%. Точність збігу ключових слів із авторськими коливається в проміжку 43,6-62,9%. Середній збіг змістовних ключових слів порівняно зі всіма знайденими системою коливається в проміжку 38,9-75,8% в залежності від етапів аналізу текстів статей. Точність збігу ключових слів порівняно зі всіма знайденими системою коливається в проміжку 34,3-71,9% в залежності від етапів аналізу текстів статей.

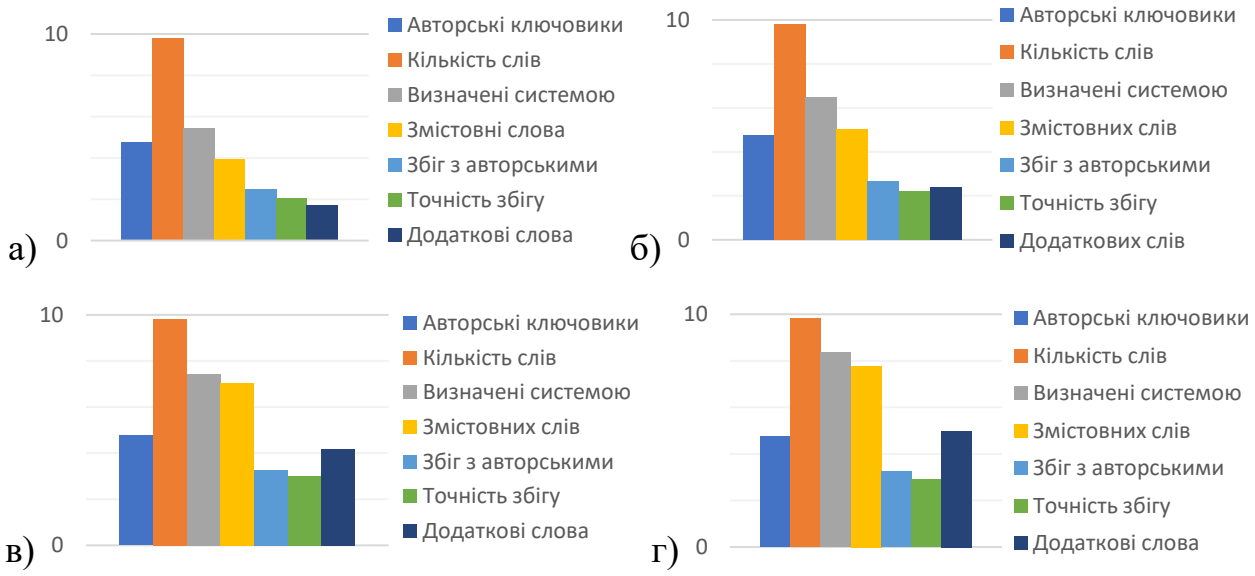


Рис. 5.10. Середньоарифметична поява значущих слів порівняно з авторськими для: а) етапу 1, крок 1, б) етапу 1, крок 2, в) етапу 2, крок 1 та г) етапу 2, крок 2

Для A_3 найчастіше модуль ідентифікував кількість ключових слів {5, 7, 3} (≥10), хоча розподіл знайдених ключових слів в межах [1;18] слів (окрім 17). Для A_4 найчастіше ІС визначила кількість ключових слів також {5, 7, 3}, хоча розподіл знайдених ключових слів є в межах [1;18] (окрім 17), але збільшилась кількість ідентифікованих слів та досягнуто найбільшого показника надійності. Для A_5 найчастіше модуль ідентифікував кількість ключових слів {7, 6, 5, 10, 8}, хоча розподіл знайдених ключових слів в межах [2;14] (значно звузився діапазон). Для A_6 найчастіше модуль ідентифікував кількість ключових слів {8, 5, 7, 10}, розподіл ідентифікованих ключових слів в межах [3;16] (покращилась точність). Точність визначення ключових слів збільшується в процесі модерації словників та ML-модуля. Різниця між кількістю ключових слів, визначених автором та ідентифікованих модулем при A_3 складає 44,39919 % (різниця у %).

Таблиця 5.4

Описові статистичні дані ідентифікації ключових слів при експериментах

Назва	A_1	A_3	A_4	A_5	A_6
Середнє	4,808081	5,515152	6,565657	7,505051	8,434343
Стандартна помилка	0,180859	0,310393	0,39035	0,301297	0,324611
Медіана	4	5	6	7	8
Мода	4	5	5	7	8
Стандартне відхилення	1,799528	3,088371	3,883932	2,997869	3,229841
Дисперсія вибірки	3,238301	9,538033	15,08493	8,987219	10,43187
Ексцес	0,652815	1,705273	0,748643	-0,45645	-0,50438
Асиметричність	0,947939	1,125305	1,065716	0,537598	0,517047
Інтервал	8	16	17	12	13
Мінімум	2	1	1	2	3
Максимум	10	17	18	14	16
Сума	476	546	650	743	835
Рахунок	99	99	99	99	99
Найбільший(1)	10	17	18	14	16
Найменший(1)	2	1	1	2	3
Рівень надійності(95,0%)	0,35891	0,615965	0,774637	0,597914	0,64418

Таблиця 5.5

Статистичні дані побудови гістограми для A_3 та $A_3 \div A_6$ (Рис. 5.11)

№	n	%	№	n	%	n	%	№	n	%	n	%	№	n	%
	A_1					A_3					A_4				
1	0	0,00	4	27	27,27	2	2,02	5	20	20,20	2	2,02	5	20	20,20
2	4	4,04	5	21	48,48	10	12,12	7	16	36,36	10	12,12	7	16	36,36
3	20	24,24	3	20	68,69	12	24,24	3	12	48,48	12	24,24	3	12	48,48
4	27	51,52	6	11	79,80	4	28,28	2	10	58,59	4	28,28	2	10	58,59
5	21	72,73	8	8	87,88	20	48,48	6	9	67,68	20	48,48	6	9	67,68

6	11	83,84	7	5	92,93	9	57,58	4	4	71,72	9	57,58	4	4	71,72
7	5	88,89	2	4	96,97	16	73,74	8	4	75,76	16	73,74	8	4	75,76
8	8	96,97	10	3	100,00	4	77,78	10	4	79,80	4	77,78	10	4	79,80
9	0	96,97	1	0	100,00	2	79,80	11	3	82,83	2	79,80	11	3	82,83
10	3	100,00	9	0	100,00	4	83,84	12	3	85,86	4	83,84	12	3	85,86
11	0	100,00	11	0	100,00	3	86,87	14	3	88,89	3	86,87	14	3	88,89
12	0	100,00	12	0	100,00	3	89,90	1	2	90,91	3	89,90	1	2	90,91
13	0	100,00	13	0	100,00	2	91,92	9	2	92,93	2	91,92	9	2	92,93
14	0	100,00	14	0	100,00	3	94,95	13	2	94,95	3	94,95	13	2	94,95
15	0	100,00	15	0	100,00	1	95,96	16	2	96,97	1	95,96	16	2	96,97
16	0	100,00	16	0	100,00	2	97,98	18	2	98,99	2	97,98	18	2	98,99
17	0	100,00	17	0	100,00	0	97,98	15	1	100,00	0	97,98	15	1	100,00
18	0	100,00	18	0	100,00	2	100,00	17	0	100,00	2	100,00	17	0	100,00
Ще	0	100,00	Ще	0	100,00	0	100,00	Ще	0	100,00	0	100,00	Ще	0	100,00

	A_5					A_6				
1	0	0,00	7	15	15,15	0	0,00	8	14	14,14
2	1	1,01	6	14	29,29	0	0,00	5	12	26,26
3	5	6,06	5	13	42,42	1	1,01	7	11	37,37
4	9	15,15	10	12	54,55	9	10,10	10	11	48,48
5	13	28,28	8	11	65,66	12	22,22	4	9	57,58
6	14	42,42	4	9	74,75	9	31,31	6	9	66,67
7	15	57,58	12	6	80,81	11	42,42	9	9	75,76
8	11	68,69	3	5	85,86	14	56,57	11	5	80,81
9	4	72,73	14	5	90,91	9	65,66	14	5	85,86
10	12	84,85	9	4	94,95	11	76,77	12	4	89,90
11	1	85,86	13	3	97,98	5	81,82	13	4	93,94
12	6	91,92	2	1	98,99	4	85,86	15	3	96,97
13	3	94,95	11	1	100,00	4	89,90	16	2	98,99
14	5	100,00	1	0	100,00	5	94,95	3	1	100,00
15	0	100,00	15	0	100,00	3	97,98	1	0	100,00
16	0	100,00	16	0	100,00	2	100,00	2	0	100,00
17	0	100,00	17	0	100,00	0	100,00	17	0	100,00
18	0	100,00	18	0	100,00	0	100,00	18	0	100,00
Ще	0	100,00	Ще	0	100,00	0	100,00	Ще	0	100,00

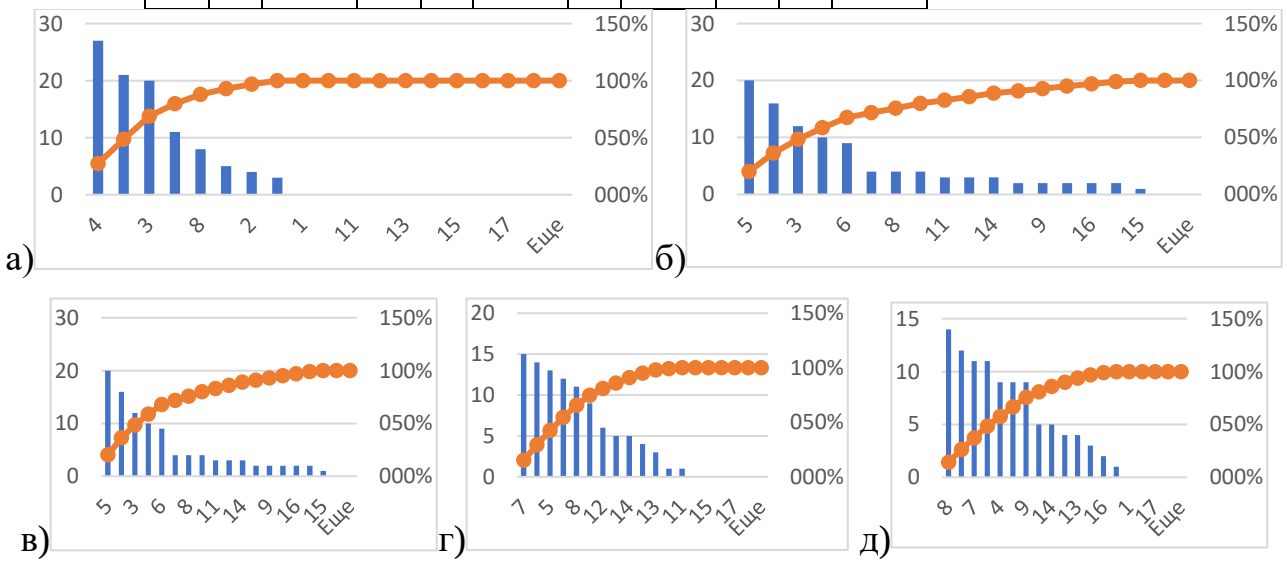


Рис. 5.11. Гістограма для вибірки а) A_1, б) A_3, в) A_4, г) A_5 та д) A_6

Точність покращується при A_4 – 33,70672 %, значно покращується при A_5 – 24,33809 %, а при A_6 складає 14,96945 % (Таблиця 5.4). Таблиця 5.5 демонструє дані дослідження статей при генеруванні множин ключових слів (Рис. 5.11).

Проведений аналіз для 100 відфільтрованих текстів без мета-даних та невідфільтрованих текстів. Отримані середні значення для 100 відфільтрованих текстів $\overline{Per_f} = 0{,}28$ та невідфільтрованих $\overline{Per_0} = 0{,}19$ показують, що така фільтрація наукових статей покращує щільність ключовиків у 1,48 раз або на 47,83 % (Рис. 5.12).

Рис. 5.12. Результати перевірки статей без уточнення тематичного словника

Отримані середні значення для 100 текстів $\overline{Per_f^v} = 0{,}34$ та $\overline{Per_0^v} = 0{,}25$ з врахуванням уточнення тематичного словника через поповнення заблокованих слів показують, що фільтрація з одночасною модерацією тематичного словника покращує щільність ключових слів у 1,35 раз або на 35,44 % (Рис. 5.13).

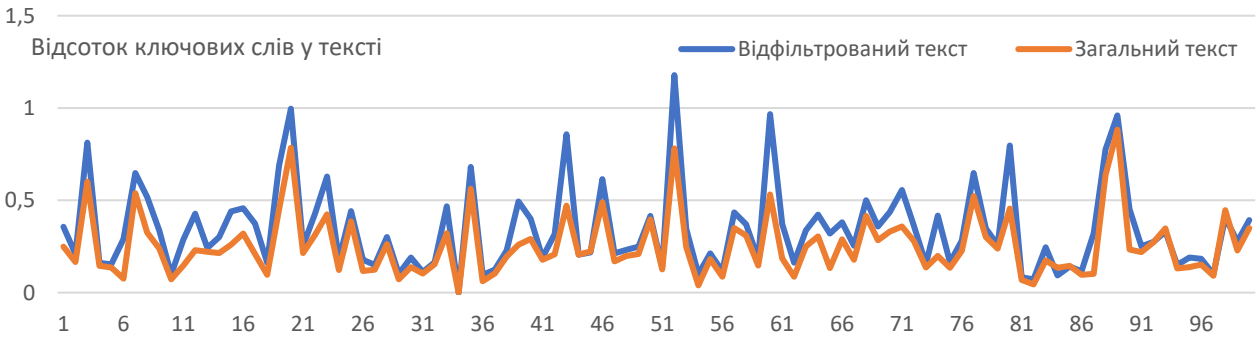


Рис. 5.13. Результати перевірки статей з уточненням тематичного словника

Порівняння значень в первинному авторському тексті $\overline{Per_0} = 0{,}19$ та $\overline{Per_0^v} = 0{,}25$ без/з уточнення тематичного словника відповідно демонструє

ефективність модерації тематичного словника у початковому тексті – щільність ключових слів збільшується у 1,34 раз або на 34,33 % (Рис. 5.14).

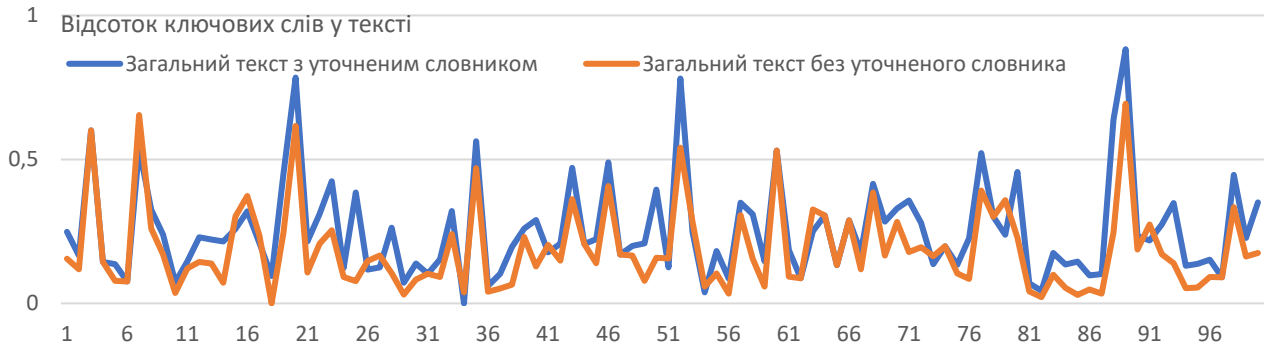


Рис. 5.14. Результати перевірки первинних статей з різними словниками

Порівняння значень в відфільтрованому авторському тексті $\overline{Per_f} = 0{,}28$ та $\overline{Per_f^v} = 0{,}34$ без/з уточнення тематичного словника відповідно демонструє ефективність модерації тематичного словника у відфільтрованому тексті – щільність ключових слів збільшується у 1,23 раз або на 23,14 % (Рис. 5.15).

Рис. 5.15. Результати перевірки відфільтрованих статей з різними словниками

5.1.4. Аналіз методів ідентифікації стійких словосполучень як ключових слів

Ідентифікація стійких словосполучень складається із таких етапів: МА, СА, виділення ключових слів та аналіз ключових словосполучень на стійкість (Рис. 5.16). Для україномовних текстів найкраще використовувати поєднання підходів процедурного, табличного та статистичного стемінгу. В процедурному підході МА акцент надається використанню при аналізі слів готових словників основ та словників готових форм (СГФ). Тоді алгоритм МА складається з етапів: пошук в СГФ, виділення основи та пошук основи в словнику. Основою більшості МА української мови є дерево (tree) або скінчений автомати без виходу (Finite State Automata, FSA) (Рис. 5.17).

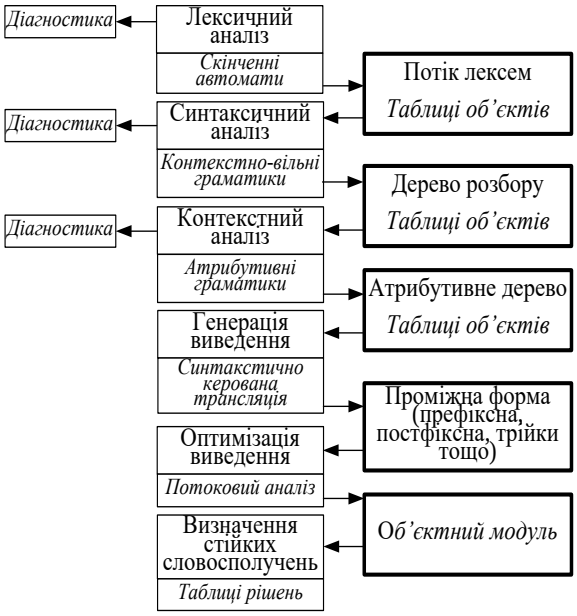
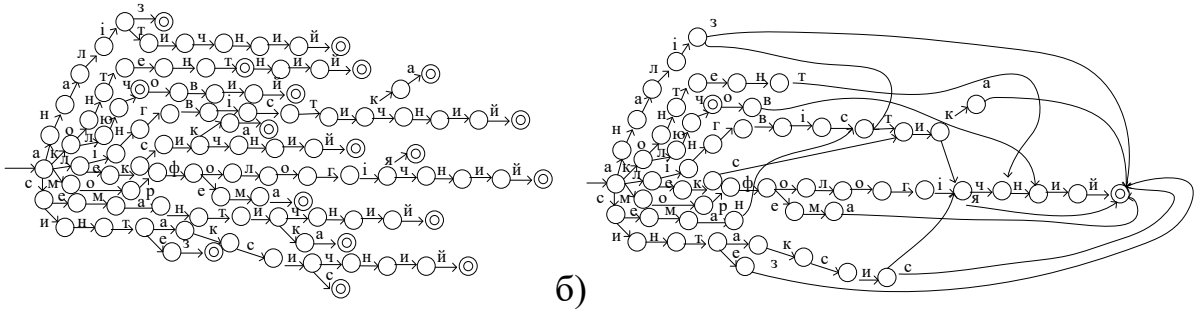


Рис. 5.16. Ідентифікація стійких словосполучень україномовних текстів



а) б)

Рис. 5.17. Методи зберігання результатів МА: а) tree та б) FSA

За формою флексій визначається тип слова (Рис. 4.16). Алгоритм працює з окремими словами, тому зміст слова не враховується. Також недоступні частини мови (прикметник, іменник тощо) та категорії морфології (основа, суфікс тощо). Варіанти правил стемінгу україномовних слів: короткі слова залишаються незмінними, змінюється при стемінгу (є виключенням), не змінюється при стемінгу (є виключенням), відповідає регулярному виразу, змінює закінчення, має незмінне закінчення або від слова відсікається флексія. Все це суттєво ускладнює алгоритм ідентифікації ключових слів. Тому спочатку треба аналізувати розповсюджені флексії.

Синтаксис – правила поєднання слів в коректні вирази – словосполучення та речення (порівн.: синтаксис мови програмування). Задача СА (синтаксичного аналізатора, parser) – побудувати синтаксичну структуру вхідного речення.

Аспектами реалізації СА є словники (інформація про індивідуальні одиниці мови); формальні правила та взаємодія із сусідніми рівнями опрацювання (морфологічний аналіз, семантичний аналіз). Найчастіше при СА використовують правила контекстно-вільної граматики (Context-free grammar, CFG): <N,T,X,R>, де N – множина не термінальних символів, T – множина термінальних символів($N \cap T = \emptyset$), X - аксіома($X \in N$), R – множина правил перетворення (підстановки) типу $Y \to \alpha$, де $Y \in N$, α - список термінальних та не термінальних символів. Приклад CFG:

$$N = \{S, NP, PP, V, N, A\},\ T = \{\text{система, рубрикувати, україномовний, контент, за, ключовий, слово}\},\ S$$

$$R = \{S \to NPVP, S \to NPVPPP, NP \to AN, PP \to PNP, VP \to VNP,$$
$$NP \to \textit{система}, V \to \textit{рубрикувати}, A \to \textit{україномовний},$$
$$A \to \textit{ключовий}, N \to \textit{контент}, N \to \textit{слово}, P \to \textit{за}\}.$$

Недоліком використання CFG є періодична поява неоднозначності при СА, наприклад, «Система рубрикує україномовний контент за ключовими словами» (Рис. 5.18).

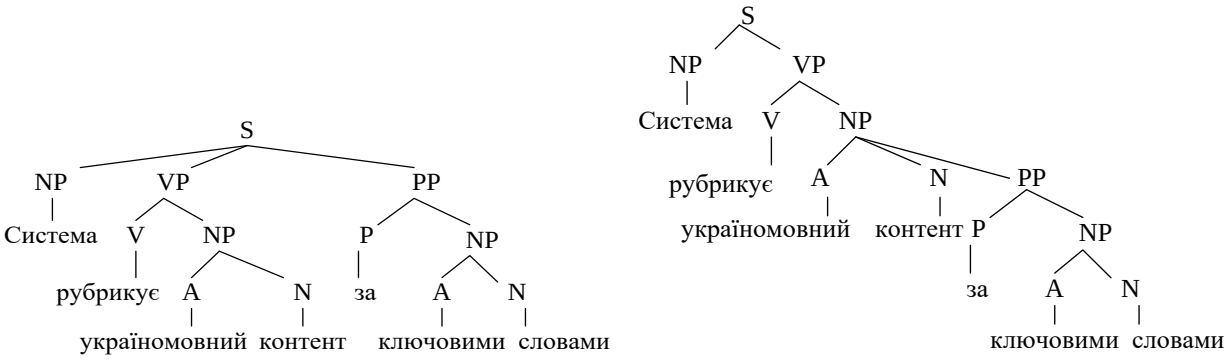


Рис. 5.18. Приклади неоднозначності в CFG.

Прикладами відомих систем СА для англомовних тестів є: «Machinese Phrase Tagger» (Рис. 5.19) та VISL. Не існує жодного online доступного інформаційного ресурсу для СА україномовних текстів.

«Ontology Matcher Demo» використовує Machinese метадані для пошуку об'єктів онтології в тексті (Рис. 5.20). На Рис. 5.21-Рис. 5.22. поданий результат СА на інформаційному ресурсі VISL.

а)

Text	Baseform	Phrase syntax and part-of-speech
The	the	premodifier, determiner
train	train	nominal head, noun, single-word noun phrase
went	go	main verb, indicative past
on	on	adverbial head, adverb
up	up	preposed marker, preposition
the	the	premodifier, determiner
track	track	nominal head, noun, single-word noun phrase
out	out	adverbial head, adverb
of	of	preposed marker, preposition
sight	sight	nominal head, noun, single-word noun phrase
,	,	
around	around	preposed marker, preposition
one	one	nominal head, pro-nominal
of	of	postmodifier, preposition

б)

0	4	This	this	PRON		
5	2	is	be	V		
8	1	a	a	DET		
10	4	test	test	N	test	V

Рис. 5.19. а) Machinese Phrase Tagger 4.9.1 analysis; б) Machinese Tokenizer

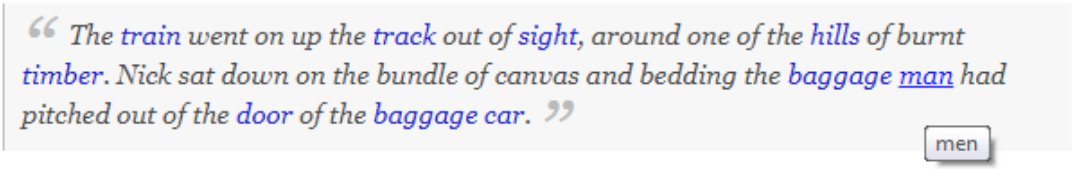


Рис. 5.20. Ontology Matcher

Для СА україномовних текстів таких інформаційних ресурсів не існує. Та і сам процес СА досить громіздкий. Для вхідного речення: «Він зробив це так незручно, що зачепив образок мого ангела, який висів на дубовій спинці ліжка, і що вбита муха впала мені прямо на голову» приклад СА з використанням предсинтаксису (або Parsing by chunks – розбиття речення на фрази, які не перетинаються, (плоска структура) ≠ повному розбору, наприклад, (the boy (with the hat)) ←→ (the boy) with (the hat)) україномовних текстів для ідентифікації стійких словосполучень при визначенні ключових слів поданий на Рис. 5.22.

Для виділення стійких словосполучень в аналізованих текстах та проведення їх порівняльного аналізу скористаємося 4-ма різними методами: FREG (частота+морфологічні шаблони, тобто прямий підрахунок кількості слів); t-тест; статистика χ^2; LR – відношення правдоподібності.

Tree structure

Enter English text to parse:

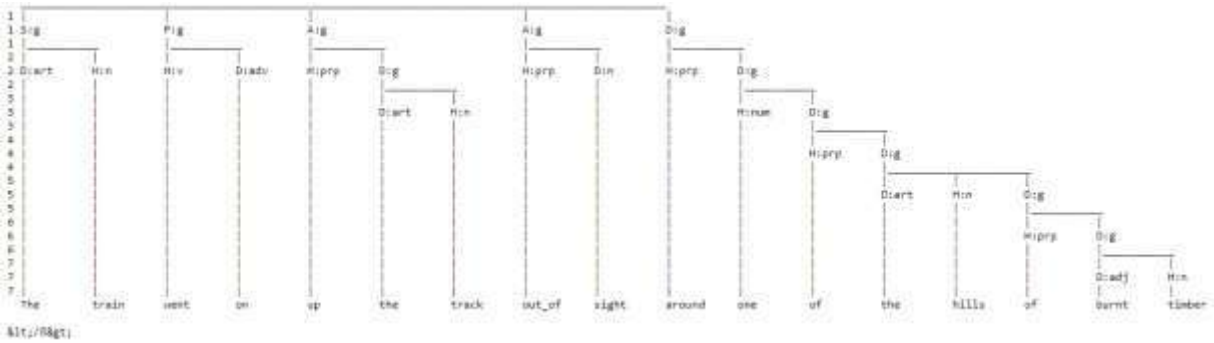
The train went on up the track out of sight, around one of the hills of burnt timber.

Parse and Show | Export and Download | Reset

Visualization: Vertical ▾ Notational convention default ▾

```
&lt;ß&gt;
&lt;s&gt;

SOURCE: Running text
1. The train went on up the track out of sight, around one of the hills of burnt timber.
A1
STA:cl(fcl)
,
.
|-S:g(np)
| |-D:art('the' S/P)    The
| |-H:n('train' S NOM)  train
|-P:g(vp)
| |-H:v('go' IMPF)      went
| |-D:adv('on') on
|-A:g(pp)
| |-H:prp('up') up
| |-D:g(np)
|   |-D:art('the' S/P)  the
|   |-H:n('track' S NOM)        track
|-A:g(pp)
  |-H:prp('out_of')     out_of
  |-D:n('sight' S NOM)  sight
|-D:g(pp)
  |-H:prp('around')     around
  |-D:g(np)
    |-H:num('one' &lt;card&gt; S)       one
    |-D:g(pp)
      |-H:prp('of')     of
      |-D:g(np)
```



</ß>

<ß>
The [the] <*> <def> **ART** S/P @>N #1->2
train [train] <DA:tog> <Vground> <def> <nhead> **N** S NOM @SUBJ> #2->3
went [go] <DA:gå> <move> <mv> **V** IMPF @FS-STA #3->0
on [on] <DA:på> **ADV** @MV< #4->3
up [up] <DA:op=ad> **PRP** @<SA #5->3
the [the] <def> **ART** S/P @>N #6->7
track [track] <DA:spor> <Lpath> <sem-l> <def> <nhead> **N** S NOM @P< #7->5
out of [out=of] <complex> <DA:ude=af> **PRP** @<ADVL #8->3
sight [sight] <DA:sigt> <percep-w> <Labs> <idf> <nhead> **N** S NOM @P< #9->8
, [,] **PU** @PU #10->0
around [around] <insertion> <DA:omkring> **PRP** @>A #11->0
one [one] <fr:78> <f:3664212> <card> **NUM** S @P< #12->11
of [of] <DA:af> <np-close> **PRP** @N< #13->12
the [the] <def> **ART** S/P @>N #14->15
hills [hill] <DA:høj> <Lmountain> <def> <nhead> **N** P NOM @P< #15->13
of [of] <DA:af> <np-close> **PRP** @N< #16->15
burnt [burnt] <DA:brændt> <SYN:cooked> <SYN:destroyed> <jppl> <tempered-2>
<SYN:treated> **ADJ** POS @>N #17->18
timber [timber] <DA:tømmer> <mat> <idf> <nhead> **N** S NOM @P< #18->16
. [.] **PU** @PU #19->0
</ß>

Рис. 5.21. Результат СА на інформаційному ресурсі VISL

Рис. 5.22. Результат СА україномовного речення

Колокація (collocations) – це словосполучення як семантично і синтаксично лінгвістична одиниця, де одна частину обирають за сенсом, а інша залежить від першої (наприклад, ставити умови – вибір дієслова *ставити* визначається традицією і залежить від іменника *умови*, при слові *пропозицію* буде інше дієслово – *вносити*). Це є обмежена (вибіркова) сполучуваність слів: фразеологізми, ідіоми, імена власні та торгові марки. До колокації часто відносять складні наіменування (наприклад, крейсер москва, руський корабль, безпілотник Байрактар, від'ємний наступ, німецькі леопарди, жест доброї волі тощо). Інше найменування того ж явища – стійкі словосполучення, N-грами. Приклади collocations –

- Грати роль, мати значення, впливати, справляти враження
- Засоби масової…, зброя масової…, вищий навчальний ….;
- глибокий старець ↔ поверхневий/мілкий невеликий юнак;
- міцний чай ↔ сильний чай;
- Кока-кола, Microsoft Windows;
- Гола Пристань, Нова Каховка, Володимир Волинський, Нью Йорк, Стив Джобс.

1. Метод FREG – це прямий підрахунок частоти вживання пар (трійок). Наприклад, FREG для речення « В літературі описано декілька підходів до автоматичного виділення стійких словосполучень.» → в літературі; літературі описано; описано декілька; декілька підходів; підходів до; до автоматичного; автоматичного виділення; виділення стійких; стійких словосполучень. Нажаль в результаті застосування цього методу на великих обсягах тексту отримуємо так зване «сміття» із-за високої частоти службових слів. Метод також вимагає враховувати частоту появи та шаблони словосполучень.

2. Метод t-тест полягає у перевірці статистичних гіпотез та використання статистичної моделі MA:

- H_0: слова зустрілись випадково;
- $P(w^1w^2) = P(w^1)P(w^2)$;
- врахування не тільки пар, але і частоти вживання окремих слів (тих, що складають пару);
- $t = \frac{\bar{x}-\mu}{\sqrt{\frac{s^2}{N}}}$, де $\bar{x}$ - емпіричне середнє, μ - теоретичне середнє, s^2 - емпірична дисперсія, N - розмір емпіричної вибірки.

Метод є не зовсім коректний для мови, але дозволяє отримати результати на практиці, наприклад, частота появи стійкого словосполучення «контент аналіз» в [14] при $P(контент) = 85/4338$ та $P(аналіз) = 53/4338$ є

$$H_0: P(аналіз) = P(контент)P(аналіз) \approx 2{,}39 \cdot 10^{-4}.$$

В схемі Бернулі $s^2 = p(1-p) \approx p$ при значеннях $\bar{x} = 18/4338$ та $t \approx 3{,}997955$.

3. Метод χ^2 Пірсона застосовують до таблиць розміром 2x2 (Таблиця 5.6). В розрахунках не очікують нормальності.

Таблиця 5.6

Приклад застосування методу χ^2 Пірсона

w_i	w_1 = контент	$w_1 \neq$ контент
w_2 = аналіз	18 (контент аналіз)	35 (e.g., статистичний аналіз)
$w_2 \neq$ аналіз	67 (в тому числі, контент моніторинг)	4218 (в тому числі, статистичний моніторинг)

Наприклад, $\chi^2 = \frac{N(O_{11}O_{22}-O_{12}O_{21})^2}{(O_{11}+O_{12})(O_{11}+O_{21})(O_{12}+O_{22})(O_{21}+O_{22})} \approx 286{,}0595.$

4. Метод LR полягає в розрахунку гіпотез ($p_1 >> p_2$)

$$H_1: P(w^2|w^1) = p = P(w^2|\neg w^1)$$

$$H_2: P(w^2|w^1) = p_1 \neq p_2 = P(w^2|\neg w^1)$$

де $p = \frac{c_2}{N}$; $p_1 = \frac{c_{12}}{c_1}$; $p_2 = \frac{c_2 - c_{12}}{N - c_1}$. Тоді, використовуючи біноміальний розподіл $b(m,n,p) = C_m^n p^m (1-p)^{n-m}$,отримаємо відношення правдоподібності LR

$$L(H_1) = b(c_{12}, c_1, p) b(c_2 - c_{12}, N - c_1, p),$$

$$L(H_2) = b(c_{12}, c_1, p_1) b(c_2 - c_{12}, N - c_1, p_2),\ log\ \lambda = {L(H_1)}\Big/{L(H_2)},$$

де $-2\ log\ \lambda$ в асимптотиці розподілено як χ^2.

Експеримент вилучення термінів провадився на 3-ох статтях із різних ПО. Шаблоном проведення експерименту є: [Прикм.+Ім.], [Дієприкм. + Ім.], [Ім. + Ім., Род. В.], [Ім. + Ім., Ор. В.], [Ім. + '-' + Ім.]. Під час експерименту використано 6 методів: визначені авторами статтей вручну (A); визначені системою Victana.lviv.ua, враховуючи закон Зіпфа (B); частота+морфологічні шаблони FREG (C); t-тест (D); відношення правдоподібності LR (F); статистика χ^2 (G). Поведено аналіз 3-ох статтей українською мовою, та перекладених на англійську (Таблиця Д.1-Таблиця Д.2 додатку Д). Жирним виділені ключові слова, які зустрічаються у результатах застосування всіх методів, курсивом – лише в методах B-G, а підкресленні – в методах A та C-G. При проведенні лінгвістичного аналізу для формування алфавітно-частотних словників по два слова використано такі особливості:

- Біграми формувалися у межах знаків пунктуації (якщо між словами існувало хоч якийсь знак пунктуації – ці слова не вважалися біграмою);
- Алафавітно-частотний словник з двох слів формувався на основі їх основ (біграм) та контент-аналізу цих біграм;
- При аналізі флексій аналізованих слів не враховувалися дієслова при формуванні алфавітно-частотного словника біграм (дієслова вважалися одним із знаків пунктуації);

- Перед лінгвістичним аналізом текстів були вилучені всі стопові слова (частки, прислівники, сполучники) та займенники.

Статистичні методи дозволяють врахувати вживання окремих слів. Тонкощі пов'язані із застосуванням методів для різних обсягів даних та діапазонів ймовірностей (кращий за t-тест для більших p, де нормальність порушується; відношення правдоподібності краще апроксимується χ^2 ніж таблиці 2x2 для малих обсягів). Частіше використовується не для прийняття/відхилення гіпотез, а для ранжування словосполучень-кандидатів.

Для порівняння з отриманими результатами використаємо бібліотеку від Google – word2vec, яка зарекомендувала себе в якості альтернативи tf-idf (A_1 – Таблиця Д.3 додатку Д). Використаємо також вбудовані методи для пошуку словосполучень на Python. Але на цих датасетах вона не дуже добре спрацювала, бо для гарної роботи їй потрібні величезні корпуси. Найцікавіше що вона дозволяє робити це після переводу кожного слова з корпусу в простір, розмірність якого задає користувач, наприклад,

'король' + 'жінка' - 'чоловік' = 'королева'

Після переводу в простір деякої розмірності кожне слово стає вектором, тому з ними можна утворювати базові операції складання, віднімання, множення, тощо. Також розглянемо аналіз через біграми (A_2 – Таблиця Д.3 додатку Д) і скіп грами (A_3 – Таблиця Д.3 додатку Д). Результати кращі за word2vec, а саме найкраще впоралися аналіз скіпграм із значенням 3 і також із очищенням від стоп-слів в англійській мові (A_4 – Таблиця Д.3 додатку Д). Але цим результатам досить далеко до отриманих в Таблиця Д.1 додатку Д. Результат погіршений за рахунок не врахування знаків пунктуації та використання стопових слів при лінгвістичному аналізі як змістовних.

5.2. Параметрична рубрикація тексту українською мовою

При класифікації тексту реалізують визначення граматичних мета-даних слова на основі графемного/морфологічного аналізу (Рис. 5.23, алг. 5.4).

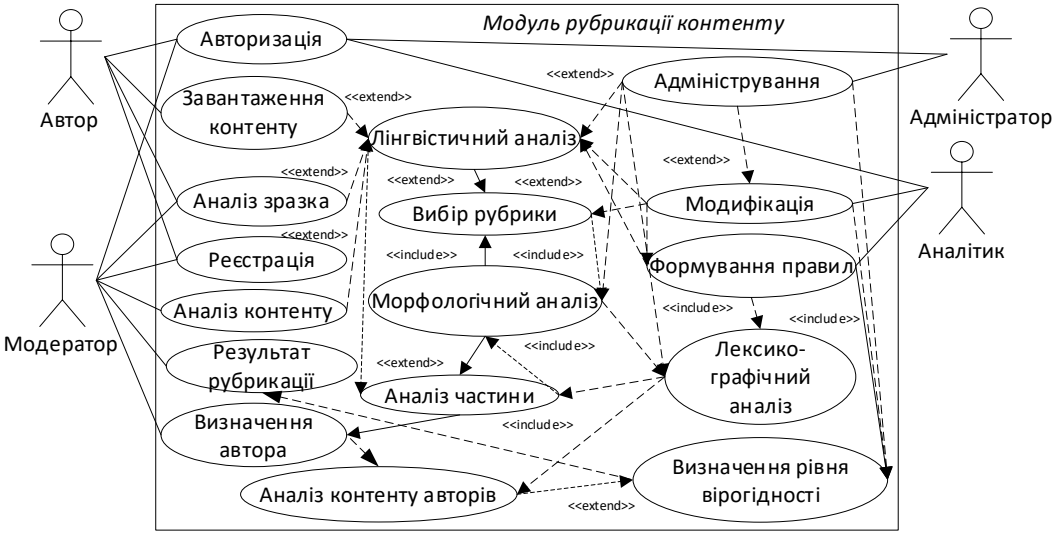


Рис. 5.23. Діаграма варіантів використання класифікації тексту

Алгоритм 5.4. Тематична класифікація україномовного контенту

Етап 1. Розбиття україномовного тексту C_r на частини (абзаци/параграфи тощо).

Крок 1. Завантаження в модуль генерування дерева частин тексту C_r.

Крок 2. Формування нового масиву стрічок в структурі.

Крок 3. Парсинг рядків символів частин тексту C_r.

Крок 4. Ідентифікація крапки як закінчення речення, а не частини скорочення та перехід до кроку 5, інакше збереження в масиві та перехід до кроку 3.

Крок 5. Ідентифікація символу закінчення тексту та перехід до кроку 6, інакше маркування закінчення частини тексту перехід до кроку 2.

Крок 6. Збереження дерева частин тексту C_r як структури $U_{CT}^{B} \in U_{CT}$.

Етап 2. Розбиття частини на вирази зі збереженням структури тексту $C_з$.

Крок 1. Аналіз нової структури частини тексту $U_{CT}^{B} \in U_{CT}$. Формування структури виразу (абзацу/речення тощо) $U_{CT}^{R} \in U_{CT}$ із ключем ID_part типу *n-to*-1 із структурою частин тексту C_r.

Крок 2. Формування нового масиву в структурі речень $U_{CT}^{R} \in U_{CT}$.

Крок 3. Парсинг символів до знаку наступного пунктуації.

Крок 4. Якщо скорочення або спецзапис (дата, гроші тощо) згідно регулярного виразу, то відповідне маркування цієї послідовності та перехід до кроку 5, інакше збереження у структурі $U_{CT}^{R} \in U_{CT}$ та перехід до кроку 2.

Крок 5. Якщо кінець частини тексту, то маркування та перехід до кроку 6, інакше перехід до кроку 2.

Крок 6. Збереження дерева речень у вигляді структури $U_{CT}^{R} \in U_{CT}$.

Крок 7. Якщо кінець тексту, то перехід до етапу 3, інакше перехід до кроку 1.

Етап 3. Розбиття речень на лексеми зі збереженням зв'язку з відповідним реченням $U_{CT}^{L} \in U_{CT}$ та відповідно номеру позиції в реченні.

Крок 1. Формування структури лексем $U_{CT}^{L} \in U_{CT}$ із полями ID_lex, ID_sent, N_lex, T_lex як опис мета-даних лексеми.

Крок 2. Аналіз лексеми речення з $U_{CT}^{R} \in U_{CT}$.

Крок 3. Формування нової лексеми в структурі лексем $U_{CT}^{L} \in U_{CT}$.

Крок 4. Парсинг символів до першого символу не з українського алфавіту або апострофа та збереження в структурі лексем.

Крок 5. Якщо символ закінчення речення, то перехід до кроку 6, інакше перехід до кроку 3.

Крок 6. Синтаксичний аналіз на основі алг. 5.2.

Крок 7. Морфологічний аналіз на основі отриманих ланцюжків лексем.

Етап 4. Ідентифікація тематики україномовного тексту $U_{CT}^{T} \in U_{CT}$.

Крок 1. Ідентифікація ієрархічної структури ознак $U_{CT}^{T} \in U_{CT}$ кожної семантично значущої лексеми із іменникової групи, окрім займенників.

Крок 2. Генерування словника з ієрархією типів властивостей лексем.

Крок 3. Уніфікація при необхідності подібних лексем.

Крок 4. Ідентифікація множини ключових слів $KeyWords$ тексту $C'_r = \alpha_r(\alpha_m(C_r, U_K), U_{CT})$ при $U_{CT} = \{U_{CT1}, U_{CT2}, U_{CT3}, U_{CT4}\}$, де U_{CT} – колекція умов класифікації, U_{CT1} – множина тематичних ключових слів, U_{CT2} – множина частот появи ключових слів, U_{CT3} – залежності появи ключових слів згідно різних тем, U_{CT4} – частоти появи тематичних ключових слів.

Крок 5. Формування $U_{Ct}^{T} \in U_{Ct}$ в множині $KeyWords$ з $TKeyWords$ (тематичні ключові слова) для $Topic$ та $Category$.

Крок 6. Розрахунок $QuantitativeryTKey$ (частоти появи тематичних ключових слів) та $FKeyWords$ (частоти появи ключових слів), а також коефіцієнтів $Static$ (статистичної важливості термів), $CofKeyWords$ (тематичних ключових слів контенту), $Comparison$ (появи ключових слів різних тем), $Addterm$ (міри наявності додаткових термів).

Крок 7. Розрахунок Якщо є збіг ключових слів контенту з ключовими поняттями тем, то перехід до кроку 9, інакше перехід до кроку 8.

Крок 8. Генерування нової рубрики з набором ключових термінів тексту C'_r.

Крок 9. Присвоєння визначеному класу теми терміни аналізованого тексту C'_r.

Крок 10. Розрахунок $Location$ – коефіцієнта ваги контенту C'_r в темі.

Етап 5. Заповнення мета-даними україномовного аналізованого тексту для атрибутів $Topic$, $Category$, $Location$, $Static$, $Addterm$, $CofKeyWords$, $TKeyWords$, $FKeyWords$, $Comparison$, $QuantitativeryTKey$.

5.3. Виявлення дублювання/плагіату/рерайту контенту

При визначенні дублювання текстового контенту (наприклад при ідентифікації плагіату/рерайту або дублів інтегрованого контенту з різних джерел) основна NLP-задача полягає у аналізі ступеня подібності рядків. Також це можна застосувати при перевірці орфографії або автокорекції введення тексту як інтуїтивному передбаченні, що саме користувач хоче ввести. Іншим прикладом слугує ідентифікації ключового значення текстового контенту або визначення, чи два рядки *Національний університет* «*Львівська політехніка*» або *НУ* «*Львівська політехніка*» є одним за значенням ключовим словом. Мінімальна редакційна відстань дозволяє кількісно оцінити припущення про подібність аналізованих рядків як обчислення мінімальної кількості операцій редагування через вставлення (i), видалення (d), заміщення (r), синомізацію (s), перестановку (p), необхідні для перетворення одного рядка в інший (Рис. 5.24). Вирівнювання на основі порожнього рядка/символу є відповідністю між підрядками двох послідовностей рядків/речень/слів.

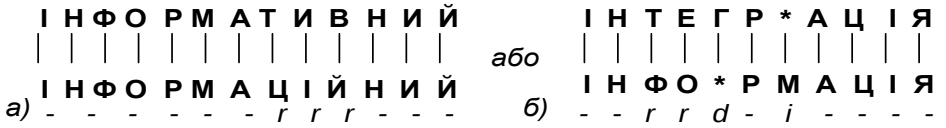


Рис. 5.24. Схема аналізу мінімальної редакційної відстані

Для кожної з таких операцій призначають певну вартість/вагу. Відстань Левенштейна між двома рядками є найпростішим ваговим коефіцієнтом, в якому кожна з п’яти операцій має вартість 1 [1018]. Відстань Левенштейна для схеми Рис. 5.24а рівна 3, а для схеми Рис. 5.24б рівна 4. Альтернативною є метрика, де кожне вставлення/видалення оцінюють в 1, а інші операції не допускаються або оцінюються по 2 (r), 3 (p) та 4 (s) відповідно. Тоді для схем Рис. 5.24 відстані Левенштейна рівні по 6 кожна. Процес знаходження мінімальної редакційної відстані (Рис. 5.25) полягає у пошуку найкоротшого шляху – послідовності редагувань від одного символьного рядка до іншого (Рис. 5.26) на основі динамічного програмування [1019-1022].

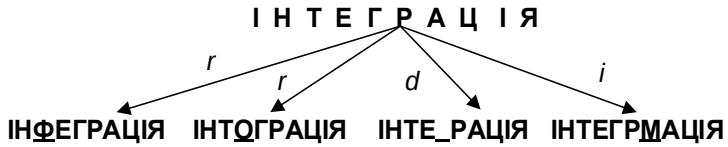


Рис. 5.25. Схема процесу знаходження редакційної відстані

Рис. 5.26. Приклад шляху редагування від одного рядка до іншого

Алгоритм 5.5. Мінімальна редакційна відстань на основі [1023].

Етап 1. Визначаємо $S[0,0] = 0$, $n = const$, $m = const$, $i = 0$, $j = 0$.

Етап 2. Парсимо текст X та виділяємо рядок довжини n для порівняння. Позначаємо вхідний рядок як A ($|A| = n$). Визначаємо $S[i, 0] = S[i - 1,0] + d - f_m(A[i])$.

Етап 3. Парсимо текст Y та виділяємо рядок довжини m для порівняння. Позначаємо цільовий рядок як B ($|B| = m$) . Визначаємо $S[0, j] = S[i, j - 1] + i - f_m(B[j])$

Етап 4. Обчислення мінімальної редакційної відстані між двома рядками.

Крок 4.1. Ідентифікуємо $S[i,j]$ як відстань редагування між $A[1 \dots i]$ та $B[1 \dots j]$, тобто відстань між A та B є $S(n,m)$.

Крок 4.2. Обчислюємо $S[i,j]$ шляхом прийняття мінімуму з п'яти можливих шляхів через матрицю реакційних відстаней (Рис. 5.27):

$$S[i,j] = min \begin{cases} S[i-1,j] + d - f_{dm}(A[i]) \\ S[i,j-1] + i - f_{im}(B[j]) \\ S[i-1,j-1] + r - f_{rm}(A[i], B[j]) \\ S[i-1,j-1] + p - f_{pm}(A[i+1], A[i]) \\ S[i-1,j] + s - f_{sm}(A[i], X[j]) \end{cases} \quad (5.1)$$

A\B	#	І	Н	Ф	О	Р	М	А	Ц	І	Я
#	0	1	2	3	4	5	6	7	8	9	10
І	1	0	1	2	3	4	5	6	7	8	9
Н	2	1	0	1	2	3	4	5	6	7	8
Т	3	2	1	2	3	4	5	6	7	8	9
Е	4	3	2	3	4	5	6	7	8	9	10
Г	5	4	3	4	5	6	7	8	9	10	11
Р	6	5	4	5	6	5	6	7	8	9	10
А	7	6	5	6	7	6	7	6	7	8	9
Ц	8	7	6	7	8	7	8	7	6	7	8
І	9	8	7	8	9	8	9	8	7	6	7
Я	10	9	8	7	8	9	10	9	8	7	6

Рис. 5.27. Приклад матриці розрахунку мінімальної редакційної відстані

Якщо наперед відомі значення ваг для визначених операції, тоді обчислюємо як:

$$S[i,j] = min \begin{cases} S[i-1,j] + 1 \\ S[i,j-1] + 1 \\ S[i-1,j-1] + \begin{cases} 0, if\ A[i] = B[j] \\ 2, if\ A[i] \neq B[j] \end{cases} \\ S[i-1,j] + \begin{cases} 0, if\ A[i] = A[i+1] \\ 3, if\ A[i] \neq A[i+1] \end{cases} \\ S[i-1,j] + \begin{cases} 0, if\ A[i] = X[j] \\ 4, if\ A[i] \neq X[j] \end{cases} \end{cases} \quad (5.2)$$

Заповнення матриці розрахунку мінімальної редакційної відстані (Рис. 5.27).

Етап 5. Якщо не кінець тексту Y, то $j = j + 1$, парсимо текст Y, виділяємо наступний рядок довжини m для порівняння та перехід до етапу 4.

Етап 6. Якщо не кінець тексту X, то $i = i + 1$, парсимо текст X, виділяємо наступний рядок довжини n для порівняння та перехід до етапу 3.

Етап 7. Визначення в матриці розрахунку мінімальної редакційної відстані (Рис. 5.27) оптимального найкоротшого шляху – послідовності редагувань від одного символьного рядка до іншого (Рис. 5.28) [1024].

A\B	#	І	Н	Ф	О	Р	М	А	Ц	І	Я
#	**0**	← 1	← 2	← 3	← 4	← 5	← 6	← 7	← 8	← 9	← 10
І	↑ 1	↖ **0**	← 1	← 2	← 3	← 4	← 5	← 6	← 7	← 8	← 9
Н	↑2	↑1	↖ **0**	← 1	← 2	← 3	← 4	← 5	← 6	← 7	← 8
Т	↑ 3	↑2	↑1	↖ **2**	← 3	← 4	← 5	← 6	← 7	← 8	← 9
Е	↑ 4	↑3	↑2	↑3	↖ **4**	← 5	← 6	← 7	← 8	← 9	← 10
Г	↑ 5	↑4	↑3	↑4	↑ **5**	← 6	← 7	← 8	← 9	← 10	← 11
Р	↑ 6	↑5	↑4	↑5	↑6	↖ **5**	← **6**	← 7	← 8	← 9	← 10
А	↑ 7	↑6	↑5	↑6	↑7	↑6	↑7	↖ **6**	← 7	← 8	← 9
Ц	↑ 8	↑7	↑6	↑7	↑8	↑7	↑8	↑7	↖ **6**	← 7	← 8
І	↑ 9	↑8	↑7	↑8	↑9	↑8	↑9	↑8	↑7	↖ **6**	↖ 7
Я	↑ 10	↑9	↑8	↖ 7	← 8	↑9	↑10	↑9	↑8	↑7	↖ **6**

Рис. 5.28. Приклад визначення шляху розрахунку мінімальної відстані

Крок 7.1. Визначаємо та зберігаємо послідовно у кожній клітинці $S[i,j]$ матриці розрахунку мінімальної редакційної відстані 1-3 зворотних вказівників (зліва, зверху і/або по діагоналі) на попередню комірку ($S[i-1,j]$, $S[i,j-1]$ і/або $S[i-1,j-1]$) з якої можливий перехід у поточну клітинку (Рис. 5.28), не порушуючи зміни редакційної відстані.

Крок 7.2. Аналізуючи з останньої комірки $S[n,m]$, рухаємося через матрицю за зворотними вказівниками до $S[0,0]$, не порушуючи зміни послідовності редагувань та визначаючи найкоротший шлях редакційної відстані.

Кожна комірка з жирним шрифтом відображає вирівнювання пари літер у двох рядках. Якщо в одному рядку виділяються дві суміжні комірки, тоді реалізована операція вставлення від джерела до цілі, наприклад літери М після Р (5→6); дві підряд смужки, розташовані в одному стовпчику, позначають видалення, наприклад, літери Г після Е (замінену перед цим на Ф, тобто 4→5).

Аналогічно алгоритм визначення мінімальної відстані можна застосувати для слів в реченні (перевірка плагіату/рерайту, розрахунок втрат мовлення, машинний переклад) замість символів в рядку (перевірка орфографії правопису, розрахунок частоти помилки слова). Наприклад, для корекції правопису, заміни

ймовірно відбудуться між літерами відповідної природної мови, розташованими поруч на клавіатурі. Алгоритм Viterbi [1025] є найкращим варіантом розрахунку мінімальної редакційної відстані, вираховуючи максимальну вірогідність вирівнювання одного рядка з іншим.

Для розпізнавання текстового контенту як дубль/плагіат або частковий рерайт достатньо порівнювати символьні ланцюжки шаблону та аналогів із знаходження мінімальної відстані. Для розпізнавання контенту із суттєвим рерайтом цього недостатньо. Тоді розпізнавання полягатиме в ідентифікації колекції концептів і термів відповідного тексту-шаблону на основі розрахунку міри подібності до ймовірних аналогів текстового контенту [1026-1032]. Колекцію ідентифікованих концептів та термів доповнюють з онтології іншими на основі узагальнених зв'язків типу IS-A на один рівень догори та іншими семантичними зв'язками, вага важливості яких перевищує порогове значення [1026-1032]. Зв'язки між концептами та термами у контенті ідентифікують для усунення неоднозначності розпізнавання для формування зв'язного графу семантичного образу відповідного контенту [1026-1032]. Порівняння подібності випливає із розрахунку семантичної відстані між відповідним контентом (Рис. 5.29) [1026-1032].

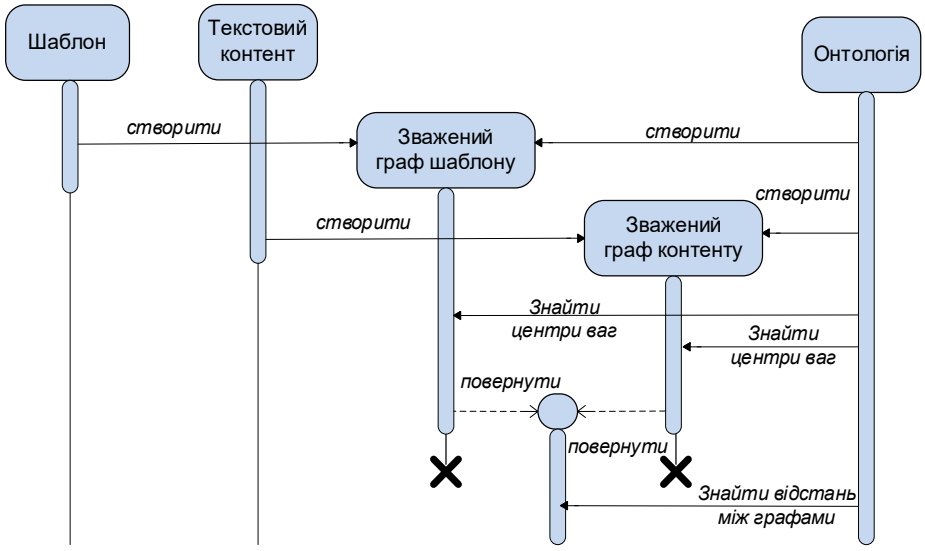


Рис. 5.29. Діаграма послідовності семантичного порівняння контенту

Процес порівняння контенту та ранжування за подібністю за допомогою онтології з пошуком текстових ланцюжків за взірцем містить [1026-1032]:

1) Зважений концептуальний граф G текстового контенту-шаблону.
2) Доповнений онтологією контенту-шаблону зважений концептуальний граф G' із знаходженням батька до кожної вершини G на основі зв'язків між концептами.
3) Зважений концептуальний граф $\hat{G} = G \cup G'$ на основі результатів СА та СЕМ.
4) Редукції надлишкових елементів зваженого концептуального графу $\hat{G}$.
5) Обчислення центрів ваг (Рис. 5.30) і семантичної відстані між G та G'.

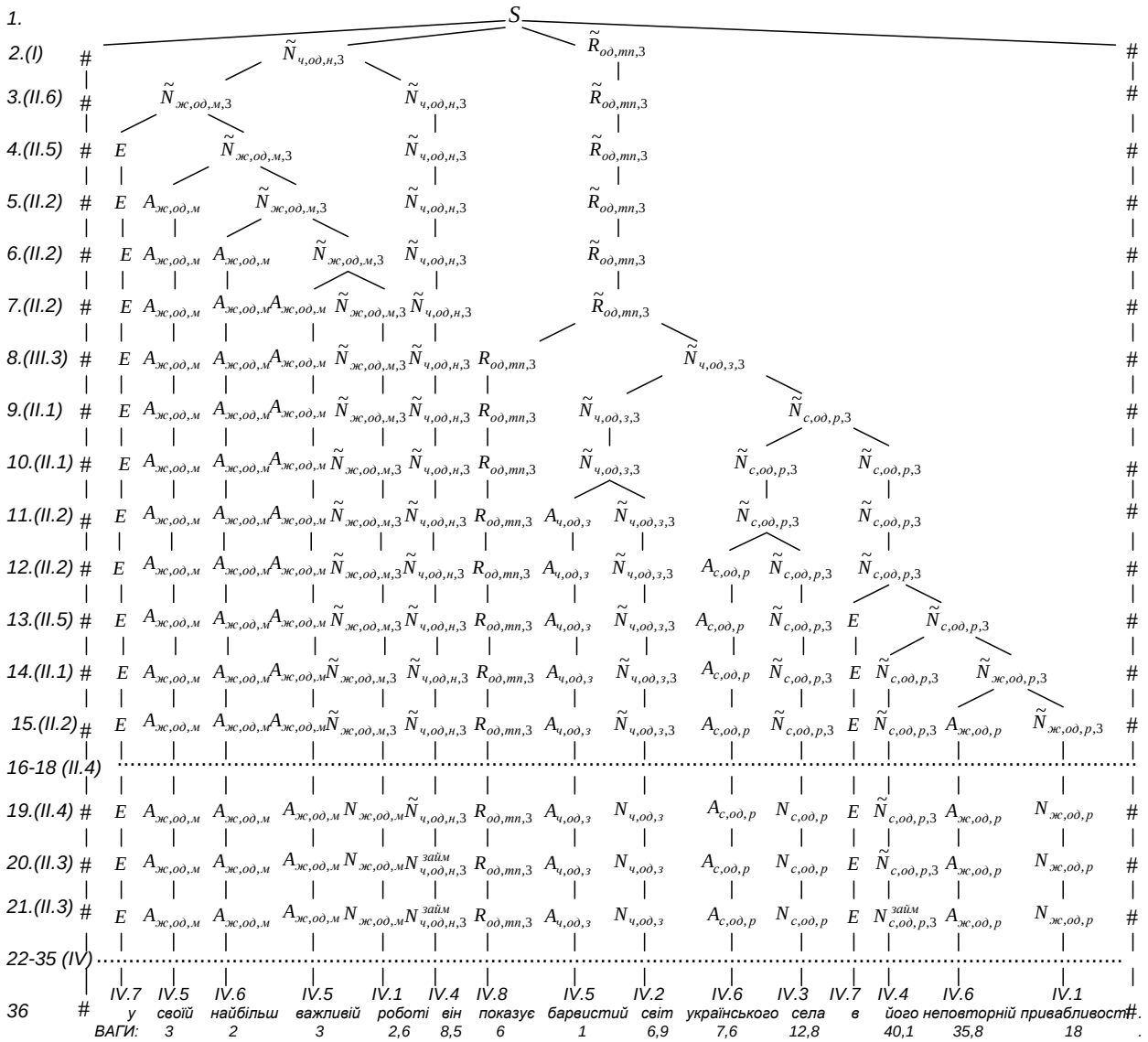


Рис. 5.30. Результат СА та СЕМ до речення українською

Згідно експериментальної апробації з анотаціями науково-технічних публікацій підхід на основі адаптивної онтології підвищує точність пошуку подібності контенту у середньому на 18 % у порівнянні з методом зважених концептуальних графів (Монтеза-Гомеса) та 27% у порівняння з методом за

коефіцієнтом Дайса (Рис. 5.30). Аналіз ефективності перелічених методів проведено за параметром – точність пошуку [1026-1032]:

$$точність = \frac{число\_знайдених\_релевантних(експерт)}{число\_усіх\_знайдених(програма)} \quad (5.3)$$

Таблиця 5.7

Порівняння методів [1026-1032]

Назва методу	Точність χ, %
на основі адаптивної онтології	90
зважених концептуальних графів (Монтеза-Гомеса)	72
за коефіцієнтом Дайса	63

Метод за коефіцієнтом Дайса ідентифікував 63% подібного контенту до шаблону лише ті анотації науково-технічних публікацій, де присутня найбільша кількість спільних слів із шаблоновими, але не завжди корелювалися із змістом прототипу. У той час метод на основі адаптивної онтології дав найкращий результат з врахуванням подібності контексту шаблону та аналогів.

5.4. Основні результати та висновки розділу

Розглянуто особливості методу синтаксичного аналізу україномовного текстового контенту, спрямованого на автоматичне виявлення значущих ключових слів вхідних текстів. Визначено роль і формальні ознаки синтаксичного аналізатора в процесі виявлення ключових слів тематики контенту, проведено декомпозицію процедур запропонованого методу на 4-х етапах. Порівняно з відомими синтаксичними аналізаторами, запропонований метод забезпечує самовдосконалення та самонавчання автоматизованої системи визначення ключових слів за рахунок механізму ідентифікації значущих статистичних параметрів у визначених модератором межах. Експериментальне дослідження підтвердило достовірність методу – для різних методик опрацювання первинного тексту середній збіг списків виявлених ключовиків з авторськими змінюється у проміжку 52,6-68,5%. Точність збігу ключових слів із авторськими коливається в проміжку 43,6-62,9%. Середній збіг змістовних ключових слів порівняно зі всіма знайденими системою коливається в проміжку

38,9-75,8% згідно із етапами аналізу текстів статей. Точність збігу ключових слів порівняно зі всіма знайденими системою коливається в проміжку 34,3-71,9% відповідно до етапів аналізу текстів статей.

Розроблено метод визначення стійких словосполучень на основі ідентифікації ключових слів україномовного тексту та аналізу коефіцієнтів лексичного мовлення автора тексту в еталонних уривках контенту, що дало можливість на основі статистичної лінгвістики покращити точність методу визначення стилю автора тексту на 9%. Метод полягає у використанні закону Зіпфа при формуванні стійких словосполучень як ключових з врахуванням наступних правил попереднього лінгвістичного опрацювання тексту: вилучення всіх стових слів; біграми формувати лише в межах знаків пунктуації; дієслово та займенник вважати знаками пунктуації; дієслова визначати за їх флексіями; біграми формувати на основі їх основ без врахування їх флексій; визначення прикметників за їх флексіями та вважати, що прикметники мають бути лише на першому місці у біграмі з україномовних текстів. Розроблено програми комплекс для визначення стійких словосполучень як ключових. Запропоновано підхід до розроблення ПЗ лінгвістичного контент-аналізу для визначення стійких словосполучень при ідентифікації ключових слів текстового україномовного та англомовного контенту. Особливість підходу полягає у адаптації лінгвостатистичного аналізу лексичних одиниць до особливостей конструкцій україномовних та англомовних слів/текстів. Досліджено результати експериментальної апробації запропонованого методу контент-аналізу англомовних та україномовних текстів для визначення стійких словосполучень при ідентифікації ключових слів технічних текстів.

Основні результати розділу опубліковані у роботах [163, 535, 958-983, 984-1008].

РОЗДІЛ 6
ТЕХНОЛОГІЯ ОПРАЦЮВАННЯ УКРАЇНОМОВНОГО ТЕКСТУ ДЛЯ ІДЕНТИФІКАЦІЇ ПЕРСОНАЛЬНИХ ОЗНАК АВТОРА КОНТЕНТУ

6.1. Особливості та типові ознаки авторського тексту

Вагомим значенням у лінгвостатистиці є аналіз зміни динаміки та частоти появи у тексті лінгвістичної одиниці. Дослідження коефіцієнтів персональних ознак авторського стилю (алг. 6.1) ґрунтується на розрахунках та аналізі [385]:

– ступеня концентрації авторського тексту ($I_{kt} = W_{10}/W$): відношення кількості слів із абсолютною частотою появи в тексті ≥10 до кількості всіх слів;

– ступеня винятковості авторського тексту ($I_{wt} = W_1/W$): відношення кількості слів із абсолютною частотою появи рівно 1 до кількості всіх слів;

– ступеня зв'язності авторського мовлення ($K_z = (Z + S)/(3P)$): частка появи службових слів в окремих реченнях україномовного текстового контенту;

– ступеня синтаксичної складності авторського мовлення ($K_s = 1 - P/W$): залежність кількості речень в тесті до кількості слів (не загальної кількості слів);

– ступеня лексичної різноманітності авторського мовлення ($K_l = W/N$): частка обсягу словникового запасу слів з тексту до загального обсягу всіх слів.

Алгоритм 6.1. Дослідження персональних ознак авторського стилю

Етап 1. Інтеграція з достовірних джерел, параметричне фільтрування (ліквідування інформаційного шуму, наприклад, тегів, рисунків тощо) та форматування україномовного тексту (наприклад, ліквідування апострофу або заміна на один тип, ліквідування). Важливим є спосіб організації відбору та обсяг текстової вибірки: для визначення характеристик він повинен складати щонайменше 18 тис. слів.

Етап 2. Лематизація україномовного текстового контенту.

Етап 3. Ліквідування неоднорідності лінгвістичних одиниць (наприклад, приведення скорочень до повного запису, або числових значень).

Етап 4. Генерування частотних словників україномовного текстового контенту на основі статистичної дистрибуції у необхідних числових метриках.

Етап 5. Ідентифікація/розрахунок коефіцієнтів/індексів персональних ознак авторського стилю на основі частотних словників, наприклад, аналіз частки та особливостей появи службових/стопових/маркованих слів, та розділових знаків, слів/речень/абзаців/параграфів/розділів різних довжин тощо.

Етап 6. Аналіз коефіцієнтів/індексів на ступень точності та достовірності.

Етап 7. Лексико-статистичне моделювання дистрибуцій ознак стилю автора.

Етап 8. Генерування шаблонів авторського стилю в межах певного жанру або тематики в певному часовому проміжку.

Етап 9. Експериментальна апробація для навчання системи оцінювання рівня приналежності україномовних текстів певного жанру/тематики до шаблону конкретного авторського стилю.

6.2. Метод визначення стилю автора україномовних текстів на основі технологій лінгвометрії, стилеметрії та глотохронології

Кожна мова характеризується множиною службових слів (частка, сполучник та прийменник - Таблиця 6.1 - понад 70 слів), а на персональний стиль автора зокрема завдяки особливості повсякденного мовлення накладаються особливості вживання цих слів. Наприклад, оди автор надає перевагу слову *однак*, інший *отже*, або не враховуючи правила української мови надають перевагу часто одному із сполучників як *і*, *та*, *й*. Хтось надає перевагу прийменнику *тобто*, а хтось його аналогам. Аналіз та порівняння появи та частоти стоп-слів як службових (є ще слова-паразити, характерні певному автору для висловлювання певної тематики, сленг, суржик тощо) дає можливість змоделювати лексико-статистичний шаблон стилю конкретного автора.

Таблиця 6.1

Службові частини української мови (стоп-слова)

Частина мови	Список стопових слів
Прийменники	без, біля, близько, в, вглиб, від, для, до, з, за, з-за, з-під, крізь, на, над, під, по, поза, при, про, проміж, у, через
Сполучники	а, або, але, й, і, коли, немов, одначе, проте, та, та й, так, також, тобто, через те що, хоча, чи, що, щоб, якщо
Частки	або, адже, аякже, би, вже, ж, же, ледве чи, лише, мов, немов, навіть, не, ні, он, ось, так, тільки, то, тобто, уже, це, чи

На Рис. 6.1–Рис. 6.4 подане графічне зображення відносної частоти появи стопових слів в чотирьох різних текстах (Уривок 1-4) та в шаблоні (Еталон) на основі статистичних даних появи службового слова (Таблиця Д.4 додатку Д).

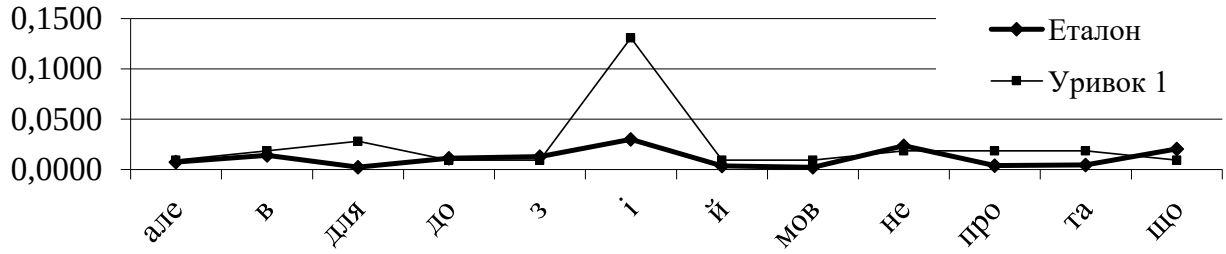


Рис. 6.1. Ймовірність появи стоп-слів (коефіцієнт кореляції – $R_{e\text{-}y1}$=0,6076)

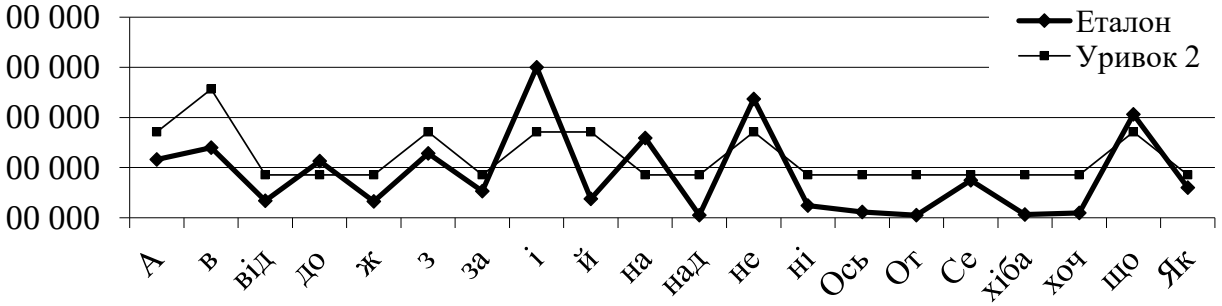


Рис. 6.2. Ймовірність появи стоп-слів (коефіцієнт кореляції – $R_{e\text{-}y2}$=0,7066)

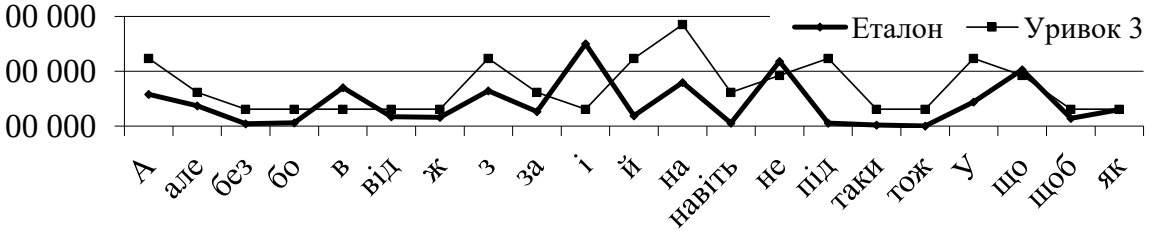


Рис. 6.3. Ймовірність появи стоп-слів (коефіцієнт кореляції – $R_{e\text{-}y3}$=0,2810)

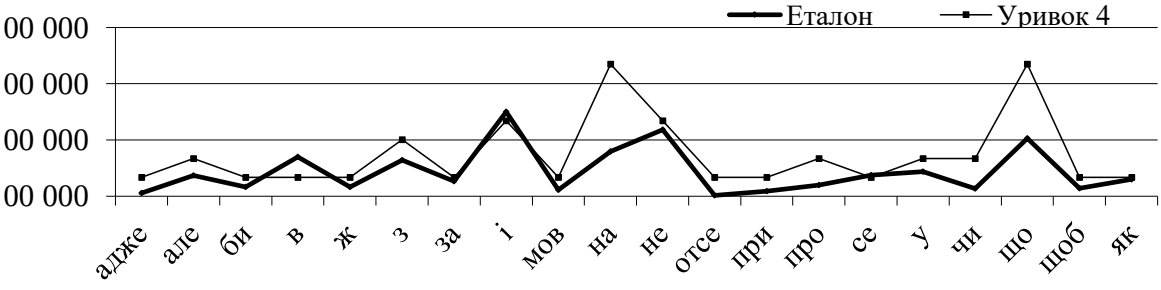


Рис. 6.4. Ймовірність появи стоп-слів (коефіцієнт кореляції – $R_{e\text{-}y4}$=0,7326)

Результати аналізу чотирьох текстів (Таблиця 6.2) демонструють, що більш ймовірно Уривок 4 належить автору шаблону (хоча між результатами дослідження текстів 4 та 2 є не суттєва різниця, але вони все таки якщо написані

в одному проміжку часу, не належать автору шаблону, якщо в різні проміжки з шаблоном – ймовірність приналежності цьому автору зростає).

Таблиця 6.2

Коефіцієнти кореляції для стоп-слів

№	R_{e-U}	Частка	Сполучник	Прийменник	R'_{e-U}
4	**0,7326**	**0,9594**	**0,9544**	**0,5639**	**0,6905**
2	0,7066	0,9580	0,5714	0,4928	0,4913
1	0,6076	1	0,79	0,72	0,6900
3	0,2810	0,8800	0,1624	0,1517	0,2254

Отже, застосування методу опорних слів дало такі результати: серед досліджуваних уривків найбільшу ймовірність належати до еталону справді отримав той уривок, автором якого є й автором еталону. Інші результати також підтверджують дієвість методу опорних слів у авторській атрибуції текстів. Так, у першому дослідженні наступну за величиною ймовірність належати до еталону має уривок з іншого твору того самого автора. Уривок 1, що теж належить до еталону, «програв» Уривку 4 лише одну десяту в коефіцієнті кореляції. Також адекватним є результат для Уривку 3, якого з еталоном розділяють близько ста років. Висунуте припущення про незначущість впливу частки як параметра методу на результати привело до зменшення коефіцієнтів кореляції. Понад усе, різниця між коефіцієнтами кореляції для Уривку 1 та Уривку 4 значно зменшилася і склала 0,0005. Проте, для підтвердження чи спростування того факту, що частки не є визначальним фактором в авторському стилі необхідно виконати ґрунтовніші дослідження. Для досягнення мети дослідження розроблено модуль з можливістю обрання мови/мов аналізованого контенту, яка реалізована на Web-ресурсі (Рис. 6.5). Експериментальна апробація функціонування модуля ідентифікації та аналізу колекції службових слів із 100 науково-технічних публікацій проведено у 3 етапи (алг. 6.2).

Алгоритм 6.2. Аналіз та інтерпретація лінгвостатистичних досліджень ідентифікації авторського стилю мовлення.

Етап І. Лексичний аналіз тексту для визначення стопових слів та розрахунку коефіцієнтів лексичного авторського мовлення (різноманітності тексту).

Крок 1. Фільтрування україномовного текстового контенту від інформаційного шуму (спеціальні символи, рисунки, теги, цифри, формули тощо).

Крок 2. Визначення розміру текстового контенту – зайве відсікається.

Крок 3. Ідентифікація обсягу речень P в україномовному текстовому контенті.

Крок 4. Ідентифікація кількості слів N в україномовному текстовому контенті.

Крок 5. Ідентифікація обсягу за частотним словником основ слів W.

Крок 6. Ідентифікація обсягу слів W_1, що вжиті в тексті рівно один раз.

Крок 7. Ідентифікація обсягу слів W_{10}, що вжиті в тексті ≥10 разів.

Крок 8. Ідентифікація обсягу прийменників Z в текстовому контенті.

Крок 9. Ідентифікація обсягу сполучників S в текстовому контенті.

Крок 10. Обчислення ступеня винятковості текстового контенту: $I_{wt}=W_1/W$.

Крок 11. Обчислення ступеня концентрації текстового контенту: $I_{kt}=W_{10}/W$.

Крок 12. Обчислення ступеня зв'язності мовлення тексту: $K_z=(Z+S)/(3*P)$.

Крок 13. Обчислення ступеня синтаксичної складності тексту: $K_s=1–P/W$.

Крок 14. Обчислення ступеня лексичної різноманітності тексту: $K_l=W/N$.

Крок 15. Табличне подання результатів на http://victana.lviv.ua/nlp/linhvometriia.

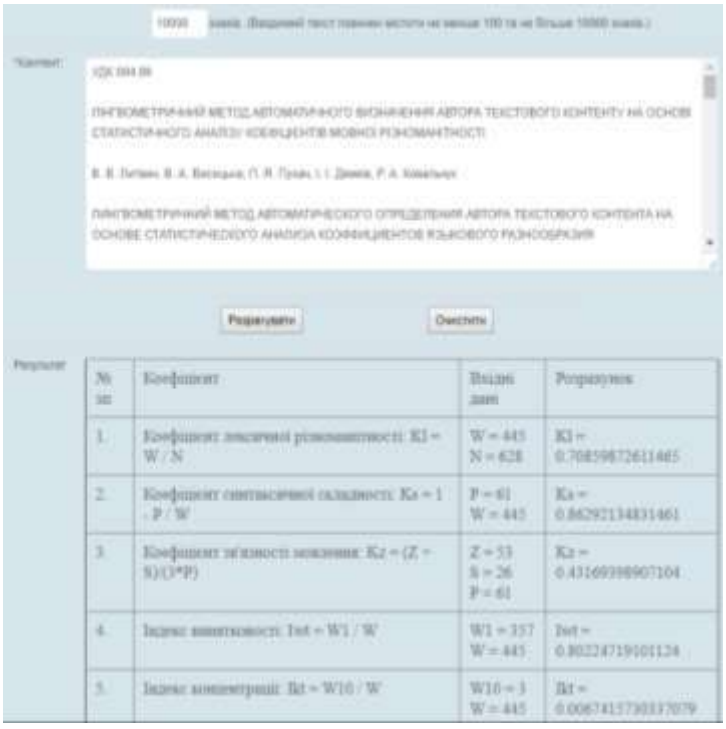


Рис. 6.5. Приклад аналізу тексту на http://victana.lviv.ua/nlp/linhvometriia

Етап II. Визначення стилю автора за методами стилеметрії.

Крок 1. Перевірка довжин еталонного тексту та вибраних уривків та приведення довжини еталонного тексту до мінімального із перевірених.

Крок 2. Очищення еталонного тексту від спецсимволів та ін.

Крок 3. Визначення кількості слів у тексті еталону.

Крок 3. Визначення кількості стоп-слів (прийменників + сполучників + часток) у тексті еталону.

Етап III. Аналіз тексту методом глотохронології згідно списку Сводеша.

Приклад результату лексичного аналізу одного україномовного текстового контенту для визначення стопових слів та розрахунку коефіцієнтів лексичного авторського мовлення (різноманітності тексту) подано в Таблиця 6.3.

Таблиця 6.3

Приклад аналізу авторського стилю мовлення

Ступень	Результат	Обчислення
винятковості: $I_{wt}=W_1/W$	W_1=141; W=184	I_{wt}=0.7663
концентрації: $I_{kt}=W_{10}/W$	W_{10}=2; W=184	I_{kt}=0.01
лексичної різноманітності: $K_l=W/N$	W=184; N=295	K_l=0.6237
синтаксичної складності: $K_s=1–P/W$	P=18; W=184	K_s=0.902
зв'язності мовлення: $K_z=(Z+S)/(3*P)$	Z=20; S=28; P=18	K_z=0.889

Для етапу III основним завданням є визначення кількості слів із 200-слівного списку Сводеша, які присутні в творах різних часових зрізів, та визначення відсоткового складу таких слів в уривках. Також дослідимо кількість спільних слів зі списку Сводеша для обраних уривків. Для розгляду підберемо фрагменти, написані з розривом у кілька років. Нехай уривки утворені, наприклад, із 250 слів, не враховуючи заголовка та власних назв. Порівняння 200-слівного списку Сводеша та Уривку 1 викладені у Таблиця 6.4 (спільні слова виділені кольором).

Таблиця 6.4

Слова зі списку Сводеша

№	Слово	АЧ	ВЧ	Слово	АЧ	ВЧ	Слово	АЧ	ВЧ
	Уривок 1			Уривок 2			Уривок 3		
1	і	19	0,2500	і	6	0,1224	і	10	0,2174
2	що	6	0,0789	що	3	0,0612	що	4	0,087
3	з	5	0,0658	з	2	0,0408	з	2	0,0435
4	все	4	0,0526	все	4	0,0816	все	3	0,0652
5	в	4	0,0526	в	7	0,1429	в	4	0,087

6	на	3	0,0395	на	1	0,0204	на	1	0,0217
7	там	3	0,0395	там	2	0,0408	там	1	0,0217
8	ні	3	0,0395	ні	1	0,0204	ні	7	0,1522
9	знати	2	0,0263	знати	2	0,0408	знати	2	0,0435
10	який	2	0,0263	який	4	0,0816	який	1	0,0217
11	ви	1	0,0132	ви	1	0,0204	вони	2	0,0435
12	what	1	0,0132	хто	1	0,0204	хто	2	0,0435
13	як	2	0,0263	якщо	1	0,0204	якщо	1	0,0217
14	он	5	0,0658	тут	2	0,0408	тут	1	0,0217
15	довго	2	0,0263	далеко	1	0,0204	довго	1	0,0217
16	я	6	0,0789	це	2	0,0408	це	1	0,0217
17	старий	2	0,0263	товстий	1	0,0204	інший	1	0,0217
18	слухати	1	0,0132	кидати	1	0,0204	казати	1	0,0217
19	чоловік	1	0,0132	потік	1	0,0204	приходити	1	0,0217
20	багато	1	0,0132	один	2	0,0408			
21	рік	1	0,0132	назад	1	0,0204			
22	ім’я	1	0,0132	інший	1	0,0204			
23	сонце	1	0,0132	білий	1	0,0204			
24				дещо	1	0,0204			
Усього		76		49			46		

В Уривку 1, обсягом 253 слова, є 23 слова з 200-слівного списку Сводеша. Ці слова складають 30,04 % від усього уривку. В Уривку 2, обсягом 262 слова, є 24 слова з 200-слівного списку Сводеша. Ці слова складають 18,7 % від усього уривку. В Уривку 3, обсягом 246 слів, є 19 слів із 200-слівного списку Сводеша. Ці слова складають 18,7 % від усього уривку. Аналізуючи отримані дані, зауважуємо, що слова зі списку Сводеша в Уривку 1 складають 30 % від уривку, що значно більше, ніж 18,7 %, як в Уривках 2 та 3 (Рис. 6.6а). Такі результати є закономірними та прозорими: з часом збагачується і словниковий запас людини. Також для цих уривків на Рис. 6.6б графічно зображено такі результати:

– у вузлах зазначено уривок та кількість слів у ньому зі списку Сводеша;

– на дугах вказано кількість спільних слів зі списку Сводеша для цих уривків та коефіцієнт кореляції для цих уривків;

– у центрі зазначена загальна кількість слів, спільних для уривків та списку Сводеша (Таблиця 6.4 - спільні слова виділені кольором).

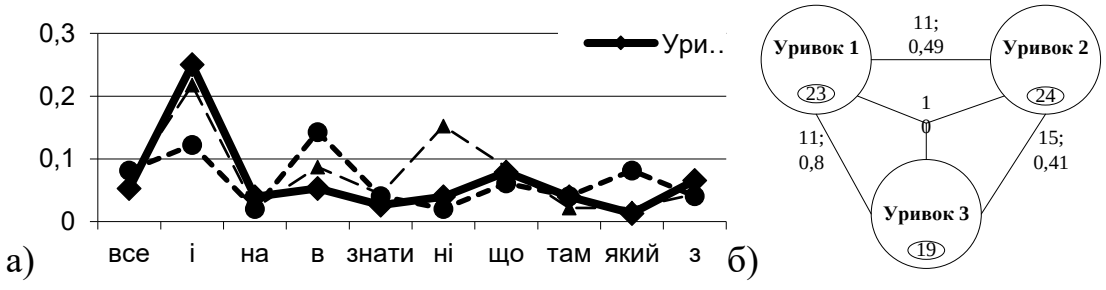


Рис. 6.6. Чисельні результати дослідження уривків

При експериментальній апробації проведено аналіз понад 300 україномовних уривків текстів (перші 10000 знаків) одноосібних (понад 100 авторів) науково-технічних публікацій Вісника НУ «Львівська політехніка» серії «Інформаційні системи та мережі» за період 2001–2021 рр. (алг. 6.3).

Алгоритм 6.3. Ідентифікація та аналіз колекції службових слів текстів

Етап 1. Дослідження публікацій на ідентифікацію діапазону оптимального обсягу аналізованого україномовного текстового контенту.

Крок 1. Аналіз україномовного текстового контенту в повному обсязі (алг. 6.2).

Крок 2. Аналіз уривків україномовного текстового контенту в діапазонах [10;1000000] знаків з початку науково-технічної публікації.

Крок 3. Аналіз отриманих результатів. Оптимальним аналізом україномовного текстового контенту є діапазон [100;10000] знаків. Якщо ≤100 знаків – значення стилістичних параметрів різних авторів подібні, а одного автора на різних уривках різних публікаціях – інколи є суттєво відмінними. Якщо ≥10000 знаків – параметри майже не змінюються, до того ж різні публікації мають ≤10000 знаків та досить мало публікацій мають ≥10000 знаків.

Крок 3. Аналіз уривків україномовного текстового контенту в діапазоні [100;10000] знаків понад 100 різних авторів для формування загальних стилістичних шаблонів автора.

Етап 2. Дослідження результатів зміни ступеня різноманітності авторського мовлення в залежності від часового проміжку на діапазоні [2001; 2021] рр. для формування періодичних стилістичних шаблонів автора.

Етап 3. Ідентифікація параметрів, які змінюються з часом та діапазон зміни, та параметрів, які не змінюються або суттєво не змінюються.

Етап 4. Дослідження публікацій для ідентифікації стилів мовлення автора за загальними та періодичними шаблонами в різні періоди часу [2001; 2021] р.

Етап 5. Дослідження обчислених параметрів мовлення для генерування підмножини потенційних авторів з подібним стилем, що і інші еталоні колективні роботи з періоду [2001; 2021] рр., серед авторів яких є автори шаблонів одноосібних науково-технічних публікацій.

Етап 6. Аналіз результатів. Якщо в згенерованих підмножинах потенційних авторів присутні справжні автори колективної роботи, то визначити параметри, які більш точно можуть це ідентифікувати. Експерименти провести на декількох алгоритмах. Обрати найкращий для ідентифікації стилю потенційного автора в текстах в різні періоди часу.

6.3. Лінгвометричний аналіз визначення автора контенту на основі статистичних параметрів різноманітності мовлення

Кожний автор з часом вдосконалює як свій словниковий запас та стиль написання публікацій. Тому необхідно дослідити чи змінюються параметри стилістичної різноманітності мовлення авторів з часом, які саме змінюються та в якому діапазоні (Рис. 6.7). З часом автори частіше застосовуються коротші слова (Рис. 6.7, *а*), а ступені лексичної різноманітності K_l та синтаксичної складності K_s суттєво не змінюються (рис. Рис. 6.7, *б–г*).

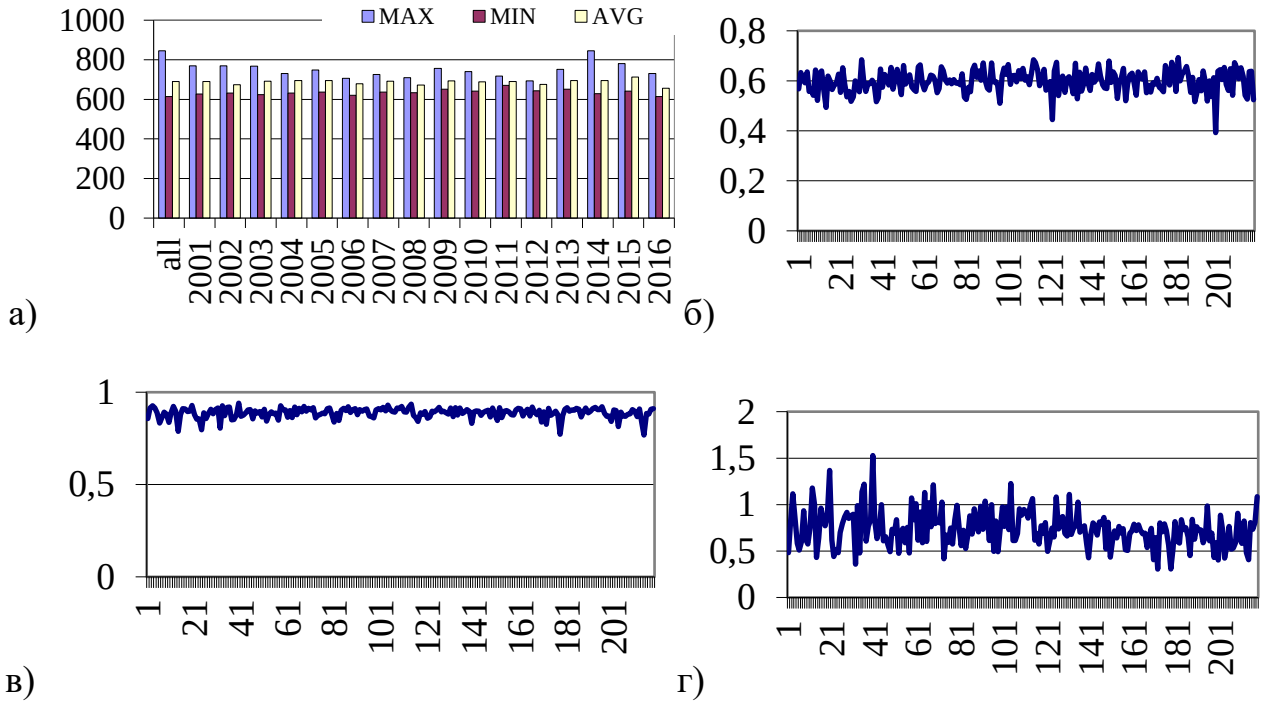


Рис. 6.7. Дистрибуція: *а* – слів та параметрів мовлення для однакових за обсягом текстів в діапазоні [2001; 2021] рр.: *б* – K_l; *в* – K_s; *г* – K_z

Ступень зв'язності мовлення K_z не суттєво зменшується. В 2001 р. змінюється в межах [0,5;1,2], а в 2021 р. – в межах [0,4; 0,9] (Рис. 6.8).

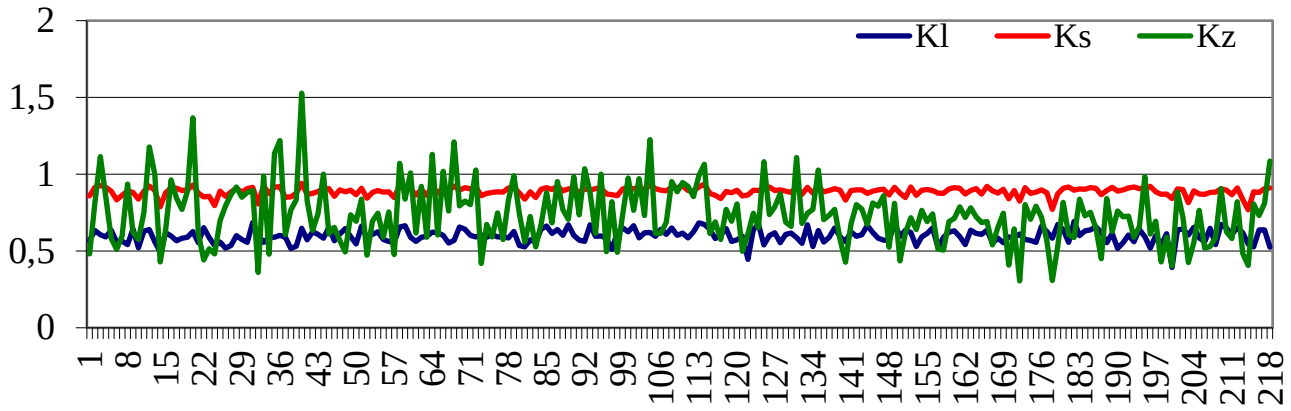


Рис. 6.8. Аналіз дистрибуції параметрів стилю мовлення K_l, K_s та K_z

Дистрибуція не змінюється суттєво в часі для параметра винятковості I_{wt}, а для параметра концентрації I_{kt} присутні суттєві зміни (Рис. 6.9). Наприклад, автори з плином часу для певної тематики дослідження частіше використовуються в своїх публікаціях деякі службові слова частіше (Рис. 6.10).

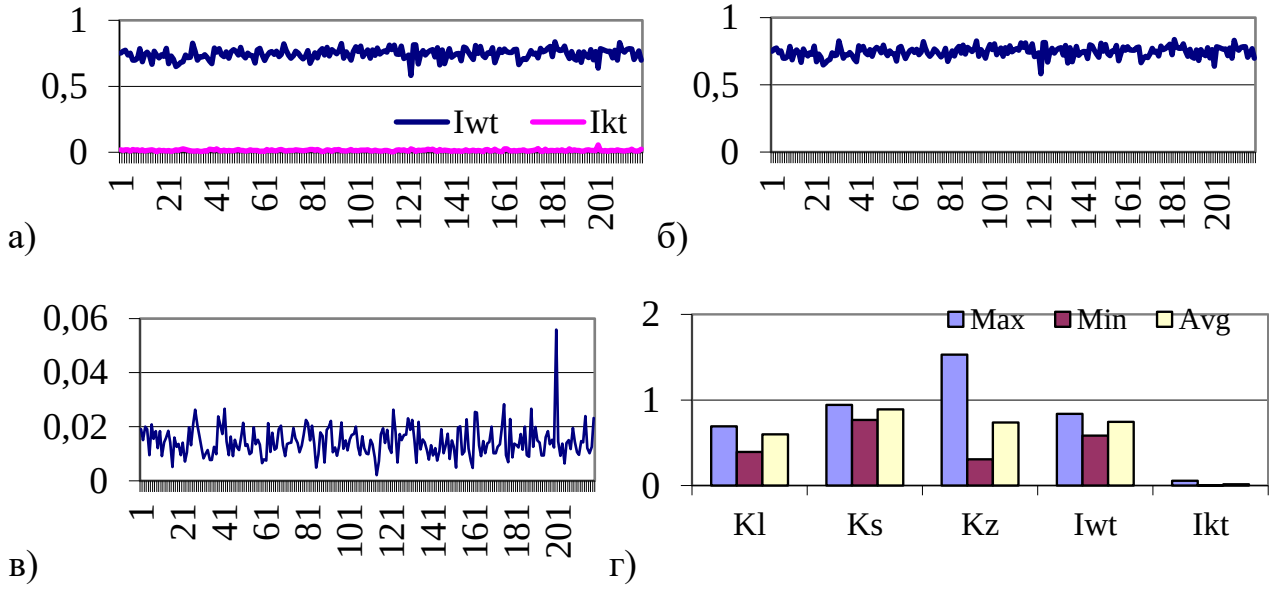


Рис. 6.9. Дистрибуція ступеня мовлення для: *а* – обох параметрів; *б* – I_{wt}; *в* – I_{kt}; *г* – мінімальне, максимальне та середнє значення для всіх параметрів

Відповідно до результатів, поданих на Рис. 6.11 з часом автори вживають коротші речення для опису власних досліджень в україномовних науково-технічних текстах. Також зменшується обсяг появи прийменників в україномовних науково-технічних текстах, але дистрибуція появи сполучників майже не зменшується з часом (Рис. 6.12).

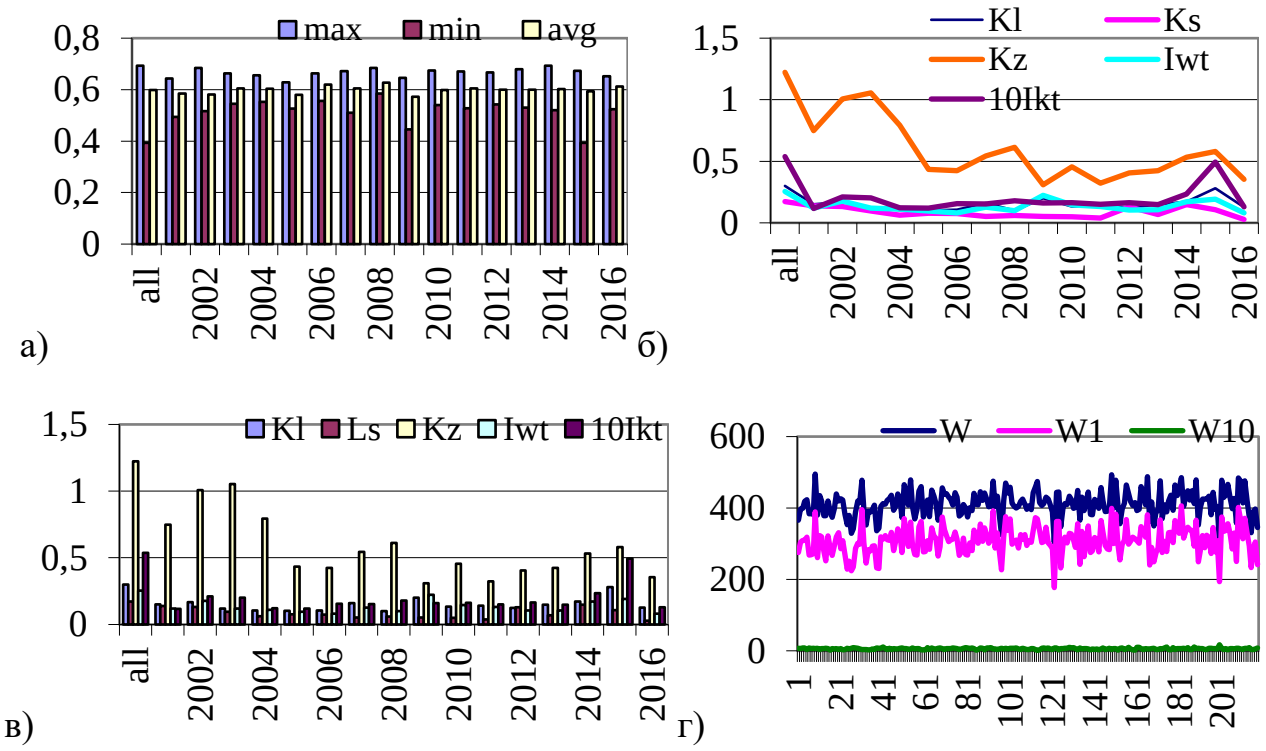


Рис. 6.10. Дистрибуція параметрів мовлення для однакових за обсягом текстів в діапазоні 2001–2017 рр.: *а* – максимальне, мінімальне та середнє значення K_l; *б,в* – зміна значень параметрів; *г* – поява словоформ (всіх, лише 1 раз та ≥10)

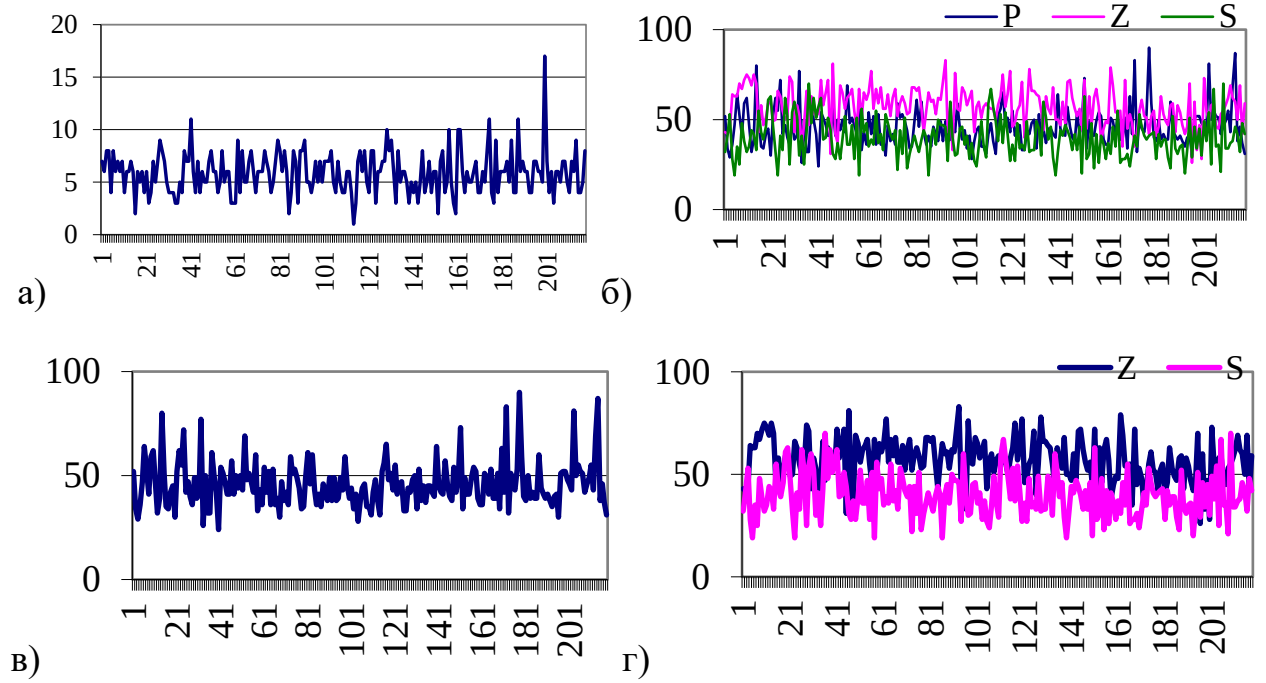


Рис. 6.11. Дистрибуція появи слів: *а* – ≥10 (W_{10}); *б* – ступеня зв'язності мовлення; *в* – речень; *г* – прийменників, та сполучників

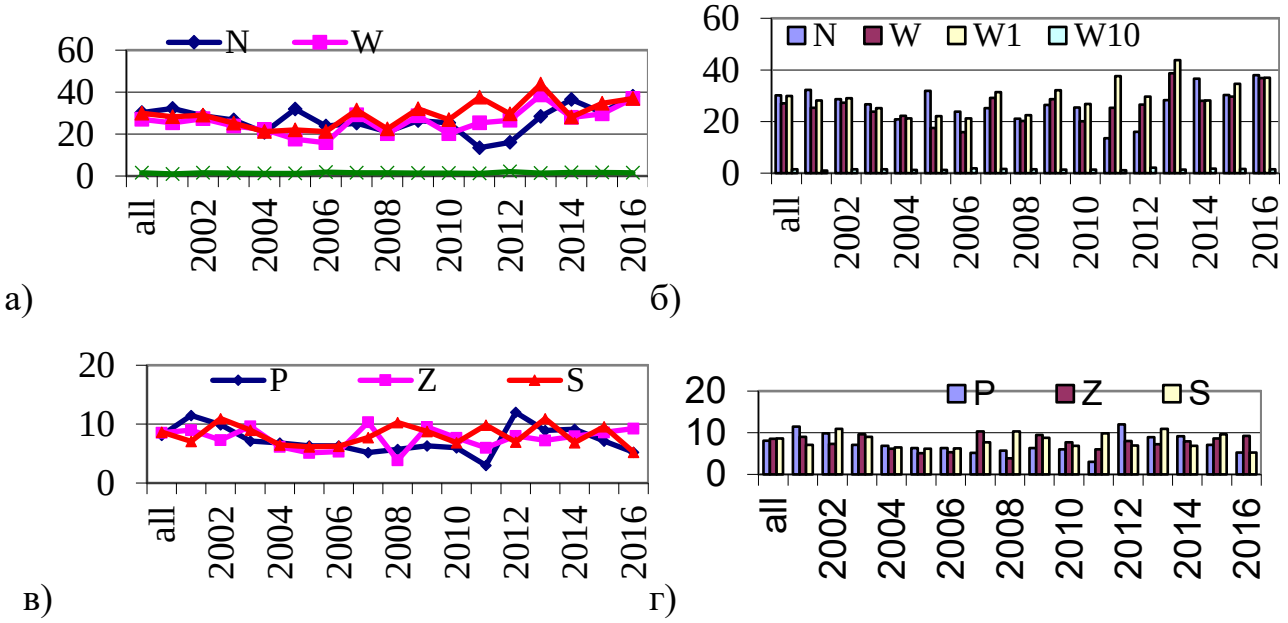


Рис. 6.12. Зміна дистрибуції ознак стилю мовлення автора в часі

Необхідно знайти діапазон приросту кожного із досліджуваного параметру (Рис. 6.13), так як існує динаміка зміни не лише ознак стилю мовлення автора за визначений період наукової діяльності, але і окремих параметрів (обсяг появи речень, сполучників та прийменників, словоформ на загальний обсяг слів, словоформ, які вживані точно 1 раз та ≥10).

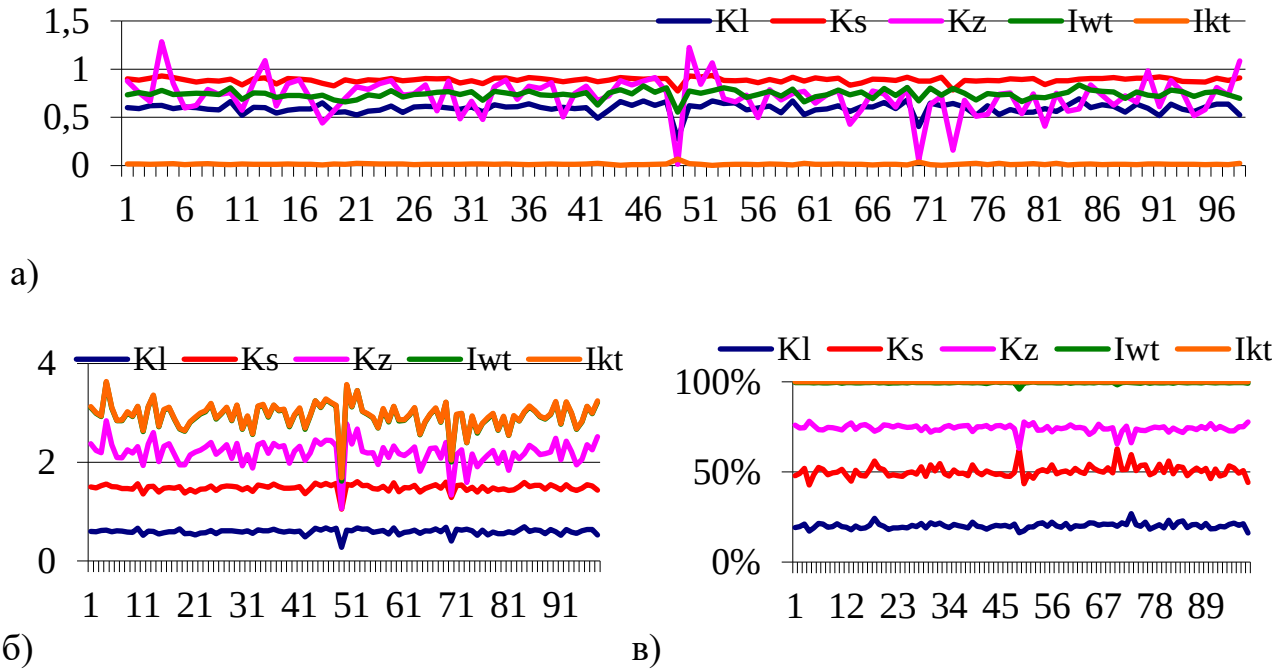


Рис. 6.13. Дослідження зміни у часі за ознаками мовлення: *а* – ідентифікації стилю автора; *б* –загальної суми; *в* – вкладення значення на основі нормування

Ознака авторського мовлення окрім *Kz* значно не змінюються. Тоді дослідимо публікації за додатковими параметрами (Рис. 6.14). Введення додаткових параметрів зменшить множину потенційних авторів, з подібними стилями мовлення (Рис. 6.15 та Таблиця Д.5 додатку Д).

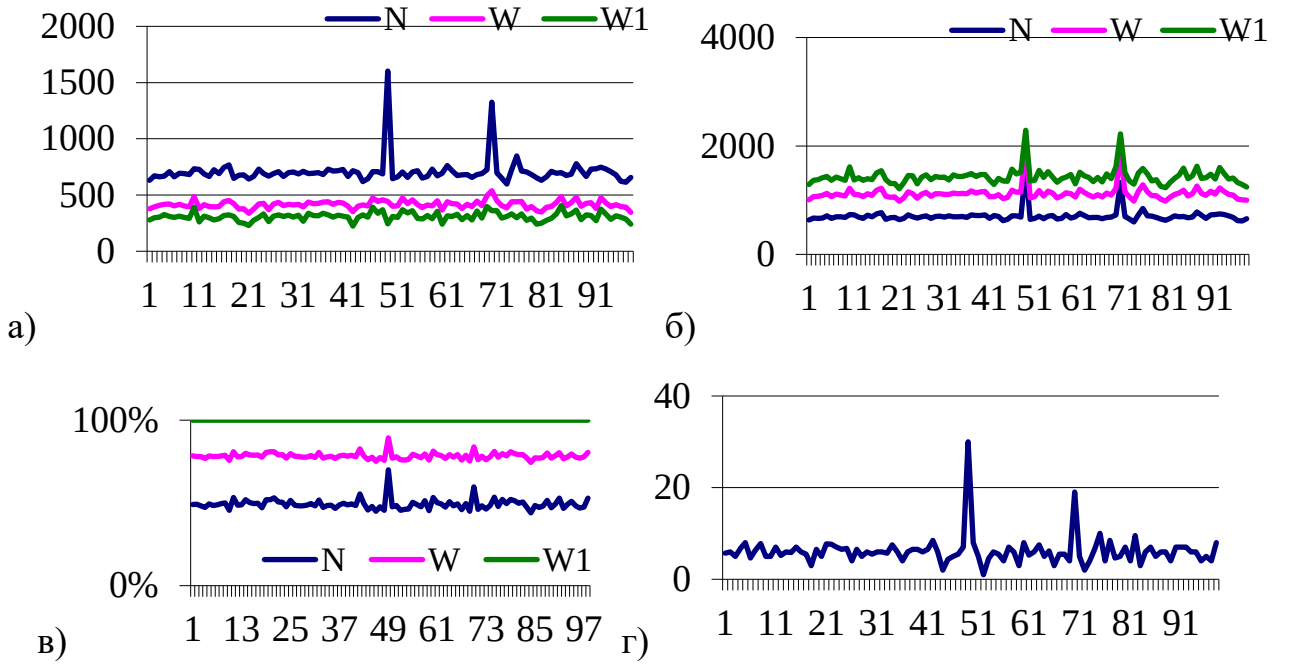


Рис. 6.14. Дослідження у часі зміни за ознаками мовлення: *а* – ідентифікації стилю автора; *б* – загальної суми; *в* –вкладення значення; *г* – зміни W10

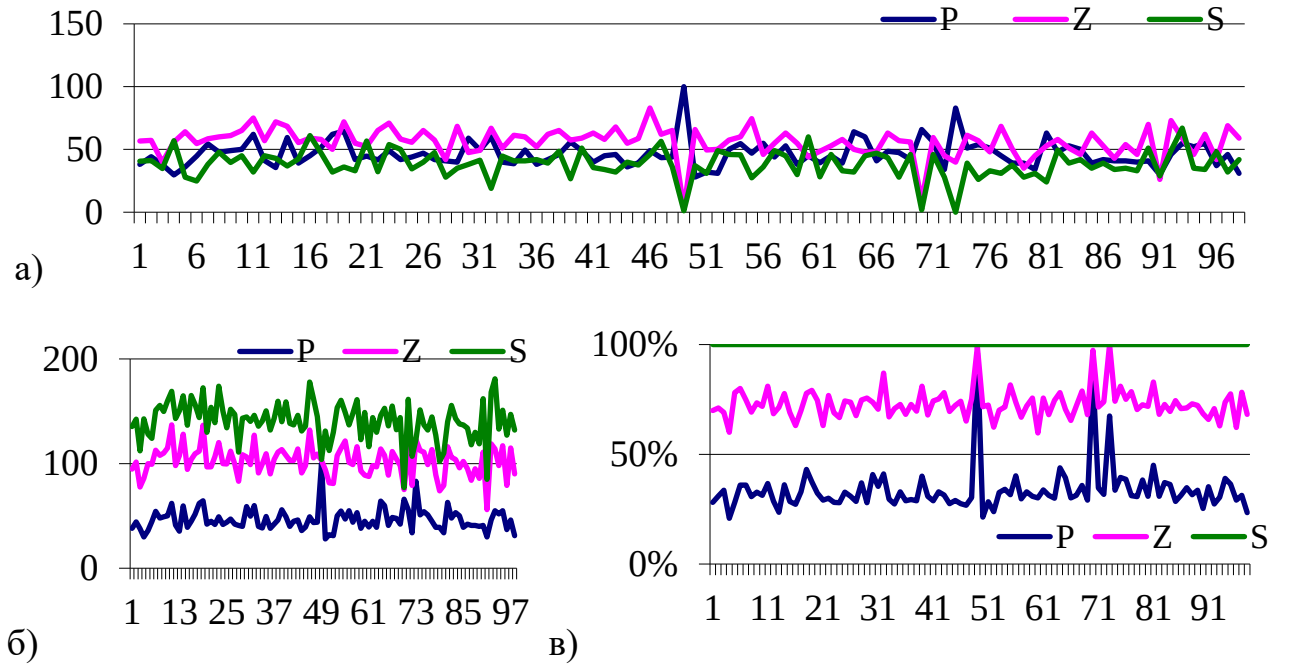


Рис. 6.15. Дослідження у часі зміни за ознаками мовлення: *а* – ідентифікації стилю автора; *б* – загальної суми; *в* – вкладення кожного значення

6.4. Метод кількісної оцінки визначення авторства текстового контенту на основі статистичного аналізу розподілу N-грам

Кожна мова має власні статистичні параметри. Наприклад, для українських текстів виявлено, що статистичними параметрами стилів можна вважати частоти голосних, приголосних, пропуски між словами, а також м'яких і сонорних груп приголосних (Рис. 4.31, Таблиця Д.6 додатку Д). Для досягнення мети дослідження розроблено систему з можливістю обрання мови/мов аналізованого контенту, яка реалізована на Web-ресурсі Victana. Для якісного та ефективного аналізу контенту при визначенні ступеня авторства конкретної людини пропонуємо аналізувати еталонного тексту та досліджуваного в декілька етапів.

– Алгоритм 1. Лінгвометричний аналіз коефіцієнтів різноманіття авторського мовлення (алг. 6.5);

– Алгоритм 2. Стилометричний аналіз (алг. 6.6);

– Алгоритм 3. Аналіз стійких словосполучень (алг. 6.7);

– Алгоритм 4. Лінгвостатистичний аналіз через N-грам (алг. 6.8).

На Web-ресурсі для лінгвометричного аналізу є такі поля (Рис. 6.16):

– Знаків. (Введений текст повинен містити не менше 100 та не більше 10000 знаків.) – виставляється максимальний розмір контенту.

– Контент – поле, куди копіюється із буфера досліджуваний текст.

– Розрахувати – запуск розрахунку.

– Очистити – очищення введених даних.

Алгоритм 6.5. Лінгвометричний аналіз тексту для визначення авторства.

Етап 1. Фільтрування україномовного текстового контенту від інформаційного шуму (спеціальні символи, рисунки, теги, цифри, формули тощо).

Етап 2. Визначення розміру текстового контенту – зайве відсікається.

Етап 3. Ідентифікація обсягу речень в україномовному текстовому контенті.

Етап 4. Ідентифікація загального обсягу слів у тексті N.

Етап 5. Ідентифікація обсягу унікальних слів W в текстовому контенті.

Етап 6. Ідентифікація обсягу прийменників Z в текстовому контенті.

Етап 7. Ідентифікація обсягу сполучників S в текстовому контенті.

Етап 8. Розрахунок коефіцієнтів авторського мовлення.

Етап 9. Вивід результатів кінцевому користувачу (Таблиця 6.5, Рис. 6.16).

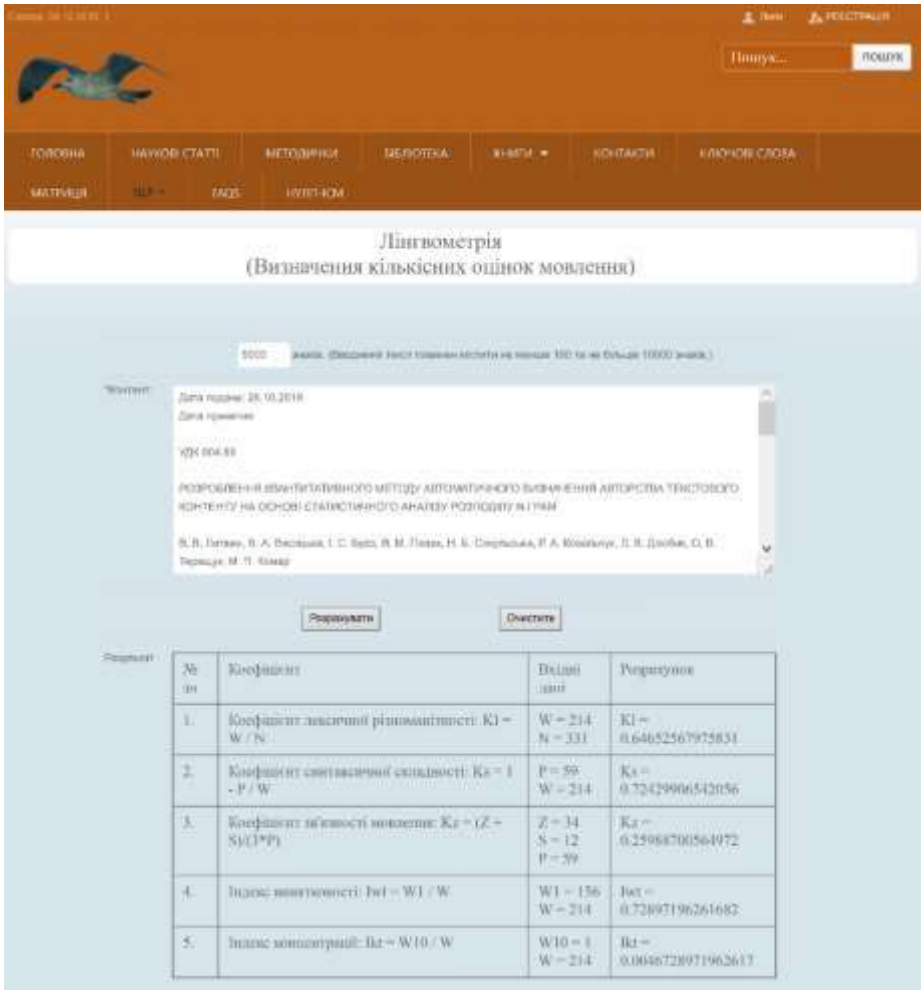


Рис. 6.16. Приклад результату застосування лінгвометричного аналізу

Таблиця 6.5

Приклад розрахунків коефіцієнтів авторського мовлення

Коефіцієнт	Вхідні дані	Розрахунок
Коефіцієнт лексичної різноманітності: $K_l=W/N$	W=184, N=295	K_l=0.62372881355932
Коефіцієнт синтаксичної складності: $K_s=1–P/W$	P=18, W=184	K_s=0.90217391304348
Коефіцієнт зв'язності мовлення: $K_z=(Z+S)/(3*P)$	Z=20, S=28, P=18	K_z=0.88888888888889
Індекс винятковості: $I_{wt}=W_1/W$	W_1=141, W=184	I_{wt}=0.76630434782609
Індекс концентрації: $I_{kt}=W_{10}/W$	W_{10}=2, W=184	I_{kt}=0.010869565217391

На Web-ресурсі для стилеметричого аналізу є такі поля (Рис. 6.17):

– Еталонний текст – поле, куди копіюється із буфера Еталонний текст.

– Вибрати Уривок 1 (2, 3) – відкриваємо доступ до уривків. Доступ до наступного уривку тільки після активації доступу до попереднього. Доступ відкривається послідовно від меншого числа до більшого.

– Уривок 1 (2, 3) – поле, куди копіюється із буфера текст уривку. Введений текст повинен містити не менше 100 знаків. (Зараз 0) – Після запуску розрахунку буде розраховано та показано реальну кількість знаків кожного уривку окремо.

– Розрахувати – запуск розрахунку.

– Очистити – очищення введених даних.

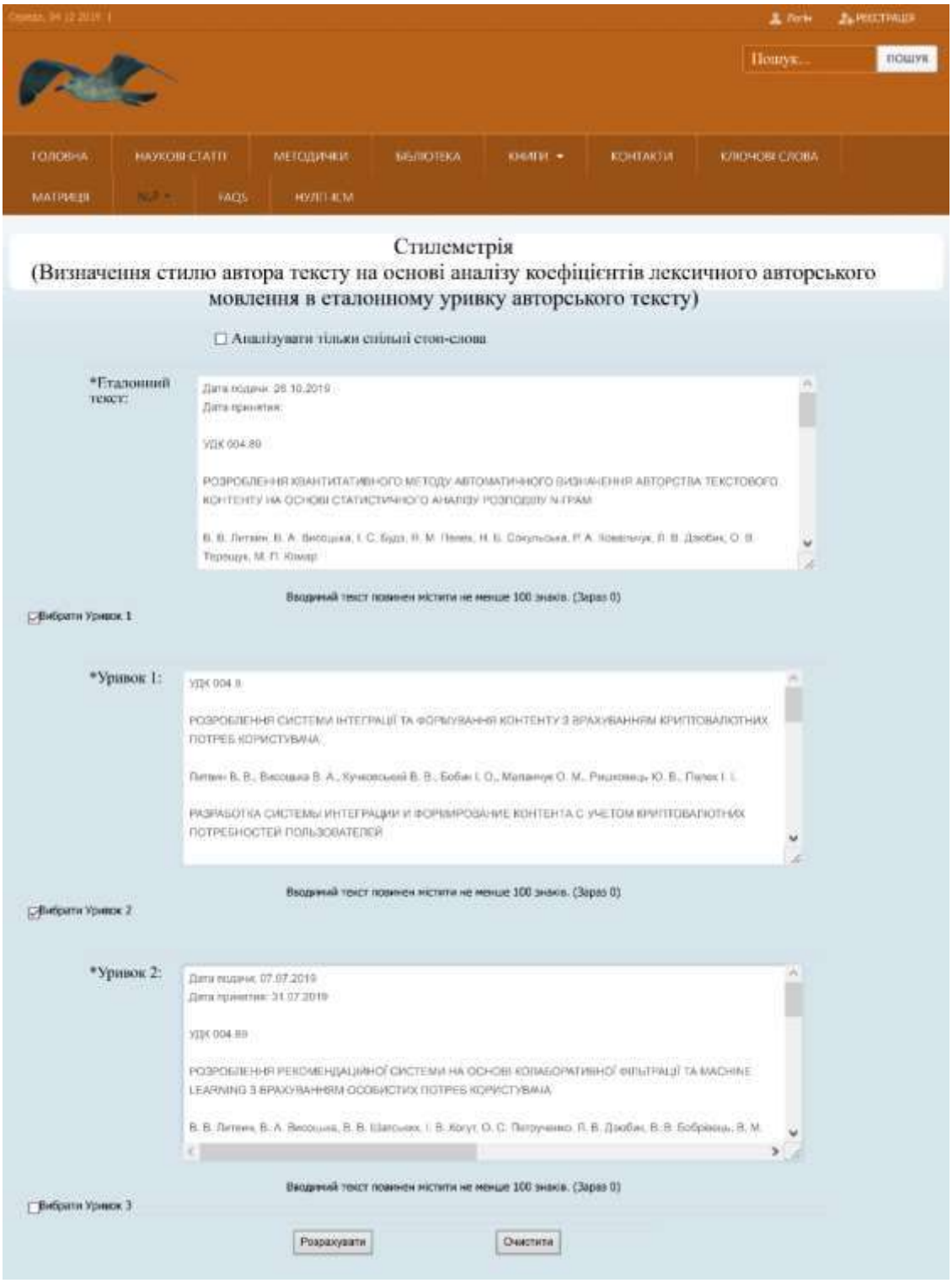


Рис. 6.17. Приклад введення даних для стилеметричого аналізу

Алгоритм 6.6. Стилеметричний аналіз тексту для визначення авторства.

Етап 1. Перевірка довжин еталонного тексту та вибраних уривків та приведення довжини еталонного тексту до мінімального із перевірених.

Етап 2. Очищення еталонного тексту від спецсимволів та інш.

Етап 3. Визначення кількості слів у тексті еталону.

Етап 4. Визначення кількості стоп-слів (прийменників + сполучників + часток) у тексті еталону (Рис. 6.18, Рис. 6.19).

Розрахувати Очистити

Уривок 1 слів: 3046. Еталонний текст слів: 2465.

Стоп-слово	АЧ	ВЧ	Частина мови	АЧ стал.	ВЧ в сталоні
та	158	0.051871306631648	Сполучник	167	0.067748478701826
з	149	0.048916611950098	Прийменник	113	0.045841784989858
в	129	0.042350623768877	Прийменник	198	0.080324543610548
а	44	0.014445173998687	Сполучник	53	0.021501014198783
і	99	0.032501641497045	Сполучник	72	0.02920892494929
for	33	0.010833880499015	Прийменник	8	0.0032454361054767
and	136	0.044648719632305	Сполучник	13	0.0052738336713996
для	166	0.054497701904137	Прийменник	183	0.074239350912779
по	33	0.010833880499015	Прийменник	9	0.0036511156186613
це	10	0.0032829940906106	Частка	29	0.011764705882353
від	14	0.0045961917268549	Прийменник	42	0.017038539553753
до	31	0.010177281680893	Прийменник	70	0.028397565922921
через	22	0.0072225869993434	Прийменник	2	0.00081135902636917
без	6	0.0019697964543664	Прийменник	2	0.00081135902636917
або	2	0.00065659881812213	Частка	38	0.015415821501014
за	48	0.015758371634931	Прийменник	37	0.01501014198783
чи	9	0.0029546946815496	Частка	16	0.0064908722109533
на	128	0.042022324359816	Прийменник	120	0.04868154158215
якщо	1	0.00032829940906106	Сполучник	10	0.0040567951318458
не	33	0.010833880499015	Частка	37	0.01501014198783
то	1	0.00032829940906106	Частка	6	0.0024340770791075
так	13	0.0042678923177938	Частка	9	0.0036511156186613
що	16	0.005252790544977	Сполучник	64	0.025963488843813
при	7	0.0022980958634274	Прийменник	23	0.0093306288032454
щоб	16	0.005252790544977	Сполучник	5	0.0020283975659229
коли	4	0.0013131976362443	Сполучник	25	0.010141987829615
лише	1	0.00032829940906106	Частка	11	0.0044624746450304

Рис. 6.18. Приклад результату застосування стилеметричого аналізу

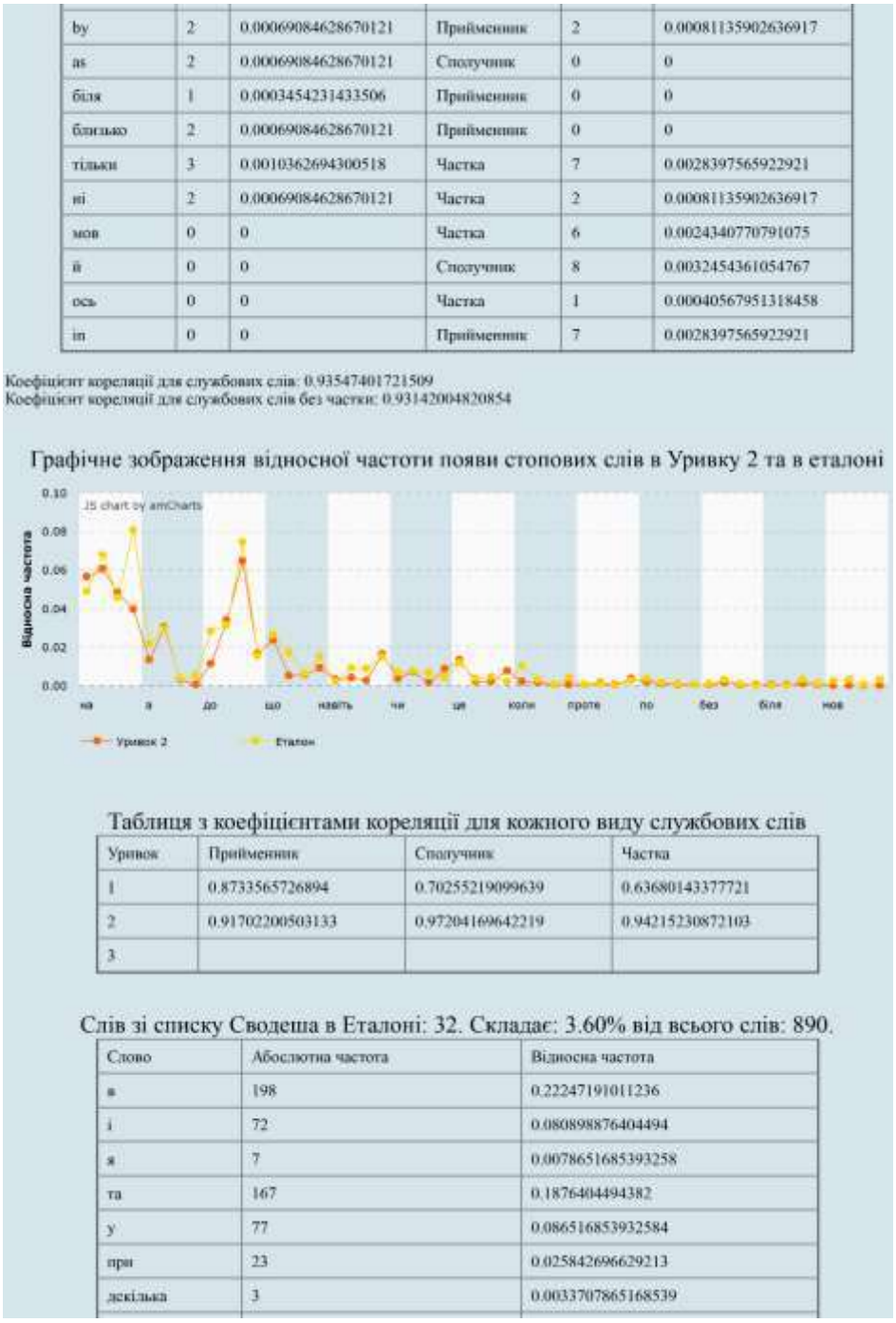
Етап 5. Довжина Уривка 1 не більше мінімально тексту.

Етап 6. Очищення Уривка 1 від спецсимволів та інш.

Етап 7. Визначення кількості слів W1 для Уривка 1.

Етап 8. Визначення кількості стоп-слів (прийменників + сполучників + часток) в тексті.

Етап 9. Підготовка окремих масивів (уривок та еталон) для розрахунку коефіцієнта кореляції (Рис. 6.19).



by	2	0.00069084628670121	Прийменник	2	0.00081135902636917
as	2	0.00069084628670121	Сполучник	0	0
біля	1	0.0003454231433506	Прийменник	0	0
близько	2	0.00069084628670121	Прийменник	0	0
тільки	3	0.0010362694300518	Частка	7	0.0028397565922921
ні	2	0.00069084628670121	Частка	2	0.00081135902636917
мов	0	0	Частка	6	0.0024340770791075
й	0	0	Сполучник	8	0.0032454361054767
ось	0	0	Частка	1	0.00040567951318458
in	0	0	Прийменник	7	0.0028397565922921

Коефіцієнт кореляції для службових слів: 0.93547401721509
Коефіцієнт кореляції для службових слів без часток: 0.93142004820854

Графічне зображення відносної частоти появи стопових слів в Уривку 2 та в еталоні

Таблиця з коефіцієнтами кореляції для кожного виду службових слів

Уривок	Прийменник	Сполучник	Частка
1	0.8733565726894	0.70255219099639	0.63680143377721
2	0.91702200503133	0.97204169642219	0.94215230872103
3			

Слів зі списку Сводеша в Еталоні: 32. Складає: 3.60% від всього слів: 890.

Слово	Абсолютна частота	Відносна частота
в	198	0.22247191011236
і	72	0.080898876404494
я	7	0.0078651685393258
та	167	0.1876404494382
у	77	0.086516853932584
при	23	0.025842696629213
декілька	3	0.0033707865168539

Рис. 6.19. Результат застосування стилеметричого аналізу для Уривку 2

Етап 10. Виклик функції для розрахунку коефіцієнта кореляції.

Етап 11. Формування масиву для формування графічного зображення відносної частоти появи стопових слів в Уривку 1 та в еталоні.

Етап 12. Виклик функції для розрахунку графіка ВЧ (Рис. 6.20).

Етап 13. Виклик функції для розрахунку коефіцієнта кореляції Уривків 2(3) для кожного зі службових слів.

Етап 14. Формуємо слова списку Сводеша із довідника, визначення кількості слів із списку Сводеша в тексті уривку (для еталонного тексту та вибраних уривків – Таблиця 6.6).

Етап 15. Формуємо спільні для Еталону, Уривків 1–3 та списку Сводеша.

Етап 16. Результати дослідження виводяться на екран (Таблиця 6.7)

Таблиця 6.6

Уривок 1 слів: 153. Еталонний текст слів: 153

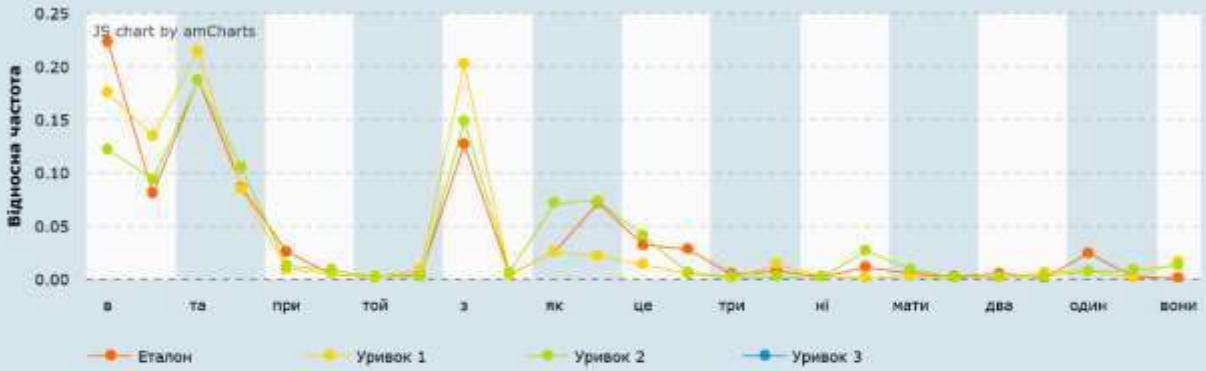
Слово	АЧ	ВЧ	Частина мови	АЧ етал	ВЧ в еталоні
в	5	0.032679738562	Прийменник	5	0.032679738562
а	2	0.0130718954248	Сполучник	2	0.0130718954248
це	1	0.0065359477124	Частка	1	0.0065359477124
та	16	0.1045751633987	Сполучник	16	0.1045751633987
для	7	0.0457516339869	Прийменник	7	0.0457516339869
з	2	0.0130718954248	Прийменник	2	0.0130718954248
ж	1	0.0065359477124	Частка	1	0.0065359477124
і	3	0.019607843137	Сполучник	3	0.019607843137
також	2	0.0130718954248	Сполучник	2	0.0130718954248
мов	2	0.0130718954248	Частка	2	0.0130718954248
у	1	0.0065359477124	Прийменник	1	0.0065359477124
що	1	0.0065359477124	Сполучник	1	0.0065359477124
за	1	0.0065359477124	Прийменник	1	0.0065359477124

Таблиця 6.7

Спільні для Еталону, Уривків 1–3 та списку Сводеша: 8 (26.67 %) від всього: 30

№	Спільні	АЧ	Еталон	Уривок 1	Уривок 2	Уривок 3
1	в	5	0.167	0.167	0.167	0.167
2	це	1	0.033	0.033	0.033	0.033
3	та	16	0.533	0.533	0.533	0.533
4	з	2	0.167	0.167	0.167	0.167
5	коло	1	0.033	0.033	0.033	0.033
6	і	3	0.1	0.1	0.1	0.1
7	у	1	0.033	0.033	0.033	0.033
8	що	1	0.033	0.033	0.033	0.033

Графічне зображення ВЧ появи слів, спільних для Уривків 1–3 з еталоном та списку Сводеша



Слова, спільні для Еталону, Уривків 1–3 та списку Сводеша

№ зп	Спільні слова	ВЧ в Еталоні	ВЧ в Уривку 1	ВЧ в Уривку 2	ВЧ в Уривку 3
1	в	0.22247191011236	0.17503392130258	0.12166488794023	
2	і	0.080898876404494	0.13432835820896	0.093916755602988	
3	та	0.1876404494382	0.21438263229308	0.18676627534685	
4	у	0.086516853932584	0.084124830393487	0.10458911419424	
5	при	0.025842696629213	0.0094979647218453	0.012806830309498	
6	те	0.0044943820224719	0.005427408412483	0.0085378868729989	
7	той	0.002247191011236	0.0013568521031208	0.0021344717182497	
8	зі	0.0044943820224719	0.0094979647218453	0.0032017075773746	
9	з	0.12696629213483	0.20217096336499	0.14834578441836	
10	цей	0.0044943820224719	0.0040705563093623	0.0064034151547492	
11	як	0.025842696629213	0.025780189959294	0.071504802561366	
12	що	0.071910112359551	0.021709633649932	0.073639274279616	
13	це	0.032584269662921	0.013568521031208	0.040554962646745	
14	коли	0.028089887640449	0.005427408412483	0.0064034151547492	
15	три	0.0044943820224719	0.0013568521031208	0.0021344717182497	
16	де	0.0078651685393258	0.013568521031208	0.0021344717182497	
17	ні	0.002247191011236	0.0027137042062415	0.0021344717182497	
18	якщо	0.01123595505618	0.0013568521031208	0.026680896478122	
19	мати	0.0056179775280899	0.0027137042062415	0.0096051227321238	
20	великий	0.002247191011236	0.0013568521031208	0.0010672358591249	
21	два	0.0056179775280899	0.0013568521031208	0.0021344717182497	
22	багато	0.01123595505618	0.0067842605156038	0.0032017075773746	

Рис. 6.20. Приклад результату стилеметричого аналізу для Уривків 1–3

Для автоматизованого опрацювання тексту має велике значення не тільки те, яка частота появи тієї чи іншої категорії, а взагалі її присутність в

досліджуваному тексті. Підводячи підсумки, слід зазначити, що використання контент-аналізу для створення інформаційних систем дозволяє вловити дистрибуцію різних ознак аналізованого текстового контенту.

Наприклад, частотні характеристики тексту (середній розмір речень) може свідчити про певну специфіку інтелектуальних здібностей особи у плані вербального подання думок. За визначенням середнього розміру речень можна дати характеристику зміни емоційного стану індивіда. Одним із найбільш значних ознак у психолінгвістичному аналізі текстового контенту є вибір аналізу словникового варіанта в контекстній залежності. Завдяки встановленню коефіцієнта словникової різноманітності мовлення (Таблиця 6.8), можна ідентифікувати, наприклад, ступень можливої наявності у автора шизофренії.

Таблиця 6.8

Коефіцієнти частотних характеристик тексту

Коефіцієнт	Формула
Словникової різноманітності	$К_{слов.різном.}$=різних слів/$2N_{всіх\ слів}$
Дієслівності (агресивності)	$К_{дієсл.}$=дієслів/$N_{всіх\ слів}$·100 %
Емоційності тексту	$К_{прикм.}$=прикм/$2N_{всіх\ слів}$
Логічної зв’язності	$К_{лог.зв'язн.}$=служб. слів/$3N_{реч}$
Емболії (засміченості)	$К_{емб.}$=ембол/$N_{всіх\ слів}$·100 %

Іншим критерієм мовної компетенції є коефіцієнт дієслівності (агресивності). Суть цього коефіцієнту полягає у співвідношенні обсягу дієслів і дієслівних форм (дієприслівників і дієприкметників) до загального обсягу слів. Високий показник агресивності констатує про наявність високого ступеня негативної емоційності автора, яка відображена в самому тексті проявами зміни динаміки подій та іншими характерними особливостями. Параметр логічної зв’язності на основі аналізу службових слів свідчить про достатньо гармонійний ступень логічної побудови тексту. Коефіцієнт засміченості мовлення – це співвідношення загального обсягу без семантичного навантаження слів до загального обсягу слів. До складу без семантичного навантаження слів належать вигуки (а-а-а, е-е-е, м-м-м, ха-ха, ну-ну, еге, ж, ой тощо), вульгаризми (ненормативна лексика), непотрібні повторення. Коефіцієнт засміченості мовлення констатує або про степінь негативного емоційного стану людини (нервовість, наляканість, некомфортність в оточенні тощо) або низький рівень

культури мовлення та інтелекту. Навіть враховуючи той факт, що художній текст в принципі вважається андрогенним та є переплетінням функцій підрядності – якостей авторського «Я», певним чином градуюються в залежності від характерологічного профілю того чи іншого автора. Іншими словами, текст оригіналу і текст перекладу знаходяться у залежності від їх авторів.

На Web-ресурсі для аналізу стійких словосполучень є такі поля (Рис. 6.21):

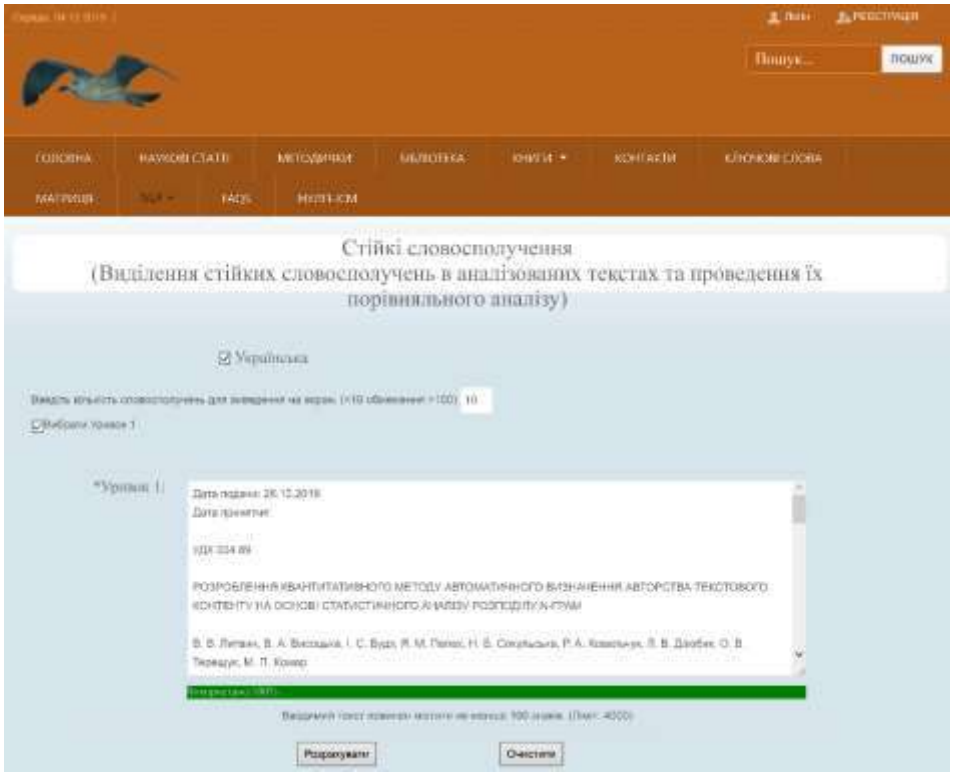


Рис. 6.21. Приклад застосування аналізу стійких словосполучень

– Введіть кількість словосполучень для виведення на екран (10;100) – скільки словосполучень буде виведено на екран після розрахунку.

– Вибрати Уривок 1 (2, 3) – відкриваємо доступ до уривків. Доступ до наступного уривку тільки після активації доступу до попереднього. Доступ відкривається послідовно від меншого числа до більшого. (Не реалізовано – аналізується тільки один уривок)

– Уривок 1 – поле, куди копіюється із буфера текст відповідного уривку.

– Використано:57 % Введений текст повинен містити не менше 100 знаків. (Ліміт: 4000) – аналіз розміру тексту.

– Розрахувати – запуск розрахунку.

– Очистити – очищення введених даних.

Алгоритм 6.7. Лінгвостатистичний аналіз стійких словосполучень.

Етап 1. Очищення отриманого контенту від спецсимволів та інш.

Етап 2. Формуємо список заблокованих слів із бази даних в залежності від вибраної мови контексту.

Етап 3. Підготовка до формування масивів подвійних словосполучень та всіх слів. На вході масив: ключ – цифри, значення – текст, розбитий по реченням (розділювач крапка). Слова звіряються з базою даних ключових слів та по правилу, описаному в базі даних, приводить дане слово до основи слова, якщо само не є основою слова.

Етап 4. Визначення стійких словосполучень за методом FREG: отримати абсолютну частоту словосполучень (Рис. 6.22).

№	FREG			t-тест		LR		X2	
1	2			3		4		5	
	Словосполучення	АЧ	ВЧ	Словосполучення	t	Словосполучення	logL	Словосполучення	X2
1	определения автор	2	0.015267	текстовий контента	1.381827	определения автор	1.68e-1	текстовий контента	86.656531
2	анализован текст	2	0.015267	анализируемого текста	1.371031	визначення автор	6.71e-2	анализируемого текста	64.484496
3	определения принадлежности	2	0.015267	ключовий слово	1.371031	конкретним автор	4.77e-3	ключовий слово	64.484496
4	анализируемого текста	2	0.015267	визначення приналежність	1.371031	конкретний автор	4.77e-3	визначення приналежність	64.484496
5	конкретним автор	2	0.015267	анализован текст	1.349440	анализован текст	1.46e-3	анализован текст	42.312661
6	ключовий слово	2	0.015267	определения принадлежности	1.349440	определения принадлежности	1.46e-3	определения принадлежности	42.312661
7	текстовий контента	2	0.015267	конкретним автор	1.306258	анализируемого текста	5.13e-4	конкретним автор	24.575194
8	визначення автор	2	0.015267	конкретний автор	1.306258	ключовий слово	5.13e-4	конкретний автор	24.575194
9	конкретний автор	2	0.015267	визначення автор	1.198303	визначення приналежність	5.13e-4	визначення автор	10.503358
10	визначення приналежність	2	0.015267	определения автор	1.090348	текстовий контента	2.16e-4	определения автор	5.890164

Рис. 6.22. Приклад результату застосування аналізу стійких словосполучень

Етап 5. Визначення стійких словосполучень за методом t-тест: P(W1)*P(W2) врахування не тільки пар, але і частоти вживання окремих слів (тих, що складають пару).

Етап 6. Визначення стійких словосполучень за методом LR.

Етап 7. Визначення стійких словосполучень за методом X2 (Таблиця 6.9).

Етап 8. Результати дослідження виводяться на екран.

Таблиця 6.9

Список за рейтингом частоти появи стійких словосполучень для статті 1, словосполучень: 45. Всього слів: 108

№	FREG			t-тест		LR		X2	
	Словосполучення	АЧ	ВЧ	Словосполучення	t	Словосполучення	logL	Словосполучення	X2
1	система електронний	4	0.088889	система електронний	1.822222	інформаційний технологія	5.03 e–1	прийняття рішення	45.000000
2	інформаційний система	4	0.088889	електронний контент-комерція	1.578091	інтелектуальний система	2.13 e–1	система електронний	45.000000
3	електронний контент-комерція	3	0.066667	розділ науковий	1.319933	інформаційний система	8.36 e–2	електронний контент-комерція	32.946429
4	розділ науковий	2	0.044444	інформаційний система	1.222222	портал науковий	5.58 e–2	розділ науковий	29.302326
5	портал науковий	1	0.022222	прийняття рішення	0.977778	курс технологія	3.31 e–2	курс технологія	21.988636
6	інтелектуальний система	1	0.022222	курс технологія	0.955556	сховище дані	3.31 e–2	сховище дані	21.988636
7	прийняття рішення	1	0.022222	сховище дані	0.955556	прийняття рішення	8.27 e–3	портал науковий	14.318182
8	курс технологія	1	0.022222	портал науковий	0.933333	розділ науковий	1.89 e–3	інформаційний система	5.848550
9	сховище дані	1	0.022222	інтелектуальний система	0.777778	електронний контент-комерція	1.55 e–4	інтелектуальний система	3.579545
10	інформаційний технологія	1	0.022222	інформаційний технологія	0.688889	система електронний	1.37 e–6	інформаційний технологія	1.890409

Якщо в базі даних відсутнє слово добавляється автоматично. Модератору необхідно для цього слова описати правило приведення слова до основи слова.

При ідентифікації автора тексту передбачається, що текст відображає індивідуальну манеру письма автора, яка дозволяє відрізнити його від інших. Щоб порівнювати тексти між собою необхідно зіставити тексту деяку числову характеристику, яка була б наближена для текстів одного і того ж автора, і суттєво різнилася б для творів різних авторів. Такою характеристикою може бути щільність розподілу літеросполучень з трьох поспіль символів (3-грам).

Визначається, як сукупність емпіричних частот вживання літер або їх поєднань. При аналізі тексту на основі щільності розподілу N-грам не враховують входження розділових знаків, пробілів і цифр. Завдання ідентифікації автора невідомого тексту в термінах щільності розподілу N-грам визначається так. Дано деякий набір текстів, в якому містяться твори Y відомих авторів. Нехай L_y – кількість контенту y-го автор. $N_{i,y}$ – кількість символів в i-му контенті y-го учасника, i=1, ..., L_y. Щільність розподілу N-грам контенту, обсяг якого дорівнює $N_{i,y}$, задається як множина значень $f_{i,y}(j)=k_j/N_{i,y}$, k_j – кількість вживання N-грами під номером j. Аргумент j=1,...,$y(n,M)$, відповідає номеру літеросполучення (N-грами) при алфавітному впорядкуванні, де M – потужність алфавіту мови написаного тексту, n – порядок N-грами, тобто кількість символів в літеросполученні. $y(n, M)=M^n$ – кількість N-грам в даному алфавіті. Кожен автор ототожнюється з його середньозваженою щільністю розподілу N-грам за формулою $p_y(j) = \frac{1}{N_y}\sum_{i=1}^{L_y} p_{i,y}N_{i,y}$. Вони є авторськими еталонами. Для порівняння двох текстів, або тексту і авторського еталону, необхідно задати відстань між відповідними функціями розподілу. Як метрики відстані застосовують норму в просторі функцій як доданків. Так, наприклад, відстань $p_{x,y}$ між щільністю розподілу N-грам невідомого тексту p_x і будь-якої авторської щільності розподілу N-грам p_y розраховують як:

$$p_{x,y} = \left|\left|p_x - p_y\right|\right| = \sum_{j=1}^{y(n,M)} \left|p_x(j) - p_y(j)\right|.$$

Текст «x» належить тому автору, відстань до щільності розподілу N-грам якого буде найменшим. При вирішенні задачі класифікації набір даних не розбивався явно на тестові і тренувальні множини. Середньозважені щільності розподілу N-грам будувалися по всій множині контенту одного автора. Відстань від контенту i до конкретного автора y обчислювалося як:

$$P_{i,y} = \frac{\left|\left|p_{i,y} - p_y\right|\right|}{1 - \frac{N_{i,y}}{N_y}}.$$

Формула дозволяє виключити участь щільності розподілу N-грам контенту

і в середній щільності розподілу N-грам конкретного автора. На Web-ресурсі для аналізу N-грам є такі поля (Рис. 6.23):

– Вибрати мову тексту – мова тексту для аналізу (дослідження). За замовчуванням «Українська».

– Число грами – кількість знаків у грамі. Можна міняти на 1, 2, 3, 4. За замовчуванням 3.

– Обмеження тексту в знаках.

– Текст – поле, куди копіюється із буфера досліджуваний текст.

– Генерувати – для запуску генерації N-грам.

– Очистити – очищення введених даних.

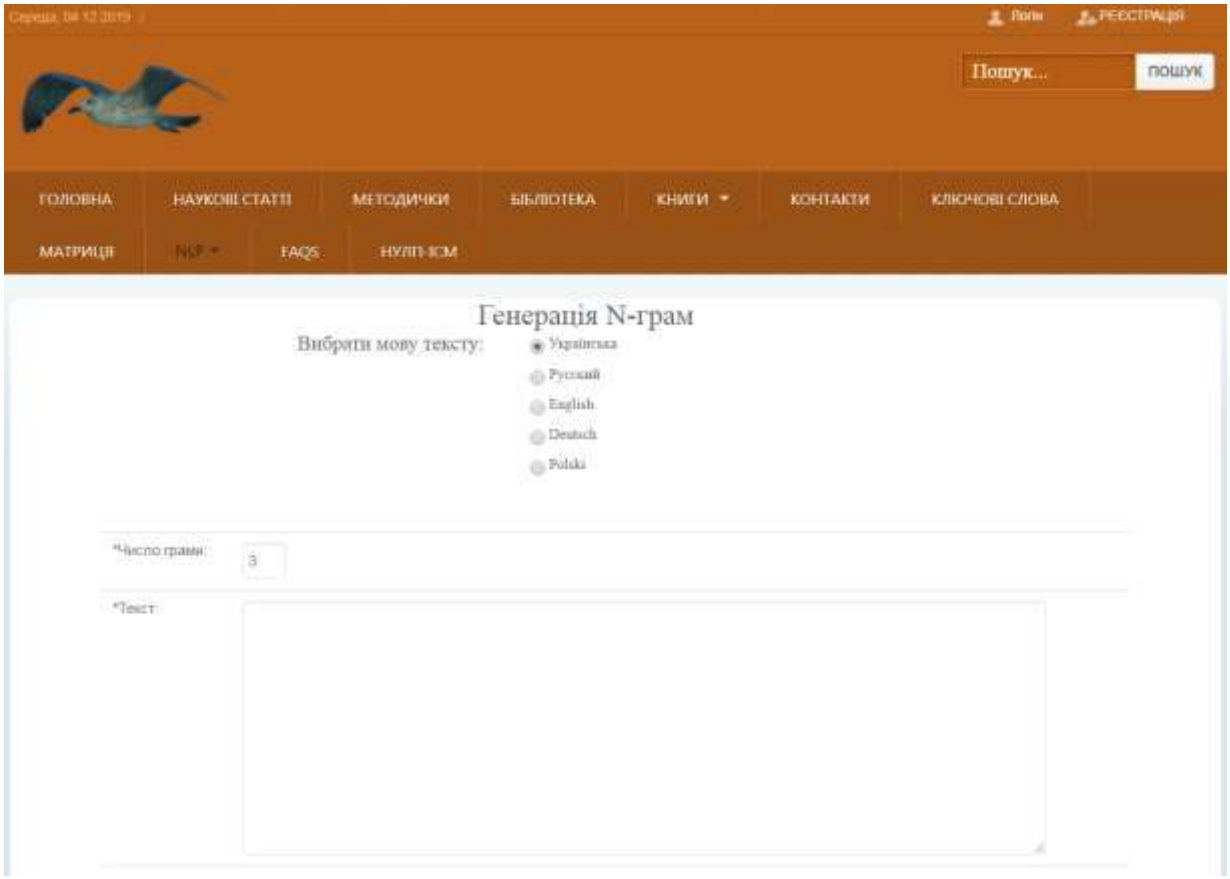


Рис. 6.23. Приклад застосування аналізу N-грам тексту

Алгоритм 6.8. Лінгвостатистичний аналіз N-грам тексту.

Етап 1. Очищення досліджуваного тексту (цифри, спецсимволи).

Етап 2. Вираховуємо кількість слів у тексті.

Етап 3. Всі слова тексту переводимо в нижній регістр.

Етап 4. Видаляємо пробіли.

Етап 5. В залежності від вибраної мови підставляється відповідний алфавіт.

Етап 6. В залежності від встановленого числа грами запускається відповідна функція, яка розраховує всі можливі варіанти грам і зберігає в масиві.

Етап 7. Далі запускається функція підрахування кількості входження слів. Тут же розраховуємо відносну частоту входження та зберігаємо в масиві: порядковий номер грами, сама грама, кількість входжень даної грами, відносна частота входження даної грами.

Етап 8. Наступна функція формує отриманий в попередній функції масив для експорту в CSV файл. Цей файл зберігається на сервері. Його можна завантажити на комп'ютер користувача (дослідника) по посиланню, доступ до якого буде після формування форми з результатами дослідження.

Етап 9. Результати дослідження виводяться на екран (тільки ті грами, які знайдено в тексті).

Етап 10. Відкривається доступ до файлу експорту.

Етап 11. Виводяться узагальненні результати:

– розмір алфавіту;

– кількість слів у тексті;

– кількість знаків в тексті з пробілами;

– кількість знаків в тексті повністю очищеному;

– всього N-грам;

– всього знайдено N-грам без повторень;

– всього знайдено N-грам з повтореннями.

Порівняємо три публікації [1, 74, 77] науково-технічного спрямування між собою на основі лінгвостатистичного аналізу 3-грам. Статті 1, 2 написані одним колективом [1, 74], Стаття 3 написана іншим автором [77] (Таблиця 6.10). Мова тексту – українська (літер в алфавіті – 33, тоді всього можливих N-грам 35937).

Таблиця 6.10

Значення параметрів для аналізованих статей 1–3

Параметри	Стаття 1	Стаття 2	Стаття 3
Всього N-gram	35937	35937	35937
Всього знайдених N-gram (без повторень)	4354	4377	3890

Всього знайдених N-gram (з повторенням)	29494	29862	36383
Всього слів	5475	5358	6060
Всього знаків в неочищеному тексті	39792	39663	47084
Всього знаків в очищеному тексті	29967	32570	37062

Але при порівняні статей будемо враховувати лише ті 3-грами, які зустрілися в тексті одночасно в трьох статтях хоча б один раз. Тому для цього конкретного прикладу всіх 3-грам є 2147. Тобто, для Статті 1 аналізуємо 78,4814 % 3-грам, для Статті 2 – 72,6332 % та для Статті 3 – 84,1271 %. Відповідно різниця вживання відповідних 3-грам між Статтями 1 та 2 є R_{12}=56,5254 %, між Статтями 2 та 3 – R_{23}=69,4271 %, між Статтями 1 та 3 – R_{13}=62.9839 %. Самі ці показники показують, що характеристики статті 1 та 2 більш подібні ($R_{23}>R_{12}$ на 12.9017 %, $R_{23} > R_{13}$ на 6.4432 %, $R_{13}> R_{12}$ на 6.4585 %, тобто $R_{23}>R_{13}>R_{12}$), ніж характеристики відповідно Статті 1–3 та 2–3. Чим менше R_{ij}, тим більша ступінь, що статті написані одним і тим же автором. Тоді в випадку Стаття 1 та 2 більш ймовірніше написана одним автором/колективом, ніж Статті 2–3 та Статті 1–3 відповідно. Але проаналізуємо для вживання окремих кластерів 3-грам у відповідних статтях та порівняємо отримані результати (Таблиця 6.11).

Таблиця 6.11

Значення параметрів появи 3-грам для аналізованих статей 1–3

3-грама	Середнє значення 1 появи			Діапазон появи all, %	Збіг для статей, %			Розбіжність для статей, %		
	1	2	3		1–2	2–3	1–3	1–2	1–3	2–3
а_ _	0,0393	0,0430	0,0392	6,112–6,709	4,2322	4,6322	4,197	0,0271	0,0297	0,0269
б_ _	0,0220	0,0415	0,0262	0,594–1,121	0,7046	0,7738	0,4884	0,0261	0,0287	0,0181
в_ _	0,0390	0,0367	0,0388	4,262–4,522	3,5581	4,1064	3,6523	0,0307	0,0354	0,0315
г_ _	0,0302	0,0234	0,0455	0,749–1,454	0,6551	1,3451	1,309	0,0205	0,0420	0,0409
д_ _	0,0292	0,0290	0,0354	2,263–2,764	1,5257	2,0978	1,8299	0,0196	0,0269	0,0235
е_ _	0,0438	0,0359	0,0555	3,197–4,941	3,0263	3,6893	4,0674	0,0340	0,0415	0,0457
є_ _	0,0189	0,0114	0,0321	0,252–0,707	0,2508	0,5443	0,6077	0,0114	0,0247	0,0276
ж_ _	0,0338	0,0243	0,0274	0,341–0,474	0,25	0,2302	0,2126	0,0179	0,0164	0,0152
з_ _	0,0273	0,0234	0,0352	1,311–1,973	1,1879	1,25	1,3259	0,0212	0,0223	0,0237
и_ _	0,0376	0,0338	0,0366	4,327–4,818	3,2931	4,0083	3,5984	0,0257	0,0313	0,0281
і_ _	0,0294	0,0277	0,0288	4,772–5,051	3,5963	3,9431	3,7918	0,0209	0,0229	0,0220
ї_ _	0,0114	0,0117	0,0168	0,038–0,125	0,2247	0,3031	0,2386	0,0102	0,0138	0,0108
й_ _	0,0180	0,0131	0,0188	0,301–0,432	0,3352	0,3469	0,3483	0,0146	0,0151	0,0151
к_ _	0,0383	0,0340	0,0415	2,791–3,400	2,4206	3,2381	2,4931	0,0295	0,0395	0,0304
л_ _	0,0539	0,0401	0,0364	2,073–3,070	2,4437	1,8021	2,0952	0,0429	0,0316	0,0368
м_ _	0,0238	0,0264	0,0343	2,168–3,123	1,7619	2,6603	1,8196	0,0194	0,0292	0,0200
н_ _	0,0468	0,0420	0,0474	6,421–7,257	3,8242	5,1327	4,0623	0,0250	0,0335	0,0266
о_ _	0,0473	0,0397	0,0540	6,473–8,795	5,3403	7,5276	6,3371	0,0328	0,0462	0,0389
п_ _	0,0476	0,0559	0,0720	1,858–2,809	1,6619	2,5456	2,1261	0,0426	0,0653	0,0545
р_ _	0,0384	0,0426	0,0456	3,690–4,380	3,1902	4,3566	3,4834	0,0332	0,0454	0,0363
с_ _	0,0541	0,0377	0,0381	3,169–4,541	3,3187	2,7052	3,4299	0,0395	0,0322	0,0408

т_ _	0,0445	0,0417	0,0429	5,174–5,518	3,5467	4,712	4,6607	0,0286	0,0380	0,0376
у	0,0286	0,0267	0,0332	2,193–2,726	1,7905	1,9852	1,9443	0,0218	0,0242	0,0237
ф	0,0384	0,0595	0,0401	0,276–0,495	0,3069	0,4759	0,3211	0,0345	0,0619	0,0374
х	0,0155	0,0180	0,0252	0,573–0,934	0,5083	0,7426	0,7957	0,0137	0,0201	0,0215
ц	0,0246	0,0345	0,0305	0,591–0,829	0,568	0,4416	0,4748	0,0237	0,0184	0,0198
ч	0,0425	0,0223	0,0559	0,513–1,324	1,0044	0,9368	0,6924	0,0437	0,0407	0,0301
ш	0,0145	0,0194	0,0457	0,194–0,657	0,2169	0,2917	0,6854	0,0130	0,0438	0,0378
щ	0,0200	0,0118	0,0201	0,064–0,100	0,1401	0,0828	0,1404	0,0097	0,0092	0,0142
ь	0,0317	0,0256	0,0329	0,998–1,285	0,6593	0,7983	0,7326	0,0169	0,0205	0,0188
ю	0,0173	0,0234	0,0309	0,277–0,494	0,1558	0,3005	0,2673	0,0097	0,0188	0,0167
я	0,0206	0,0216	0,0201	1,444–1,554	0,9522	1,0555	0,9361	0,0132	0,0147	0,0130

Згідно з даними Таблиця 6.12 та Рис. 6.24 частина літер в українській мові найчастіше вживані, інші – набагато рідше. Для найчастіше вживаних літер частота появи 3-грам з такими початковим літерами буде розподіл майже однаковий (пікові значення на графіку Рис. 6.24), а для інших літер – ні.

Таблиця 6.12

Розподіл частот появи 1-грами в Статтях 1–3

1-грама	Стаття 1		Стаття 2		Стаття 3	
	Кількість	ВЧ	Кількість	ВЧ	Кількість	ВЧ
о	2824	0.094240	2472	0.075898	3870	0.103601
н	2471	0.082460	2370	0.072766	2888	0.077312
а	2255	0.075252	2698	0.082837	2491	0.066685
т	2102	0.070146	1956	0.060055	2141	0.057315
і	1789	0.059701	1967	0.060393	2250	0.060233
и	1732	0.057799	1852	0.056862	2036	0.054504
в	1654	0.055196	1590	0.048818	1915	0.051265
с	1549	0.051692	1327	0.040743	1384	0.037050
е	1404	0.046853	1453	0.044612	2090	0.055950
р	1335	0.044550	1722	0.052871	1893	0.050676
к	1279	0.042682	1110	0.034080	1453	0.038897
л	1116	0.037242	927	0.028462	906	0.024254
у	987	0.032937	960	0.029475	1195	0.031990
д	859	0.028666	939	0.028830	1319	0.035310
м	808	0.026964	976	0.029966	1399	0.037451
п	647	0.021591	825	0.025330	1138	0.030464
я	647	0.021591	681	0.020909	864	0.023129
з	623	0.020790	644	0.019773	946	0.025325
ь	498	0.016619	418	0.012834	613	0.016410
ч	459	0.015317	289	0.008873	574	0.015366
г	408	0.013615	373	0.011452	651	0.017427
х	355	0.011847	384	0.011790	482	0.012903
б	284	0.009477	569	0.017470	428	0.011458
ж	246	0.008209	210	0.006448	176	0.004712
й	239	0.007976	260	0.007983	265	0.007094
ц	224	0.007475	334	0.010255	299	0.008004
є	188	0.006274	165	0.005066	347	0.009289
ф	179	0.005973	209	0.006417	137	0.003668
ї	174	0.005807	217	0.006663	270	0.007228
ю	156	0.005206	277	0.008505	289	0.007737
ш	117	0.003904	169	0.005189	281	0.007522
щ	95	0.003170	52	0.001597	128	0.003427

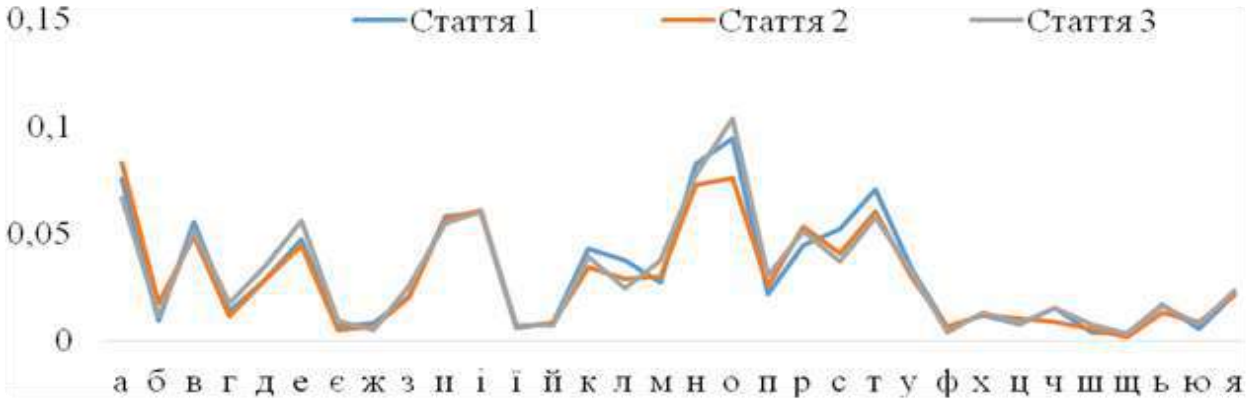


Рис. 6.24. Графік розподілу частот появи 1-грами в Статтях 1–3

Тому доцільно досліджувати лише триграми для початкових літер, що рідше зустрічаються в текстах конкретної мови для визначення ступеня належності тексту відповідному автору (наприклад, Рис. 6.25-Рис. 6.26).

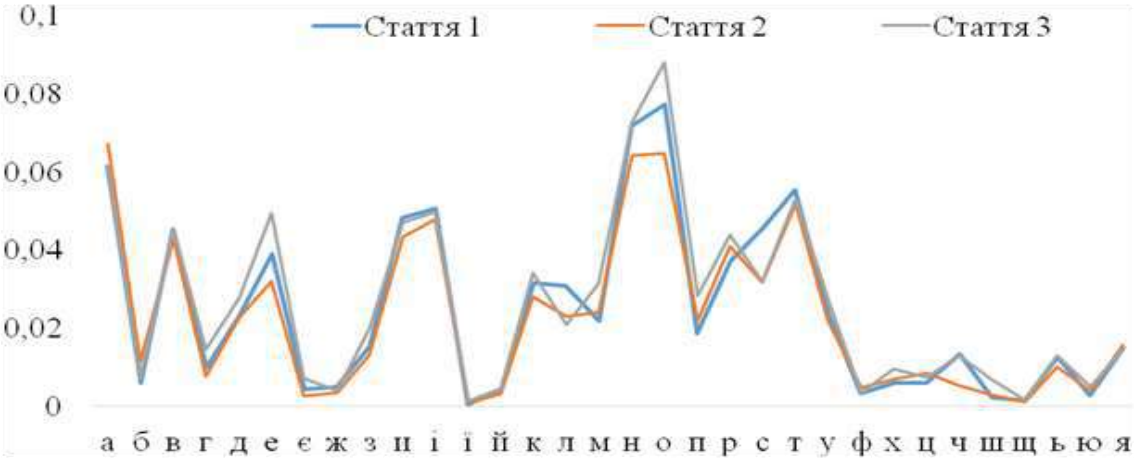


Рис. 6.25. Графік вживання 3-грам, які починаються з конкретної літери

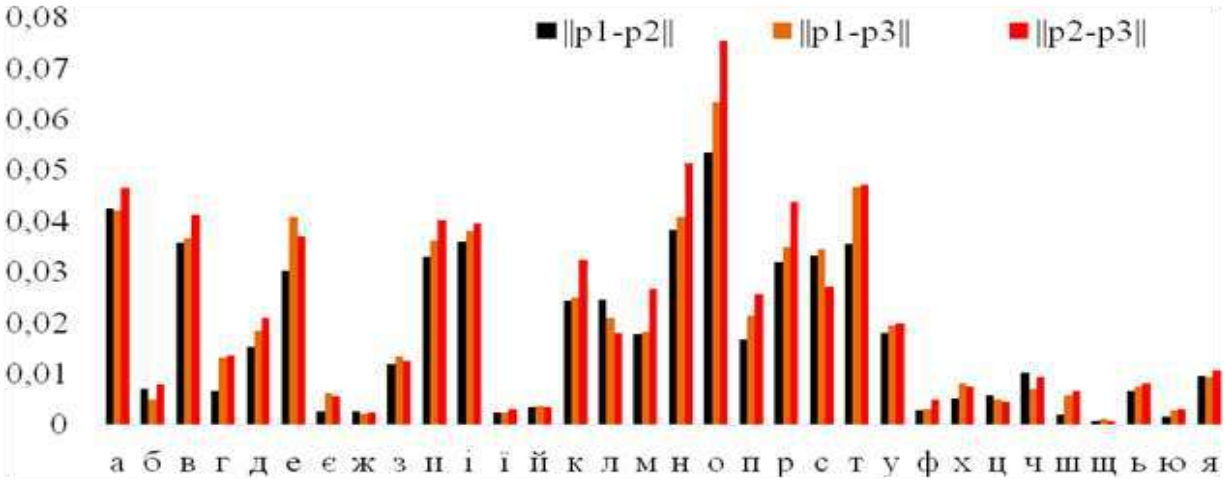


Рис. 6.26. Графік різниці вживання 3-грам, які починаються з конкретної літери

Згідно цих графіків випливає, що Стаття 1 та Стаття 2 ймовірніше були

написані одним автором, хоча Стаття 1 та Стаття також могли бути написані одним автором (але це не є істиною). А ось статті 2–3 точно були написані різними авторами. Застосування лінгвостатистичного аналізу 3-грам до множини статей дозволить сформувати підмножину подібних за лінгвістичними характеристиками публікацій. Накладання на цю підмножину додаткових умов у вигляді проведення лінгвостатистичних аналізів (множини ключових слів, стійких словосполучень, стилеметричного, лігвометричного тощо) дозволить значно скоротити цю підмножину, уточнивши список ймовірніших авторських робіт. Так, аналіз змісту та частоти появи лише службових слів відокремить статті 1та 3 в різні підмножини, статті 1 та 2 залишить в одній.

6.5. Аналіз розробленого методу кількісної оцінки ідентифікації потенційного автора науково-технічної публікації

Метод складається з шести алгоритмів аналізу україномовних текстів.

Алгоритм I. Попереднє опрацювання даних на основі контент-аналізу (парсинг, сегментація та токенізація тексту, а також лінгвістичний аналіз тексту).

Алгоритм II. Обчислення та аналіз ознак стилю мовлення автора (частота вживання слів, обсяг знаків пунктуації, речень, символів, слів і співвідношення кількості знаків і речень).

Алгоритм III. Розрахунок та аналіз параметрів стилю мовлення автора (зв’язність мовлення, синтаксична складність, лексична різноманітність, ступінь концентрації та винятковості тексту).

Алгоритм IV. Класифікація за параметрами та лексичним ознаками текстового контенту інших публікацій (застосування класифікаторів як нечіткі, SVM і комбінація попередніх двох).

Алгоритм V. Аналіз продуктивності на основі отриманих результатів для визначення точності кожного класифікатора.

Алгоритм VI. Визначення підмножини потенційних авторів на основі фільтрування з множини всіх досліджуваних через аналіз ознак та параметрів стилю (алгоритми VIII–XI).

Розроблено систему типу лексер (токенізатор, сегментатор) як частини аналізатора тексту на основі токенізації (Рис. 6.27). Токени витягуються під час роботи правил парсера і негайно перевіряються на відповідність умов в синтаксичних правилах для уникнення генерування абсурду (Рис. 6.28).

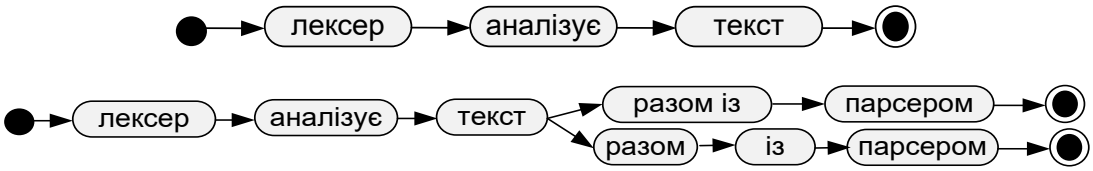


Рис. 6.27. Ілюстрація графів токенізації

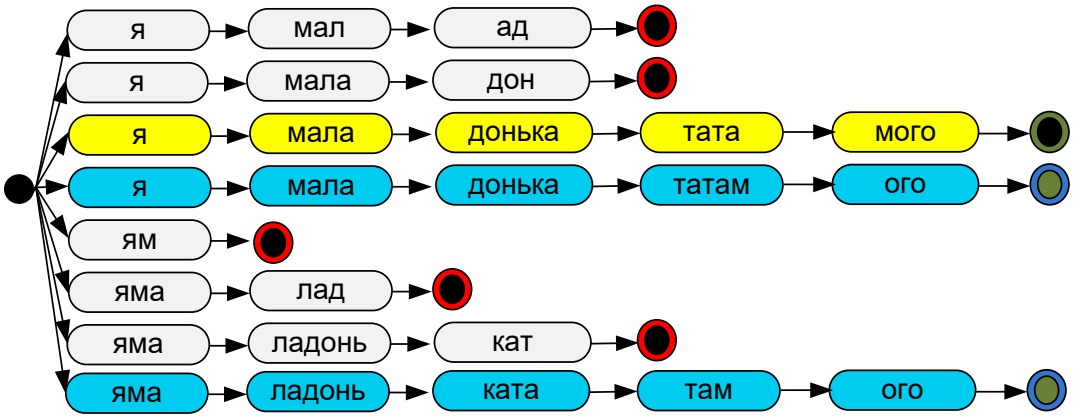


Рис. 6.28. Ілюстрація графу токенізації без синтаксичних правил

Правила допомагають вирішити кілька завдань, збільшуючи ефективність роботи граматичного движка, який в ході розбору тексту завантажує скомпільовані правила, не витрачаючи час на розбір синтаксису (алгоритм VII).

Алгоритм VII. Сегментатор текстового контенту

Крок 1. Розпізнаванні слова.

Крок 2. Визначення меж лексем.

Крок 3. Визначення повних словоформ.

Крок 4. Ідентифікація неподільних токенів, в яких є точки, пропуски і т.д.

Крок 5. Розбиття тексту на речення.

Крім визначення меж лексем, лексер також виконує попереднє розпізнавання морфологічних атрибутів слів, перетворюючи лексеми в токени.

Розрізняють при побудові україномовних речень з прямим порядком слів іменну групу $\tilde{N}$ та дієслівну групу $\check{R}$ (Рис. 6.29, Рис. 6.30).

I) $S \to \# \tilde{N}_{РД,ЧЛ,н,ОС} \tilde{R}_{ЧЛ,тп,ОС} \#$.

II) $\tilde{N} = \{AN\}$ or $\tilde{N} = N^p$

1) $\tilde{N}_{РД,ЧЛ,ВД,3} \to \tilde{N}_{РД,ЧЛ,ВД,3} \tilde{N}_{РД',ЧЛ',р,ОС}$;

2) $\tilde{N}_{РД,ЧЛ,ВД,3} \to A_{РД,ЧЛ,ВД} \tilde{N}_{РД,ЧЛ,ВД,3}$;

3) $K_1 \tilde{N}_{РД,ЧЛ,ВД,ОС} K_2 \to K_1 N^{займ}_{РД,ЧЛ,ВД,ОС} K_2$;

4) $\tilde{N}_{РД,ЧЛ,ВД,3} \to N_{РД,ЧЛ,ВД}$;

5) $\tilde{N}_{РД,ЧЛ,ВД,3} \to E\tilde{N}_{РД,ЧЛ,ВД,3}$;

6) $\tilde{N}_{РД,ЧЛ,ВД,3} \to \tilde{N}_{РД,ЧЛ,ВД,3} \tilde{N}_{РД,ЧЛ,м,3}$.

III) $\tilde{R} = R\tilde{N}$ or $\tilde{R} = \tilde{N}R$

1) $\tilde{R}_{ЧЛ,тп,ОС} \to R_{ЧЛ,тп,ОС} \tilde{N}_{РД',ЧЛ',з,ОС'} \tilde{N}_{РД'',ЧЛ'',о,ОС''}$;

2) $\tilde{R}_{ЧЛ,тп,ОС} \to R_{ЧЛ,тп,ОС} \tilde{N}_{РД',ЧЛ'С,о,ОС'} \tilde{N}_{РД'',ЧЛ'',з,ОС''}$;

3) $\tilde{R}_{ЧЛ,тп,ОС} \to R_{ЧЛ,тп,ОС} \tilde{N}_{РД',ЧЛ',з,ОС'}$;

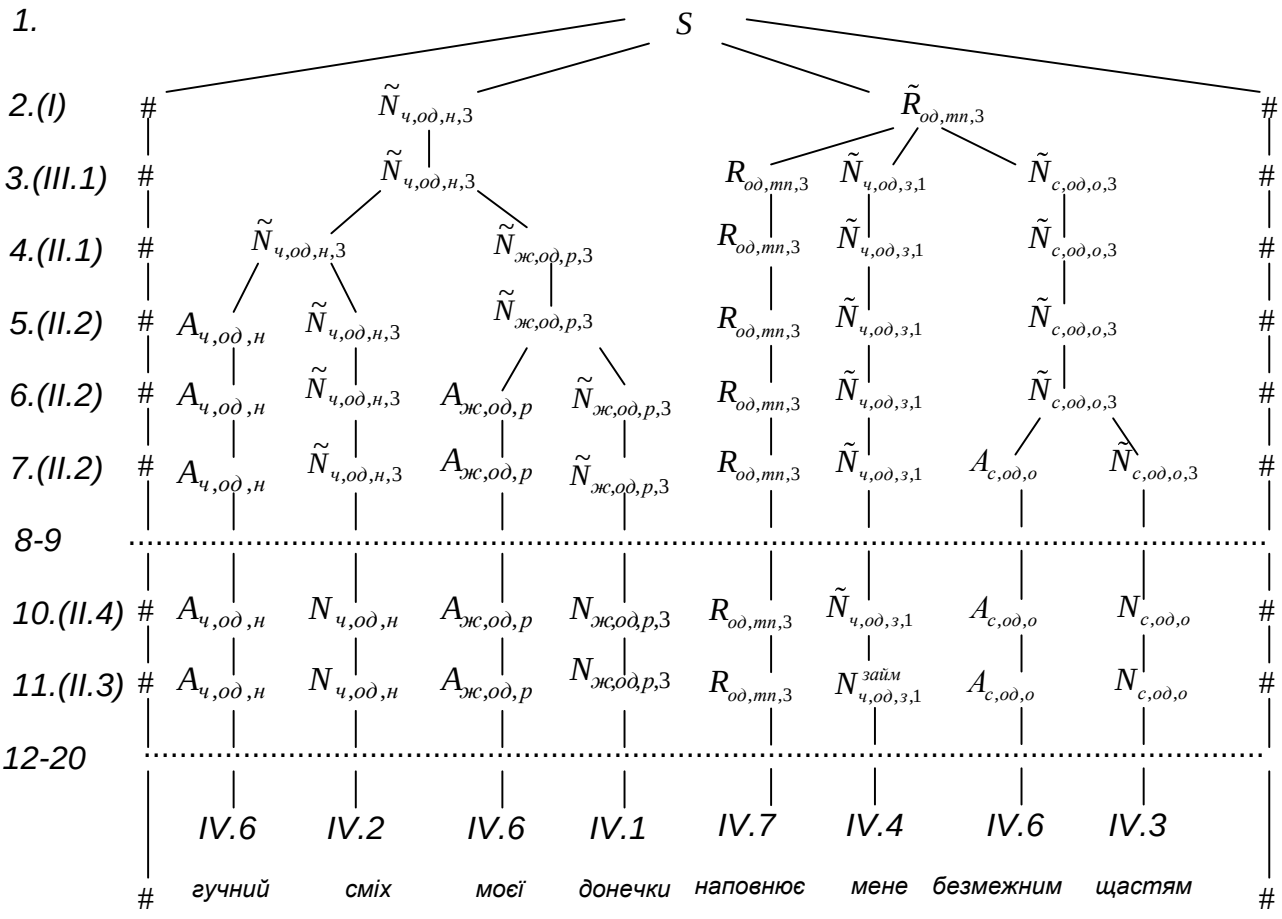
4) $\tilde{R}_{ЧЛ,тп,ОС} \to R_{ЧЛ,тп,ОС} \tilde{N}_{РД',ЧЛ',о,ОС'}$;

5) $\tilde{R}_{ЧЛ,тп,ОС} \to R_{ЧЛ,тп,ОС} E\tilde{N}_{РД,ЧЛ,м,3}$:

6) $\tilde{R}_{ЧЛ,тп,ОС} \to E\tilde{N}_{РД,ЧЛ,м,3} R_{ЧЛ,тп,ОС}$.

IV) $Words = \{x_1, x_2, x_3, \ldots, x_n\}$

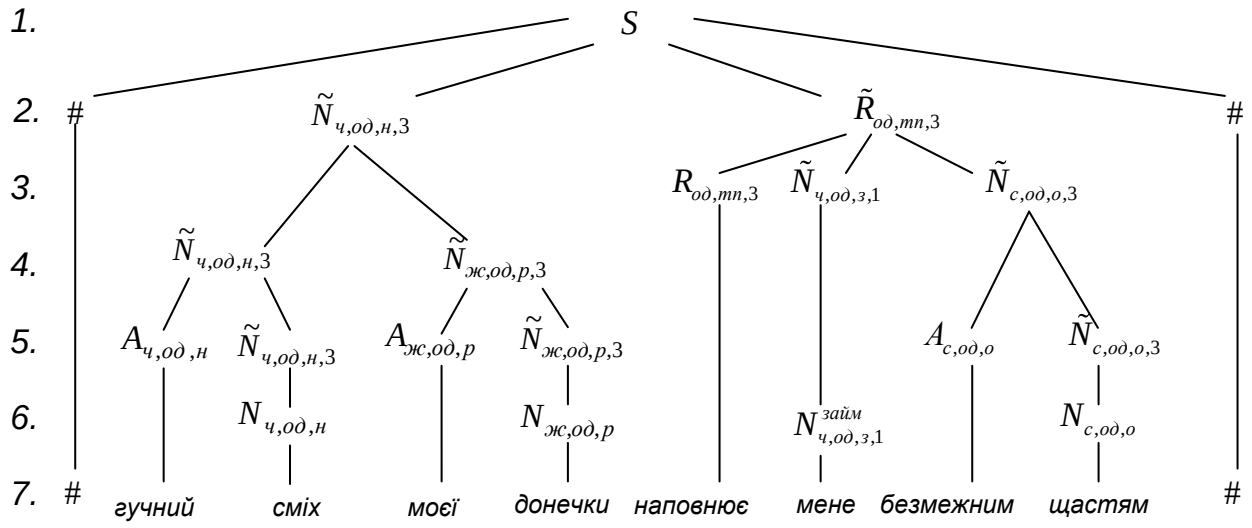
Рис. 6.29. Продукційні правила аналізу україномовного речення, де N – іменник, A – прикметник, $N^{займ}$ – займенник; число/*ЧЛ* (*од*, *мн*); рід/*РД* (*ч*, *ж*, *с*;), особа/ОС (1, 2, 3); відмінок/*ВД* (*н*, *р*, *д*, *з*, *о*, *м*, *к*); час/*ЧС* (*тп*, *мн*, *мб*)



1. S
2. (I) $\# \ \tilde{N}_{ч,од,н,3} \ \tilde{R}_{од,тп,3} \#$
3. (III.1) $\# \tilde{N}_{ч,од,н,3} \ R_{од,тп} \ \tilde{N}_{ч,од,з,1} \ \tilde{N}_{с,од,о,3} \#$
4. (II.1) $\# \tilde{N}_{ч,од,н,3} \ \tilde{N}_{ж,од,р,3} \ R_{од,тп} \ \tilde{N}_{ч,од,з,1} \ \tilde{N}_{с,од,о,3} \#$
5. (II.2) $\# A_{ч,од,н} \ \tilde{N}_{ч,од,н,3} \ \tilde{N}_{ж,од,р,3} \ R_{од,тп} \ \tilde{N}_{ч,од,з,1} \ \tilde{N}_{с,од,о,3} \#$
6. (II.2) $\# A_{ч,од,н} \ \tilde{N}_{ч,од,н,3} \ A_{ж,од,р} \ \tilde{N}_{ж,од,р,3} \ R_{од,тп} \ \tilde{N}_{ч,од,з,1} \ \tilde{N}_{с,од,о,3} \#$
7. (II.2) $\# A_{ч,од,н} \ \tilde{N}_{ч,од,н,3} \ A_{ж,од,р} \ \tilde{N}_{ж,од,р,3} \ R_{од,тп} \ \tilde{N}_{ч,од,з,1} \ A_{с,од,о} \ \tilde{N}_{с,од,о,3} \#$
8. (II.4) $\# A_{ч,од,н} \ N_{ч,од,н} \ A_{ж,од,р} \ \tilde{N}_{ж,од,р,3} \ R_{од,тп} \ \tilde{N}_{ч,од,з,1} \ A_{с,од,о} \ \tilde{N}_{с,од,о,3} \#$
9. (II.4) $\# A_{ч,од,н} \ N_{ч,од,н} \ A_{ж,од,р} \ N_{ж,од,р} \ R_{од,тп} \ \tilde{N}_{ч,од,з,1} \ A_{с,од,о} \ \tilde{N}_{с,од,о,3} \#$
10. (II.4) $\# A_{ч,од,н} \ N_{ч,од,н} \ A_{ж,од,р} \ N_{ж,од,р} \ R_{од,тп} \ \tilde{N}_{ч,од,з,1} \ A_{с,од,о} \ N_{с,од,о} \#$
11. (II.3) $\# A_{ч,од,н} \ N_{ч,од,н} \ A_{ж,од,р} \ N_{ж,од,р} \ R_{од,тп} \ N^{займ}_{ч,од,з,1} \ A_{с,од,о} \ N_{с,од,о} \#$
12. (IV.6) $\#$*гучний* $N_{ч,од,н} \ A_{ж,од,р} \ N_{ж,од,р} \ R_{од,тп} \ N^{займ}_{ч,од,з,1} \ A_{с,од,о} \ N_{с,од,о} \#$
13. (IV.2) $\#$*гучний сміх* $A_{ж,од,р} \ N_{ж,од,р} \ R_{од,тп} \ N^{займ}_{ч,од,з,1} \ A_{с,од,о} \ N_{с,од,о} \#$
14. (IV.6) $\#$*гучний сміх моєї* $N_{ж,од,р} \ R_{од,тп} \ N^{займ}_{ч,од,з,1} \ A_{с,од,о} \ N_{с,од,о} \#$
15. (IV.1) $\#$*гучний сміх моєї* $N_{ж,од,р} \ R_{од,тп} \ N^{займ}_{ч,од,з,1} \ A_{с,од,о} \ N_{с,од,о} \#$
16. (IV.7) $\#$*гучний сміх моєї донечки* $R_{од,тп} \ N^{займ}_{ч,од,з,1} \ A_{с,од,о} \ N_{с,од,о} \#$
17. (IV.7) $\#$*гучний сміх моєї донечки наповнює* $N^{займ}_{ч,од,з,1} \ A_{с,од,о} \ N_{с,од,о} \#$
18. (IV.4) $\#$*гучний сміх моєї донечки наповнює мене* $A_{с,од,о} \ N_{с,од,о} \#$
19. (IV.6) $\#$*гучний сміх моєї донечки наповнює мене безмежним* $N_{с,од,о} \#$
20. (IV.3) $\#$*гучний сміх моєї донечки наповнює мене безмежним щастям*#

Рис. 6.30. Ілюстрація аналізу структури україномовного речення

Отримаємо дерево складових, або синтаксичну структуру аналізованого речення (Рис. 6.31).



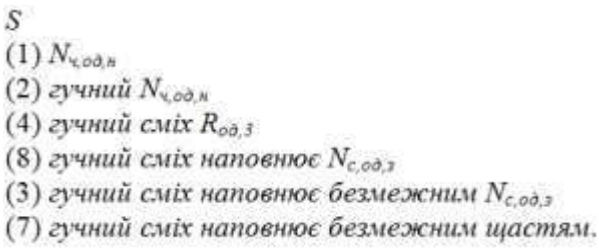


Рис. 6.31. Ілюстрація аналізу україномовного речення

Для словникових лексем також визначається словникова стаття, формою якої є лексема. В алфавітно-частотних словниках через / для слова визначені його характеристики (Рис. 6.32-Рис. 6.33).

Уривок 1	Уривок 2	Уривок 3
буферизувати/**ABGH**	клавіатурний/**V**	консоль/ij
відформатувати/**AB**	Кобол/**e**	конфігуратор/**efg**
декодувати/**ABGH**	кодек/**efg**	копілефт/**e**
кешувати/**ABGH**	кодер/**efg**	копірайт/**e**
кирилічний/**V**	кодогенератор/**efg**	криптографічний/**V**
кілобайтовий/**V**	кодосумісний/**V**	криптозахищений/**V**
кілобайт/**efg**	комбосписок/**ab**	крос-асемблер/**efg**
кілобітовий/**V**	комутований/**V**	крос-компілятор/**efg**
кілобіт/**efg**	конкатенація/**ab**	кука/**ab**
кілобод/**efg**	консольний/**V**	курсорний/**V**

Рис. 6.32. База правил алфавітно-частотного словника частин мови), де A – дієслово, інші великі літери– додаткові характеристики дієслова, V – прикметник, маленькі літери англійського алфавіту – характеристики іменника

```
Файл  Правка  Вид  Справка
##################################################################
# Групи a b c d o
#
# -- Перша відміна: іменники жіночого та чоловічого та середнього роду
#
# -- Друга відміна: іменники чоловічого роду із закінченнян на -ар -ир
#                   наголошені (Мішана група на -ар -ир)
#
# -- Друга відміна: іменники чоловічого роду з чергуванням -і -о
#
# -- Числівники -ять, -сят, -сто
#
#

SFX a V 235

#
# ОДНИНА (множина перенесена в гр. b)
#
# Спочатку перша відміна
#
# тверда група в Називному відмінку однини з закінченням на -а
# однина
SFX a   а и [^жчшщ]а  # хата  хати  (Р.)
SFX a   а і [^ггкх]а  # хата  хаті  (Д.М.)
SFX a   а у а    # хата  хату  (З.)
SFX a   а ою [^жчшщ]а  # хата  хатою  (О.)
```

Рис. 6.33. Регулярні вирази морфологічного аналізу іменників

В базі даних збережені регулярні вирази приведення до основи слова (Рис. 6.34), де flag правило ідентифікації типу слова (наприклад, іменникова група, однина), mask – флексії слова (в квадратних дужках – виключення), find – флексії слова в називному відмінку, repl – флексії слова при відмінюванні (Рис. 6.35).

а)

id	ordering	state	flag	type	lang	mask	find	repl
26	26	1	a	SFX	uk	ін	ін	оном
27	27	1	a	SFX	uk	ін	ін	оні
28	28	1	a	SFX	uk	ір	ір	огу
29	29	1	a	SFX	uk	ір	ір	огові
30	30	1	a	SFX	uk	ір	ір	огом
31	31	1	a	SFX	uk	ір	ір	озі
32	32	1	a	SFX	uk	[^л]ід	ід	оду
33	33	1	a	SFX	uk	[^л]ід	ід	одові
34	34	1	a	SFX	uk	[^л]ід	ід	одом
35	35	1	a	SFX	uk	[^л]ід	ід	оді
36	36	1	a	SFX	uk	[^пг]лід	ід	ьоду
37	37	1	a	SFX	uk	[^пг]лід	ід	ьодові
38	38	1	a	SFX	uk	[^пг]лід	ід	ьодом
39	39	1	a	SFX	uk	[^пг]лід	ід	ьоді
40	40	1	a	SFX	uk	[пг]лід	ід	оду
41	41	1	a	SFX	uk	[пг]лід	ід	одові
42	42	1	a	SFX	uk	[пг]лід	ід	одом
43	43	1	a	SFX	uk	[пг]лід	ід	оді
44	44	1	a	SFX	uk	іб	іб	обу

б)

id	ordering	state	word	lang
1	1	1	після	uk
2	2	1	між	uk
3	3	1	are	en
4	4	1	and	en
5	5	7	між	uk
6	6	1	been	en
7	7	1	has	en
8	8	1	their	en
9	9	1	any	en
10	10	1	the	en
11	11	1	with	en
12	12	1	таких	uk
13	13	1	їхніми	uk
14	14	1	как	ru
15	15	1	такої	uk

Рис. 6.34. База правил визначення: *а* – основи слова; *б* – службових слів

```
# Іменники із закінченням на –ін з чергуванням -і -о
SFX a ін    ону   ін        # загін загону          (Д.Р.)
SFX a ін    онові ін        # загін загонові  (Д.)
SFX a ін    оном  ін        # загін загоном   (О.)
SFX a ін    оні   ін        # загін загоні           (М.)
третій рядок описує
# Іменники із закінченням на –ір з чергуванням -і -о
SFX a ір    огу   ір        # батір батогу           (Д.Р.)
SFX a ір    огові ір        # батір батогові  (Д.М.)
SFX a ір    огом  ір        # батір батогом   (О.)
SFX a ір    озі   ір        # батір батозі           (М.)
дев'ятий рядок описує
# Іменники із закінченням на –ід з чергуванням -і -о
SFX a ід    оду   [^л]ід    # провід проводу (Д.Р.)
SFX a ід    одові [^л]ід    # провід проводові  (Д.)
SFX a ід    одом  [^л]ід    # провід проводом   (О.)
SFX a ід    оді   [^л]ід    # провід проводі  (М.)
```

Рис. 6.35. Ілюстрація правил визначення основи слова

Також в базі даних (Рис. 6.34б) є словник службових слів, тобто слів, які є додатковими параметрами для аналізу особливостей стилю мовлення автора, та врахування при аналізі текстів впливає суттєво на кінцевий результат.

Визначимо оптимальний розроблений алгоритм з чотирьох (VIII-XI) для ідентифікації стилю автора публікації на основі аналізу його колективних робіт.

Алгоритм VIII. Фільтрація множини аналізованих авторських стилів

```
int i=0, j=0;
while (i<4){
   int c1=0, c2=0, cc2=0;
   while (j<94){
        int s=0;
        while (l<12){
           if ((K[i][l]+abs(F[l]-K[i][l]))>A[j][l]) &&
                      ((K[i][l]-abs(F[l]-K[i][l]))< A[j][l])
              s+=1;
           if (l>6) && ((K[i][l]+abs(F[l]-K[i][l]))>A[j][l]) &&
                       ((K[i][l]-abs(F[l]-K[i][l]))< A[j][l])cc2+=s;
           l+=1;
           }
        A2[j]=s;
        A3[j]=cc2;
        c1+=s;
        c2+=s;
        j+=j;
        }
   float t1=c1/94, t2=c2/94;
   int filtr1=0, filtr2=0, filtr3=0
   while (j<94){
        if(A2[j]>=t1) filtr1+=1;
        if(A3[j]>=t2) filtr2+=1;
        if (A2[j]>=t1)&&(A3[j]>=t2)filtr3+=1;
        j+=1;
   i+=1;
   }
```

Масив K[i][l] – параметри та коефіцієнти стилю для 4-ох колективних робіт (Таблиця 6.13 та Таблиця Д.5 додатку Д – виділено жовтим кольором), частина авторів яких є під № 6 та 30 (виділено синім кольором). Масив A[j][l] – ознаки стилю для 94-ох авторів. Масив F[l] – середні значення ознак стилю для 94-ох авторів. Алгоритм визначає, чи значення параметрів та коефіцієнтів мовлення стилю *j*-того автора попадає в межі $[x_i+x_{сер}; x_i-x_{сер}]$ відхилення значень параметрів та коефіцієнтів мовлення стилю *i*-тої колективної роботи. Заповнюються через фільтри масиви A2 (автори, значення більшості параметрів та коефіцієнтів подібні на стиль колективу *i*) та A3 (автори, значення більшості лише коефіцієнтів подібні на стиль колективу i). Далі з отриманих попередніх масивів

накладанням нового фільтру формується нова підмножина авторів (стилі яких більш подібні на колективні – *i*-ту роботу).

Таблиця 6.13

Результат роботи алгоритму аналізу стилю автора публікації на Victana [16] 94 авторів на понад 300 одноосібних публікаціях за період 2001–2021 рр.

№	*N*	*W*	W_1	W_{10}	*P*	*Z*	*S*	K_l	K_s	K_z	I_{wt}	I_{kt}
1	622	397	305	5	37	42	48	0,64	0,91	0,81	0,77	0,013
2	614	391	287	4	46	69	32	0,64	0,88	0,73	0,73	0,01
3	658	345	241	8	31	59	42	0,52	0,91	1,07	0,7	0,023
4	631,3	377,7	277,7	5,7	38	56,7	40,7	0,6	0,9	0,88	0,73	0,015
5	661,1	402,7	299,7	4,7	44,7	54,7	24,8	0,61	0,89	0,6	0,74	0,012
6	694,5	417,4	313,1	6,4	54,3	58,5	38,1	0,6	0,87	0,62	0,75	0,015
7	691,8	403,4	301,6	7,8	47,8	60	47,8	0,58	0,88	0,79	0,75	0,019
8	682,5	394,2	291	5	49	61	39,7	0,58	0,88	0,74	0,74	0,013
9	733,5	486,5	392	5	50	65	45	0,66	0,9	0,76	0,8	0,01
…….												
29	704,5	412	303,5	5,5	59	47,5	38	0,58	0,86	0,49	0,74	0,013
30	688,8	416,8	321,9	6	49,7	49,3	41,3	0,6	0,88	0,67	0,77	0,016
…….												
94	680	414	314	4	55	62	34	0,6	0,87	0,58	0,76	0,01

В результаті отримаємо значення, подані в Таблиця 6.14 (алгоритм VIII). Стовпці A – це результат аналізу всіх значень векторів коефіцієнтів та параметрів мовлення авторів з Таблиця 6.13. Стовпці B – це результат аналізу лише останніх 5 стовпців в Таблиця 6.13. Нажаль цей алгоритм надав такі результати, що наведені автори цих робіт малоймовірно самі написали (найкращі результати виділені червоним кольором – і замало, щоб стверджувати, що вони є авторами понад 50% цих колективних робіт). Хоча з іншого боку цей алгоритм дає гарні результати – зменшуючи на першому етапі визначення авторства кількість авторів (до 34,04% із загальної кількості учасників проекту). Це необхідно для подальшої фільтрації через аналіз стопових слів (прийменників та сполучників) та ключових слів, особливості семантики та лексики при побудові речень тощо.

Таблиця 6.14

Експериментальна апробація алгоритмів I–IV на Web-ресурсі Victana [16]

Алгоритм	Колектив	Середнє значення		Автор				Фільтр			%
				6		30		1	2	3	
		A	B	A	B	A	B	A	B		
VIII	1	5.55319	2.3617	3	2	**6**	2	48	39	35	37,2
	2	7.361702	3.21277	6	3	6	3	40	37	25	26,6
	3	7.521277	3.925532	**8**	**5**	5	**5**	58	35	35	37,2
	4	4.148936	1.457447	3	2	3	0	41	43	33	35,1
	$\bar{x}_i$	6,15	2,74	5,0	3,0	5,0	2,5	46,8	38,5	32,0	34,0

IX	1	5.85106	2.75532	5	2	**8**	**3**	53	53	46	48,9
	2	5.6383	2.7234	**6**	**4**	4	**3**	53	56	43	45,7
	3	3.45745	1.04255	3	0	2	0	40	21	15	15,9
	4	6.2766	2.90426	6	**3**	5	2	44	54	41	43,6
	$\bar{x}_i$	5,31	2,36	5,0	2,3	4,8	2,0	47,5	46,0	36,3	38,6
X	1	6.44681	2.6383	**9**	**3**	**6**	**3**	46	55	42	44,7
	2	7.23404	3.39362	**8**	**4**	**8**	3	45	46	34	36,2
	3	6.46809	2.55319	**8**	**4**	**9**	**4**	48	46	39	41,5
	4	7.8516	3.54255	**9**	3	**9**	**5**	53	51	43	45,7
	$\bar{x}_i$	7,00	3,03	8,5	3,5	8,0	3,8	48,0	49,5	39,5	42,0
XI	1	6.31915	2.11702	**3**	2	**8**	**3**	45	35	29	30,9
	2	4.82979	2.14894	**6**	**3**	**6**	2	51	36	30	31,9
	3	5.89362	2.5	**8**	**4**	**9**	**4**	56	42	41	43,6
	4	5.53191	2.58511	**8**	**3**	**7**	2	49	53	43	45,7
	$\bar{x}_i$	5,64	2,34	6,3	3,0	7,5	2,8	50,3	41,5	35,8	38,0

В результаті отримаємо значення, подані в Таблиця 6.14 (алгоритм IX). Тоді проаналізуємо алгоритм IX. Суттєво не відрізняється від попереднього, лише умовою в третьому циклі:

```
if ((K[i][l]+V[l])>A[j][l]) && ((K[i][l]- V[l])< A[j][l]) s+=1
```

де V[l] – масив середніх абсолютних значень відхилень точок даних від середнього значення. Отримані результати трохи покращились, але не настільки, щоб стверджувати, що автори під номером 6 та 30 є справжніми авторами колективних робіт 1–4, хоча вони їх точно писали. З іншого боку, трохи збільшилась кількість авторів (до 38,56 % із загальної кількості учасників проекту) з подібністю в стилі мовлення. Тепер проаналізуємо алгоритм X. В алгоритмі 1 також замінимо в третьому циклі умову на таку:

```
if (abs(A[j][l]- K[i][l])>abs(K[i][l]-F[l])) s+=1
```

В результаті отримаємо значення, подані в Таблиця 6.14 (алгоритм X). Як бачимо – отримані значення гарантовано дають зрозуміти, що стиль авторів під номерами 6 та 30 досить наближений (понад 75–100 %) на стиль колективних робіт 1-4 відповідно (червоним кольором виділені позитивні результати). Хоча значно зросла кількість авторів (до 42,02 % із загальної кількості учасників проекту) з подібністю в стилі мовлення. З іншого боку, в той список багато увійшло тих, то не попав на попередніх етапах дослідження, і випали з множини ті, що увійшли також на попередніх двох етапах дослідження. Тепер спробуємо все таких зменшити ту загальну кількість, застосувавши алгоритм XI до отриманих початкових даних – параметрів та коефіцієнтів мовлення 94-ох учасників проекту. В алгоритмі X вдосконалимо в третьому циклі умову:

```
if ((abs(A[j][l]- K[i][l])>abs(K[i][l]-F[l])) && (abs(A[j][l]-
F[l])>abs(K[i][l]-F[l])))|| ((abs(A[j][l]- K[i][l])<abs(K[i][l]-F[l])) &&
(abs(A[j][l]- F[l])<abs(K[i][l]-F[l]))) s+=1
```

В результаті отримаємо значення, подані в Таблиця 6.14 (алгоритм XI). Отримані значення також підтверджують, що стиль авторів під номерами 6 та 30 досить наближений (понад 75–100 %) на стиль колективних робіт 1–4 відповідно (червоним кольором виділені позитивні результати). Також значно зменшили кількість авторів (до 38,03 % із загальної кількості учасників проекту) з подібністю в стилі мовлення. На Рис. 6.36 подані детальні графіки отриманих результатів при застосуванні алгоритмів VIII–XI (під номерами 1–4 відповідно) для аналізу розробленого нами методу визначення стилю автора.

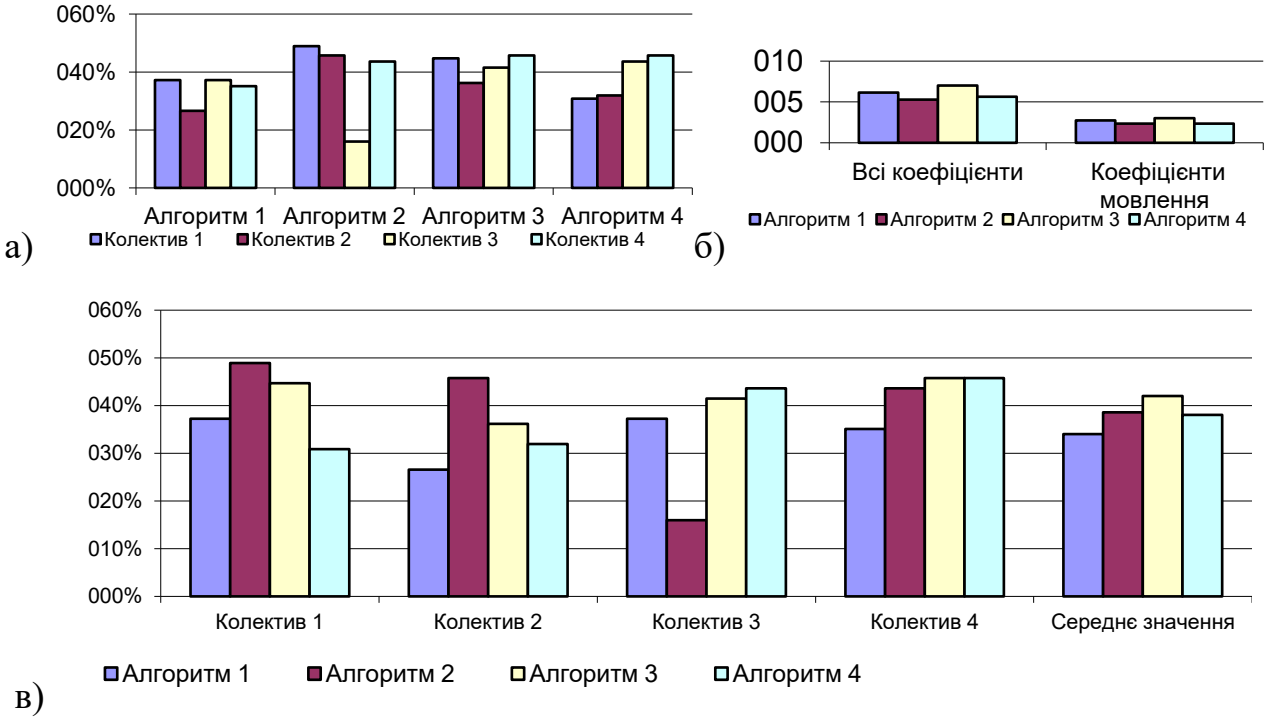


Рис. 6.36. Дослідження ідентифікації стилю: *а* – за розробленими алгоритмами; *б* – з врахуванням ознак мовлення; в – для аналізованих колективних робіт

Далі для визначення стилю автора використано аналіз стопових слів (прийменників та сполучників) та ключових слів творів авторів, як потрапили до тих 38,03 %. Кожна особистість має свій особливий словниковий запас для передачі своєї думки, в тому числі так званих «паразитичних» (тобто, отже, хоча тощо) та службових слів (і, та, й, але, хоч би тощо).

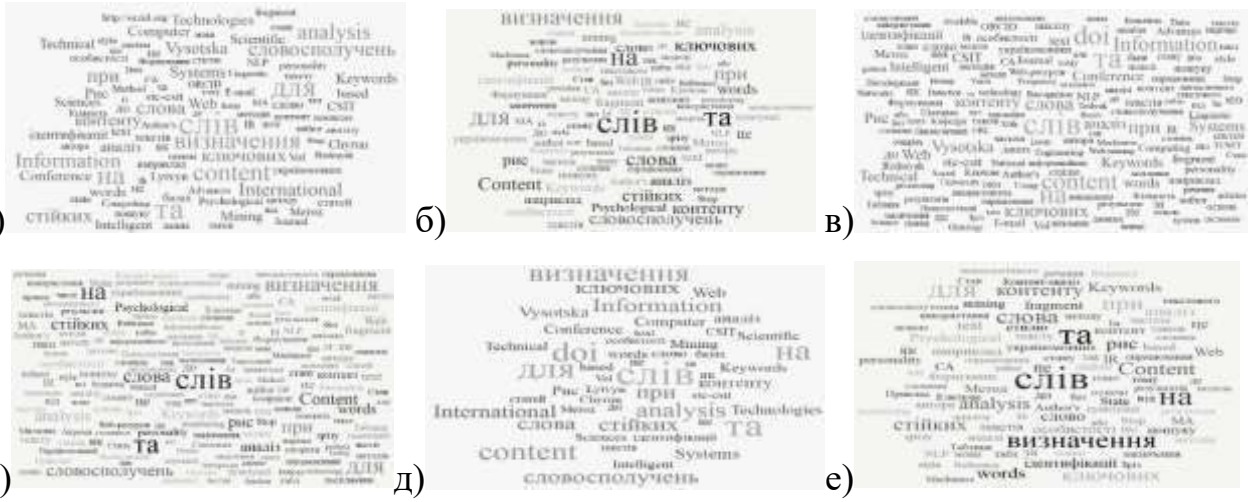


Рис. 6.37. Дослідження стилю на етапі 2 для тексту з побудовою частотного словника: *а* – повного зі 100 слів; *б* – основного зі 100 слів; *в* – повного з 200 слів; *г* – основного з 200 слів; *д* – повного з 50 слів; *е* – основного з 50 слів

На Рис. 6.37 поданий приклад аналізу стилю автора на другому етапі – через аналіз частоти появи службових та ключових слів з врахування різних фільтрів аналіз повних текстів з списком літератури та анотаціями на різних мова, та аналіз лише інформативної частини публікації, тобто основного тексту з побудовою відповідно частотного словника на 200, 100 та 50 слів).

6.6. Основні результати та висновки розділу

Розроблено метод визначення автора в україномовних текстах на основі аналізу коефіцієнтів лексичного авторського мовлення в еталонному уривку авторського тексту, який ґрунтується на аналізі колекції ключових слів, стійких словосполучень, показників лінгвометрії, стилеметрії, а також результатів аналізу N-грам на основі порівнянь різниць вживання 2-грам та 3-грам для подібних за стилем публікацій в межах [6;7]%, а для точно не подібних – >12%), що забезпечило можливість визначити множину потенційних авторів публікацій з більш ніж одного автора (до [9;34]% із загальної кількості учасників проекту) та розробити метод ідентифікації авторського стилю.

Розроблено метод ідентифікації стилю автора тексту на основі аналізу ознак стилю мовлення автора в шаблонному уривку авторського тексту. Метод полягає в порівняльному аналізі авторської атрибуції в статистично опрацьованому

доробку автора (еталоні) з довільним аналізованим уривком. Метод оцінює ступінь приналежності тексту до шаблону авторського стилю із аналізом відповідних коефіцієнтів лексичного авторського мовлення. Причому метод працює при умові, що шаблон авторського стилю згенерований на достовірних даних. Для атрибуції використано аналіз опорних слів, отримані результати подано у вигляді коефіцієнтів кореляції. Окремо згадаємо про еволюцію значущості одного із параметрів тексту – в авторській атрибуції текстів.

Розроблено алгоритм ідентифікації службових слів на основі лінгвістичного аналізу текстового контенту. Для кожного з уривків проаналізовані та порівняні із еталонним значеннями абсолютні та відносні частоти появи стопових слова. Отже, застосування методу опорних слів дає такі результати: знаходження серед досліджуваних уривків того, що найбільш ймовірно належить до еталону. Інші результати підтверджують дієвість методу опорних слів у авторській атрибуції текстів. Висунуте припущення про незначущість впливу частки як параметра методу на результати привело до зменшення коефіцієнтів кореляції. Проте, для підтвердження чи спростування того факту, що частки не є визначальним фактором в авторському стилі необхідно виконати ґрунтовніші дослідження.

Розроблено алгоритм лексичного аналізу україномовних текстів та алгоритм синтаксичного аналізатора текстового контенту. Особливостями алгоритмів є адаптація морфологічного та синтаксичного аналізу словоформ до особливостей побудови україномовних слів/текстів. Враховувалась належність до частини мови та відмінювання в межах цієї частини мови на основі аналізу флексій та основ слів за регулярними виразами.

Проведено порівняння результатів контент-моніторингу на множині 300 одноосібних робіт технічного спрямування біля 100 різних авторів за період 2001–2021 рр. для визначення чи змінюються і як коефіцієнти різноманітності тексту цих авторів в різні проміжки часу. Найкращих результатів за критерієм щільності досягає метод аналізу статті без початкової обов’язкової інформації як анотації та ключові слова різними мовами, а також списку літератури.

Проведено декомпозицію методу ідентифікації потенційного автора на

основі аналізу параметрів стилю мовлення як зв'язність мовлення, ступінь синтаксичної складності, лексична різноманітність, ступінь концентрації та винятковості. Проаналізовані також ознаки авторського стилю, як загальний обсяг слів тексту, обсяг унікальних слів, обсяг сполучників/прийменників, обсяг речень, обсяг слів із частотою 1 та ≥10. Для прикладу проаналізовано 3-грами 3-х статей. Для Статті 1 проаналізовано 78,4814 % 3-грам, для Статті 2 – 72,6332 % та для Статті 3 – 84,1271 %. Відповідно різниця вживання відповідних 3-грам між Статтями 1–2 є R_{12}=56,5254 %, між 2 і 3 – R_{23}=69,4271 %, між 1 і 3 – R_{13}=62.9839 %. Самі ці показники показують, що характеристики статті 1 і 2 більш подібні (R_{23}>R_{12} на 12.9017 %, R_{23} > R_{13} на 6.4432 %, R_{13}> R_{12} на 6.4585 %, тобто R_{23}>R_{13}>R_{12}), ніж характеристики відповідно Статті 1–3 і 2–3. Чим менше R_{ij}, тим більша ступінь, що статті написані одним автором. Тоді в випадку Стаття 1–2 більш ймовірніше написана одним автором, ніж Статті 2–3 і 1–3 відповідно.

Достовірність наукових та практичних результатів підтверджується відповідними матеріалами про впровадження дисертаційних досліджень, а також за рахунок порівняння одержаних практичних результатів на різних вибірках достовірних вхідних даних. Розроблено КЛС на інформаційному ресурсі http://victana.lviv.ua засобами CMS Joomla! (для розроблення е-каркасу статей), PHP (для реалізації методів опрацювання текстового контенту), HTML (для реалізації розмітки сторінок), CSS (для опису стилів сторінок), MySQL (для зберігання даних та словників). Експериментальне дослідження підтвердило достовірність методу визначення ключових слів – для різних алгоритмів опрацювання первинного тексту середній збіг списків виявлених ключовиків з авторськими змінюється у проміжку 52,6-68,5%. Точність збігу ключових слів із авторськими коливається в проміжку 43,6-62,9%. Середній збіг змістовних ключових слів порівняно зі всіма знайденими системою коливається в проміжку 38,9-75,8% в залежності від етапів аналізу текстів статей. Точність збігу ключових слів порівняно зі всіма знайденими системою коливається в проміжку 34,3-71,9% в залежності від етапів аналізу текстів статей. Основні результати розділу опубліковані у роботах [163, 535, 958-983, 984-1008].

ВИСНОВКИ

У дисертаційній роботі вирішено важливу науково-прикладну проблему аналізу та синтезу КЛС для розв’язання різних задач опрацювання україномовного текстового контенту на основі розроблення нових та удосконаленні відомих моделей, методів та засобів NLP.

Під час виконання роботи одержано такі результати:

1. Проведено аналіз сучасного стану та перспективи розвитку ІТ опрацювання природної мови, що дало змогу визначити проблему та задачі дослідження, а також сформувати загальні напрями дослідження при відсутності некомерційних КЛС з відкритим кодом для опрацювання україномовного текстового контенту та стандартизованого підходу проектування.

2. Обґрунтовано актуальність розв’язання проблеми аналізу та синтезу КЛС на основі розроблення загальної структури системи опрацювання україномовного текстового контенту, яка за рахунок взаємодії основних процесів/компонентів ІС та адаптованих до української мови методів лінгвістичного опрацювання текстового контенту на основі графемного, морфологічного, лексичного, синтаксичного, семантичного, структурного, онтологічного та прагматичного аналізу дозволила вдосконалити ІТ інтелектуального аналізу текстового потоку для розв’язку конкретної задачі NLP. Це забезпечило адаптацію процесів NLP для аналізу україномовного текстового контенту та на їх основі підвищити точність отриманих результатів на 6-48% в залежності від конкретної задачі NLP. Наприклад, для задачі NLP визначення ключових слів україномовного тексту щільність ключових слів збільшується в діапазоні [1,23; 1,48] раз або на [23,14; 47,83]% в залежності від якості/точності поповнення тематичного словника через машинне навчання.

3. Вдосконалено методи опрацювання інформаційних ресурсів як інтеграція, управління та супровід україномовного контенту, що дозволило адаптувати процес інтелектуального аналізу текстового потоку та розробити метрики ефективності функціонування КЛС для до розв’язку різних задач NLP. Розроблені методи та засоби дають можливість будувати КЛС опрацювання

україномовного текстового контенту згідно потреб постійної/потенційної цільової аудиторії на основі аналізу історії дій користувачів веб-сайту.

4. Удосконалено методи NLP на основі регулярних виразів узгодження з шаблонами, що дало змогу адаптувати методи токенізації та нормалізації тексту каскадами простих підстановок регулярних виразів та кінцевих автоматів.

5. Удосконалено метод МА україномовного тексту на основі сегментації та нормування слова, сегментації речення та модифікованого алгоритму стемінгу Портера як ефективного засобу ідентифікації афіксів лем для можливості розмічування аналізованого слова, що дало змогу підвищити точність пошуку ключових слів на 9%.

6. Удосконалено ІТ інтелектуального аналізу текстового потоку на основі опрацювання інформаційних ресурсів, що дало змогу адаптувати загально типову структуру модулів інтеграції, управління та супроводу контенту для розв’язку різних задач NLP та підвищити ефективність функціонування КЛС на 6-9%. Це стало можливим завдяки поєднанню адаптованих до української мови методів лінгвістичного аналізу, вдосконаленої ІТ опрацювання інформаційних ресурсів, МН та множини метрик оцінювання ефективності функціонування КЛС. Основний принцип побудови таких КЛС полягає на модульності, що полегшує їх побудову згідно вимог щодо наявності відповідних процесів для розв’язку конкретної задачі NLP.

7. Розроблено метод визначення автора в україномовних текстах на основі аналізу коефіцієнтів лексичного авторського мовлення в еталонному уривку авторського тексту, який ґрунтується на аналізі колекції ключових слів, стійких словосполучень, показників лінгвометрії, стилеметрії, а також результатів аналізу N-грам на основі порівнянь різниць вживання 2-грам та 3-грам для подібних за стилем публікацій в межах [6;7]%, а для точно не подібних – >12%), що забезпечило можливість визначити множину потенційних авторів публікацій з більш ніж одного автора (до [9;34]% із загальної кількості учасників проекту) та розробити метод ідентифікації авторського стилю.

8. Розроблено метод визначення стійких словосполучень на основі ідентифікації ключових слів україномовного тексту та аналізу коефіцієнтів лексичного мовлення автора тексту в еталонних уривках контенту, що дало можливість на основі статистичної лінгвістики покращити точність методу визначення стилю автора тексту на 9%.

9. Достовірність наукових та практичних результатів підтверджується відповідними матеріалами про впровадження дисертаційних досліджень, а також за рахунок порівняння одержаних практичних результатів на різних вибірках достовірних вхідних даних. Розроблено КЛС на інформаційному ресурсі http://victana.lviv.ua засобами CMS Joomla! (для розроблення е-каркасу статей), PHP (для реалізації методів опрацювання текстового контенту), HTML (для реалізації розмітки сторінок), CSS (для опису стилів сторінок), MySQL (для зберігання даних та словників). Експериментальне дослідження підтвердило достовірність методу визначення ключових слів – для різних алгоритмів опрацювання первинного тексту середній збіг списків виявлених ключовиків з авторськими змінюється у проміжку 52,6-68,5%. Точність збігу ключових слів із авторськими коливається в проміжку 43,6-62,9%. Середній збіг змістовних ключових слів порівняно зі всіма знайденими системою коливається в проміжку 38,9-75,8% в залежності від етапів аналізу текстів статей. Точність збігу ключових слів порівняно зі всіма знайденими системою коливається в проміжку 34,3-71,9% в залежності від етапів аналізу текстів статей.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Неретин О. П. Глобализация и информатизация как факторы становления современного культурного пространства. Вестник КазГУКИ. 2012. №1. URL: https://cyberleninka.ru/article/n/globalizatsiya-i-informatizatsiya-kak-faktory-stanovleniya-sovremennogo-kultumogo-prostranstva.
2. Цитульский А. М., Иванников А. В., Рогов И. С. NLP - обработка естественных языков. StudNet. 2020. №6. URL: https://cyberleninka.ru/article/n/nlp-obrabotka-estestvennyh-yazykov.
3. The free dictionary by Farlex. Linguistic System, https://encyclopedia2.thefreedictionary.com/Linguistic+System
4. De Saussure F. Course in general linguistics. Columbia University Press, 2011, https://www.academia.edu/download/59483321/Literary_Theory_-_An_Anthology_Blackwell20190601-94544-fbkrbp.pdf#page=78
5. Von Humboldt W., von Humboldt W. F. Humboldt: 'On Language': On the Diversity of Human Language Construction and Its Influence on the Mental Development of the Human Species. Cambridge, 1999. URL: https://books.google.com.ua/books?hl=uk&lr=&id=_UODbGlD4WUC&oi=fnd&pg=PR7&dq=W.+von+Humboldt&ots=liuOK2dUQb&sig=SL9GEwLRuko9gQgRoBDH8WfbqX0&redir_esc=y#v=onepage&q=W.%20von%20Humboldt&f=false
6. Baudouin de Courtenay J. Kilka ogólników o obiektywnej i subiektywnej odrębności „Ukrainy" pod względem językowym, narodowym i państwowym. URL: http://www.plansprachen.ch/Baudouin_de_Courtenay_Propaganda.pdf
7. Yamada M., Fujita A., Yamamoto M., Miyata R., Onish N., Kageura K. Metalanguage for the translation process. Translation in Transition, 46. 2020. URL: https://devrobgilb.com/Files/TT5_Oct_2020_BookOfAbstracts.pdf#page=51
8. Galitsky B. A. Using Extended Tree Kernel to Recognize Metalanguage in Text. Uncertainty Modeling. Springer, Cham. 2017. P. 71-96.
9. Glottopedia. Linguistic information system. URL: http://www.glottopedia.org/index.php/Linguistic_information_system
10. Lamb Sydney M. Linguistic and Cognitive Networks. 1969. URL: https://files.eric.ed.gov/fulltext/ED031694.pdf
11. Lenhart Schubert. Computational linguistics. Stanford Encyclopedia of Philosophy. URL: https://plato.stanford.edu/entries/computational-linguistics/
12. Гольдберг Й. Нейросетевые методы в обработке естественного языка : руководство. М: ДМК Пресс, 2019. 282 с. ISBN 978-5-97060-754-1
13. Досин Д.Г. Методологічні засади розроблення інтелектуальних інформаційно-пошукових систем на основі визначення корисності знань : дисертація на здобуття наукового ступеня доктора технічних наук : 05.13.06 – "Інформаційні технології" / Дмитро Григорович Досин ; Українська академія друкарства. Львів, 2021. 391 с.
14. Величко В.Ю. Науково-технологічні основи знання-орієнтованої обробки природномовних текстів та її застосування : автореферат дисертації на здобуття наукового ступеня доктора технічних наук : 05.13.06 – "Інформаційні технології" / Величко Віталій Юрійович: Інституті кібернетики імені В.М. Глушкова НАН Інституті кібернетики імені В.М. Глушкова НАН. Київ, 2021. 46 с.
15. Досин Д.Г. Методологічні засади розроблення інтелектуальних інформаційно-пошукових систем на основі визначення корисності знань : автореферат дисертації на здобуття наукового ступеня доктора технічних наук : 05.13.06 – "Інформаційні технології" / Дмитро Григорович Досин ; Українська академія друкарства. Львів, 2021. 45 с.
16. Feduhko S. Development of a software for computer-linguistic verification of socio-demographic profile of web-community member. Webology, 2014. Vol. 11(2). URL: http://www.webology.org/2014/v11n2/a126.pdf

17. Demydov I. Architecture of the Computer-linguistic System for Processing of Specialized Web-communities' Educational Content. 2020. URL: http://ceur-ws.org/Vol-2616/paper1.pdf
18. DeKeyser R. M. Learning second language grammar rules: An experiment with a miniature linguistic system. Studies in second language acquisition, 1995. Vol. 17(3), P. 379-410.
19. Bisikalo O., Vysotska V. Linguistic analysis method of Ukrainian commercial textual content for data mining. CEUR Workshop Proceedings. 2020. Vol. 2608. P. 224–244.
20. Lytvyn V., Pukach P., Vysotska,V., Vovk M., Kholodna N. Identification and Correction of Grammatical Errors in Ukrainian Texts Based on Machine Learning Technology. Mathematics 2023. Vol. 11. 904. ISSN 2227-7390.
21. Bisikalo O., Vysotska V., Burov Y., Kravets P. Conceptual model of process formation for the semantics of sentence in natural language. CEUR Workshop Proceedings. 2020. Vol. 2604. P. 151–177.
22. Bisikalo O., Vysotska V., Lytvyn V., Brodyak O., Vyshemyrska S., Rozov Y. Experimental investigation of significant keywords search in Ukrainian content. Advances in Intelligent Systems and Computing. 2021. Vol. 1293. P. 3–29.
23. Chyrun L., Kis Ia., Vysotska V., Chyrun L. Content monitoring method for cut formation of person psychological state in social scoring. CSIT-2018, 11–14 вересня 2018 р., Львів. 2018. Т. 2. С. 106–112.
24. Liu L. Foreign Linguistic System Construction Based on Computer-aided Technology. In Journal of Physics: Conference Series. 2020, April. Vol. 1533, No. 3, p. 032018. IOP Publishing.
25. Herzog O. Text understanding in LILOG: integrating computational linguistics and artificial intelligence: final report on the IBM Germany LILOG-Project. C. R. Rollinger (Ed.). Berlin: Springer. 1991.
26. Paris C. L., Swartout W. R., Mann W. C. Natural language generation in artificial intelligence and computational linguistics. Vol. 119. Springer Science & Business Media. 2013.
27. McCarthy P., Chutima Boonthum-Denecke. Applied natural language processing. Information Science Reference, 2011.
28. Vajjala S. Machine Learning and Applied Linguistics. arXiv preprint arXiv:1803.09103. 2018.
29. Meurers D. Natural language processing and language learning. Encyclopedia of applied linguistics, 2012. P. 4193-4205.
30. Pazienza M. T., Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology. Vol. 1299. Springer, 2006.
31. Jones Karen Sparck. How much has information technology contributed to linguistics? arXiv preprint cmp-lg/9702011. 1997.
32. Chowdhury G. Natural language processing. Annual review of information science and technology. Vol 37.1. 2003. P. 51-89.
33. Reshamwala Alpa, Dhirendra Mishra, Prajakta Pawar. Review on natural language processing. IRACST Engineering Science and Technology: An International Journal (ESTIJ). Vol. 3.1. 2013. P. 113-116.
34. Jusoh Shaidah. A study on nlp applications and ambiguity problems. Journal of Theoretical & Applied Information Technology. Vol. 96.6. 2018.
35. Alieksieieva K., Berko A., Vysotska V. Technology of commercial web-resource processing. CADSM, 24–27 лют. 2015, Львів, Поляна. Львів, 2015. С. 340–344.
36. Andrunyk V., Chyrun L., Vysotska L. Electronic content commerce system development. CADSM., 24–27 лют. 2015, Львів, Поляна. Львів, 2015. С. 434–438.
37. Batiuk T., Vysotska V., Lytvyn V. Intelligent system for socialization by personal interests on the basis of SEO technologies and methods of machine learning. CEUR Workshop Proceedings. 2020. Vol. 2604. P. 1237–1250.

38. Berko A., Vysotska V., Lytvyn V., Naum O. Planning the activities of intellectual agents in the electronic commerce systems. Радіоелектроніка. Інформатика. Управління. 2018. № 4. С. 143–158.
39. Bublyk M., Lytvyn V., Vysotska V., Sokulska N., Chyrun L., Matseliukh Y. The decision tree usage for the results analysis of the psychophysiological testing. CEUR Workshop Proceedings. 2020. Vol. 2753. P. 458–472.
40. Bublyk M., Vysotska V., Panasyuk V., Brodyak O., Chyrun L. Assessing security risks method in e-commerce system for IT portfolio management. CEUR Workshop Proceedings. 2021. Vol. 2853. P. 462–479.
41. Bisikalo, O., Danylchuk, O., Kovtun, V., Kovtun O., Nikitenko O., Vysotska V. Modeling of Operation of Information System for Critical Use in the Conditions of Influence of a Complex Certain Negative Factor. International Journal of Control, Automation and Systems. 2022. Vol. 20. P. 904–1913.
42. Chyrun L., Leshchynskyy E., Lytvyn V., Rzheuskyi A., Vysotska V., Borzov Y. Intellectual analysis of making decisions tree in information systems of screening observation for immunological patients. CEUR Workshop Proceedings. 2019. Vol. 2488. Vol. 1. P. 281–296.
43. Дьомкін В. Анонс lang-uk: створюємо умови для повноцінної обробки україномовних текстів. URL: https://dou.ua/lenta/columns/lang-uk/
44. Texty.org.ua. Волонтери створюють умови для повноцінної обробки україномовних текстів. URL: https://texty.org.ua/fragments/68164/Volontery_stvorut_umovy_dla_povnocinnoji_obrobky_ukrajinomovnyh-68164/
45. Кунанець Н., Козак І. Інтелектуальна система опрацювання природномовних текстів українською мовою. URL: http://ena.lp.edu.ua:8080/bitstream/ntb/38388/1/104_218-219.pdf
46. Лозицький О. А. Прикладна програмна система опрацювання україномовних технічних текстів для людей з вадами зору. SISN. 2015; Випуск 832(3). С. 315–331. URL: http://science.lpnu.ua/sites/default/files/journal-paper/2018/jun/12943/22-315-331.pdf
47. Кулинський О. С. Автоматизоване опрацювання природомовних текстів з використанням засобів штучного інтелекту. Науковий вісник НЛТУ України . 2011. №6. URL: https://cyberleninka.ru/article/n/avtomatizovane-opratsyuvannya-prirodomovnih-tekstiv-z-vikoristannyam-zasobiv-shtuchnogo-intelektu.
48. Chyrun L., Vysotska V., Chyrun L., Gozhyj A., Kalinina I. SEO technology for web resource processing. COLINS 2018. Workshop. P. 40–52.
49. Chyrun L., Vysotska V., Lytvyn V. Specifics informational resources processing for textual content linguistic analysis. MEMSTECH 2016, 20–24 Apr., 2016, Lviv, Polyana, Ukraine, 2016. P. 214–219.
50. Інструменти для роботи з текстом. URL: https://studway.com.ua/robota-z-tekstom/
51. Romanyshyn M. Intro to Natural Language Processing. Grammarly, Inc., 2017.
52. Romanyshyn, M.: Grammatical Error Correction: why commas matter. COLINS, 2017. URL: http://colins.in.ua/wp-content/uploads/2017/04/Grammatical-Error-Correction-whycommas- matter.pdf
53. Skopyk, K.: Language modelling and its use cases. In: Computational Linguistics and Intelligent Systems, COLINS, 2018. Vol. 2. P. 1-11. URL: http://colins.in.ua/wp-content/uploads/2018/07/Languagemodelling-and-its-use-cases.pdf
54. Kravets P., Burov Y., Oborska O., Vysotska V., Dzyubyk L., Lytvyn V. Stochastic Game Model of Data Clustering. CEUR Workshop Proceedings. 2021. Vol. 2853. P. 198–213.
55. Kovalchuk V., Lytvyn V., Vysotska V., Hrendus M., Naum O. The information system for identification of content set based on analysis of similar texts. COLINS 2018. Workshop. P. 122–127.
56. Kanishcheva O., Vysotska V., Chyrun L., Gozhyj A. Method of integration and content management of the information resources network. Advances in Intelligent Systems and Computing (AISC). 2018. Vol. 689. P. 204–216.

57. Gozhyj A., Kalinina I., Vysotska V., Gozhyj V. The method of web-resources management under conditions of uncertainty based on fuzzy logic. CSIT-2018 (Львів,11–14 вересня 2018 р.). 2018. Т. 1. С. 343–346.
58. Gozhyj A., Vysotska V., Yevseyeva I., I Kalinina., Gozhyj V. Web resources management method based on intelligent technologies. Advances in Intelligent Systems and Computing (AISC). 2019. Vol. 871. P. 206–221.
59. Lytvyn V., Vysotska V., Shatskykh V., Kohut I., Petruchenko O., Dzyubyk L., Bobrivetc V., Panasyuk V., Sachenko S., Komar M. Design of a recommendation system based on collaborative filtering and machine learning considering personal needs of the user. Eastern-European Journal of Enterprise Technologies. 2019. № 4/2 (100). С. 6–28.
60. Balush I., Vysotska V., Albota S. Recommendation System Development Based on Intelligent Search, NLP and Machine Learning Methods. CEUR Workshop Proceedings. 2021. Vol. 2917. P. 584-617. E-ISSN: 1613-0073.
61. Kholodna N., Vysotska V., Albota S. A Machine Learning Model for Automatic Emotion Detection from Speech. CEUR Workshop Proceedings. 2021. Vol. 2917. P. 699-713. E-ISSN: 1613-0073.
62. Husak V., Lozynska O., Karpov I., Peleshchak I., Chyrun S., Vysotskyi A.: Information System for Recommendation List Formation of Clothes Style Image Selection According to User's Needs Based on NLP and Chatbots. CEUR workshop proceedings. 2020. Vol. 2604, P. 788-818.
63. Meleshko Y., Yakymenko M., Semenov S. A Method of Detecting Bot Networks Based on Graph Clustering in the Recommendation System of Social Network. CEUR Workshop Proceedings. 2021. Vol. 2870. P. 1249-1261.
64. Artemenko O., Pasichnyk V., Kunanets N., Shunevych K. Using sentiment text analysis of user reviews in social media for e-tourism mobile recommender systems. CEUR workshop proceedings. 2020. Vol. 2604. P. 259-271.
65. Makara S., Chyrun L., Burov Y., Rybchak Z., Peleshchak I., Peleshchak R., Holoshchuk R., Kubinska S., Dmytriv A. An Intelligent System for Generating End-User Symptom Recommendations Based on Machine Learning Technology. CEUR workshop proceedings. 2020. Vol. 2604. P. 844-883.
66. Boyko N., Telishevskyi P., Kushka B. Analysis of Recommendation System Methods for Accuracy of Predicted Estimates. CEUR Workshop Proceedings. 2021. Vol. 2870. P. 1878-1888.
67. Shakhovska N., Fedushko S., Greguš ml. M., Shvorob I., Syerova Yu. Development of Mobile System for Medical Recommendations. The 15th International Conference on Mobile Systems and Pervasive Computing (MobiSPC). 2019. Vol. 155. P. 43-50.
68. Jurafsky Dan, Martin James H. Speech and Language Processing. URL: https://web.stanford.edu/~jurafsky/slp3/ed3book_sep212021.pdf
69. Weizenbaum J. ELIZA –A computer program for the study of natural language communication between man and machine. CACM. 1966. Vol. 9. P. 36–45.
70. Weizenbaum J. Computer Power and Human Reason: From Judgement to Calculation. W.H. Freeman and Company. 1976.
71. ElizaBot. URL: https://www.masswerk.at/elizabot/
72. ELIZA: a very basic Rogerian psychotherapist chatbot. URL: https://web.njit.edu/~ronkowit/eliza.html
73. Dan Jurafsky, James H. Martin. Regular Expressions, Text Normalization, Edit Distance. URL: https://web.stanford.edu/~jurafsky/slp3/2.pdf
74. Watson Personality Insights. URL: https://dataplatform.cloud.ibm.com/docs/content/wsj/landings/personality-insights.html
75. IBM Personality Insights. URL: https://www.symanto.com/ibm-personality-insights
76. Balogh Z. Analysis of public data on social networks with IBM Watson. Sciences, 2018. Vol. 12.12. P. 455-460.

77. McGetrick C. Investigation into the Application of Personality Insights and Language Tone Analysis in Spam Classification. 2017. URL: https://arrow.tudublin.ie/scschcomdis/120/
78. Hrazdil K., Mahmoudian F., Nazari J. A. Executive personality and sustainability: Do extraverted chief executive officers improve corporate social responsibility? Corporate Social Responsibility and Environmental Management. 2021. Vol. 28.6. P. 1564-1578.
79. Zhang H. CrossCheck: an effective tool for detecting plagiarism. Learned publishing. 2010. Vol. 23(1). P. 9-14.
80. Lin W.-Y. C. Self-plagiarism in academic journal articles: From the perspectives of international editors-in-chief in editorial and COPE case. Scientometrics. 2020. Vol. 123(1). P. 299-319.
81. Swapna M., Manish G. S., Rachana B. Automatic Text Summarization Using NLTK. Think India Journal. 2019. Vol. 22(35). P. 828-833.
82. Karimov R. N., Samedov F. R., Yunisov J. K. Rewriting, academic fraud and falsification: analyzing cases of academic practice in St. Petersburg, Russia. Application of Information and Communication Technologies, AICT, Baku; Azerbaijan; 12-14 October 2016. Institute of Electrical and Electronics Engineers Inc. P. 774-778.
83. Gregory A., Leeman J. On the Perception of Plagiarism in Academia: Context and Intent. arXiv preprint arXiv:2104.00574, 2021. URL: https://arxiv.org/abs/2104.00574
84. Lytvyn V., Vysotska V., Burov Ye., Bobyk I., O Ohirko. The linguometric approach for co-authoring author's style definition. IDAACS-SWS 2018, Lviv, 20–21 September 2018. P. 29–34.
85. Prasad A., Jyothi P. How accents confound: Probing for accent information in end-to-end speech recognition systems. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020. P. 3739-3753.
86. Lytvyn V., Vysotska V., Burov Y., Demchuk A. Defining author's style for plagiarism detection in academic environment. Data stream mining and processing, August 21–25, 2018, Lviv, Ukraine. 2018. P. 128–133.
87. Vysotska V., Kanishcheva O., Hlavcheva Y. Authorship identification of the scientific text in Ukrainian with using the lingvometry methods. CSIT-2018, 11–14 вересня 2018 р., Львів. 2018. Т. 2. С. 34–38.
88. Stuart L. M., Tazhibayeva S., Wagoner A. R., Taylor J. M. On identifying authors with style. 2013 IEEE International Conference on Systems, Man, and Cybernetics. 2013. P. 3048-3053.
89. Gonzalez-Lopez J. A., Gomez-Alanis A., Doñas J. M. M., Pérez-Córdoba J. L., Gomez A. M. Silent speech interfaces for speech restoration: A review. IEEE Access. 2020. Vol. 8. P. 177995-178021.
90. Rodríguez-Tapia B., Soto I., Martínez D. M., Arballo N. C. Myoelectric interfaces and related applications: Current state of EMG signal processing–A systematic review. IEEE Access. 2020. Vol. 8. P. 7792-7805.
91. Zhang S., Zang J., Zhang X., Chen H., Mikami B., Zhao G. "Silent" amino acid residues at key subunit interfaces regulate the geometry of protein nanocages. ACS nano. 2016. Vol. 10(11). P. 10382-10388.
92. Wadhawan A., Kumar P. Sign language recognition systems: A decade systematic literature review. Archives of Computational Methods in Engineering. 2021. Vol. 28(3). P. 785-813.
93. Al-Ahdal M. E., Nooritawati Md T. Review in sign language recognition systems. IEEE Symposium on Computers & Informatics (ISCI). 2012. P. 52-57.
94. Davydov M., Nikolski I., Pasichnyk V. Real-time Ukrainian sign language recognition system. International Conference on Intelligent Computing and Intelligent Systems. 2010. P. 875-879.
95. Davydov M., Lozynska O. Information system for translation into Ukrainian sign language on mobile devices. CSIT. 2017. P. 48-51.

96. Davydov M., Lozynska O. Mathematical method of translation into Ukrainian sign language based on ontologies. Conference on Computer Science and Information Technologies. Springer, Cham. 2017. P. 89-100.
97. Davydov M., Lozynska O. Linguistic models of assistive computer technologies for cognition and communication. CSIT. IEEE. 2016. P. 171-174.
98. Lozynska O., Davydov M. Information technology for Ukrainian Sign Language translation based on ontologies. Econtechmod. 2015. Vol. 4(2). P. 13-18.
99. Davydov M., Nikolski I., Pasichnyk O. System of finger movement identification for sign language recognition. Central European Student Conference in Linguistics. 2006. P. 29-31.
100. Hawking S. URL: https://scholar.google.com.ua/citations?user=-AEEg5AAAAAJ&hl=uk&oi=ao
101. Kewley-Port D., M. Nearey T. Speech synthesizer produced voices for disabled, including Stephen Hawking. The Journal of the Acoustical Society of America. 2020. Vol. 148(1). R1-R2.
102. Hawking S. Dirac Memorial. Paul Dirac: The man and his work, 2005.
103. Lytvyn V., Vysotska V., Osypov M., Slyusarchuk O., Slyusarchuk Y. Development of intellectual system for data de-duplication and distribution in cloud storage. Webology. 2019. Vol. 16(2). P. 1-42.
104. Wu X., Wang H., Wei D., Shi M. ANFIS with natural language processing and gray relational analysis based cloud computing framework for real time energy efficient resource allocation. Computer communications. 2020. Vol. 150. P. 122-130.
105. Christoph J., Griebel L., Leb I., Engel I., Köpcke F., Toddenroth D., Prokosch H.-U., Laufer J., Marquardt K., Sedlmayr M. Secure secondary use of clinical data with cloud-based NLP services. Methods of information in medicine. 2015. Vol. 54(03). P. 276-282.
106. Ghorbani M., Bahaghighat M., Xin Q., Özen, F. ConvLSTMConv network: a deep learning approach for sentiment analysis in cloud computing. Journal of Cloud Computing. 2020. Vol. 9(1). P. 1-12.
107. Papanikolaou N., Pearson S., Mont M. C., Ko R. K. A toolkit for automating compliance in cloud computing services. International Journal of Cloud Computing. 2014. Vol. 3(1). P. 45-68.
108. Strubell E., Ganesh A., McCallum A. Energy and policy considerations for deep learning in NLP. arXiv preprint arXiv:1906.02243. 2019. URL: https://arxiv.org/abs/1906.02243
109. Du Z., Qian Y., Liu X., Ding M., Qiu J., Yang Z., Tang, J. All NLP tasks are generation tasks: A general pretraining framework. arXiv preprint arXiv:2103.10360. 2021.
110. Park K., Lee J., Jang S., Jung D. An empirical study of tokenization strategies for various Korean NLP tasks. arXiv preprint arXiv:2010.02534. 2020.
111. Lewis P., et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems. 2020. Vol. 33. P. 9459-9474.
112. Oliinyk V., Vysotska V., Burov Y., Mykich K., Basto-fernandes V. Propaganda detection in text data based on NLP and machine learning. CEUR Workshop Proceedings. 2020. Vol. 2631. P. 132–144. E-ISSN: 1613-0073.
113. Shu C., Dosyn D., Lytvyn V., Vysotska V., Sachenko A., Jun S. Building of the predicate recognition system for the NLP ontology learning module. IDAACS, September 18–21, 2019, Metz, France. 2019. P. 802–808.
114. Lytvyn V., Vysotska V., Uhryn D., Hrendus M., Naum O. Analysis of statistical methods for stable combinations determination of keywords identification. Eastern-European Journal of Enterprise Technologies. 2018. № 2/2. P. 23–37.
115. Allen J. F. Natural language processing. Encyclopedia of computer science. 2003. P. 1218-1222.
116. Hirschberg J., Manning C. D. Advances in natural language processing. Science. 2015. Vol. 349(6245). P. 261-266.

117. Narendra L. W., Setyaningsih E. R. Designing a Transactional Smart Assistant in Indonesian using Rasa Framework. 7th International Conference on Electrical, Electronics and Information Engineering (ICEEIE). IEEE. 2021. P. 1-6.
118. Miertschin E., et al. Smart Assistant Guided Flowback Data Analysis. SPE Hydraulic Fracturing Technology Conference and Exhibition. OnePetro, 2020.
119. Manchanda S. Automation of Bid Proposal Preparation Through AI Smart Assistant. In Data Management, Analytics and Innovation. Springer, Singapore. 2021. P. 45-57.
120. Sarakhman K., Kempnyk R., Chyhura V. ChatBot using NLP. Computational linguistics and intelligent systems: proceedings of the 4nd International conference. 2020. P. 429-432.
121. Nakazawa T., et al. Example-based machine translation based on deeper NLP. Third International Workshop on Spoken Language Translation: Evaluation Campaign. 2006.
122. Khan N. S., Abid A., Abid K. A novel natural language processing (NLP) based machine translation model for English to Pakistan sign language translation. Cognitive Computation. 2020. Vol. 12(4). P. 748-765.
123. Kumar S., Anastasopoulos,A., Wintner S., Tsvetkov Y. Machine translation into low-resource language varieties. arXiv preprint arXiv:2106.06797. 2021/
124. Grabar N., Kanishcheva O., Hamon T. Multilingual aligned corpus with Ukrainian as the target language. SLAVICORP. 2018.
125. Grabar N., Hamon T. Creation of a multilingual aligned corpus with Ukrainian as the target language and its exploitation. Computational linguistics and intelligent systems (COLINS 2017). National Technical University «KhPI», 2017.
126. Cherednichenko O., Kanishcheva O., Yakovleva O., Arkatov D. Collection and Processing of a Medical Corpus in Ukrainian. COLINS. 2020. P. 272-282.
127. Khaburska A., Tytyk I. Toward Language Modeling for the Ukrainian. Advances in Data Mining, Machine Learning, and Computer Vision. Proceedings. 2019. P. 71.
128. Sharoff S. Toward Pan-Slavic NLP: Some Experiments with Language Adaptation. The 6th Workshop on Balto-Slavic Natural Language Processing. 2017. P. 1-2.
129. Olkhovska A., Frolova I. Using Machine Translation Engines in the Classroom: A Survey of Translation Students' Performance. Advanced Education. 2020. Vol. 15. P. 47-55.
130. Su J., Lytvyn V., Vysotska V., Sachenko A., Dosyn D. Model of touristic information resources integration according to user needs. CSIT-2018, 11–14 вересня 2018 р., Львів. 2018. Т. 2. С. 113–116.
131. Mishra A.J., Sanjay K. A survey on question answering systems with classification. Journal of King Saud University-Computer and Information Sciences. 2016. Vol. 28(3). P. 345-361.
132. Lytvyn V., Gozhyj A., Kalinina I., Vysotska V., Shatskykh V., Chyrun L., Borzov Yu. An intelligent system of the content relevance at the example of films according to user needs. CEUR Workshop Proceedings. 2019. Vol. 2516. P. 1–23.
133. Loper E., Bird S. NLTK: The natural language toolkit. arXiv preprint cs/0205028, 2002.
134. Sarker I. H., Hoque M. M., Uddin M., Alsanoosy, T. Mobile data science and intelligent apps: concepts, ai-based modeling and research directions. Mobile Networks and Applications. 2021. Vol. 26(1). P. 285-303.
135. Lytvyn V., Vysotska V., Shakhovska N., Mykhailyshyn V., Medykovskyy M., Peleshchak I., Fernandes V. B., Peleshchak R., Shcherbak S. A smart home system development. Advances in Intelligent Systems and Computing (AISC). 2020. Vol. 1080. P. 804–830.

136. Lytvyn V., Burov Y., Kravets P., Vysotska V., Demchuk A., Berko A., Ryshkovets Y., Shcherbak S., Naum O. Methods and models of intellectual processing of texts for building ontologies of software for medical terms identification in content classification. CEUR Workshop Proceedings. 2019. Vol. 2488. P. 354–368. E-ISSN: 1613-0073.
137. Rzheuskyi A., Kutyuk O., Vysotska V., Burov Y., Lytvyn V., Chyrun L. The architecture of distant competencies analyzing system for IT recruitment. CSIT (Львів, 17–20 вересня 2019 р.). 2019. Т. 3. С. 254–261.
138. Lytvyn V., Kowalska-Styczen A., Peleshko D., Rak T., Voloshyn V., Noennig J. R., Vysotska V., Nykolyshyn L., Pryshchepa H. Aviation aircraft planning system project development. Advances in Intelligent Systems and Computing. 2020. Vol. 1080. P. 315–348.
139. Lytvyn V., Dosyn D., Vysotska V., Demchuk A., Demkiv L., Lytvyn I. Intellectual agent construction method based on the subject field ontology. CSIT. 2020. P. 40-46.
140. Chyrun L., Vysotska V., Kis Ia., Chyrun L. Content analysis method for cut formation of human psychological stat. Data stream mining and processing, August 21–25, 2018, Lviv, Ukraine. 2018. P. 139–144.
141. Lytvyn V., Vysotska V., Peleshchak I., Basyuk T., Kovalchuk V., Kubinska S., Chyrun L., Rusyn B., Pohreliuk L., Salo T. Identifying textual content based on thematic analysis of similar texts in big data. CSIT. Львів, 17–20 вересня 2019. Т. 2. С. 84–91.
142. Su J., Vysotska V., Sachenko A., Lytvyn V., Burov Ye. Information resources processing using linguistic analysis of textual content. IDAACS., Bucharest, Sept. 21–23, 2017. 2017. P. 573–578.
143. Vysotska V., Lytvyn V., Bublyk M., Demchuk A., Demkiv L., Shpak Y. Method of ontology quality assessment for knowledge base in intellectual systems based on ISO\IEC 25012. CSIT, Збараж, 23–26 вересня, 2020. P. 109–113.
144. Lytvyn V., Vysotska V., Demchuk A., Demkiv I., Ukhans'ka O., Hladun V. R., Kovalchuk R., Petruchenko O., Dzyubyk L., Sokulska N. Design of the architecture of an intelligent system for distributing commercial content in the internet space based on SEO-technologies, neural networks, and machine learning. Eastern-European Journal of Enterprise Technologies. 2019. № 2/2 (98). С. 15–34.
145. Vysotska V., Berko A., Bublyk M., Chyrun L., Vysotsky A., Doroshkevych K. Methods and tools for web resources processing in e-commercial content systems. CSIT, Збараж, 23–26 вересня, 2020. P. 114–118.
146. Burov Y., Lytvyn V., Vysotska V., Shakleina I. The basic ontology development process automation based on text resources analysis. CSIT, Збараж, 23–26 вересня, 2020. P. 280–284.
147. Lytvyn V., Vysotska V., Bublyk M., Nanivskyi R., Grudowski P., Matseliukh Y. Developing methods for building intelligent systems of information resources processing using an ontological approach. Advances in Intelligent Systems and Computing. 2021. Vol. 1293. P. 345–370.
148. Dejong G. F. Raymond Joseph Mooney. Genome Biology. 2005. Vol. 6(5). P. r40.
149. Falessi D., Layman L. Automated classification of NASA anomalies using natural language processing techniques. IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW). 2013. P. 5-6.
150. Lally A., Fodor P. Natural language processing with prolog in the IBM Watson system. The Association for Logic Programming (ALP) Newsletter. 2011. A. 9.
151. Holzinger A., et al. Combining HCI, natural language processing, and knowledge discovery-potential of IBM content analytics as an assistive technology in the biomedical field. International Workshop on Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data. Springer, Berlin, Heidelberg. 2013. P. 13-24.

152. Wang X., Chen X., Li H., Zhang F., Zhao X., Han Y., Wang X., Hao, Y. Genome-wide identification and expression pattern analysis of NLP (Nin-like protein) transcription factor gene family in apple. Scientia Agricultura Sinica. 2019. Vol. 52(23). P. 4333-4349.
153. Hitzler P. A review of the semantic web field. Communications of the ACM. 2021. Vol. 64(2). P. 76-83.
154. Wang K., Thrasher C., Viegas E., Li X., Hsu B. J. P. An overview of Microsoft Web N-gram corpus and applications. Proceedings of the NAACL HLT. 2010. P. 45-48.
155. Richardson S. The evolution of an NLP System. NLP Group Microsoft Research, Presentation at the LREC. 2000.
156. Guerini M., Strapparava C., Stock O. Evaluation metrics for persuasive NLP with google AdWords. International Conference on Language Resources and Evaluation (LREC'10). 2010.
157. Sosnina E. Yandex Parsing Software in a Data Extraction Project. Interactive Systems: Problems of Human-Computer Interaction. 2015. P. 98-101.
158. Polyakov V., et al. Bigrams and chunking: Advantages for using in automatic spelling correction in Russian and English. International Journal. 2017. Vol. 73(10). P. 108-120.
159. Azunre P. Transfer learning for natural language processing. Simon and Schuster. 2021.
160. Lytvyn V., Vysotska V., Dosyn D., Holoschuk R., Rybchak Z. Application of sentence parsing for determining keywords in Ukrainian texts. CSIT, 5–8 Sept., 2016, Lviv, Ukraine. Lviv: Publishing Lviv Polytechnic. 2017. P. 326–331.
161. Бойчук М. Ontology Parsing Ukrainian Language. Technology Audit and Production Reserves. 2012. Vol. 5. P. 13-14.
162. Dmytriv A., Vysotska V., Bublyk M. The Speech Parts Identification for Ukrainian Words Based on VESUM and Horokh Using. CSIT, 22-25 Sept., Lviv, Ukraine. 2021. – Vol. 2. P. 21–33.
163. Lytvyn V., Vysotska V., Maria H. Method of data expression from the Ukrainian content based on the ontological approach. Радіоелектроніка. Інформатика. Управління. 2018. № 3 (46). P. 144–157.
164. Teletska A. Automatic parsing system of user stories. Topical Issues of Humanities, Technical and Natural Sciences. 2019. P. 253-255.
165. Davydov M., Lozynska O., Pasichnyk V. Effective algorithm for parsing sentences using semantically attributed weighted affix context free. Радіоелектроніка, інформатика, управління. 2017. Vol. 4 (43). P. 124-130.
166. Darchuk N. Automatic parsing of texts of the corpus of the Ukrainian language. Ukrainian linguistics. 2013. Vol. 43. P. 11-19.
167. Kotsyba N., Moskalevskyi B. Using transitivity information for morphological and syntactic disambiguation of pronouns in Ukrainian. Вісник Національного університету "Львівська політехніка". 2019. Vol. 5. P. 101-115.
168. Kudin A. Implicit Agents in Ukrainian: Evidence from Retrieval Interference in Sentence Processing. University of California, Santa Cruz, 2021.
169. Senyk M. The Porter Stemming Algorithm for Ukrainian. URL: http://www.senyk.poltava.ua/projects/ukr_stemming/stemming_about.html
170. Moseichuk V. Porter stemming algorithm for Ukrainian languages. 2020.
171. Berko A., Matseliukh Y., Ivaniv Y., Chyrun L., Schuchmann V. The Text Classification Based on Big Data Analysis for Keyword Definition Using Stemming. CSIT 2021. Vol. 1. P. 184-188.
172. Golub T. V., Zeleneva I. Y., Parkhomenko A. V., Hrushko S. S. Stemming algorithm modification for acceleration of Ukrainian texts processing. Інформатика, кібернетика та обчислювальна техніка. 2020. Vol. 1(30). P. 34-41.
173. Andrusyak B., Rimel M., Kern R. Detection of Abusive Speech for Mixed Sociolects of Russian and Ukrainian Languages. RASLAN. 2018. P. 77-84.

174. Baláž P., Zábojník S., Hričovský M. EU fossil fuel imports and changes after Ukrainian crisis. SHS Web of Conferences. EDP Sciences. 2020. P. 05005.

175. Brykczyñska G. Prisons and prisoners: Some observations, comments and ethical reflections based on a visit to a prison hospital in the Ukrainian Republic. Nursing ethics. 2002. Vol. 9(4). P. 361-372.

176. Сеник Д. А. Использование регулярных выражений в разработке контента правовых баз данных большого объема. Технические науки. 2009. Vol. 6. URL: https://cyberleninka.ru/article/n/ispolzovanie-regulyarnyh-vyrazheniy-v-razrabotke-kontenta-pravovyh-baz-dannyh-bolshogo-obema.

177. Lobur M., Romaniuk A., Romanyshyn M. Defining an approach for deep sentiment analysis of reviews in Ukrainian. Вісник Нац. ун-т "Львів. політехніка". 2012. № 747. С. 124-130.

178. Lobur M., Romanyshyn M., Romaniuk A. Sentiment-annotated corpus of reviews in Ukrainian. Вісник Нац. ун-т "Львів. політехніка". 2012. № 747. С. 131-138.

179. Dmytrash O., Romanyuk A., Melnyk M. Scheme for Named Entities Annotation in Ukrainian Texts. MEMSTECH, 16-20 квітня 2013, Поляна, Україна. 2013. С.176-178.

180. Dmytrash O., Romanyuk A. Annotated Corpus of Named Entities for Ukrainian Language. CADSM 2013, 19-23 лютого 2013, Поляна. 2013. С.80-81.

181. Bekesh, R., et al. Structural Modeling of Technical Text Analysis and Synthesis Processes. COLINS. 2020. P. 562-589.

182. Lytvyn V., Vysotska V., Pukach P., Nytrebych Z., Demkiv I., Senyk A., Malanchuk O, Sachenko S., Kovalchuk R., Huzyk N. Analysis of the developed quantitative method for automatic attribution of scientific and technical text content written in Ukrainian. Eastern-European Journal of Enterprise Technologies. 2018. № 6/2. С. 19–31.

183. Vysotska V., Berko A., Bublyk M., Chyrun L., Vysotsky A., Doroshkevych K. Methods and tools for web resources processing in e-commercial content systems. CSIT, Збараж, 23–26 вересня, 2020. P. 114–118.

184. Kutuzov A., Kopotev M., Sviridenko T., Ivanova L. Clustering comparable corpora of Russian and Ukrainian academic texts: Word embeddings and semantic fingerprints. arXiv preprint arXiv:1604.05372. 2016.

185. Babych B. Graphonological Levenshtein edit distance: Application for automated cognate identification. Baltic Journal of Modern Computing. 2016. Vol. 4(2). P. 115-128.

186. Katrenko S., Adriaans P. Named entity recognition for Ukrainian: a resource-light approach. Proceedings of the Workshop on Balto-Slavonic Natural Language Processing. 2007. P. 88-93.

187. Burdick L., et al. Building a Flexible Knowledge Graph to Capture Real-World Events. TAC. 2019.

188. Buk S., Rovenchak A. Simple Definition of Distances between Texts from Rank frequency Distributions. A Case of Ukrainian Long Prose Works by Ivan Franko. 2019. Vol. 46. P. 1-11.

189. Buk S. Lexical base as a compressed language model of the world (on the material of the Ukrainian language). Psychology of Language and Communication. 2009. Vol. 13(2). P. 35-44.

190. Buk S. Quantitative comparison of texts (on the material of the 1884 and 1907 editions of the novel «Boa Constrictor» by Ivan Franko). Ukrainian Literary Studies 2012. 76. P. 179–192.

191. Buk S., Rovenchak, A. Probing the 'temperature' approach on Ukrainian texts: Long-prose fiction by Ivan Franko. Studies in Quantitative linguistics, 2016. P. 160–175.

192. Kharkevych G., et al. Usage of Fourier transformation theory in machine translation. 2nd International Conference on Advanced Trends in Information Theory (ATIT). 2020. P. 196-199.

193. Irvine A. Statistical machine translation in low resource settings. Proceedings of the 2013 NAACL HLT Student Research Workshop. 2013. P. 54-61.

194. Och F. J., et al. Syntax for statistical machine translation. Johns Hopkins University 2003 Summer Workshop on Language Engineering, Center for Language and Speech Processing, Baltimore, MD, Tech. Rep. 2003.

195. Lozynska O., Davydov M., Pasichnyk V., Veretennikova N. Rule-based machine translation into Ukrainian sign language using concept dictionary. ICTERI. 2019. P. 191-201.

196. Lozynska O., Savchuk V., Pasichnyk V. Individual sign translator component of tourist information system. Conference on Computer Science and Information Technologies. Springer, 2019. P. 593-601.

197. Babych S., Eberle K., Babych Development of hybrid machine translation systems for under-resourced languages: automated creation of lexical and morphological resources for MT. Applied and Literary Translation and Interpreting: Theory, Methodology, Practice, 5.

198. Antonsen L., Trosterud T., Tyers F. M. A North Saami to South Saami machine translation prototype. Northern European Journal of Language Technology. 2016. Vol. 4. P. 11-27.

199. Гайденко Ю. О. Збереження кількісної семантики морфологічних одиниць англійської мови в україномовному перекла. URL: https://ela.kpi.ua/bitstream/123456789/8042/1/Gaydenko.pdf

200. Троцюк М. А. Применение модели дистрибутивной семантики Word2vec к анализу текстовой информации. https://rep.bstu.by/bitstream/handle/data/3170/60-62.pdf?sequence=1

201. Іванова Л. Семантико-синтаксичнаструктура речень із тривалентними дистрибутивними предикатами. Наукові записки. Сер.: Філологічні науки. 2008. 80. P. 55-59.

202. Іванова Л. До проблеми дослідження речень із дистрибутивними предикатами. Наукові записки 254. URL: https://www.cuspu.edu.ua/download/nauk_zapiski/2005_vipusk_59_zamovlennya_3866.pdf#page=254

203. Рогушина Ю., Гладун А. Застосування онтологічного аналізу для обробки метаданих при інтерпретації Big Data на семантичному рівні. Проблеми програмування, 2020. http://dspace.nbuv.gov.ua/handle/123456789/180494

204. Кутня Г. Онтологічні властивості статичності/динамічності як диференційні семантичні ознаки предикатів. Вісник Львівського університету (72). http://publications.lnu.edu.ua/bulletins/index.php/philology/article/view/10857

205. Андон П. І., Рогушина Ю. В., Гришанова І. Ю., Резніченко В. А., Киридон А. М., Арістова А. В., Тищенко А. О. Досвід використання семантичних технологій для створення інтелектуальних ВЕБ-енциклопедій (на прикладі розробки порталу E-BYE). Проблеми програмування. 2020. http://dspace.nbuv.gov.ua/handle/123456789/180470

206. Ястремська Т. Формальна і семантична структури дериватів у говорах української мови: проблема кореляції. Studia Slavica Academiae Scientiarum Hungaricae. 2021. Vol. 65(1). P. 209-230.

207. Межов О. Семантична та формальна інтерпретація ускладнення синтаксичних одиниць. Тенденції розвитку української лексики. URL: http://ukraina.uw.edu.pl/sites/default/files/pliki/tendencje%20gramatyki.pdf#page=302

208. Личук М. І. Формальний та семантичний принципи впорядкування нечленованих реченнєвих побудов. 2018. URL: http://dglib.nubip.edu.ua:8080/handle/123456789/8049

209. Висоцька В. Метод авторифікації тексту науково-технічних публікацій на основі лінгвістичного аналізу коефіцієнтів мовної різноманітності. Радіоелектроніка. Інформатика. Управління. 2020. № 1(52). С. 108–124.

210. Берко А. Ю., Висоцька В. А., Чирун Л. В. Лінгвістичний аналіз текстового комерційного контенту. Вісник Національного університету "Львівська політехніка". 2015. № 814. С. 203–227.

211. Бісікало О. В., Висоцька В. А. Метод лінгвістичного аналізу україномовного комерційного контенту. Вісник Національного університету "Львівська політехніка". 2016. № 854. С. 185–204.

212. Пасічник В. В., Щербина Ю. М., Висоцька В. А., Шестакевич Т. В. Математична лінгвістика. Книга 2. Комбінаторна лінгвістика: навчальний посібник. Львів: Вид-во Львів. політехніки, 2019. 250 с.

213. Литвин В. В. Висоцька В.А., Оливко Р. М. Метод визначення семантичної метрики на основі тезаурусу предметної області. Інтелектуальні системи та прикладна лінгвістика, Харків, 14 квіт. 2016 р. С. 10–12.

214. Панченко Т. В. Еквівалентність двох систем паралельного виконання. Проблеми програмування. URL: http://dspace.nbuv.gov.ua/handle/123456789/144587

215. Билінська О. The structural-semantic models of slogans in Ukrainian political discourse. Наукові записки Національного університету «Острозька академія». Серія «Філологічна». 2015. Вип. 59, С. 260-262.

216. Береза, А. М. Основи створення інформаційних систем. К.: КНЕУ, 2001.

217. Войтенко Я. С. Дослідження методів семантичного аналізу текстів для інтелектуалізації Web-сайтів. 2020. URL: https://openarchive.nure.ua/handle/document/12609

218. Врачинська А. Модифікований алгоритм попередньої обробки нечітких логічних правил. 2021. URL: https://ela.kpi.ua/bitstream/123456789/42147/1/Vrachynska_bakalavr.pdf

219. Шестакевич Т. В., Висоцька В. А., Чирун Л. В., Чирун Л. Б. Моделювання семантики речення природною мовою за допомогою породжувальних граматик. Вісник НУЛП. 2015. № 814. С. 335–352.

220. Катренко А. В., Пастернак О. В. Системні аспекти інвестування в галузі інформаційних технологій. Вісник Національного університету Львівська політехніка. 2014. Vol. 805. P. 402-411.

221. Катренко А. В. Системний аналіз об'єктів та процесів комп'ютеризації. Львів: Новий світ, 2003.

222. Верес О. М., Верес Ю. О., Катренко А. В. СППР з керування розподілом обмежених ресурсів. Вісн. Нац. ун-ту "Львів. політехніка". 2008. № 610. С .52-62.

223. Катренко А. В., Верес Ю. О. Координація у системах підтримання прийняття рішень з розподілу обмежених ресурсів. Вісн. НУ "Львів. Політехніка". 2009. № 653 С. 117-128.

224. Катренко А. В., Грімнак О. В. Система підтримання прийняття рішень для багатокрокових процесів з використанням ланцюгів маркова. Вісн. Нац. ун-ту "Львів. Політехніка". 2008. № 631. С. 148-155.

225. Верес О. М., Катренко А. В., Рішняк І. В., Чаплига В. М. Управління ризиками в проектній діяльності. Вісн. Нац. ун-ту "Львів. Політехніка". 2003. № 48. С. 38-49.

226. Катренко А. В., Антоняк Т. І. Розв'язання задач оптимального розміщення об'єктів методом імітаційного моделювання. Вісник НУЛП. 2011. № 715. С. 150-162.

227. Катренко А. В., Савка І. В. Механізми координації у складних ієрархічних системах. Вісник Національного університету "Львівська політехніка?. 2008. № 631. С. 156-166.

228. Krivenko S., Rotaniova N., Lazarevska Y. Автоматизована система виявлення нестандартних дій за допомогою сценарного аналіза тексту. Кібербезпека: освіта, наука, техніка. 2021. Vol. 1(13). P. 92-101.

229. Шило Р. С. Дослідження експертної системи діагностування стану людини в умовах надзвичайних ситуацій. 2019. URL: https://openarchive.nure.ua/handle/document/11912

230. Іванов М. Є. Дослідження методів NLP для розробки сервісу голосового управління. 2019. URL: https://openarchive.nure.ua/handle/document/12069

231. Озерова Н., Широков В. Перший том словника української мови у 20 томах. Мовознавство. 2011. С. 3-13.

232. Широков В. А. та ін. Лінгвістичні та технологічні основи тлумачної лексикографії : Монографія;;НАН України, Укр. мовно-інформаційний фонд НАН України. К. : Довіра, 2010.

233. Широков В. А. та ін. Лінгвістичні та технологічні основи тлумачної лексикографії/ НАН України, Укр. мовно-інформаційний фонд НАН України. К. : Довіра, 2010. ISBN 978-966-507-283-6

234. Широков В. А. та ін. Лінгвістично-інформаційні студії: праці Українського мовно-інформаційного фонду НАН України: у 5 т. / Т. 1: Наукова парадигма та основні мовно-інформаційні структури. Київ. Український мовно-інформаційний фонд НАН України. 2018. 271 с.

235. Широков В. А. та ін. (2018). Лінгвістично-інформаційні студії : праці Українського мовно-інформаційного фонду НАН України : у 5 т.. Т. 2. Граматичні системи. Київ : УМІФ НАНУ, 2018. 300 с. URL: http://repository.kpi.kharkov.ua/bitstream/KhPI-Press/43220/1/Book_2018_Shyrokov_Linhv-inform_studii_2.pdf

236. Широков В. А. та ін. Лінгвістично-інформаційні студії : праці Українського мовно-інформаційного фонду НАН України : у 5 т. Т. 3 : Тлумачна лексикографія. Кн. 1 : Словник української мови у двадцяти томах. Київ: Український мовно-інформаційний фонд НАН України, 2018. 276 с.

237. Широков В. А. та ін. Лінгвістично-інформаційні студії : праці Українського мовно-інформаційного фонду НАН України : у 5 т. Т. 3 : Тлумачна лексикографія. Кн. 2 : Системна семантика тлумачних словників. Київ : УМІФ НАНУ, 2018. URL: http://repository.kpi.kharkov.ua/bitstream/KhPI-Press/43222/1/Book_2018_Shyrokov_Linhv-inform_studii_3_2.pdf

238. Широков В. А. та ін. Праці Українського мовно-інформаційного фонду НАН України: у 5 т. Т. 3: Тлумачна лексикографія. Кн. 3: Динаміка лексико-семантичного складу Словника української мови у 20 томах. Київ: Український мовно-інформаційний фонд НАН України, 2018. 230 с.

239. Широков В. А. та ін. Лінгвістично-інформаційні студії: праці Українського мовно-інформаційного фонду НАН України : у 5 т. / Т. 4 : Корпусна та когнітивна лінгвістика. Київ. Український мовно-інформаційний фонд НАН України. 2018. 246 с.

240. Широков В. А. та ін. Лінгвістично-інформаційні студії: праці Українського мовно-інформаційного фонду НАН України : у 5 т. / Т. 5 : Віртуалізація лінгвістичних технологій. Київ. Український мовно-інформаційний фонд НАН України. 2018. 239 с.

241. Широков В. А. Лексикографічне представлення семантичних станів. Математические машины и системы. 1999. № 3. С. 21.

242. Широков В. А., et al. Застосування Українського національного лінгвістичного корпусу в лексикографії та лінгвістичних експертизах. Не все сплива рікою часу. 2011. 285.

243. Широков В. А., Остапова И. В. Парадигма цифрового лексикографического пространства и метод виртуальных лексикографических лабораторий. ББК. 81.1 С48, 110.

244. Широков В. А., et al. Онтологізовані лексикографічні системи в сучасній термінографії. Наукова термінологія нового століття: теоретичні і прикладні виміри: матеріали, ББК 63.3 я2. 220.

245. Старко В., Чейлитко Н. Концепція створення Браунського корпусу української мови «Комп'ютерна лінгвістика: сучасне та майбутнє». Матеріали Міжнародної науково-практичної конференції. – К.: КНЛУ, 2012. – С. 45-46. URL: http://www.mova.info/zbirnyk.pdf

246. Старко В., Чейлитко Н. Параметризація корпусу як спосіб підвищення його репрезентативності та збалансованості "Українське мовознавство", випуск 43, 2013, С. 87-94. URL: http://philology.knu.ua/library/zagal/Ukr_movoznavstvo_2013_43/87-94.pdf

247. Cheilytko, N., Starko, V., Galkin, A. The Ukrainian Brown Corpus and Dependency Tree Modeling. CADSM 2013. – Львів: НУЛП, 2013. С. 58- 60. URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6543167

248. Старко В. Формування Браунського корпусу української мови Мовні і концептуальні картини світу. 2014. Вип. 48. С. 415-421. URL: http://philology.knu.ua/files/library/movni_i_konceptualni/48/40.pdf

249. Гордієнко Н. Сучасна лексикографія як об'єкт лінгвістики. Українська мова, 2011. URL: http://dspace.nbuv.gov.ua/bitstream/handle/123456789/42872/07-Gordienko.pdf?sequence=1

250. Бодик О. П., Рудакова Т. М. Сучасна українська літературна мова. Лексикологія. Фразеологія. Лексикографія. Центр учбової літератури, 2011.

251. Шипнівська О. Створення бази даних для розробки автоматичного синтаксичного аналізатора україномовних текстів (на матеріалі простого речення).Вісник Київського НУ імені Тараса Шевченка. Військово-спеціальні науки. 2012. № 28. С. 28-31.

252. Купріянов Є. В. Комп'ютерна лексикографія як проблема сучасного мовознавства. 2008. URL: http://repository.kpi.kharkov.ua/bitstream/KhPI-Press/2480/1/2008_Kupriyanov_Kompiuterna%20leksykohrafiia.pdf

253. Рудь Н. А. Textanalisator у системі досліджень ідіолекту мовної особистості. URL: https://library.sspu.edu.ua/wp-content/uploads/2018/04/12-7.pdf#page=154

254. Ніколаєвський О. Автоматизація укладання компонентів лінгвістичного забезпечення модуля автоматичного морфологічного аналізу різномовних текстів. Вісник Київського НУ імені Тараса Шевченка. Військово-спеціальні науки. 2012. № 28. С. 24-28.

255. Дарчук Н. Автоматичний синтаксичний аналіз текстів корпусу української мови. Українське мовознавство. 2013. № 43. С. 11-19.

256. Бісікало О. В., Висоцька В. А. Виявлення ключових слів на основі методу контент-моніторингу україномовних текстів. Радіоелектроніка. Інформатика. Управління. 2016. № 1 (36). С. 74–83.

257. Бісікало О. В., Висоцька В. А. Застосування методу синтаксичного аналізу речень для визначення ключових слів україномовного тексту. Радіоелектроніка. Інформатика. Управління. 2016. № 3 (38). С. 54–65.

258. Бісікало О. В., Висоцька В. А. Експериментальне дослідження пошуку значущих ключових слів україномовного контенту. Вісник НУ "Львівська політехніка". 2015. № 829. С. 255–272.

259. Чирун Л. Б., Чирун Л. В., Висоцька В. А. Метод визначення авторства текстового україномовного контенту. ISDMCI, 21–27 трав. 2018 р., Залізний Порт, Україна. С. 287–289.

260. Сірук О. Б. Опрацювання діалектних даних у Корпусі українських діалектних текстів. Філологічні студії. Науковий вісник Криворізького державного педагогічного університету. 2011. № 6 (2). С. 107-110.

261. Лозицький О. А., Кунанець Н. Е. Система опрацювання технічних текстів українською мовою з метою їх адаптації для людей з вадами зору. Вісник НУ }Львівська політехніка", 2014, 805: 316-324.

262. Кульчицький І. М. Технічні аспекти опрацювання комп'ютером природномовної інформації. Вісник Національного університету "Львівська політехніка". Інформаційні системи та мережі, 2014, 783: 344-353.

263. Rusyn B., et al. The mobile application development based on online music library for socializing in the world of bard songs and scouts' bonfires. Conference on Computer Science and Information Technologies. Springer, Cham, 2019. P. 734-756.

264. Korobov M. Morphological analyzer and generator for Russian and Ukrainian languages. International conference on analysis of images, social networks and texts. 2015. P. 320-332.

265. Perkhach R.-Y., et al. Method of Structural Semantic Analysis of Dental Terms in the Instructions for Medical Preparations. COLINS. 2020. P. 662-669.

266. Shakhovska K., Shakhovska N. Veselý P. The sentiment analysis model of services Providers' feedback. Electronics, 2020, 9.11: 1922.

267. Bekesh R., et al. Structural Modeling of Technical Text Analysis and Synthesis Processes. COLINS. 2020. P. 562-589.

268. Sazhok M., Robeiko V. Bidirectional Text-To-Pronunciation Conversion with Word Stress Prediction for Ukrainian. UkrObraz, 2012.
269. Онлайн-бібліотека Горох. URL: https://goroh.pp.ua/
270. Рисін А., Старко В. Великий електронний словник української мови (ВЕСУМ). Вебверсія 5.6.0. 2005-2022. URL: https://r2u.org.ua/vesum/
271. Рисін А., Старко В. Корпус сучасної української мови (БрУК). URL: https://github.com/brown-uk
272. Рисін А., Старко В. Корпусна група БрУК. URL: https://r2u.org.ua/corpus
273. Rysin, A., Starko, V. Project to generate POS tag dictionary for Ukrainian language. https://github.com/brown-uk/dict_uk
274. Словник української мови (СУМ-11). URL: http://sum.in.ua/
275. Словник української мови online (СУМ-20). URL: https://sum20ua.com/Entry/index?wordid=1&page=0
276. Словник. URL: https://slovnyk.ua/
277. Голянич М. І., et al. Лінгвістичний аналіз тексту. 2012. URL: http://lib.pnu.edu.ua:8080/handle/123456789/2809
278. Комарницька О. Моделювання процедур лінгвістичного аналізу тексту в інтелектуальній системі оцінювання знань. Науковий вісник Чернівецького університету: Германська філологія. 2015. Vol. 740-741: 85-88.
279. Зубань О. Параметризована база даних як інструмент дослідження корпусу текстів. Лексикографічний бюлетень, 2006. URL: http://dspace.nbuv.gov.ua/handle/123456789/72851
280. Палагін О. В., et al. Про один підхід до аналізу та розуміння природномовних об'єктів. 2008. URL: http://dspace.nbuv.gov.ua/bitstream/handle/123456789/6503/15-Palagin.pdf?sequence=1
281. Литвиненко Л. О. Особливості побудови лінгвістичного процесора паралельної обробки природномовного тексту. Збірник наукових праць Військового інституту Київського НУ ім. Т. Шевченка, 2013, 39: 176-180.
282. Zubań O. Корпус української мови-комп'ютерна експертна система лінгвістичного аналізу українськомовного тектсу. TEKA Komisji Polsko-Ukraińskich Związków Kulturowych, 2018, 6.13: 191-206.
283. Reese R M. Natural language processing with Java. Packt Publishing Ltd, 2015.
284. Kumar A., et al. Ask me anything: Dynamic memory networks for natural language processing. In: International conference on machine learning. PMLR, 2016. p. 1378-1387.
285. Sun S., Luo C., Chen J. A review of natural language processing techniques for opinion mining systems. Information fusion, 2017, 36: 10-25.
286. Kao A., Poteet S. R. Natural language processing and text mining. Springer Science & Business Media, 2007.
287. Lytvyn V., Sharonova N., Hamon T., V Vysotska., Grabar N., Kowalska-Styczen A. Computational linguistics and intelligent systems. CEUR Workshop Proceedings. 2018. Vol. 2136. 390 p. E-ISSN: 1613-0073
288. Vysotska V. Computer linguistics for online marketing in information technology. Saarbrücken: LAP, 2018. 396 p.
289. Vysotska V., Chyrun L. Linguistic analysis and modelling semantics of textual content for digest formation. MEST Journal. 2015. Vol. 3, № 1. P. 127–148. ISSN: 2334-7171.
290. Chyrun L., Vysotska V., Kozak I. Informational resources processing intellectual systems with textual commercial content linguistic analysis usage constructional means and tools development. Econtechmod :2016. Vol. 5, № 2. P. 85–94.
291. Vysotska V. Linguistic analysis of textual commercial content for information resources processing. TCSET'2016 : proc. of the XIIIth Intern. conf, Feb. 23–26, 2016, Lviv, Slavske, Ukraine. Lviv, 2016. P. 709–713.
292. Lytvyn V., Vysotska V., Veres O., Rishnyak I., Rishnyak H. Content linguistic analysis methods for textual documents classification. CSIT'2016, 6–10 Sept., 2016, Lviv, Ukraine. Lviv, 2016. P. 190–192.

293. Lytvyn V., Vysotska V., Chyrun L., Smolarz A., Naum O. Intelligent system structure for web resources processing and analysis. COLINS'2017, 21 Apr. 2017, Kharkiv. P. 56–74.

294. Lytvyn V., Vysotska V., Wojcik W., Dosyn D. A method of construction of automated basic ontology. 1st International conference computational linguistics and intelligent systems, COLINS'2017, 21 Apr. 2017, Kharkiv. P. 75–83.

295. V Lytvyn., Vysotska V., Chyrun L., Hrendus M., O Naum. Content analysis of text-based information in E-commerce systems. Proceedings of the 2nd International conference, COLINS 2018. Workshop. P. 81–94. ISSN 2523-4013

296. Rusyn B., Vysotska V., Pohreliuk L. Methods of information resources processing in virtual library. Proceedings of the 2nd International conference, COLINS 2018. Workshop. P. 28–39.

297. Lytvyn V., Oborska O., Vysotska V., Dosyn D., Demchuk A. Ontology using for decision making in a competitive environment. Computational Linguistics and Intelligent Systems. Proceedings. 2018. Vol. 2: proceedings of the 2nd International conference, COLINS 2018. Workshop. P. 17–27. ISSN 2523-4013

298. Селіванова О. О. Проблеми класифікації методів лінгвістики. In: Методи лінгвістичних досліджень: матер. міжнар. наук.-практ. конф. Слов'янськ: СДПУ. 2010. p. 180-186.

299. Глущенко В. А. Лінгвістичний Метод: Гомогенні Й Гетерогенні Теорії В Українському Мовознавстві. 160-річчю від дня народження Івана Яковича Франка, 2016, 20.

300. Глущенко В. А. Лінгвістичний метод і його структура. Мовознавство, 2010, 6: 32-44.

301. Глущенко В. А. Лінгвістичний метод: онтологічний, телеологічний та операційний компоненти. Лінгвістичний вісник. 2016. Вип. 6. С. 3-9.

302. Глущенко В. Теорії лінгвістичного методу в українському мовознавстві кінця XX ст.–початку XXI ст. Теоретичні й прикладні проблеми сучасної філології. 2015. Vol. 2: 6-17.

303. Шуменко О. Основи наукових досліджень. 2020. URL: https://essuir.sumdu.edu.ua/handle/123456789/77695

304. Щербина Ю., Висоцька В., Шестакевич Т. Методи та розділи математичної лінгвістики в структурі курсу комп'ютерна лінгвістика. Наукові праці Чорноморського ДУ ім.П. Могили. 2009. Вип. 104: 203-208.

305. Пасічник В. В., Висоцька В., Щербина Ю., Шестакевич Т. Математична лінгвістика. Львів: Новий світ. 2012.

306. Краснобаєва-Чорна Ж. В. Квантитативні методи в лінгвістиці: новітні тенденції. Науковий вісник Дрогобицького державного педагогічного університету імені Івана Франка. 2018. Vol. 9: 93-97.

307. Помірко Р. Тропи та тропеїзація англомовного масмедійного дискурсу. 2016. PhD Thesis. Львівський національний університет ім. І. Франка. https://www.lnu.edu.ua/wp-content/uploads/2016/02/dis_ostapchuk.pdf

308. Федорова О. В. Оказіональні номінації у сучасному англійськомовному кінодискурсі як проблема перекладу українською мовою. 2021. URL: http://rep.knlu.edu.ua/xmlui/handle/787878787/2554

309. Мелещенко О. О. Дискурсивні стратегії англомовного політичного твіттінгу Дональда Трампа: когнітивний мультимодальний аналіз. 2021. URL: http://ekhnuir.univer.kharkov.ua/handle/123456789/17445

310. Lytvyn V., Vysotska V., Rzheuskyi A. Technology for the psychological portraits formation of social networks users for the IT specialists recruitment based on Big Five, NLP and Big Data Analysis. CEUR Workshop Proceedings. 2019. Vol. 2392. P. 147–171. E-ISSN: 1613-0073

311. Chyrun L., Andrunyk V., Vysotska V. Content analysis peculiarities of user internet activities for personality psychological state slice formation. MEST Journal. 2017. Vol. 6, № 2. P. 26–46.

312. Гасько Р. В., Висоцька В. А., Чирун Л. Б. Інформаційна система аналізу психологічного стану особистості. Вісник Національного університету "Львівська політехніка". № 829. С. 102–128.

313. Гасько Р. В., Чирун Л. В., Висоцька В. А. Особливості контент-аналізу користувацької Інтернет-діяльності для формування зрізу психологічного стану особистості. Вісник НУЛП 2017. № 864. С. 221–238.
314. Shuotian Bai, Tingshao Zhu, Li Cheng: Big-Five personality prediction based on user behaviors at social network sites. URL: http://arxiv.org/pdf/1204.4809v1.pdf
315. Shuotian Bai: List of computer science publications by Shuotian Bai. URL: http://dblp.uni–trier.de/pers/hd/b/Bai:Shuotian
316. Liu Dong, Campbell W. Keith. The Big Five personality traits, Big Two metatraits and social media: A meta-analysis. Journal of Research in Personality. 2017. Vol. 70. P. 229-240.
317. Chorley Martin J., Whitaker Roger M., Allen Stuart M. Personality and location-based social networks. Computers in Human Behavior. 2015. Vol. 46. P. 45-56.
318. Souri A., Hosseinpour S., Rahmani A. M. Personality classification based on profiles of social networks' users and the five-factor model of personality. Human-centric Computing and Information Sciences. 2018. Vol. 8.1. P. 1-15.
319. Fang Ruolian, et al. Integrating personality and social networks: A meta-analysis of personality, network position, and work outcomes in organizations. Organization science. 2015. Vol. 26.4. P. 1243-1260.
320. Zhu Xiumei, et al. Pathways to happiness: From personality to social networks and perceived support. Social networks. 2013. Vol. 35.3. P. 382-393.
321. Leprі B., Staiano J., Shmueli E., Pianesi F., Pentland A. The role of personality in shaping social networks and mediating behavioral change. User Modeling and User-Adapted Interaction. 2016. Vol. 26(2). P. 143-175.
322. Buettner R. Predicting user behavior in electronic markets based on personality-mining in large online social networks. Electronic Markets. 2017. Vol. 27.3. P247-265.
323. Цимбал Н. А. Методологічні аспекти дослідження явища мотивації. 2013.
324. Тодорова Н. Ю. Фразеологічні одиниці просторової семантики в українській та англійській мовах. PhD Thesis. Львів. 2018.
325. Kocherha H. V., Martynovska Y. O. мотиваційна маркованість відіменних дієслів у національно-мовному ландшафті писемних пам'яток староукраїнської мови XIV–XVII ст. Publishing House "Baltija Publishing", 2021.
326. Ярмоленко Г. А. Предикатно-аргументна мотивація віддієслівних прикметників сучасної української мови. Linguistic Bulletin. 2015. Vol. 20.
327. Ярмоленко Га. Віддієслівні іменники з локативним значенням: когнітивно-ономасіологічний аспект. Мовознавчий вісник. 2010. Vol. 10. P. 269-272.
328. Малевич Л. Д. Когнітивно-ономасіологічний аналіз українських термінів-суфіксальних девербативівна позначення дій і процесів. Актуальні проблеми філології та перекладознавства. 2016. Vol. 10 (2). P. 134-138.
329. Глуховська М. С. Обсяг поняття асоціативної мотивації у словотворі української мови. Лінгвістичні дослідження. 2014. Vol. 37. P. 65-69.
330. Гаращенко Л. Б., Діброва О. В. Когнітивно-Ономасіологічний Аналіз Аналітичних Номінацій Науково-Технічної Термінології. Актуальні питання сучасної педагогіки: творчість, майстерність. 2020. 450 с.
331. Ярмоленко Г. Пропозиційно-диктумна мотивація фразеологічних одиниць української мови. 2020. URL: https://dspace.uzhnu.edu.ua/jspui/handle/lib/36237
332. Мацюк З. Лінгвістичні основи методики викладання граматики української мови як іноземної. Теорія і практика викладання української мови як іноземної. Львів: Вид. центр ЛНУ ім. Івана Франка. 2007. С. 31-39.
333. Вакуленко М. Методологічні засади вивчення наукової термінології. Термінологічний вісник. 2013. № 2 (2). С. 16-21.

334. Глущенко В. А. Порівняльно-історичний метод в українському мовознавстві 20-х–60-х рр. XIX ст. Наукові праці [Чорноморського державного університету імені Петра Могили комплексу Києво-Могилянська академія]. Серія: Філологія. Мовознавство. 2016. № 272, Вип. 260. С. 21-25.

335. Черхава О. О. Реконструкція Теолінгвістичної Матриці Релігійно-Популярного Дискурсу (на матеріалі англійської, німецької та української мов). Вид. центр КНЛУ, 2017.

336. Кочерган М. П. Загальне мовознавство. К.: Академія, 1999. № 288: 9.

337. Голубовська І. Класичні мови у контексті сучасного мовознавства: епістеми, метамова, інструментарій. Studia linguistica. 2014. Vol. 8. P. 3-10.

338. Глущенко В. А., Тищенко К. А. Порівняльно-історичний метод в українському та російському мовознавстві 20-х – 60-х рр. XIX ст. Слов'янськ : Вид-во Б. І. Маторіна. 2016. 98 с.

339. Бялик В. Д. Внутрішня форма слова та умотивованість англійського неологізму. Наукові праці Кам'янець-Подільського національного університету імені Івана Огієнка. Філологічні науки. 2010. Vol. 22 (1) С. 51-55.

340. Кислюк Л. П. Словотвірна номінація в сучасній українській мові. 2018. PhD Thesis. Київ, 39.

341. Редько Є. О. Типи і способи номінування осіб в українських арготичних системах. PhD Thesis. Дніпропетровський національний університет ім. Олеся Гончара. Дніпропетровськ, 2016.

342. Lefer M.-A., Grabar N. Super-creative and overbureaucratic: A cross-genre corpusbased study on the use and translation of evaluative prefixation in ted talks and EU parliamentary debates. Across Languages and Cultures. 2015. Vol. 16(2). P. 187–208.

343. Hrytsiv N., Shestakevych T., Shyyka J. Corpus Technologies in Translation Studies: Fiction as Document. CEUR Workshop Proceedings. 2021. Vol. 2917. P. 327-343.

344. Lutskiv A., Lutsyshyn R. Corpus-Based Translation Automation of Adaptable Corpus Translation Module. CEUR Workshop Proceedings. 2021. Vol. 2870. P. 511-527.

345. I Bekhta., Hrytsiv N. Computational Linguistics Tools in Mapping Emotional Dislocation of Translated Fiction. CEUR Workshop Proceedings. 2021. Vol. 2870. P. 685-699.

346. Kopp A., Orlovskyi D., Orekhov S. An Approach and Software Prototype for Translation of Natural Language Business Rules into Database Structure. CEUR Workshop Proceedings. 2021. Vol. 2870. P. 1274-1291.

347. Anokhina, T., Kobyakova, I., Shvachko, S.: Going parallel: using earlier translations as background for facilitating re-translation technique. CEUR workshop proceedings. 2020. Vol. 2604. P. 249-258.

348. Kubinska S., Vysotska V., Matseliukh Y. User Mood Recognition and Further Dialog Support. CSIT, 22-25 Sept., Lviv, Ukraine. 2021. Vol. 2. P. 34–39.

349. Bisikalo O.V., Dovgalets S.M., Pijarski P., Lisovenko A.I., Development of dialog system powered by textual educational content. Proceedings of SPIE - The International Society for Optical Engineering, 2016, 10031, 100314E.

350. Aksonov D., Gozhyj A., Kalinina I., Vysotska V. Question-Answering Systems Development Based on Big Data Analysis. CSIT, 22-25 Sept., Lviv, Ukraine. 2021. – Vol. 1. P. 113–118.

351. Stasiuk, L. Computer Sampling and Quantitative Analysis in Exploring Secondary Functions of Questions in Speech Genres of Intimate Communication. CEUR workshop proceedings. 2020. Vol. 2604. P. 227-238.

352. Breja M., Jain S. K. Causality for Question Answering. CEUR workshop proceedings. 2020. Vol. 2604. P. 884-893.

353. Hamon T., Grabar N., Mougin F. Querying biomedical Linked Data with natural language questions. Semantic Web. 2017. Vol. 8(4). P. 581–599.

354. Bublyk M., Kalynii T., Varava L., Vysotska V., Chyrun L., Matseliukh Y. Decision support system design for low voice emergency medical calls at smart city based on chatbot management in social networks. Webology. 2022. Vol. 19, iss. 2. P. 2135-2178. E-ISSN: 1735-188X. URL: https://www.webology.org/abstract.php?id=1430
355. Bublyk M., Zahreva Y., Vysotska V., Matseliukh Y., Chyrun L., Korolenko O. Information system development for recording offenses in smart city based on cloud technologies and social networks. Webology. 2022. Vol. 19, iss. 2. P. 1870-1898. E-ISSN: 1735-188X. URL: https://www.webology.org/abstract.php?id=1412
356. Husak V., Lozynska O., Karpov I., Peleshchak I., Chyrun S., Vysotskyi A. Information system for recommendation list formation of clothes style image selection according to user's needs based on NLP and Chatbots. CEUR workshop proceedings. 2020. Vol. 2604. P. 788-818.
357. Romanovskyi O., Pidbutska N., Knysh A. Elomia Chatbot: The effectiveness of artificial intelligence in the fight for mental health. CEUR Workshop Proceedings. 2021. Vol. 2870. P. 1215-1224.
358. Yarovyi A., Kudriavtsev D. Method of multi-purpose text analysis based on a combination of knowledge bases for intelligent chatbot. CEUR Workshop Proceedings. 2021. Vol. 2870. P. 1238-1248.
359. Shakhovska N., Basystiuk O., Shakhovska K. Development of the speech-to-text chatbot interface based on Google API. CEUR Workshop Proceedings. 2019. Vol. 2386. P. 212-221.
360. Фаліна Е. О. Порівняльна типологія граматичних категорій дієслова у корейській і українській мовах. 2020. URL: http://rep.knlu.edu.ua/xmlui/handle/787878787/820
361. Денисова С. П. Словотвірні гнізда з вершинами-онімами в українській та англійській мовах: контрастивний аспект. 2017. PhD Thesis. Національний педагогічний університет імені МП Драгоманова.
362. Цимбал Н. А. Мета і завдання зіставного лінгвістичного аналізу української і туркменської мов. 2017. URL: https://dspace.udpu.edu.ua/bitstream/6789/8457/1/Tsymbal_N_A_%20D_%20Tokliev.pdf
363. Волянюк І. О. Особливості застосування структурного методу в мовознавчих дослідженнях. Publishing House "Baltija Publishing", 2021.
364. Янковець О. В. The peculiarities and stages of the word-compounding analysis during the English border guard terms study. Філологічні Студії. 2019. 46 с.
365. Завальська Л. Теоретичні засади дослідження українського політичного дискурсу в лінгвопрагматичному аспекті. Вісник Одеського національного університету. Філологія. 2018. Vol. 23.2 (18). P. 42-48.
366. Єфремова Н. В., Гончарук С. В. Методи дослідження лексичної синонімії. 2015. URL: https://evnuir.vnu.edu.ua/handle/123456789/7695
367. Гладкий А. В. Формальные грамматики и языки. М.: Наука, 1973. 368 с.
368. Гладкий А. В. Синтаксические структуры естественного языка в автоматизированных системах общения. М.: Наука, 1985. 144 с.
369. Гладкий А., Мельчук И. Элементы математической лингвистики. М.: Наука, 1969. 192 с.
370. Гладкий А. Алгоритмическая природа инвариантных свойств грамматик непосредственно составляющих. Алгебра и логика. 1964. Vol. 3.2. С. 17.
371. Гладкий А. В. Алгоритмическая нераспознаваемость существенной неопределенности контекстно-свободных языков. Алгебра и логика. 1965. Vol. 4.4. С. 53-63.
372. Гладкий А. В. О точных и математических методах в лингвистике и других гуманитарных науках. Вопросы языкознания. 2007. Vol. 5. С. 22-38.

373. Гладкий А. Математические методы изучения естественных языков. Труды Математического института имени ВА Стеклова. 1973. Vol. 133.0. С. 95-108.
374. Гладкий А., Мельчук И. Элементы математической лингвистики. Наука, 1969.
375. Гладкий А. В. Язык, математика и лингвистика. Математика в школе. 1994. Vol. 1. С. 2-9.
376. Гладкий А. Размышления о взаимодействии лингвистики и математики. URL: http://elementy. ru/lib/164549.
377. Гладкий А. В. О формальных методах в лингвистике. Вопр. языкознання. 1966. Vol. 3. С. 52.
378. Гладкий А. В. Математическая лингвистика. Лингвистический энциклопедический словарь. М.: Сов. энциклопедия. 1990. С. 287-289.
379. Гладкий А. В. Лингвистика и математика. Всесоюзная научная, 1974.
380. Ткаченко А. І. Лексико-граматична трансформаційна матриця перекладу синтаксичних конструкцій у англо-українських текстах художнього дискурсу. 2021.
381. Рослицька М. В. Прецедентне ім'я в політичному дискурсі: формально-семантичні ознаки і соціопрагматичний потенціал. 2018. PhD Thesis. Львів, 2019.
382. Таценко Н. В. Степанов В. В., Ущаповська І. В. Когнітивно-прагматична корелятивність семантичного простору мови. Сумський державний університет, 2020.
383. Гавриш О., Гавриш Е. Методи дослідження мовних контактів. 2017. URL: https://ir.kneu.edu.ua/handle/2010/21758
384. Фабіан М. П. Застосування процедури формалізованого аналізу лексичної семантики в зіставних дослідженнях. Проблеми зіставної семантики. 2011. Vol. 10 (1). С. 206-211.
385. Бук С. Основи статистичної лінгвістики. Видавничий центр ЛНУ імені Івана Франка, 2008. 124 с.
386. Бук С. Квантитативна параметризація текстів Івана Франка: проект та його реалізація. Вісник Львівського університету. Серія: Філологічна. 2013. Vol. 58. С. 290-307.
387. Бук С. Сучасні методи дослідження мови письменника у слов'янознавстві. Проблеми слов'янознавства. 2012. Vol. 61. С. 86-95.
388. Бук С. Структурне анотування у корпусі текстів (на прикладі прози Івана Франка). 2009. URL: http://dspace.nbuv.gov.ua/handle/123456789/6052
389. Бук С., Ровенчак А. Онлайн-конкорданс роману Івана Франка "Перехресні стежки". Іван Франко: дух, наука, думка, воля: Матеріали Міжнародного наукового конгресу, присвяченого. 2006. С. 203-211.
390. Бук С. Корпус текстів у лінгводидактиці (на матеріалі омонімії у корпусі великої прози Івана Франка). Вісник Львівського університету. Серія філологічна. 2012. Vol. 57. С. 106-116.
391. Бук С. Слов'янський досвід укладання частотних словників мови письменника. Проблеми слов'янознавства. 2011. Vol. 60. С. 217-224.
392. Бук С. Пряма й авторська мова великої прози Івана Франка: лінгвостатистичне дослідження у контексті корпусної лінгвістики. Вісник Львівського університету. Серія філологічна. 2011. Vol. 52. С. 199-209.
393. Бук С. Кількісне зіставлення текстів (на матеріалі редакцій 1884 та 1907 років повісті Івана Франка Boa constrictor"). Українське літературознавство. 2012. Vol. 76. С. 179.
394. Gordiienko-Mytrofanova I., Sauta S. Психологічний зміст фугітивності як компонента грайливості / ігрової компетентності. Psychological Journal. 2021. Vol. 7.2. С. 88-104.
395. Забурко М. П. Дослідження методів організації API для інформаційної системи обчислення вагових коефіцієнтів та інтегрального показника опінії об'єктів. 2013. URL: http://elartu.tntu.edu.ua/handle/123456789/2678

396. Масенко Л. Нариси з соціолінгвістики. Нац. ун-т "Києво-Могилянська академія". Київ: Києво-Могилян. акад., 2010. 242 с.

397. Загнітко А. П. Теорія граматики і тексту. 2014. URL: http://r.donnu.edu.ua/handle/123456789/160

398. Брідко Т. В. Методичні аспекти емпиричного дослідження варіативності вимови (з позиції соціолінгвістики). URL: http://dspace.nbuv.gov.ua/bitstream/handle/123456789/55920/60-Bridko.pdf?sequence=1

399. Холод О. М. Соціальні комунікації: соціо-і психолінгвістичний аналіз. 2010. URL: http://lib.pnu.edu.ua:8080/handle/123456789/5864

400. Shakhovska N., Vysotska V., Chyrun L. Intelligent systems design of distance learning realization for modern youth promotion and involvement in independent scientific researches. Advances in Intelligent Systems and Computing (AISC). 2017. Vol. 512. P. 175–198. ISSN 2194-5357, E-ISSN: 2194-5365.

401. Lytvyn V., Vysotska V., Burov Ye., Veres O., Rishnyak I. The contextual search method based on domain thesaurus. Advances in Intelligent Systems and Computing (AISC). 2018. Vol. 689. P. 310–319.

402. Vysotska V., Basto F. V., Lytvyn V., Emmerich M., Himyak M. Method for determining linguometric coefficient dynamics of Ukrainian text content authorship. Advances in Intelligent Systems and Computing (AISC). 2019. Vol. 871. P. 132–151. ISSN 2194-5357, E-ISSN: 2194-5365

403. Vysotska V., Burov Y., Lytvyn V., Oleshek O. Automated monitoring of changes in web resources. Advances in Intelligent Systems and Computing (AISC). 2020. Vol. 1020. P. 348–363. ISSN 2194-5357, E-ISSN: 2194-5365.

404. Vysotska V. Ukrainian participles formation by the generative grammars use. CEUR Workshop Proceedings. 2020. Vol. 2604. P. 407–427. E-ISSN: 1613-0073.

405. Tymoshenko K., Vysotska V., Kovtun O., Holoshchuk R., Holoshchuk S. Real-time Ukrainian text recognition and voicing. CEUR Workshop Proceedings. 2021. Vol. 2870. P. 357–387. E-ISSN: 1613-0073.

406. Lytvyn V., Pukach P., Bobyk I., Vysotska V. The method of formation of the status of personality understanding based on the content analysis. Eastern-European Journal of Enterprise Technologies. 2016. № 5/2 (83). С. 4–12.

407. Литвин В. В., Бобик І. О., Висоцька В. А. Застосування системи алгоритмічних алгебр для граматичного аналізу символьних обчислень виразів логіки висловлювань. Радіоелектроніка. Інформатика. Управління. 2016. № 4 (39). С. 77–89.

408. Романова Н. В. Психолінгвістичні методи дослідження емотивної лексики. Наукові записки [Національного університету Острозька академія]. Сер.: Філологічна. 2012. Vol. 29. С. 176-179.

409. Іванюк Г., Горошко О., Мельник І. Psychosemantic Meaning of the concept of «teacher» in the linguistic consciousness of students of pedagogical specialties. Psycholinguistics. Pereiaslav-Khmelnytskyi Hryhorii Skovoroda State Pedagogical University. 2020. Vol. 28 (1). 288 p.

410. Коваль Л. М. Психолінгвістика. методичні рекомендації для здобувачів вищої освіти магістерського рівня освітньо-професійної програми "Українська мова та література". Вінниця : ФОП Кушнір Ю.В., 2021. 68 с.

411. Марчук Л. М. Категорія градації в сучасній українській літературній мові. Українська мова, 2008.

412. Минзак О. В. Антонімічні пари: огляд метонімічних характеристик в англійській мові. Наукові записки [Національного університету Острозька академія]. Сер.: Філологічна. 2010. Vol. 13. С. 468-473.

413. Векуа Н. В. Антонімія якісних прикметників у сучасній українській мові. 2006 http://enpuir.npu.edu.ua/handle/123456789/1663

414. Марчук Л. М. Проблема умовності категорій градації (кількості та якості). Наукові праці Кам'янець-Подільського національного університету імені Івана Огієнка. Філологічні науки. 2011. Vol. 28. С. 274-277.

415. Ковалевська А. В. Модальнісне редагування як інструмент оптимізації сугестивного ефекту реклами. Слов'янський збірник. 2014. Vol. 18. С. 147-152.
416. Chomsky N. Morphophonemics of Modern Hebrew (Routledge Revivals). Routledge, 2013.
417. Chomsky N. Transformational analysis. University of Pennsylvania, 1955.
418. Хомский Н. О некоторых формальных свойствах грамматик. Кибернетический сборник. М.: Мир, 1962. № 5. С. 279–311.
419. Хомски Н., Миллер Дж. Формальный анализ естественных языков. Кибернетический сборник. М.: Мир, 1965. № 1. С. 231-290.
420. Хомски Н. Язык и мышление. Публикации ОСиПЛ. 1972. № 2. 122 с.
421. Хомски Н. Синтаксические структуры. Сборник «Новое в лингвистике». М.: ИЛ, 1962. № 2. С. 412-527.
422. Chomsky N. Three models for the description of language. I. R. E. Trans. PGIT 2, 1956. P. 113-124.
423. Chomsky N. On certain formal properties of grammars, Information and Control 2. A note on phrase structure grammars, Information and Control. 1959. Vol. 2, P. 137-267, 393-395.
424. Chomsky N. On the notion "Rule of Grammar". Proc. Symp. Applied Math., 12. Amer. Math. Soc., 1961.
425. Chomsky N. Context-free grammars and pushdown storage. Quarterly Progress Reports, № 65, Research Laboratory of Electronics, M. I. T., 1962.
426. Chomsky N. Formal properties of grammars. Handbook of Mathemati-Mathematical Psychology, 2, ch. 12, Wiley, 1963. P. 323-418.
427. Chomsky N. The logical basis for linguistic theory. Proc. IX-th Int. Cong. Linguists, 1962.
428. Chomsky N., Miller G. A. Finite state languages. Information and Control. 1958. Vol. 1. P. 91-112.
429. Chomsky N., Miller G. A. Introduction to the formal analysis of natural languages. Handbook of Mathematical Psychology. Vol. 2, Ch. 12, Wiley, 1963. P. 269-322.
430. Chomsky N., Schützenberger M. P. The algebraic theory of context-free languages. Computer programming and formal systems, North-Holland, MR152391. Amsterdam 1963. P.118–161.
431. Johnson M. Lakoff G. Why cognitive linguistics requires embodied realism. 2002. URL: https://www.degruyter.com/document/doi/10.1515/cogl.2002.016/html
432. Lakoff G. Ten lectures on cognitive linguistics. Brill, 2019. URL: https://brill.com/view/title/54941
433. Lakoff G. Cognitive semantics. 1988. URL: https://escholarship.org/content/qt04086580/qt04086580.pdf
434. Lakoff G. Women, fire, and dangerous things. What categories reveal about the mind. Chicago, 1987. 631p.
435. Лакофф Дж., Джонс М. Метафоры, которыми мы живем. Теория метафоры. 1990. С. 387–415.
436. Lakoff G. The contemporary theory of metaphor. Metaphor and thought. Cambridge, 1993. 245 p.
437. Geeraerts Dirk (ed.). Cognitive linguistics: Basic readings. Walter de Gruyter, 2006.
438. Langacker Ronald W. Cognitive grammar. Basic Readings, 2008, 29. URL: https://www.degruyter.com/document/doi/10.1515/9783110199901/pdf#page=37
439. Langacker R. Foundations of cognitive grammar: In 2 volumes. Stanford: Stanford Univ. Press, 1987. Vol.1. 516 p.
440. Lakoff G. Cognitive versus generative linguistics: How commitments influence results. Language and communication, 1990, 1.1. URL: https://escholarship.org/content/qt2tj4t3cw/qt2tj4t3cw.pdf
441. Rehani Manu, Wolf Warren L. Methods and systems for measuring semantics in communications. U.S. Patent No 9,269,353, 2016. URL: https://patentimages.storage.googleapis.com/00/d2/da/886c00fc2dce4b/US9269353.pdf

442. Ковбасюк Л. А. Корпусна лінгвістика та германістика: теоретичні засади і перспективи. Наукові записки [Ніжинського державного університету ім. Миколи Гоголя]. Філологічні науки. 2017. Vol. 1. С. 9-14.

443. Ковбасюк Л. А., Романова Н. В. Сучасні лінгвістичні теорії. Херсон: ХДУ, 2008. 96 с.

444. Филлмор Ч. Фреймы и семантика понимания. Новое в зарубежной лингвистике. М.: Прогресс, 1988. Вып. 23. С. 52-92.

445. Штерн І. Б. Вибрані топіки та лексикон сучасної лінгвістики. К.: Артек, 1998. 336 с.

446. Жаботинская С. А. Концептуальний анализ: типы фреймов. Вісник Черкаського університету. Сер. Філологічні науки. 1999. Вип. 11. С. 12-25.

447. Маслова В. А. Лингвокультурология. М.: Академия, 2001. 208 с.

448. Кубрякова Е., Демьянков В., Панкрат Ю., Лузина Л. Словарь когнитивных терминов. М.: МГУ, 1997. 245 с.

449. Шевченко О. М.; Шевченко Н. С. Когнітивна лінгвістика як напрям мовознавчого дослідження. 2020.

450. Варчук Л. Когнітивна лінгвістика здобутки та напрями досліджень. Наукові записки. 2017. Вип. 24. С. 96-102.

451. Гончарук Н. М. Дослідження психологічних аспектів комунікації у прикладних лінгвістичних теоріях. Проблеми сучасної психології. 2018. Vol. 39. С. 90-100.

452. Зеленько О. А. Мемна структура внутрішньої форми членороздільної звукової мови–функційний аналог генного коду. 2021. URL: http://lib.ndu.edu.ua/dspace/bitstream/123456789/2006/1/18.pdf

453. Зеленько А. С. Лінгво-гносеологічний аналіз соціального компонента свідомості. 2021. URL: http://lib.ndu.edu.ua/dspace/handle/123456789/2118

454. Остапчук І. І. Тропи та тропеїзація англомовного масмедійного дискурсу. URL: http://www. lnu. edu. ua/wp-Content/Uploads/2016/02/Dis_Ostapchuk. Pdf, 2016.

455. Помірко Р. С. Тропи та тропеїзація англомовного масмедійного дискурсу. 2016. PhD Thesis. Львівський національний університет імені Івана Франка. URL: https://www.lnu.edu.ua/wp-content/uploads/2016/02/dis_ostapchuk.pdf

456. Петренко О. В. Метафоризація як механізм найменування німецьких понять робототехніки. Південний архів (філологічні науки). 2020. Vol. 82. С. 92-95.

457. Галапчук-Тарнавська О. М. Лінгвістика тексту та функціональна лінгвістика: програма вибіркової навчальної дисципліни. 2019. URL: http://95.217.214.133/handle/123456789/16647

458. Зеленько А. С. Про лінгвістичну парадигму та функціональну лінгвістику в україністиці. Вісник. 2008. С. 140.

459. Овсієнко Л. Текст як об'єкт вивчення лінгвістики й лінгводидактики. Теоретична і дидактична філологія. 2014. Vol. 17. С. 114-131.

460. Декало В. Конструктивний метод моделюваннння синтаксичних структур у рамках теорії принципів і параметрів. Наукові записки. Випуск 118. Серія: Філологічні науки (мовознавство). Кіровоград: РВВ КДПУ ім. В. Винниченка, 2013. 600 с.

461. Дудко І. В. Особливості аналізу дієслова: формальний і функціональний аспекти. Науковий часопис Національного педагогічного університету імені МП Драгоманова. Серія 10: Проблеми граматики і лексикології української мови. 2012. Vol. 9. С. 62-65.

462. Нечипоренко Б. Когнітивний аспект дослідження сугестивної функції синтаксису в політичному дискурсі китайських ЗМІ. Вісник Львівського університету. Серія філологічна. 2011. Vol. 54. P. 181-187.

463. Андрійшина К. І. Конструювання індивідуальної авторизації в англомовних журнальних статтях. Одеський лінгвістичний вісник. 2017. Vol. 9 (1). С. 8-11.

464. Пилипак В. Егоцентрична ‘там’-семантика грамеми середнього роду в українській мові. Лінгвістичні студії, 2013. Vol. 27. С. 55-61.

465. Мельник Т. Лінгвістичні праці О. Синявського в історії стандартизації лексичного та граматичного рівнів мовної системи. Тернопільський національний педагогічний університет ім. В. Гнатюка. 2013. С. 181.

466. Anderson J. R., Bothell D., Byrne M. D., Douglass S., Lebiere C., Qin Y. An integrated theory of the mind. Psychological review. 2004. Vol. 111(4). P. 1036.

467. Tobinski D. Cognitive Architectures in times of Life 3.0: Human Intelligence or Artificial Intelligence? Sterben 2. 0:(Trans) Humanistische Perspektiven Zwischen Cyberspace, Mind Uploading und Kryonik,.2022. P. 161.

468. Kim N., Nam C.S. Adaptive Control of Thought-Rational (ACT-R): Applying a Cognitive Architecture to Neuroergonomics. Neuroergonomics. Cognitive Science and Technology. Springer, Cham. 2020.

469. Yengin I., Ince I. Applying the adaptive control of thought-rational theory into the design of mobile worked examples applications. International Journal of Robots, Education and Art. 2014. Vol. 4.2. С. 21.

470. Cao S., Liu Y. Queueing network-adaptive control of thought rational. International Journal of Human Factors Modelling and Simulation 55. 2013. Vol. 4.1. P. 63-86.

471. Vysotska V. Internet systems design and development based on Web Mining and NLP. Saarbrücken: LAP, 2018. 316 p.

472. Vysotska V. Computer linguistics for online marketing in information technology : monograph. Saarbrücken: LAP Lambert Academic Publishing, 2018. 396 p.

473. Vysotska V. Linguistic analysis of textual commercial content for information resources processing. TCSET : proc. of the XIII Intern. conf, Feb. 23–26, Lviv, Slavske, Ukraine, 2016. P. 709–713.

474. Lytvyn V., Vysotska V., Chyrun L., Dosyn D. Methods based on ontologies for information resources processing : monograph. Saarbrücken: LAP Lambert Academic Publishing, 2016. 324 p.

475. Литвин В. В., Висоцька В. А., Досин Д. Г. Методи та засоби опрацювання інформаційних ресурсів на основі онтологій: монографія. Львів: ЛА “Піраміда”, 2016. 404 с.

476. Висоцька В. А. Технології електронної комерції та Інтернет-маркетингу. Saarbrücken: LAP, 2018. 285 с.

477. Vysotska V., Lytvyn V. Web resources processing based on ontologies. Saarbrücken: LAP, 2018. 232 с.

478. Vysotska V., Shakhovska N. Information technologies of gamification for training and recruitment : monograph. Saarbrücken: LAP Lambert Academic Publishing, 2018. 248 p.

479. Висоцька В. А., Досин Д. Г., Микіч Х. І., Завущак І. І., Рибчак З. Л. Методи та засоби функціонування систем підтримки прийняття рішень на основі онтологій: монографія. Львів: Новий світ – 2000, 2019. 334 с.

480. Lytvyn V., Vysotska V., Peleshchak I., Rishnyak I., Peleshchak R. Time dependence of the output signal morphology for nonlinear oscillator neuron based on Van der Pol model. International Journal of Intelligent Systems and Applications. 2018.Vol. 10, № 4. P. 8–17.

481. Peleshchak R., Peleshchak I., Vysotska V. Methods for recognizing multispectral images based on neural networks: monograph. Beau Bassin: LAP Lambert Academic Publishing, 2020. 153 с.

482. Пелещак Р. М., Литвин В. В., Пелещак І. Р., Висоцька В. А. Розробка штучної нейронної мережі з осциляторними нейронами для розпізнавання спектральних образів. Вісник НУЛП. 2020. Вип. 7. С. 16–23.

483. РПелещак. М., Литвин В. В., Пелещак І. Р., Висоцька В. А., Черняк О. І. Побудова оптимізованої багатошарової нейронної мережі в межах нелінійної моделі узагальненої похибки. Вісник НУЛП 2021. Вип. 9. С. 53–60.

484. Литвин В. В., Пелещак Р. М., Висоцька В. А. Глибинне навчання. Львів: Видавництво Львівської політехніки, 2021. 264 с.

485. Gudivada Venkat N., Rao Dhana, Raghavan Vijay V. Big data driven natural language processing research and applications. Handbook of Statistics. Elsevier, 2015. P. 203-238.

486. Kim Y., et al. Application of natural language processing and text-mining of big-data to engineering-procurement-construction (EPC) bid and contract documents. Data Science and Machine Learning Applications. 2020. P. 123-128.

487. Agerri R., et al. Big data for natural language processing: a streaming approach. Knowledge-Based Systems. 2015. Vol. 79. P. 36-42.

488. Alblawi Amal S., Alhamed Ahmad A. Big data and learning analytics in higher education: Demystifying variety, acquisition, storage, NLP and analytics. Big Data and Analytics (ICBDA). IEEE, 2017. P. 124-129.

489. Castillo-Zúñiga Iván, et al. Internet data analysis methodology for cyberterrorism vocabulary detection, combining techniques of big data analytics, NLP and semantic web. International Journal on Semantic Web and Information Systems (IJSWIS). 2020. Vol. 16.1. P. 69-86.

490. Lytvyn V., Vysotska V., Veres O., Brodyak O., Oryshchyn O. Big Data analytics ontology. Технологічний аудит та резерви виробництва. 2018. Vol. 1, № 2 (39). С. 16–27.

491. Lytvyn V., Vysotska V., Veres O. Ontology of big data analytics. MEST Journal. 2018. Vol. 6, № 1. P. 41–60.

492. Tchynetskyi S., Peleshchak R., Peleshchak I., Vysotska V. A neural network development for multispectral images recognition. CSIT, 22-25 Sept., Lviv, Ukraine. 2021. Vol. 2. P. 278–284.

493. Ivanchyshyn D., Vysotska V., Albota S. The film script generation analysis based on the fiction book text using machine learning. CSIT, 22-25 Sept., Lviv, Ukraine. 2021. Vol. 2. P. 68–80.

494. Sartiukova A., Peleshchak R., Peleshchak I., Vysotska V. The multiclass classification of objects based on multispectral images recognition. CSIT, 22-25 Sept., Lviv, Ukraine. 2021. Vol. 1. P. 52–60.

495. Voloshynskyi O., Vysotska V., Bublyk M. Cardiovascular disease prediction based on machine learning technology. CSIT, 22-25 Sept., Lviv, Ukraine. 2021. Vol. 1. P. 69–75.

496. Mykytiuk A., Vysotska V., Albota S. Spam filtration system with the use of machine learning technology. CSIT, 22-25 Sept., Lviv, Ukraine. 2021. Vol. 1. P. 124–130.

497. Zanchak M., Vysotska V., Albota S. The sarcasm detection in news headlines based on machine learning technology. CSIT, 22-25 Sept., Lviv, Ukraine. 2021. Vol. 1. P. 131–137.

498. Voloshyn S., Peleshchak R., Peleshchak I., Vysotska V. Big data analysis for multispectral images recognition based on deep learning. CSIT, 22-25 Sept., Lviv, Ukraine. 2021. Vol. 1. P. 160–170.

499. Lytvyn V., Vysotska V., Bublyk M., Gozhyj A., Schuchmann V. solving scheduling issues methods analysis in computational Grid. CSIT, 22-25 Sept., Lviv, Ukraine. 2021. Vol. 1. P. 267–273.

500. Андруник В. А., Висоцька В. А., Пасічник В. В., Чирун Л. Б., Чирун Л. В. Чисельні методи в комп'ютерних науках. Львів: Новий Світ – 2000, 2017. Т. 1. 470 с.

501. Андруник В. А., Висоцька В. А., Пасічник В. В., Чирун Л. Б., Чирун Л. В. Чисельні методи в комп'ютерних науках. Львів: Новий Світ – 2000, 2017. Т. 2. 536 с.

502. Ришковець Ю. В., Висоцька В. А. Алгоритмізація та програмування. Ч. 1. Львів: Новий Світ – 200, 2018. 337 с.

503. Ришковець Ю. В., Висоцька В. А. Алгоритмізація та програмування. Ч. 2. Львів: Новий Світ – 2000, 2018. 316 с.

504. Висоцька В. А., Литвин В. В., Лозинська О. В. Дискретна математика: практикум (Збірник задач з дискретної математики), Львів: Новий Світ – 2000, 2019. 575 с.

505. Висоцька В. А., Оборська О. В. Python: алгоритмізація та програмування: навчальний посібник. Львів: Новий Світ – 2000, 2020. 516 с. ISBN 978-617-7519-74-3

506. Бенджамин Б., Ребекка Б., Тони О. Прикладной анализ текстовых данных на Python. Машинное обучение и создание приложений обработки естественного языка. Издательский дом Питер, 2018.
507. Jurafsky D., Martin J. H. Deep Learning Architectures for Sequence Processing. URL: https://web.stanford.edu/~jurafsky/slp3/9.pdf
508. Jurafsky D., Martin J. H. Naive Bayes and Sentiment Classification. URL: https://web.stanford.edu/~jurafsky/slp3/4.pdf
509. Jurafsky D., Martin J. H. Logistic Regression. URL: https://web.stanford.edu/~jurafsky/slp3/5.pdf
510. Jurafsky D., Martin J. H. Neural Networks and Neural Language Models. https://web.stanford.edu/~jurafsky/slp3/7.pdf
511. Bengfort B., Bilbro R., Ojeda T. Applied text analysis with python: Enabling language-aware data products with machine learning. O'Reilly Media, Inc., 2018.
512. Khan Wahab, et al. A survey on the state-of-the-art machine learning models in the context of NLP. Kuwait journal of Science. 2016. Vol. 43.4.
513. François T., Miltsakaki, E. Do NLP and machine learning improve traditional readability formulas? Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations. 2012. p. 49-57.
514. Socher R., Bengio Y. Manning, C.D. Deep learning for NLP. Tutorial Abstracts of ACL. 2012. P. 5-5.
515. AlchemyAPI. URL: http://www.alchemyapi.com/
516. Thomson R. Artificial intelligence research. URL: https://www.thomsonreuters.com/en/artificial-intelligence/research.html
517. IBM. Watson Natural Language Understanding Replaces Alchemy Language. URL: https://www.ibm.com/watson/services/alchemy-language-migration/
518. The Alchemist Within. URL: https://www.thealchemistwithin.com
519. Batrinca B., Treleaven P. C. Social media analytics: a survey of techniques, tools and platforms. Ai & Society, 2015, 30.1: 89-116.
520. Cvetković L., Milašinović B., Fertalj K. A tool for simplifying automatic categorization of scientific paper using Watson API. Information and Communication Technology, Electronics and Microelectronics, 2017. p. 1501-1505.
521. Goldberg Simon B., et al. Machine learning and natural language processing in psychotherapy research: Alliance as example use case. Journal of counseling psychology, 2020, 67.4: 438.
522. Aylien. Global news as structured data feeds with AYLIEN News API. URL: https://aylien.com/
523. Jalal Mona, et al. Performance comparison of crowdworkers and nlp tools on named-entity recognition and sentiment analysis of political tweets. arXiv preprint arXiv:2002.04181, 2020.
524. Mahi Md, et al. Sentrac: A novel real time sentiment analysis approach through twitter cloud environment. In: Advances in Electrical and Computer Technologies. Springer, 2020. p. 21-32.
525. Dale Robert. NLP meets the cloud. Natural Language Engineering, 2015, 21.4: 653-659.
526. Vieira N., Simoes A., Carvalho N. R. SplineAPI: A REST API for NLP Services. In: International Symposium on Languages, Applications and Technologies. Springer, Cham, 2015. p. 205-215.
527. Di Martino Beniamino, et al. A Q&A tool to produce an Ad-Hoc OpenAPI specification to identify equivalent REST Api Services. IEEE International Conference on Smart Computing (SMARTCOMP). IEEE, 2018. p. 375-380.
528. Jain Harshit. Caprecipes: a context-aware personalized recipes recommender for healthy and smart living. 2018. PhD Thesis. URL: http://dspace.library.uvic.ca/handle/1828/9583
529. Alarcon R., et al. REST web service description for graph-based service discovery. International Conference on Web Engineering. Springer, Cham, 2015. p. 461-478.

530. Aylien. AI-powered News API. URL: https://aylien.com/product/news-api
531. Syvokon O., Nahorna O. UA-GEC: Grammatical error correction and fluency corpus for the ukrainian language. arXiv preprint arXiv:2103.16997, 2021.
532. Kotsyba N., Mykulyak A., Shevchenko I. UGTag: morphological analyzer and tagger for the Ukrainian language. Proceedings of the international conference Practical Applications in Language and Computers (PALC 2009). URL: http://cynk.rockmetal.art.pl/~natko/papers/PALC-2009_UGTag.pdf
533. Livinska H., Makarevych O. Feasibility of Improving BERT for Linguistic Prediction on Ukrainian corpus. In: CEUR Workshop Proceedings. 2020. p. 552-561.
534. Висоцька В. А., Наум О. М. Порівняння складності автоматичного опрацювання англійських та українських текстів з врахуванням семантики та синтаксису природних мов. Вісник НУЛП 2017. № 872. С. 149–162.
535. Vysotska V., Holoshchuk S., Holoshchuk R. A comparative analysis for English and Ukrainian texts processing based on semantics and syntax approach. CEUR Workshop Proceedings. 2021. Vol. 2870. P. 311–356. E-ISSN: 1613-0073.
536. Huck M., Dutka D., Fraser A. Cross-lingual annotation projection is effective for neural part-of-speech tagging. Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects. 2019. p. 223-233.
537. Romanyshyn M. Rule-based sentiment analysis of Ukrainian reviews. International Journal of Artificial Intelligence & Applications, 2013, 4.4: 103.
538. Lande D., Dmytrenko O. Using Part-of-Speech Tagging for Building Networks of Terms in Legal Sphere. In: CEUR Workshop Proceedings. 2021. p. 87-97.
539. Babych B., Sharoff S. Ukrainian part-of-speech tagger for hybrid MT: Rapid induction of morphological disambiguation resources from a closely related language. In: Fifth Workshop on Hybrid Approaches to Translation (HyTra). Universitat Pompeu Fabra, 2016.
540. Hamon T., Grabar N. Unsupervised acquisition of morphological resources for Ukrainian. In: Computational linguistics andintelligent systems (COLINS 2017). National Technical University «KhPI», 2017.
541. Kotov M. NLP resources for a rare language morphological analyzer: danish case. In: Computational linguistics andintelligent systems (COLINS 2017). National Technical University «KhPI», 2017.
542. Romanyshyn N. Application of computer technologies in conceptual analysis. In: 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). IEEE, 2018. p. 55-57.
543. Bobichev V., Kanishcheva O., Cherednichenko O. Sentiment analysis in the Ukrainian and Russian news. Electrical and Computer Engineering (UKRCON). IEEE, 2017. P. 1050-1055.
544. Mykhaylenko V. Exploring English-Ukrainian contrastive phraseology. Науковий вісник Ужгородського університету. Серія Філологія, 2019, 2.42: 68-72.
545. Covington M. A. Concise encyclopedia of syntactic theories Ed. by Keith Brown and Jim Miller, and: Concise encyclopedia of philosophy of language Ed. by Peter V. Lamarque. Language, 2000, 76.1: 231-232.
546. Kotsyba N., Moskalevskyi B. Syntactic and morphological ambiguity of the deverbal nouns' arguments in Ukrainian and ways of its resolution. Prace Filologiczne, 2018, 193-210.
547. Прокоф'єва К. О., et al. Застосування методів НЛП із метою подолання впливу посттравматичного синдрому на особистість. Сучасні інформаційні технології у сфері безпеки та оборони, 2010, 2: 47-52.
548. Волчек Д. Г., Романов А. А. Создание и обучение онтологий на основе анализа контекста и метаданных слабоструктурированного контента. Экономика: вчера, сегодня, завтра, 2020, 10.1-1: 303-312.

549. Малахова В. Л. Основные этапы формирования прагма-семантического смысла и методы его анализа. Вестник Самарского университета. История, педагогика, филология, 2021, 27.4: 114-121.

550. Савостьянов А. Н., Пальчунов Д. Е. Когнитивные исследования и нейролингвистика: современное состояние и перспективы дальнейших исследований. Вестник Томского государственного унив., 2013, 368: 133-140.

551. Вилинбахова Е. Л. «Как говорится, статья есть статья»: некоторые аспекты функционирования тавтологий в коммуникации. Компьютерная лингвистика и интеллектуальные технологии, 2016, 15 (22): 817-829.

552. Kano Y., et al. U-Compare: A modular NLP workflow construction and evaluation system. IBM Journal of Research and Development, 2011, 55.3: 11: 1-11: 10.

553. Pollak S., Vavpetic A., Kranjc J., Lavrac N., Vintar S. NLP workflow for on-line definition extraction from English and Slovene text corpora. In: KONVENS. 2012. p. 53-60.

554. Mcentire Robin, et al. Application of an automated natural language processing (NLP) workflow to enable federated search of external biomedical content in drug discovery and development. Drug discovery today, 2016, 21.5: 826-835.

555. De Castilho R. E., Gurevych I. A broad-coverage collection of portable NLP components for building shareable analysis pipelines. Open Infrastructures and Analysis Frameworks for HLT. 2014. p. 1-11.

556. Chard K., Russell M., Lussier Y. A., Mendonça E. A., Silverstein J. C. A cloud-based approach to medical NLP. In: AMIA Annual Symposium Proceedings. American Medical Informatics Association, 2011. p. 207.

557. Chiru C.-G., Rebedea T., Ciotec S. Comparison between LSA-LDA-Lexical Chains. In: WEBIST (2). 2014. p. 255-262.

558. Louwerse M., Cai Z., Hu X., Ventura M., Jeuniaux P. Cognitively inspired NLP-based knowledge representations: Further explorations of Latent Semantic Analysis. Int.Journal on Artificial Intelligence Tools, 2006, 15.06: 1021-1039.

559. Jung N., Lee G. Automated classification of building information modeling (BIM) case studies by BIM use based on natural language processing (NLP) and unsupervised learning. Advanced Engineering Informatics, 2019, 41: 100917.

560. Moser J. R., Gütl C., Liu W. Refined distractor generation with LSA and stylometry for automated multiple choice question generation. In: Australasian Joint Conference on Artificial Intelligence. Springer, Berlin, Heidelberg, 2012. p. 95-106.

561. Ландэ Д.В. Основы интеграции информационных потоков: Монография. К.: Инжиниринг, 2006. 240 с.

562. Ланде Д.В., Фурашев В.М. Основи інформаційного і соціально-правового моделювання: монографія. К.: ТОВ "ПанТот", 2012. 144 с. ISBN 978-966-1531-22-1

563. Фурашев В.Н., Ландэ Д.В., Брайчевский С.М. Моделирование информационно-электоральных процессов: Монография. К.: НИЦПИ АпрН Украины, 2007. 182 с. ISBN 978-966-96927-2-6

564. Ландэ Д.В., Фурашев В.Н., Брайчевский С.М., Григорьев А.Н. Основы моделирования и оценки электронных информационных потоков: Монография. К.: Инжиниринг, 2006. 176 с. ISBN 966-95147-6-2

565. Ланде Д. Елементи комп'ютерної лінгвістики в правовій інформатиці. К.: НДІІП НАПрН України, 2014.

566. Додонов А.Г., Ландэ Д.В., Цыганок В.В., Андрейчук О.В., Каденко С.В., Грайворонская А.Н. Распознавание информационных операций: Монография. К.: Инжиниринг, 2017. 282 с. ISBN 978-966-2344-60-8

567. Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика. М.: МИЭМ, 2011. 272 с.

568. Додонов О.Г., Ланде Д.В., Путятін В.Г. Інформаційні потоки в глобальних комп'ютерних мережах - К: Наукова думка, 2009, - 295 с. ISBN 978-966-00-0973-9

569. Ланде Д.В., Субач І.Ю., Бояринова Ю.Є. Основи теорії і практики інтелектуального аналізу даних у сфері кібербезпеки: навчальний посібник. Київ: ІСЗЗІ КПІ ім. Ігоря Сікорського, 2018. 300 с.

570. Ландэ Д.В., Снарский А.А., Безсуднов И.В. Интернетика: Навигация в сложных сетях: модели и алгоритмы - М.: Либроком (Editorial URSS), 2009. 264 с. ISBN 978-5-397-00497-8

571. Ландэ Д.В. Поиск знаний в Internet. Профессиональная работа. М.: Диалектика, 2005. 272 с.

572. Ланде Д.В., Фурашев В.М., Григор'єв О.М. Програмно-апаратний комплекс інформаційної підтримки прийняття рішень: Науково-методичний посібник. К.: Іжиніринг, 2006. 48 с. ISBN 966-95147-4-6

573. Досин Д. Г., Висоцька В. А., Литвин В. В. Побудова системи підтримки прийняття рішень на базі адаптивної онтології. Обчислювальні методи і системи перетворення інформації: зб. пр. V-ї наук.-техн. конф., (Львів, 4–5 жовтня 2018 р.). Львів, 2018. С. 135–138.

574. Литвин В. В., Висоцька В. А., Досин Д. Г., Гірняк М. Г. Розроблення методів та засобів побудови інтелектуальних систем опрацювання інформаційних ресурсів з використанням онтологічного підходу. Вісник Національного університету «Львівська політехніка». 2015. № 832. С. 295–314.

575. Буров Є. В. Методи та засоби побудови програмних систем на основі онтологічних моделей задач : автореферат дисертації на здобуття наукового ступеня доктора технічних наук : 01.05.03 – математичне та програмне забезпечення обчислювальних машин і систем / Євген Вікторович Буров , Міністерство освіти та науки України, Національний університет «Львівська політехніка». – Львів, 2015. – 42 с. – Бібліографія: с. 34–37.

576. Буров Є. В. Методи та засоби побудови програмних систем на основі онтологічних моделей задач : дисертація на здобуття наукового ступеня доктора технічних наук : 01.05.03 – математичне та програмне забезпечення обчислювальних машин і систем / Євген Вікторович Буров , НУ «Львівська політехніка». – Львів, 2015. – 400 с.

577. Верес О. М., Оливко Р. М. Класифікація методів аналізу Великих даних. Вісник Національного університету «Львівська політехніка». 2017. № 872. С. 84–92.

578. Яценко А. О., Досин Д. Г. Порівняння ефективності алгоритмів планування, реалізованих для Марківської моделі клієнта пошукової системи. Відбір і обробка інформації. 2013. Вип. 38 (114). С. 118–124.

579. Chen J., Dosyn D., Lytvyn V., Sachenko A. Smart data integration by goal driven ontology learning. Advances in Intelligent Systems and Computing. 2017. Vol. 529. P. 283–292.

580. Досин Д. Г. Архітектура системи оцінювання пертинентності, що базується на навчанні онтології планування у вибраній предметній області. Відбір і обробка інформації. 2018. № 46 (122). С. 61–67.

581. Басюк Т. М., Досин Д. Г., Литвин В. В. Онтологічний інжиніринг. Нац. ун-т "Львів. політехніка". Львів : Вид-во Львів. політехніки, 2017. 222 с.

582. Досин Д. Г. Пертинентність інформації як цінність знань для інтелектуального агента. Вісник Національного університету «Львівська політехніка». Серія: Інформаційні системи та мережі. 2018. № 901. С. 111–117.

583. Досин Д. Г., Даревич Р. Р., Литвин В. В., Никитюк Н. В. Метод оцінювання подібності текстових документів, доповнених контекстом з онтології. Відбір і обробка інформації. 2007. Вип. 27 (103). С. 109–115.

584. Клифтон Б. Google Analytics: профессиональный анализ посещаемости веб-сайтов . – М. : Вильямс, 2009.

585. Берко А. Системи електронної контент-комерції / А. Берко, В. Висоцька, В. Пасічник. – Л. : Вид-во Нац. ун-ту “Львівська політехніка”, 2009. – 612 с.

586. Висоцька В. А. Методи і засоби опрацювання інформаційних ресурсів в системах електронної контент-комерції : автореферат дисертації на здобуття наукового ступеня кандидата технічних наук : 05.13.06 – інформаційні технології / Вікторія Анатоліївна Висоцька , НУ «Львівська політехніка». Львів, 2014. 27 с.

587. Висоцька Вікторія Анатоліївна. Методи і засоби опрацювання інформаційних ресурсів в системах електронної контент-комерції.- Дисертація на здобуття наукового ступеня кандидата технічних наук : 05.13.06 – інформаційні технології / Вікторія Анатоліївна Висоцька , НУ«Львівська політехніка». Львів, 2014. 240 с.
588. Алєксєєва К. А., Берко А. Ю., Висоцька В. А. Технологія управління комерційним web-ресурсом на основі нечіткої логіки. Радіоелектроніка. Інформатика. Управління. 2015. № 3 (34). С. 71–79.
589. Алєксєєва К. А., Берко А. Ю., Висоцька В. А. Управління Web-ресурсами за умов невизначеності. Технологічний аудит та резерви виробництва. 2015. № 2 (2). С. 4–7.
590. Алєксєєва К. А., Берко А. Ю., Висоцька В. А. Особливості процесу управління web-ресурсом комерційного контенту на основі нечіткої логіки. Вісник НУ "Львівська політехніка". 2015. № 826. С. 201–211.
591. Алєксєєва К. А., Берко А. Ю., Висоцька В. А. Інформаційна технологія управління Web-ресурсом на основі нечіткої логіки. Вісник НУ "Львівська політехніка". 2015. № 829. С. 7–28.
592. Алєксєєва К. А., Берко А. Ю., Висоцька В. А. Аналіз процесу опрацювання web-ресурсу інформаційного продукту на основі нечіткої логіки та контент-аналізу. Вісник Національного університету "Львівська політехніка". Серія: Комп'ютерні науки та інформаційні технології : зб. наук. пр. 2016. № 843. С. 122–134.
593. Берко А. Ю. Методи та засоби інтеграції даних у відкритих інформаційних системах : автореферат дисертації на здобуття наукового ступеня доктора технічних наук : 01.05.03 – математичне та програмне забезпечення обчислювальних машин і систем / Андрій Юліанович Берко , НУ "Львівська політехніка". – Львів, 2011. – 36 с. – Бібліографія: с. 27–33 (66 назв).
594. Берко Андрій Юліанович. Методи та засоби інтеграції даних у відкритих інформаційних системах : дисертація на здобуття наукового ступеня доктора технічних наук : 01.05.03 – математичне та програмне забезпечення обчислювальних машин і систем / Андрій Юліанович Берко , НУ "Львівська політехніка". – Львів, 2011.
595. Mike Loukides, What is data science? 2010. URL: https://oreil.ly/2GJBEo
596. Yelpy Insights. URL: https://blog.yelp.com/news/yelpy-insights/
597. Market Watch. 2018. URL: https://on.mktw.net/2suTk24
598. Косолапов К. Введение в рекомендательные системы. URL: https://habr.com/ru/post/476222/
599. Создание успешной SEO-стратегии для видео на YouTube. https://www.affde.com/ru/video-seo-strategies.html
600. Тамм Ян-Мартін. Рекомендательные системы: идеи, подходы, задачи. URL: https://habr.com/ru/company/jetinfosystems/blog/453792/
601. Ройзнер М. Как работают рекомендательные системы. URL: https://habr.com/ru/company/yandex/blog/241455/
602. Как очистить историю просмотра и приостановить ее запись в аккаунте YouTube. URL: https://support.google.com/youtube/answer/6342839?hl=ru&co=GENIE.Platform%3DAndroid
603. Как настроить события пикселя Facebook. URL: https://amzko.pro/kak-nastroit-sobytiya-pikselya-facebook/
604. Искусственный интеллект в Salesforce. URL: https://zvyazok.com/iskusstvennyy-intellekt-v-salesforce/
605. Топ тегов на Stack Overflow с 2010 по 2017 год. URL: https://tproger.ru/articles/stackoverflow-top-2010-2017/
606. Stack Overflow. URL: https://habr.com/ru/company/productivity_inside/blog/553890/
607. StackOverflow: 560 млн показов в месяц, 25 серверов. URL: https://habr.com/ru/post/230677/
608. Google Smart Reply. URL: https://developers.google.com/ml-kit/language/smart-reply
609. All-in-one email outreach platform for your growth teams. URL: https://smartreach.io/

610. MediaSapienS. ШВидкі відповіді у Gmail допомагають роботу краще розуміти людей. Чому це погано? URL: https://ms.detector.media/it-kompanii/post/22096/2018-11-19-shvydki-vidpovidi-u-gmail-dopomagayut-robotu-krashche-rozumity-lyudey-chomu-tse-pogano/
611. Siri. URL: https://www.apple.com/ru/siri/
612. Alex - Your personal assistant at work. URL: https://devpost.com/software/alex-your-personal-assistant-at-work
613. Hi I'm Alex, your INSEAD virtual assistant. URL: https://www.insead.edu/dialogflowsdd
614. Google Assistant. URL: https://assistant.google.com/intl/ru_ru/
615. Cortana. URL: https://www.microsoft.com/en-us/cortana
616. Siri та Netflix «заговорять» українською. URL: https://nzl.theukrainians.org/siri-ta-netflix-ukrayinska.html
617. Textra SMS. URL: https://play.google.com/store/apps/details?id=com.textra&hl=uk&gl=US
618. SMS-мессенджер для android-устройств. URL: https://overclockers.ru/lab/show/84799_4/vybiraem-sms-messendzher-dlya-android-ustrojstv-google-messages-textra-i-mood-messenger
619. iMessage. URL: https://support.apple.com/ru-ru/HT206906
620. Reverb. URL: https://reverb.com/
621. Wordnik. URL: https://www.wordnik.com/
622. ChatBot Slack. URL: https://slack.com/apps/A7FTEHPEG-chatbot?tab=more_info, https://slack.com/intl/ru-ru/
623. El Abdouli Abdeljalil, Hassouni Larbi, Anoun Houda. Mining tweets of Moroccan users using the framework Hadoop, NLP, K-means and basemap. Intelligent Systems and Computer Vision (ISCV). IEEE, 2017. p. 1-7.
624. Thavareesan S., Mahesan S. Sentiment analysis in Tamil texts: a study on machine learning techniques and feature representation. Industrial and Information Systems (ICIIS). IEEE, 2019. p. 320-325.
625. Al-Azzawy D. S., Al-Rufaye F. M. L. Arabic words clustering by using K-means algorithm. New Trends in Information & Communications Technology Applications (NTICT). IEEE, 2017. p. 263-267.
626. Spiteri Janica. Automatic crime information gathering and data analytics from online news reports. 2020. Bachelor's Thesis. University of Malta.
627. Jain A., Chakrabarti B., Upmon Y., Rout, J. K. Exploring Historical Stock Price Movement from News Articles Using Knowledge Graphs and Unsupervised Learning. In Intelligent Data Engineering and Analytics, 2022. pp. 511-519.
628. Hong Y., Xie H., Bhumbra G., Brilakis I. Comparing Natural Language Processing Methods to Cluster Construction Schedules. Journal of Construction Engineering and Management, 2021. 147(10), 04021136.
629. Lebret Rémi, Collobert Ronan. Word emdeddings through hellinger pca. arXiv preprint arXiv:1312.5542, 2013.
630. Christian H., Agus M. P., Suhartono D. Single document automatic text summarization using term frequency-inverse document frequency. ComTech: Computer, Mathematics and Engineering Applications, 2016. 7(4), 285-294.
631. Thara S., Sidharth S. Aspect based sentiment classication: Svd features. In 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI) 2017, September (pp. 2370-2374). IEEE.
632. Zinnatullin Vadim, Koledin Sergey. Analysis of scientists work directions based on natural language processing and clustering. CEUR Workshop Proceedings. 2020. p. 57-61.
633. Walia H., Rana A., Kansal V. A Supervised Approach on Gurmukhi Word Sense Disambiguation Using K-NN Method. Cloud Computing, Data Science & Engineering (Confluence). IEEE, 2018. p. 743-746.
634. Aborisade O., Anwar M. Classification for authorship of tweets by comparing logistic regression and naive bayes classifiers. IEEE International Conference on Information Reuse and Integration (IRI). IEEE, 2018. p. 269-276.

635. Rameshbhai C. Jashubhai, Paulose Joy. Opinion mining on newspaper headlines using SVM and NLP. International Journal of Electrical and Computer Engineering (IJECE), 2019, 9.3: 2152-2163.
636. Grönroos Stig-Arne, Virpioja Sami, Kurimo Mikko. Morfessor EM+ Prune: Improved subword segmentation with expectation maximization and pruning. arXiv preprint arXiv:2003.03131, 2020.
637. Giorgos Orphanos, et al. Decision trees and NLP: A case study in POS tagging. Proceedings of annual conference on artificial intelligence (ACAI). 1999.
638. Mou L., Meng Z., Yan R., Li G., Xu Y., Zhang L., Jin Z. How transferable are neural networks in nlp applications? arXiv preprint arXiv:1603.06111, 2016.
639. Rajman Martin, Besançon Romaric. Text mining: natural language techniques and text mining applications. Data mining and reverse engineering. Springer, Boston, MA, 1998. p. 50-64.
640. Puri Shalini. A fuzzy similarity based concept mining model for text classification. arXiv preprint arXiv:1204.2061, 2012.
641. Alcantud J. C. R., Varela G., Santos-Buitrago B., Santos-García G., Jiménez M. F. Analysis of survival for lung cancer resections cases with fuzzy and soft set theory in surgical decision making. PloS one, 2019. 14(6), e0218283.
642. Teng Zhi, Ren Fuji, Kuroiwa Shingo. Emotion recognition from text based on the rough set theory and the support vector machines. Natural Language Processing and Knowledge Engineering. IEEE, 2007. p. 36-41.
643. Yu S., et al. Mil-vt: Multiple instance learning enhanced vision transformer for fundus image classification. Medical Image Computing and Computer-Assisted Intervention. Springer, 2021. p. 45-54.
644. Kamath U., Liu J., Whitaker J. Deep learning for NLP and speech recognition. Cham, Switzerland: Springer, 2019.
645. Honkela Timo, Hyvärinen Aapo, Väyrynen Jaakko J. WordICA—emergence of linguistic representations for words by independent component analysis. Natural Language Engineering, 2010, 16.3: 277-308.
646. Wang Yikai, Li Weijian. Blind signal decomposition of various word embeddings based on join and individual variance explained. arXiv preprint arXiv:2011.14496, 2020.
647. Barlier Merwan, Laroche Romain, Pietquin Olivier. Learning dialogue dynamics with the method of moments. IEEE Spoken Language Technology Workshop (SLT). IEEE, 2016. p. 98-105.
648. Khan Z., Iltaf N., Afzal H., Abbas H. Enriching non-negative matrix factorization with contextual embeddings for recommender systems. Neurocomputing, 2020. 380, 246-258.
649. Nesi Paolo, Pantaleo Gianni, Tenti Marco. Geographical localization of web domains and organization addresses recognition by employing natural language processing, Pattern Matching and clustering. Engineering Applications of Artificial Intelligence, 2016, 51: 202-211.
650. Kashyap V., Ramakrishnan C., Thomas C., Sheth A. TaxaMiner: an experimentation framework for automated taxonomy bootstrapping. International Journal of Web and Grid Services, 2005. 1(2), 240-266.
651. Anita R., Subalalitha C. N. An approach to cluster Tamil literatures using discourse connectives. IEEE 1st International Conference on Energy, Systems and Information Processing (ICESIP). IEEE, 2019. p. 1-4.
652. Nakache Didier, Metais Elisabeth, Timsit Jean François. Evaluation and NLP. In: International Conference on Database and Expert Systems Applications. Springer, Berlin, Heidelberg, 2005. p. 626-632.
653. Tikhonova M., Gavrishchuk A. NLP methods for automatic candidate's cv segmentation. International Conference on Engineering and Telecommunication (EnT). IEEE, 2019. p. 1-5.
654. Li X., Sun X., Meng Y., Liang J., Wu F., Li J. Dice loss for data-imbalanced NLP tasks. arXiv preprint 2019. arXiv:1911.02855.

655. Ryu Keun Ho. BioBERT Based Efficient Clustering Framework for Biomedical Document Analysis. Genetic and Evolutionary Computing: Proceedings of the Fourteenth International Conference on Genetic and Evolutionary Computing, October 21–23, 2021, Jilin, China. Springer Nature. p. 179.
656. Rayzmann N., Aponso H., Markgraf C. Y., Chappell P. E. SUN-238 Estrogen Modulates Expression Levels of Gonadotropin-Releasing Hormone Receptor (GNRHR) in Immortalized Kisspeptin Neurons in Vitro. Journal of the Endocrine Society, 2020. 4(Supplement_1), SUN-238.
657. Tan Y., Bacchi S., Casson R. J., Selva D., Chan W. Triaging ophthalmology outpatient referrals with machine learning: a pilot study. Clinical & experimental ophthalmology, 2020. 48(2), 169-173.
658. Kim Ju-Ri. Using Markedness Principle for Abstraction of Dependency Relations of Natural Languages. Eurasian Journal of Applied Linguistics, 2021, 7.2: 58-67.
659. Heo D., Lee W., Jung B., Lee J. H. Quality estimation using dual encoders with transfer learning. Proceedings of the Sixth Conference on Machine Translation. 2021. p. 920-927.
660. Ayre K., Bittar A., Kam J., Verma S., Howard L. M., Dutta R. Developing a natural language processing tool to identify perinatal self-harm in electronic healthcare records. PloS one, 2021. 16(8), e0253809.
661. Jungmann Florian, et al. Towards data-driven medical imaging using natural language processing in patients with suspected urolithiasis. International Journal of Medical Informatics, 2020, 137: 104106.
662. 256 тисяч слів. Топ-7 фактів про українську мову. URL: https://bigkyiv.com.ua/256-tisyach-sliv-top-7-faktiv-pro-ukrayinsku-movu/
663. Сеник М. Проект: Статичне дерево закінчень для української мови. URL: http://www.senyk.poltava.ua/projects/ukr_stemming/ukr_endings.html
664. Halliday M.A.K. Categories of the Theory of Grammar. Word. 17(3). pp241-92. Reprinted in Full in On Grammar: Volume 1 of the Collected Works of M.A.K. Halliday. London and New York: Continuum. 1961. p 40
665. Halliday M.A.K. Categories of the Theory of Grammar. Word. 17(3). pp241-92. Reprinted in Full in On Grammar: Volume 1 of the Collected Works of M.A.K. Halliday. London and New York: Continuum. 1961. p 52
666. Halliday M.A.K. Categories of the Theory of Grammar. Word. 1961. 17(3). pp241-92. Reprinted in Full in On Grammar: Volume 1 of the Collected Works of M.A.K. Halliday. London and New York: Continuum.
667. Halliday M.A.K. Systemic Grammar and the Concept of a “Science of Language”. Waiguoyu (Journal of Foreign Languages), 1992. No. 2 (General Series No. 78), pp. 1-9.
668. Halliday M.A.K. Systemic Background. In "Systemic Perspectives on Discourse, Vol. 1: Selected Theoretical Papers" from the Ninth International Systemic Workshop, James D. Benson and William S. Greaves (eds). Ablex. Reprinted in Full in Volume 3 in The Collected Works of M.A.K. Halliday. London: Continuum. 1985. p. 186.
669. Halliday M.A.K. Introduction: How Big is a Language? On the Power of Language. In The Language of Science: Volume 5 in the Collected Works of M.A.K. Edited by J.J.Webster. London and New York: Continuum. p. xv. 2004.
670. Halliday M.A.K. Introduction: On the "architecture" of human language. In On Language and Linguistics. Volume 3 in the Collected Works of M.A.K. Halliday. Edited by Jonathan Webster. London and New York: Continuum. 2003.
671. Halliday M.A.K. Text as semantic choice in social contexts. Reprinted in full in Linguistic Studies of Text and Discourse. Volume 2 in the Collected Works of M.A.K. Halliday. London and New York: Continuum. 1977. pp. 23–81.
672. Halliday M.A.K. Introduction: How Big is a Language? On the Power of Language. In The Language of Science: Volume 5 in the Collected Works of M.A.K. Edited by J.J.Webster. London and New York: Continuum. p. xi. 2004.
673. Firth J.R. Selected Papers of J.R. Firth 1952-1959. London: Longman. 1968. p183.

674. Peterson James Lyle. Petri net theory and the modeling of systems. Prentice Hall PTR, 1981.
675. Питерсон Дж. Теория сетей Петри и моделирование систем. М. Мир, 1984. 264 с.
676. Liu H. C., Luan X., Li Z., Wu J. Linguistic Petri nets based on cloud model theory for knowledge representation and reasoning. IEEE Transactions on Knowledge and Data Engineering, 2017. 30(4), 717-728.
677. Liu H. C., You J. X., You X. Y., Su Q. Linguistic reasoning Petri nets for knowledge representation and reasoning. Transactions on Systems, Man, and Cybernetics: Systems, 2015. 46(4), 499-511.
678. Liu H. C., Luan X., Zhou M., Xiong Y. A new linguistic Petri net for complex knowledge representation and reasoning. IEEE Transactions on Knowledge and Data Engineering. 2020.
679. Srinivasan Padmini, Gracanin Denis. Approximate reasoning with fuzzy Petri nets. Second IEEE International Conference on Fuzzy Systems. IEEE, 1993. p. 396-401.
680. Пентус А. Е. Теория формальных языков: учеб. пособие / А. Е. Пентус, М. Р. Пентус. – М.: Изд-во ЦПИ при механикоматематическом ф-те МГУ, 2004. – 80 с.
681. Фомичев В. С. Формальные языки, грамматики и автоматы. URL: http://www.proklondike.com/ books/thproch/.
682. Попов Э. В. Общение с ЭВМ на естественном языке. М.: Наука, 1982. 360 с.
683. Мартыненко Б. К. Языки и трансляции. СПб.: Изд-во СПб. 350ун-та, 2008. 257 с.
684. Герасимов А. С. Лекции по теории формальных языков. URL: http://gasteach.narod.ru/au/tfl/tfl01.pdf.
685. Волкова И. А., Руденко Т. В. Формальные грамматики и языки. Элементы теории трансляции. М.: МГУ им. М. В. Ломоносова, 1999. 62 с.
686. Билыгаева Н. Ц. Теория алгоритмов, формальных языков, грамматик и автоматов. Улан-Удэ: Изд-во ВСГТУ, 2000. 51 с.
687. Апресян Ю. Д. Непосредственно составляющих метод. Лингвистический энциклопедический словарь под ред. В. Н. Ярцевой. М.: Советская энциклопедия, 1990. URL: http://tapemark.narod.ru/les/332a.html.
688. Апресян Ю. Д. Идеи и методы современной структурной лингвистики. М.: Просвещение, 1966. 305 с.
689. Анисимов А. Компьютерная лингвистика для всех: мифы, алгоритмы, язык. К.: Наукова думка, 1991. 208 с.
690. Анісімов А. В., Марченко О. О., Никоненко А. О. Алгоритмічна модель асоціативносемантичного контекстного аналізу текстів природною мовою. Пробл. програмув. 2008. № 2, 3. С. 379–384.
691. Гросс М., Лантен А. Теория формальных грамматик. Пер. с фр. И. А. Мельчука под ред. А. В. Гладкого. М.: Мир, 1971. 294 с.
692. Арсентьева Н. Г. О двух способах порождения предложений русского языка. Проблемы кибернетики. 1965. Вып. 14. С. 189–218.
693. Ингве В. Гипотеза глубины. Новое в лингвистике. М., 1965. Вып. IV. С. 126–138. 22.
694. Yngve V. H. A model and a hypothesis for language structure. Proceedings of American phylosophical society. 1960. 104, № 5. P. 444–466.
695. Varga D. Yngve's hypothesis and some problems of the mechanical analysis. Computational Linguistics. III. 1964. P. 47–74.
696. Yngve V. H. Random generation of English sentences. Teddington (National physical laboratory. Paper 6). 1961.
697. Падучева Е. В. О связях глубины по Ингве со структурой дерева починений. Научно-техническая информация. 1967. № 6. С. 38–43.
698. Шаров С. А. Средства компьютерного представления лингвистической информации. URL: http://www.ksu.ru/eng/science/ittc/vol000/002/.

699. Шрейдер Ю. А. Характеристики сложности структуры текста. Научно-техническая информация. № 7. 1966. С. 34–41.

700. Tesniere L. Elements de syntaxe structurale. P. 1959.

701. Postal P. M. Limitations of phrase structure grammars. The structure of language. Readings in the philosophy of language, Englewood Cliffs (N. J.). 1964. P. 137–151.

702. Hays D. G. Automatic language data processing. Computer applications in behavioral sciences, Englewood Cliffs (N. J.). 1962. P. 394–421.

703. Tosh L. W. Syntactic translation, The Hague. 1965.

704. Bar-Hillel Y., Shamir E. Finite state languages: formal representation and adequacy problems. Bulletin of the Research Council of Israel. 8F, № 3. 1960. P. 155–166.

705. Bobrow D. G. Syntactic analysis of English by computer – a survey. AFIPS conference proceedings. 24, Baltimore – London. 1963. P. 365–387.

706. Багмут А. Й. Порядок слів. Українська мова: Енцикл. К.: В-во "Укр. енциклопедія" ім. М. П. Бажана, 2007. С. 675–676.

707. Щербина Ю. М., Шестакевич Т. В., Висоцька В. А. Науковий напрям та навчальна дисципліна "Математична лінгвістика". Вісник Нац. ун-ту "Львівська політехніка" 2010. № 673. С. 384–392.

708. Щербина Ю. М. Предмет математичної лінгвістики. Вісник НУЛП. 2002. № 464. С. 340–349.

709. Висоцька В. А., Шестакевич Т. В. Генерування речень українською за допомогою породжувальних граматик. ISDMIT, Євпаторія. 27–31 травня 2012. С. 48–50.

710. Шестакевич Т. В., Висоцька В. А. Застосування породжувальних граматик для генерування речень українською мовою. Східно-Європейський журнал передових технологій. Харків, 2012. № 3/2 (57). С. 51–53.

711. Шульжук К. Синтаксис української мови. К.: Академія, 2004. 397 с.

712. Чепурна З. В. Трансформація порядку слів у простому реченні при перекладі з німецької мови українською. Філологічні науки : у 5 ч. Кіровоград: РВВ КДПУ ім. В. Винниченка, 2010. Вип. 89 (1). С. 232–236.

713. Гакман О. В. Генеративно-трансформаційна лінгвістика Н. Хомського як вираження його лінгвістичної філософії. Мультиверсум. Філософський альманах. К.: Центр духовної культури, 2005. № 45. С. 98–114.

714. Дарчук Н. П. Комп'ютерна лінгвістика (автоматичне опрацювання тексту). К.: ВПЦ "Київський університет", 2008. 351 с.

715. Демешко І. Типологія морфонологічних моделей у віддієслівному словотворенні сучасної української мови. Збірник наукових праць "Лінгвістичні студії". Розділ V. Словотвір: напрями, аспекти дослідження. Морфонологія. Донецьк, 2009. № 19. С. 162–167.

716. Зубков М. Українська мова: Універсальний довідник. К.: ВД "Школа", 2004. 496 с.

717. Любченко Т. П. Лексикографічні системи граматичного типу та їх застосування в засобах автоматизованого опрацювання мови: автореф. дис. канд. техн. наук: спец. 10.02.21. К., 2011. 19 с.

718. Марченко О. О. Алгоритми семантичного аналізу природномовних текстів: автореф. дис. на здобуття наук. ступеня канд. фіз. – мат. наук: спец. 01.05.01. О. О. Марченко. К., 2005. 15 с.

719. Партико З. В. Прикладна і комп'ютерна лінгвістика. Л.: Афіша, 2008. 224 с.

720. Потапова Г. М. Морфонологія віддієслівного словотворення (на матеріалі словотвірних гнізд з вершинами – дієсловами та віддієслівних словотвірних зон): Дис. канд. наук: 10.02.02. Г. М. Потапова. 2008. 19 с.

721. Русаченко Н. П. Морфонологічні процеси у словозміні та словотворі староукраїнської мови другої половини XVI – XVIII ст.: автореф. дис. на здобуття наук. ступеня канд. філол. наук: спец. 10.02.01. К., 2004. 24 с. URL: http://auteur.comeillemoliere.com/? p=history&m=comeille_moliere&l=rus.
722. Торосян О. М. Функціональні характеристики прислівників міри та ступеня в сучасній англійській мові: автореф. дис. на здобуття наук. ступеня канд. філол. наук. URL: http://disser.com.ua/contents/6712.html.
723. Туришева О. О. Порушення рамкової конструкції в сучасній німецькій мові: функціональний аспект, нормативний статус: автореф. дис. канд. філол. наук: спец. 10.02.04. Одеса, 2012. 20 с.
724. Український правопис. Ін-т мовознавства ім. О. О. Потебні НАН України, Ін-т укр. мови НАН України. К.: Наук. думка, 2007. 288 с.
725. Van Dijk Teun A. Pragmatic connectives. Journal of pragmatics, 1979, 3.5: 447-456.
726. Van Dijk Teun A. The semantics and pragmatics of functional coherence in discourse. Speech act theory: Ten years later, 1980, 49-65.
727. Van Dijk Teun A. Pragmatic macro-structures in discourse and cognition. CC, 1977, 77: 99-113.
728. Van Dijk Teun A. Action, action description, and narrative. New literary history, 1975, 6.2: 273-294.
729. Van Dijk Teun A. Pragmatics, presuppositions and context grammars. Pragmatics, 1976, 2.
730. Van Dijk Teun A. Context theory and the foundation of pragmatics. Studies in Pragmatics, 2008, 10: 1-13.
731. Van Dijk Teun A. Context and cognition: Knowledge frames and speech act comprehension. Journal of pragmatics, 1977, 1.3: 211-231.
732. Van Dijk Teun A. Towards an empirical pragmatics: some social psychological conditions of speech acts. Philosophica, 1981, 27.1: 127-138.
733. Van Dijk,Teun A. Macrostructures: An interdisciplinary study of global structures in discourse, interaction, and cognition. Routledge, 2019.
734. Van Dijk Teun A. Semantic macro-structures and knowledge frames in discourse comprehension. Cognitive processes in comprehension, 1977, 332: 3-31.
735. Buyssens Eric. Mise au point de quelques notions fondamentales de la phonologie. Cahiers Ferdinand de Saussure, 1949, 8: 37-60.
736. Buyssens Eric. Le signe linguistique. Revue belge de philologie et d'histoire, 1960, 38.3: 705-717.
737. Pottier Bernard. Boletim de filología, Centro de Estudos Filológicos (Lisboa). Romania, 1954, 75.298: 267-274.
738. Bloomfield Leonard Language, New York: Henry Holt, 1933, pp. 166–169.
739. Pike K.L. On tagmemes, née grammemes. International Journal of American Linguistics. 1958, Vol. 24(4):273ff.
740. Marcus,Solomon. Algebraic Linguistics, Analytical Models by Solomon Marcus. Elsevier, 1966.
741. Marcus Solomon. From structural to algebraic distributional analysis in Linguistics. 1980.
742. Mamali Cătălin, Marcus, Solomon. A psycho-linguistic approach to development. 1987.
743. Marcus Solomon. Mathematics through the glasses of Hjelmslevs semiotics. 2003.
744. Marcus Solomon. Formal languages: Foundations, prehistory, sources, and applications. In: Formal Languages and Applications. Springer, Berlin, Heidelberg, 2004. p. 11-54.
745. Marcus Solomon. Proofs and mistakes: Their syntactics, semantics, and pragmatics. Semiotica, 2012, 2012.188: 139-155.
746. Marcus Solomon. Semiotics of theatre: a mathematical-linguistic approach. 1980.
747. Schaefer Marcus. The graph crossing number and its variants: A survey. The electronic journal of combinatorics, 2012, DS21: May 21-2021.

748. Trubetzkoy Nikolai Sergeevich. Principles of phonology. 1969.
749. Трубецкой Н. С. Основы фонологии. – М.: Аспект Пресс, 2000. – 352 с.
750. Trubetzkoy N. S. Introduction to the principles of phonological descriptions. Springer Science & Business Media, 2012.
751. Bernard Pottier. Mental activities and linguistic structures. AL-Lisaniyyat, 2007, 12.13: 5-7.
752. Greimas A.J. Éléments pour une théorie de l'interprétation du récit mythique. Communications, 1966, 8.1: 28-59.
753. Greimas Algirdas Julien. Conditions d'une sémiotique du monde naturel. Langages, 1968, 10: 3-35.
754. Greimas Algirdas Julien. Sémantique structurale: recherche de méthode. Presses universitaires de France, 2015.
755. Jakobson Roman. Towards a linguistic typology of aphasic impairments. Disorders of language, 1964, 21: 42.
756. Jakobson R. The metaphoric and metonymic poles. Metaphor and metonymy in comparison and contrast, 2003, 20: 41-47.
757. Jakobson R. Metalanguage as a linguistic problem (pp. 113-121). Akadémiai Nyomda. 1976.
758. Jakobson R. On linguistic aspects of translation. In On translation (pp. 232-239). Harvard University Press. 2013.
759. Caton Steven C. Contributions of Roman Jakobson. Annual Review of Anthropology, 1987, 16.1: 223-260.
760. Hjelmslev Louis. Résumé of a Theory of Language. Travaux du Cercle linguistique de Copenhague, vol. XVI. Copenhague: Nordisk Sprog- og Kulturforlag. 1975.
761. Hjelmslev Louis. Structural analysis of language. Studia linguistica, 1947, 1.1-3: 69-78.
762. Hjelmslev Louis. Sur l'indépendance de l'épithète. Copenhague: Historisk-filologiske Meddelelser udgivet af Det Kongelige Danske Videnskabernes Selskab, i kommission hos Ejnar Munksgaard. 1956.
763. Hjelmslev Louis. Prolegomena to a Theory of Language. 1953[1943]. Baltimore: Indiana University Publications in Anthropology and Linguistics (IJAL Memoir, 7) : Madison: University of Wisconsin Press, 1961. Dt.: Hjelmslev 1974.
764. Hjelmslev Louis. Catégorie des cas (2 volumes). Acta Jutlandica VII, IX. 1935/37.
765. Hjelmslev Louis. Principes de grammaire générale. Copenhague: Bianco Lundo. 1928.
766. De Saussure Ferdinand. Cours de linguistique générale. Otto Harrassowitz Verlag, 1989.
767. Де Соссюр Фердинанд. Курс общего языкознания/Перевод на азерб. Язык НГ Джа-фарова, Изд. БГУ, 2003.
768. Benítez Pamela Faber, Usón Ricardo Mairal. The paradigmatic and syntagmatic structure of the lexical field of feeling. Cuadernos de Investigación Filológica, 1998, 23: 35-60.
769. Diller,Hans-Jürgen. Emotions in the English lexicon: A historical study of a lexical field. Amsterdam studies in the theory and history of linguistic science series 4, 1994, 219-219.
770. Kraif Olivier, Diwersy Sascha. Exploring combinatorial profiles using lexicograms on a parsed corpus: a case study in the lexical field of emotions. P. Blumenthal, I. Novakova, D. Siepmann (éd.) Les émotions dans le discours. Emotions in discourse. Bern: Peter Lang, 2014, 381-394.
771. Lutzeier P. The notion of lexical field and its application to English nouns of financial income. Lingua, 1982, 56.1: 1-42.
772. Lyons John, Linguistic semantics: An introduction. Cambridge University Press, 1995.
773. Lyons John. Semantics: Volume 2. Cambridge university press, 1977.
774. Lyons John. A note on possessive, existential and locative sentences. Foundations of language, 1967, 390-396.
775. Lyons John. Language and linguistics. Cambridge university press, 1981.
776. Lyons John. Introduction to theoretical linguistics. Cambridge university press, 1968.
777. Coseriu Eugenio. Linguistic competence: what is it really?. The Modern Language Review, 1985, 80.4: xxv-xxxv.
778. Coseriu Eugenio, Geckeler Horst. Trends in structural semantics. Gunter Narr Verlag, 1981.
779. Goddard C. Componential analysis. Culture and Language Use, 2009, Vol. 2, 58.

780. Lipka Leonhard. Semantic components of English nouns and verbs and their justification. Angol Filológiai Tanulmányok. Hungarian Studies in English, 1979, 12: 187-202.
781. Geckeler Horst. Structural semantics. Hans-Jürgen Eikmeyer & Hannes Rieser, 1981, 381-413.
782. Балдингер Курт. Семантическая теория: к современной семантике. 1980 г.
783. Балдингер Курт. Семасиология . Де Грюйтер, 2022.
784. Baldinger Kurt. Semantic theory: towards a modern semantics. 1980.
785. Heger Klaus. Noematic grammar. Prospects for a New Structuralism, 1992, 96: 91.
786. Heger Klaus. Pluricentric Languages-Differing Norms in Different Nations. 1993: 211-213.
787. Benveniste Emile. The semiology of language. 1981: 5-24.
788. Benveniste Emile. Subjectivity in language. Problems in general linguistics, 1971, 1: 223-30.
789. Hempel Carl G. A note on semantic realism. Philosophy of Science, 1950, 17.2: 169-173.
790. Hempel Carl G. The theoretician's dilemma: A study in the logic of theory construction. 1958.
791. Hempel C. G., Oppenheim, P. Studies in the Logic of Explanation. Philosophy of science, 1948, Vol. 15(2), 135-175.
792. Hempel Carl G. Vagueness and logic. Philosophy of Science, 1939, 6.2: 163-180.
793. Hempel Carl G. The philosophy of Carl G. Hempel: studies in science, explanation, and rationality. Oxford University Press, 2001.
794. Quine Willard Van Orman. Word and object. MIT press, 2013.
795. Quine Willard V., Quine, Willard Van Orman. Confessions of a confirmed extensionalist and other essays. Harvard University Press, 2008.
796. Van Orman Quine Willard. The roots of reference. 1973.
797. Quine Willard Van Orman. Word and object. Cambridge, Mass.: MIT Press, 1960.–XV, 294 p.
798. Quine Willard V., Quine Willard Van Orman. Pursuit of truth. Harvard University Press, 1990.
799. Quine Willard V., Quine Willard Van Orman. Theories and things. Harvard University Press, 1981.
800. Popper K.R. Evolutionary epistemology, rationality, and the sociology of knowledge. Open Court Publishing, 1987.
801. Popper Karl. The poverty of historicism, III. Economica, 1945, 12.46: 69-89.
802. Popper Karl. Some philosophical comments on Tarski's theory of truth. In: Proceedings of Tarski Symposium. Rhode Island: American Mathematical Society, 1974. p. 397-410.
803. Lytvyn V., Vysotska V., Veres O., Rishnyak I., Rishnyak H. Classification methods of text documents using ontology based approach. Advances in Intelligent Systems and Computing (AISC). 2017. Vol. 512. P. 229–240.
804. Vysotska V., Lytvyn V., Burov Y., Berezin P., Emmerich M., Basto F. V. Development of information system for textual content categorizing based on ontology. CEUR Workshop Proceedings. 2019. Vol. 2362. P. 53–70.
805. Lytvyn V., Vysotska V., Rusyn B., Pohreliuk L., Berezin P., Naum O. Textual content categorizing technology development based on ontology. CEUR Workshop Proceedings. 2019. Vol. 2386. P. 234–254. E-ISSN: 1613-0073
806. Литвин В. В., Ремешило-Рибчинська О. І., Висоцька В. А. Побудова онтології архітектурних термінів. Відбір і обробка інформації : міжвід. зб. наук. пр. 2017. Вип. 44 (120). С. 90–96.
807. Lytvyn V., Vysotska V., Burov Y., Demchuk A. Architectural ontology designed for intellectual analysis of e-tourism resources. CSIT (Львів,11–14 вересня 2018 р.). 2018. Т. 1. С. 335–338.
808. Lytvyn V., Burov Y., Vysotska V., Pukach Y., Tereshchuk O., Shakleina I. Abstracting Text Content Based on Weighing the TF-IDF Measure by the Subject Area Ontology. SIST, 28-30 April 2021, Nur-Sultan, Kazakhstan. Art. 9465978.

809. Кравець П., Литвин В., Висоцька В. Ігрова модель онтологічної підтримки проектів. Радіоелектроніка, інформатика, управління. 2021. № 1(56). С. 172–183.
810. Литвин В. В., Висоцька В. А., Оливко Р. М., Черна Т. М. Особливості рубрикації текстових документів з використанням онтології. ISDMIT, Залізний Порт, Україна, 25–28 трав. 2016. С. 292–295.
811. Lytvyn V., Vysotska V., Dosyn D., Burov Y. Method for ontology content and structure optimization, provided by a weighted conceptual graph. Webology. 2018. Vol. 15, iss. 2. P. 66–85. E-ISSN: 1735-188X
812. Burov Y., Vysotska V., Kravets P. Ontological approach to plot analysis and modeling. CEUR Workshop Proceedings. 2019. Vol. 2362. P. 22–31. E-ISSN: 1613-0073
813. Kravets P., Burov Y., Lytvyn V., Vysotska V. Gaming method of ontology clusterization. Webology. 2019. Vol. 16, iss. 1. P. 55–76. ISSN: 1735-188X
814. Kravets P., Lytvyn V., Vysotska V., Burov Y., Andrusyak I. Game task of ontological project coverage. CEUR Workshop Proceedings. 2021. Vol. 2851. P. 344–355. E-ISSN: 1613-0073.
815. Pashchetnyk O., Lytvyn V., Zhyvchuk V., Polishchuk L., Vysotska V., Rybchak Z., Pukach Y. The ontological decision support system composition and structure determination for commanders of Land Forces formations and units in Ukrainian Armed Force. CEUR Workshop Proceedings. 2021. Vol. 2870. P. 1077–1086. E-ISSN: 1613-0073.
816. Lytvyn V., Vysotska V., Pukach P., Vovk M., Ugryn D. Method of functioning of intelligent agents, designed to solve action planning problems based on ontological approach. Eastern-European Journal of Enterprise Technologies. 2017. № 3/2 (87). P. 11–17.
817. Lytvyn V., Vysotska V., Lozynska O., Oborska O., Dosyn D. Methods of building intelligent decision support systems based on adaptive ontology. DSMP, August 21–25, 2018, Lviv, Ukraine. 2018. P. 145–150.
818. Lytvyn V., Vysotska V., Burov Y., Hryhorovych V. Knowledge novelty assessment during the automatic development of ontologies. DSMP: proceedings of the IEEE Third international conference, Lviv, Ukraine. 2020. P. 372–377.
819. Lytvyn V., Dosyn D., Vysotska V., Hryhorovych A. Method of ontology use in OODA. Data stream mining & processing (DSMP) : proceedings of the IEEE 3rd international conference, Lviv, Ukraine. 2020. P. 409–413.
820. Lytvyn V., Vysotska V., Burov Y., Brodyak O. Approach to automatic construction of interpretation functions during ontology learling. CSIT, Збараж, 23–26 вересня, 2020. P. 267–271.
821. Lytvyn V., Bublyk M., Vysotska V., Panasyuk V., Brodyak O., Luchkevych M. Modelling of the Intelligent Agent's Behavior Scheduler Based on Petri Nets and Ontological Approach. IEEE International Conference on Smart Information Systems and Technologies (SIST), 28-30 April 2021, Nur-Sultan, Kazakhstan. Art. 9465994.
822. Lytvyn V., Oborska O., Vysotska V., Dosyn D., Demchuk A., Burov Y., Kravets P., Oleksiv N. Problems of ontology structure and meaning optimization and theirs solution methods. Proceedings of the 4th International conference "Computational linguistics and intelligent systems" COLINS, Lviv, Ukraine, June 23-24, 2020. P. 21–40.
823. Copestake A., Flickinger D., Pollard C., Sag I. A. Minimal recursion semantics: An introduction. Research on language and computation, 2005, Vol. 3(2), 281-332.
824. Copestake A., Flickinger D., Malouf R., Riehemann S., Sag I. Translation using minimal recursion semantics. Sixth Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages. 1995.
825. Copestake A. Semantic composition with (robust) minimal recursion semantics. ACL 2007 Workshop on Deep Linguistic Processing. 2007. p. 73-80.
826. Kasper Robert T. A logical semantics for feature structures. In: 24th Annual Meeting of the Association for Computational Linguistics. 1986. p. 257-266.

827. Estival D., Ballim A., Russell G., Warwick S. A syntax and semantics for feature-structure transfer. In Proceedings of the 3rd International Conference on theoretical and methodological issues in MT of NLs (TMI-90), Austin, Texas 1990, June. P. 131-143.
828. Nerbonne John. A feature-based syntax/semantics interface. Annals of Mathematics and Artificial Intelligence, 1993, 8.1: 107-132.
829. Copestake Ann, et al. Minimal recursion semantics: An introduction. Research on language and computation, 2005, 3.2: 281-332.
830. Pollard Carl, Sag Ivan A. Head-driven phrase structure grammar. University of Chicago Press, 1994.
831. Levine Robert D., Meurers Walt Detmar. Head-Driven Phrase Structure Grammar: Linguistic approach, formal foundations, and computational realization. The encyclopedia of language and linguistics,, 2006, 237-252.
832. Nerbonne John A., Netter Klaus, Pollard Carl Jesse (ed.). German in head-driven phrase structure grammar. Stanford, Cal.: Center for the Study of Language and Information, 1994.
833. Müller Stefan, Y Priemer Antonio Machicao. 12 Head-Driven Phrase Structure Grammar. Current approaches to syntax: A comparative handbook, 2019, 3: 317.
834. Miyao Yusuke, Ninomiya Takashi, Tsujii Jun'ichi. Corpus-oriented grammar development for acquiring a head-driven phrase structure grammar from the penn treebank. In: International Conference on Natural Language Processing. Springer, Berlin, Heidelberg, 2004. p. 684-693.
835. Dalrymple Mary. Lexical functional grammar. Brill, 2001.
836. Falk Yehuda. Lexical-functional grammar. Oxford University Press, 2011.
837. Dalrymple M., Kaplan R. M., Maxwell III J. T., Zaenen A. E. (Eds.). Formal issues in lexical-functional grammar (No. 47). Center for the Study of Language (CSLI). 1995.
838. Kaplan Ronald M. The formal architecture of lexical-functional grammar. Formal issues in lexical-functional grammar, 1995, 47: 7-27.
839. Neidle Carol. Lexical Functional Grammar. Encyclopedia of Language and Linguistics, 1994, 5: 2147-2153.
840. Pollard Carl, Sag Ivan A. Information-based syntax and semantics: Vol. 1: fundamentals. Center for the Study of Language and Information, 1988.
841. Copestake A., Flickinger D., Pollard C., Sag I. A. Minimal recursion semantics: An introduction. Research on language and computation, 2005, Vol. 3(2), 281-332.
842. Copestake A., Flickinger D., Malouf R., Riehemann S., Sag I. Translation using minimal recursion semantics. Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, 1995.
843. Klein Ewan, Sag Ivan A. Type-driven translation. Linguistics and Philosophy, 1985, 163-201.
844. Fodor J. D., Sag I. Referential and quantificational indefinites. Linguistics and philosophy, 1982, 5.3: 355-398.
845. Sag Ivan A., Pollard Carl. An integrated theory of complement control. Language, 1991, 63-113. SAG, Ivan A. Sign-based construction grammar: An informal synopsis. Sign-based construction grammar, 2012, 193: 69-202.
846. Copestake Ann. Robust minimal recursion semantics. Unpublished draft, 2006.
847. Dridan Rebecca, Bond Francis. Sentence comparison using Robust Minimal Recursion Semantics and an ontology. Proceedings of the Workshop on Linguistic Distances. 2006. p. 35-42.
848. Horvat Matic, Copestake Ann, Byrne Bill. Hierarchical statistical semantic realization for Minimal Recursion Semantics. Proceedings of the 11th International Conference on Computational Semantics. 2015. p. 107-117.

849. Lytvyn V., Vysotska V., Pukach P., Bobyk I., Pakholok B. A method for constructing recruitment rules based on the analysis of a specialist's competences. Eastern-European Journal of Enterprise Technologies. 2016. № 6/2 (84). С. 4–14.

850. Lytvyn V., Vysotska V., Kuchkovskiy V., Pelekh I., Bobyk I., Malanchuk O., Ryshkovets Y., Brodyak O., Bobrivetc V., Panasyuk V. Development of the system to integrate and generate content considering the cryptocurrent needs of users. Eastern-European Journal of Enterprise Technologies. 2019. № 1/2 (97). С. 18–39.

851. Шаховська Н. Б., Висоцька В. А., Чирун Л. Б. Методи та засоби дистанційної освіти для заохочення і залучення сучасної молоді до проведення самостійних наукових досліджень. Вісник НУ «Львівська політехніка». 2015. № 832. С. 254–284.

852. Шаховська Н. Б. Висоцька В. А., Скотар О. О. Розроблення архітектури інтелектуальної системи на основі інноваційних методів навчання студентів. Вісник НУ "Львівська політехніка". 2017. № 872. С. 220–229.

853. Русин Б. П., Погрелюк Л. В., Висоцька В. А., Осипов М. М., Варецький Я. Ю., Капшій О. В. Архітектура системи дедублікації та розподілу даних у хмарних сховищах під час резервного копіювання. Інформаційні технології та комп'ютерна інженерія. 2019. Т. 2, № 45. С. 40–63.

854. Русин Б., Погрелюк Л. В., Висоцька В. А., Осипов М. М. Метод дедублікації та розподілу даних у хмарних сховищах під час резервного копіювання даних. Вісник НУ «Львівська політехніка» 2019. Вип. 6. С. 1–12.

855. Shakhovska N., Vysotska V., Chyrun L. Features of e-learning realization using virtual research laboratory. CSIT'2016, 6–10 Sept., 2016, Lviv, Ukraine. Lviv: Publishing Lviv Polytechnic, 2016. P. 143–148.

856. Lytvyn V., Vysotska V., Chyrun L., Chyrun L. Distance learning method for modern youth promotion and involvement in independent scientific researches. DSMP 2016 : proc. Aug. 23–27, 2016, Lviv, Ukraine. Lviv, 2016. P. 269–274.

857. Lytvyn V., Vysotska V., Veres O., Rishnyak I., Rishnyak H. The risk management modelling in multi project environment. CSIT, 5–8 Sept., 2016, Lviv, Ukraine. Lviv: Publishing Lviv Polytechnic, 2017. P. 32–35.

858. Gozhyj A., Kalinina I., Nechakhin V., Gozhyj V., Vysotska V. Modeling an Intelligent Solar Power Plant Control System Using Colored Petri Nets. IDAACS : proceedings of the IEEE 11th International Conference, 22-25 Sept., Cracow, Poland. 2021. P. 626–631. ISSN: 2770-4262, Electronic ISSN: 2770-4254

859. А Берко. Ю., Висоцька В. А., Чирун Л. В., Чирун Л. Б. Особливості формування критеріїв оцінювання знань студентів згідно їх компетентності у ІТ-сфері. Тези доповідей «Математика. Інформаційні технології. Освіта». V Міжнародна науково-практична конференція, 5–7 черв. 2016 р., Луцьк. С. 117–118.

860. Висоцька В. А. Методика аналізу компетентностей для рекрутингу. International scientific and practical conference "Scientific Research Priorities. – 2017", 22–23 June 2017, Nowy Sanz, Poland. P. 60–62.

861. Kutyuk O., Lytvyn V., Oborska O., Vysotska V., Dosyn D., Demchuk A., Burov Y., Kravets P. Intelligent system development of distant matrix analysis for recruitment in the IT sector. Proceedings of the 4th International conference "Computational linguistics and intelligent systems" COLINS, Lviv, Ukraine, June 23-24, 2020. P. 41–78.

862. Бех П.О. Англійська мова: Навч. посібник. К. : Либідь, 1992. 272 с.

863. Английская грамматика в доступном изложении. URL: http://realenglish.ru/crash/lesson3.htm.

864. English Verbs (Part 1) – Basic Terms. URL: http://sites.google.com/site/englishgrammarguide/Home/ english-verbs–part-1—basic-terms.

865. Носков С. А. Самоучитель немецкого языка. Deutsch für sie. К.: Наука, 1999. 400 с.

866. Постнікова О. М. Німецька мова. Розмовні теми: лексика, тексти, діалоги, вправиюК.: А.С.К, 2001. Т. 1-2.

867. Кушнір О. С., Брик О. С., Дзіковський В. Є., Іваницький Л. Б., Катеринчук І. М., Кісь Я. П. Статистичний розподіл і флуктуації довжин речень в українських, російських і англійських корпусах. Вісник Національного університету Львівська політехніка. Серія: Інформаційні системи та мережі, 2016, №854, С. 228-239.

868. World Health Organization, et al. The health and well-being of men in the WHO European Region: better health through a gender approach. World Health Organization. Regional Office for Europe, 2018. URL: https://apps.who.int/iris/handle/10665/329686

869. Melton Jordan. Framing the NFL national anthem protest: an analysis of news media in post-racist America. 2019. Master's Thesis. URL: https://scholarworks.calstate.edu/downloads/x059c765v

870. Хомицька І.Ю. Методи та засоби диференціації фоностатистичних структур функціональних стилів англійської мови : дисертація на здобуття наукового ступеня кандидата технічних наук : 10.02.21 – "Структурна, прикладна та математична лінгвістика" / Ірина Юріївна Хомицька, НУ «Львівська політехніка». Львів, 2020. 262 с. URL: https://test-new.lpnu.ua/sites/default/files/2021/dissertation/10062/disertaciya-khomickoi-i-yu-17032021.pdf

871. Хомицька І.Ю. Методи та засоби диференціації фоностатистичних структур функціональних стилів англійської мови: автореферат дисертації на здобуття наукового ступеня кандидата технічних наук : 10.02.21 – "Структурна, прикладна та математична лінгвістика" / Ірина Юріївна Хомицька, НУ «Львівська політехніка». Львів, 2020. 23 с. URL: https://lpnu.ua/sites/default/files/2021/dissertation/10062/arefkhomytskaiyu-pdf.pdf

872. Бражник Н. В. Конструктивний зіставний аналіз складних речень в англомовних і україномовних інтернет-блогах. Науковий часопис Національного педагогічного університету імені МП Драгоманова. Серія 9: Сучасні тенденції розвитку мов, 2013, 10: 60-65.

873. Зернецький П. В., Бражник Н. В. Конструктивний зіставний аналіз простих речень в англомовних і україномовних інтернет-блогах. Система і структура східнослов'янських мов, 2012, 6: 251-257.

874. Коцюба З. Г. Семантична структура номінативних речень англійської та української мов. Мовознавство, 2002. URL: http://dspace.nbuv.gov.ua/handle/123456789/182824

875. Rusyn B., Lytvyn V., V Vysotska., Emmerich M., Pohreliuk L. The virtual library system design and development. Advances in Intelligent Systems and Computing (AISC). 2019. Vol. 871. P. 328–349.

876. Русин Б., Висоцька В., Погрелюк Л. Особливості проектування та розроблення інформаційної системи Virtual Library. Оптико-електронні інформаційно-енергетичні технології. 2017. Т. 34, № 2. С. 18–33.

877. Rusyn B., Vysotska V., Pohreliuk L. Model and architecture for virtual library information system. CSIT: 11–14 вересня 2018 р., Львів. 2018. Т. 1. С. 34–41.

878. Русин Б. П., Висоцька В. А., Погрелюк Л. В. Модель інформаційної системи Virtual Library. Матеріали міжнародної наукової конференції «Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту» (ISDMCI'2018), 21–27 трав. 2018 р., Залізний Порт, Україна. С. 100–102.

879. Kolhan O., Kolgan T., Padalka R. Перекладні Термінологічні Е-Словники Як Нагальна Вимога Сьогодення В Процесі Мовної Підготовки Студентів ЗВО. Наукові записки Національного університету «Острозька академія»: Серія «Філологія», 2021, 11 (79): 233-236.

880. Лучик О. І., Романова Т. О. Переваги електронних словників при вивченні німецької мови. Вісник Чернівецького торговельно-економічного інституту. Економічні науки, 2018, 1-2: 315-320.

881. Гасюк Г. Організація Роботи Студентів З Різними Видами Словників. Actual trends of modern scientific research"(March 14-16, 2021) MDPC Publishing, Munich, Germany. 2021. 805 p. 2021. p. 351.

882. Щербина Ю.М.. Нікольський Ю.В., Висоцька В.А., Шестакевич Т.В. Утворення українських дієприкметників за допомогою породжувальних граматик. Вісник Національного університету "Львівська політехніка". – Львів 2011. – № 715. – Стор. 354-369.

883. Vysotska V., Lytvyn V., Burov Y., Gozhyj A., Makara S. The consolidated information web-resource about pharmacy networks in city. CEUR Workshop Proceedings. 2018. Vol. 2255. P. 239–255. E-ISSN: 1613-0073

884. Lytvyn V., Hryhorovych A., Hryhorovych V., Vysotska V., Bublyk M., Chyrun L. Medical content processing in intelligent system of district therapist. CEUR Workshop Proceedings. 2020. Vol. 2753. P. 415–429.

885. Коробчинський М., Чирун Л., Висоцька В., Кондратьєв Є. Особливості формування та аналізу контенту інтернет-газети музичних новин. Радіоелектроніка. Інформатика. Управління. 2017. № 4. С. 139–150.

886. Вінтоняк С. М., Коробчинський М. В., Чирун Л. Б., Висоцька В. А. Аналіз особливостей Інтернет-порталу аматорських спортивних ігор. Вісник Національного університету "Львівська політехніка Серія: Інформаційні системи та мережі : зб. наук. пр. 2016. № 854. С. 21–41.

887. Naum O., Chyrun L., Kanishcheva O., Vysotska V. Intellectual system design for content formation. Computer science and information technologies : proc. of the XIIth Intern. conf. CSIT'2017, 5–8 Sept., 2016, Lviv, Ukraine. Lviv: Publishing Lviv Polytechnic, 2017. P. 131–138.

888. Clifton B. Advanced web metrics with Google Analytics. Indianapolis : John Wiley & Sons, 2012. 589 p.

889. Sulova S. A system for e-commerce website evaluation. URL: https://www.researchgate.net/profile/Snezhana-Sulova/publication/334734832_A_System_for_E-Commerce_Website_Evaluation/links/5d4549a0299bf1995b60d51f/A-System-for-E-Commerce-Website-Evaluation.pdf

890. Saura J. R. Understanding the digital marketing environment with KPIs and web analytics / J. R. Saura, P. Palos-Sánchez, L. M. Cerdá Suárez. Future Internet. 2019. Vol. 9(4). P. 76. doi: https://doi.org/10.3390/fi9040076

891. García M. D. M. R. An ontology-based data integration approach for web analytics in e-commerce / M. D. M. R. García, J. García-Nieto, J. F. Aldana-Montes. Expert Systems with Applications. 2016. Vol. 63. P. 20-34. doi: https://doi.org/10.1016/j.eswa.2016.06.034

892. Heller D. Web analytics: functions, KPIs and reports in SMEs. URL: https://kola.opus.hbz-nrw.de/opus45-kola/frontdoor/deliver/index/docId/1295/file/BachelorThesisDominikHeller.pdf

893. Golyash I. The performance audit of a corporate website as a tool for its internet marketing strategy / I. Golyash, V. Panasiuk, S. Sachenko. EUREKA: Social and Humanities. 2017. Vol. 5. P. 57-66.

894. Rodello I. A. Evaluation of the impact of promotional campaign through a social networks on the key performance indicators of website for online of group-buying in Brazil / I. A. Rodello, V. Dândolo, M. M. Grande. European Journal of Management Issues. 2016. Vol. 7. P. 244-249.

895. Shaytura S. V., Kozhayev Y. P., Ordov K. V., Antonenkova A. V., Zhenova N. A. Performance evaluation of the electronic commerce systems. Performance evaluation. 2017. Vol. 38. P. 1-11.

896. Висоцька В. А., Чирун Л. В. Формальна модель опрацювання інформаційних ресурсів в системах електронної контент-комерції. Вісник НУ «Львівська політехніка» 2015. № 814. С. 44–54.

897. Висоцька В. А., Чирун Л. В. Концептуальна модель опрацювання інформаційних ресурсів в системах електронної контент-комерції. Математичні машини і системи. 2015. № 3. С. 179–190.

898. Висоцька В. А., Чирун Л. В. Опрацювання інформаційних ресурсів у системах електронної контент-комерції. Відбір і обробка інформації : міжвід. зб. наук. пр. 2015. Вип. 42 (118). С. 84–92.

899. Висоцька В. А. Аналітичні методи опрацювання інформаційних ресурсів в системах електронної контент-комерції. Вісник НУ "Львівська політехніка". 2015. № 829. С. 76–101.

900. Vysotska V., Chyrun L. The means structure of information resources processing in electronic content commerce systems. Journal of Information Sciences and Computing Technologies (JISCT). 2015. Vol 3, № 3. P. 241–248.

901. Vysotska V., Chyrun L. Methods and means of processing information resources in electronic content commerce systems. Applied Computer Science. 2015. Vol. 11, № 2. 2015. P. 68–85.

902. Chyrun L. Vysotska V., Laba R. Information resources analysis in electronic content commerce systems. Applied Computer Science. 2016. Vol. 12, № 1. P. 48–66.

903. Vysotska V. Chyrun L. Methods of information resources processing in electronic content commerce systems. Досвід розробки та застосування приладо-технологічних САПР в мікроелектроніці : матеріали XIII Міжнар. наук.-техн. конф., 24–27 лют. 2015, Львів, Поляна. Львів, 2015. С. 328–332.

904. Vysotska V., Chyrun L. Analysis features of information resources processing. CSIT'2015, 14–17 Sept., 2015, Lviv, Ukraine. Lviv: Publishing Lviv Polytechnic, 2015. P. 124–128.

905. Vysotska V., Chyrun L., Chyrun L. Information technology of processing information resources in electronic content commerce systems. CSIT, 6–10 Sept., 2016, Lviv, Ukraine. Lviv: Publishing Lviv Polytechnic, 2016. P. 212–222.

906. Walker James A. Variation in linguistic systems. Routledge, 2012. URL: https://www.taylorfrancis.com/books/mono/10.4324/9780203854204/variation-linguistic-systems-james-walker

907. Ekdahl Bertil. Anticipatory systems as linguistic systems. In: AIP Conference Proceedings. American Institute of Physics, 2000. p. 131-140.

908. Pederson Eric. Geographic and manipulable space in two Tamil linguistic systems. In: European conference on spatial information theory. Springer, Berlin, Heidelberg, 1993. p. 294-311.

909. Pierrehumbert Janet B., Stonedahl Forrest, Daland Robert. A model of grassroots changes in linguistic systems. arXiv preprint arXiv:1408.1985, 2014.

910. Stephens R. A., Wood J. R. G. Information systems as linguistic systems: a constructivist perspective. In: Systems thinking in Europe. Springer, Boston, MA, 1991. p. 469-474.

911. Hersh William R. Linguistic Systems. Information Retrieval: A Health and Biomedical Perspective, 2003, 310-355.

912. Vysotsky A., Vysotska V., V Lytvyn., Burov Y., A Demchuk., Lyudkevych I. Consolidated information web resource for online tourism based on data integration and geolocation. CSIT (Львів, 17–20 вересня 2019 р.). С. 15–20.

913. Vysotsky A., Lytvyn V., Vysotska V., Dosyn D., Lyudkevych I., Antonyuk N., Naum O., Vysotskyi A., Chyrun L., Slyusarchuk O. Online tourism system for proposals formation to user based on data integration from various sources. CSIT-2019 (Львів, 17–20 вересня 2019 р.). 2019. Т. 2. С. 92–97.

914. Gozhyj A., Kalinina I., Gozhyj V., Vysotska V. Web service interaction modeling with colored petri nets. IDAACS: proceedings, September 18–21, 2019, Metz, France. 2019. P. 319–323.

915. Висоцька В. А., Чирун Л. Б., Чирун Л. В. Контент-моніторинг текстової інформації Web-ресурсів. ISDMIT, Залізний Порт, Україна, 25–28 трав. 2015. С. 36–38.

916. Кондратєв Є., Висоцька В Контент-аналіз текстових масивів даних. 4 Міжнародна наукова конференція ІКС-2015 «Інформація, комунікація, суспільство 2015», 20–23 трав. 2015, Україна, Львів, Славське. С. 170–171.

917. Литвин В. В., Наум О. М., Висоцька В. А. Метод інтеграції та управління контентом мережі інформаційних ресурсів туризму згідно потреб користувача. Міжнародна наукова конференція «Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту», 22–26 трав. 2017, Залізний Порт. С. 78–80.

918. Antonyuk N., Medykovskyy M., Chyrun L., Dverii M., Oborska O., Krylyshyn M., Vysotsky A., Tsiura N., Naum O. Online Tourism System Development for Searching and Planning Trips with User's Requirements. Advances in Intelligent Systems and Computing. 2020. Vol. 1080. P. 831–863.
919. Kuchkovskiy V., Andrunyk V., Krylyshyn M., Chyrun L., Vysotskyi A., Chyrun S., Sokulska N., Brodovska I. Application of Online Marketing Methods and SEO Technologies for Web Resources Analysis within the Region. CEUR workshop proceedings. Aachen: CEUR-WS.org, 2021. Vol. 2870. P. 1652–1693.
920. Rusyn B., Pohreliuk L., Kapshii O., Varetskyy J., Demchuk A., Karpov I., Gozhyj A., Gozhyj V., Kalinina I. An Intelligent System for Commercial of Information Products Distribution Based SEO and Sitecore CMS. CEUR workshop proceedings. Aachen: CEUR-WS.org, 2020. Vol. 2604. P. 760–777.
921. S Orekhov., Malyhon H., Liutenko I., Goncharenko T. Using Internet News Flows as Marketing Data Component. CEUR workshop proceedings. Aachen: CEUR-WS.org, 2020. Vol. 2604. P. 358–373.
922. Pavlenko O. , Tymofieieva I. Search Query Data Analysis: Challenges and Opportunities. CEUR workshop proceedings. Aachen: CEUR-WS.org, 2020. Vol. 2604. P. 452–461.
923. Kliuiev O., Vnukova N., S Hlibko., Brynza N., Davydenko D. Estimation of the Level of Interest and Modeling of the Topic of Innovation Through Search in Google. CEUR workshop proceedings. 2020. Vol. 2604. P. 523–535.
924. Radiuk P. , Hrypynska N. A Framework for Exploring and Modelling Neural Architecture Search Methods. CEUR workshop proceedings. Aachen: CEUR-WS.org, 2020. Vol. 2604. P. 1060–1074.
925. Veres O., Rusyn B., Sachenko A., Rishnyak I. Choosing the Method of Finding Similar Images in the Reverse Search System. CEUR workshop proceedings. Aachen: CEUR-WS.org, 2018. Vol. 2136. P. 99–107.
926. Basyuk T. , Vasyliuk A., Lytvyn V. Mathematical Model of Semantic Search and Search Optimization. CEUR workshop proceedings. Aachen: CEUR-WS.org, 2019. Vol. 2362. P. 96–105.
927. Adamuthe A. , Nitave T. Adaptive harmony search for optimizing constrained resource allocation problem. International Journal of Computing. 2018. Vol. 17(4). P. 260–269.
928. Aswani R., Kar A. K., Ilavarasan P. V., Dwivedi Y. K. Search engine marketing is not all gold: Insights from Twitter and SEOClerks. International Journal of Information Management. 2018. Vol. 38(1). P. 107-116.
929. Vysotska V., Berko A., Lytvyn V., Kravets P., Dzyubyk L., Bardachov Y., Vyshemyrska S. Information resource management technology based on fuzzy logic. Advances in Intelligent Systems and Computing (AISC). 2020. Vol. 1246. P. 164–182. Electronic ISSN 2194-5365, Print ISSN 2194-5357
930. Висоцька В. А. Гопяк М. В., Козлов П. Ю. Особливості технології управління web-ресурсом. Інженерія програмного забезпечення. 2015. № 1 (21). С. 25–35.
931. Козлов П. Ю., Висоцька В. А., Чирун Л. Б. Сучасні технології управління Web-ресурсами в інформаційній системі аналізу сервісу цифрової дистрибуції. Вісник НУ «Львівська політехніка». 2015. № 832. С. 103–128.
932. Козлов П., Висоцька В. Особливості технології управління web-ресурсом. V Міжнародна науково-практична конференція «Обробка сигналів і негаусівських процесів», 20-22 травня, 2015, Черкаси. С. 38–40.
933. Козлов П., Висоцька В. Аналіз процесу управління комерційним контентом. ISDMIT Залізний Порт, Україна, 25–28 трав. 2015. С. 36–38.
934. Козлов П., Висоцька В. Технологія управління комерційними контентом в системах електронного бізнесу. ІКС-2015 «Інформація, комунікація, суспільство 2015», 20–23 трав. 2015, Україна, Львів, Славське. С. 48–49.
935. Висоцька В. А., Козлов П. Ю. Управління Web-ресурсом. Особливості технології. «Математика. Інформаційні технології. Освіта». V Міжнародна конференція, 5–7 черв. 2016 р., Луцьк. С. 62–63.

936. Vysotska V., Basto F. V., Emmerich M. Web content support method in electronic business systems. CEUR Workshop Proceedings. 2018. Vol. 2136. P. 20–41. E-ISSN: 1613-0073

937. Висоцька В. Інформаційна технологія просування інтернет-ресурсів в пошукових системах на основі контент-аналізу ключових слів web-сторінок. Радіоелектроніка, інформатика, управління. 2021 № 3 (58). С. 133-151.

938. Висоцька В. А. Чирун Л. Б., Чирун Л. В. Аналіз процесу супроводу текстового комерційного контенту. ISDMIT, Залізний Порт, Україна, 25–28 трав. 2016. С. 42–44.

939. Берко А. Ю., Висоцька В. А., Чирун Л. В., Чирун Л. Б. Аналітичний метод супроводу текстового контенту інформаційних ресурсів. Збірник статей «Математика. Інформаційні технології. Освіта». Східноєвропейський НУ ім. Лесі Українки, кафедра вищої математики та інформатики. Луцьк, 2016. С. 11–20.

940. Pfeil Ulrike, Zaphiris Panayiotis. Applying qualitative content analysis to study online support communities. Universal access in the information society, 2010, 9.1: 1-16.

941. Gritter Mark, Cheriton David R. An architecture for content routing support in the internet. In: 3rd USENIX Symposium on Internet Technologies and Systems (USITS 01). 2001.

942. Bender Jacqueline L. Jimenez-Marroquin Maria-Carolina, Jadad Alejandro R. Seeking support on facebook: a content analysis of breast cancer groups. Journal of medical Internet research, 2011, 13.1: e16.

943. Zellars Kelly L., Perrewé Pamela L. Affective personality and the content of emotional social support: coping in organizations. Journal of Applied Psychology, 2001, 86.3: 459.

944. Андруник В. А., Висоцька В. А., Чирун Л. Б. Проект розроблення та впровадження системи електронної контент-комерції. Вісник НУ "Львівська політехніка". 2015. № 829. С. 321–348.

945. Висоцька В. А. Нога А. Ю., Козлов П. Ю. Управління Web-проектами електронного бізнесу для реалізації комерційного контенту. Вісник Національного університету "Львівська політехніка Серія: Інформаційні системи та мережі : зб. наук. пр. 2015. № 814. С. 421–434.

946. Kalinina I., Vysotska V., Sachenko S., Kovalchuk R., Gozhyj A. Qualitative and quantitative characteristics analysis for information security risk assessment in e-commerce systems. CEUR Workshop Proceedings. 2020. Vol. 2762. P. 177–190. E-ISSN: 1613-0073.

947. Vysotska V., Bublyk M., Korolenko O., Matseliukh Y., Kopach T. Network modelling of resource consumption intensities in human capital management in digital business enterprises by the critical path method. CEUR Workshop Proceedings. 2021. Vol. 2851. P. 366–380. E-ISSN: 1613-0073.

948. Bublyk M., A Kowalska-Styczeń., Lytvyn V., Vysotska V. The Ukrainian economy transformation into the circular based on fuzzy-logic cluster analysis. Energies. 2021. Vol. 14(18). Art. 5951.

949. Chyrun L., Andrunyk V., Vysotska V. Electronic content commerce system development. MEST Journal. 2015. Vol. 3, № 2. P. 10–33. ISSN: 2334-7171, ISSN (Online): 2334-7058

950. Vysotska V. Chyrun L., Kozlov P. Analysis of business processes in electronic content-commerce systems. Econtechmod : an international quarterly journal on economics in technology, new technologies and modelling processes. 2016. Vol. 5, № 1. P. 111–125. ISSN: 2084-5715

951. Vysotska V., Chyrun L., Kozlov P. Design and analysis features of generalized electronic content-commerce systems architecture. Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska. 2016. № 6 (2). P. 48–59.

952. Lytvyn V., Vysotska V. Designing architecture of electronic content commerce system. Intern. conf. CSIT'2015, 14–17 Sept., 2015, Lviv, Ukraine. Lviv: Publishing Lviv Polytechnic, 2015. P. 115–119.

953. Vysotska V., Hasko R., Kuchkovskiy V. Process analysis in electronic content commerce system. Intern. conf. CSIT'2015, 14–17 Sept., 2015, Lviv, Ukraine. Lviv: Publishing Lviv Polytechnic, 2015. P. 120–123.

954. Kravets P., V Lytvyn., Dobrotvor I., Sachenko O., Vysotska V., Sachenko A. Matrix Stochastic Game with Q-learning for Multi-agent Systems. Lecture Notes on Data Engineering and Communications Technologies. Vol. 83. 2021. P. 304–314. ISSN 23674512

955. Kravets P., Y Burov., Lytvyn V., Vysotska V., Y Ryshkovets., Brodyak O. Vyshemyrska S. Markovian Learning Methods in Decision-Making Systems. Lecture Notes on Data Engineering and Communications Technologies. Vol. 77. 2022. P. 423-437. ISSN 23674512.

956. Kravets P., Lytvyn V., Vysotska V., Ryshkovets Y., Vyshemyrska S., Smailova S. Dynamic coordination of strategies for multi-agent systems. Advances in Intelligent Systems and Computing (AISC). 2020. Vol. 1246 : Lecture notes in computational intelligence and decision making. 2020 International scientific conference "Intellectual systems of decision-making and problems of computational intelligence" ISDMCI'2020. – P. 653–670.

957. Kravets P., Lytvyn V., Burov Y., Vysotska V., Chyrun L., Panasyuk V. Making Optimal Decisions with Learning Method Based on Fuzzy Logic. Advanced Information and Communication Technologies (AICT) : proceedings of the 4th International Conference, 21-25 Sept., Lviv, Ukraine. 2021. P. 183–188.

958. Demchuk A., Lytvyn V., V Vysotska., Dilai M. Methods and means of web content personalization for commercial information products distribution. Advances in Intelligent Systems and Computing 2020. Vol. 1020. P. 332–347.

959. Lytvyn V., Vysotska V., Pukach P., O Brodyak., Ugryn D. Development of a method for determining the keywords in the slavic language texts based on the technology of web mining. Eastern-European Journal of Enterprise Technologies. 2017. № 2/2 (86). P. 14–23.

960. Lytvyn V., Vysotska V., Pukach P., Bobyk I., Uhryn D. Development of a method for the recognition of author's style in the Ukrainian language texts based on linguometry, stylemetry and glottochronology. Eastern-European Journal of Enterprise Technologies. 2017. № 4/2 (88). С. 10–18.

961. V Lytvyn., V Vysotska., Pukach P., Nytrebych Z., Demkiv I., Kovalchuk R., Huzyk N. Development of the linguometric method for automatic identification of the author of text content based on statistical analysis of language diversity coefficients. Eastern-European Journal of Enterprise Technologies. 2018. № 5/2 (95). С. 16–28.

962. Lytvyn V., Vysotska V., Budz I., Pelekh Y., Sokulska N., Kovalchuk R., Dzyubyk L., Tereshchuk O., Komar M. Development of the quantitative method for automated text content authorship attribution based on the statistical analysis of N-grams distribution. Eastern-European Journal of Enterprise Technologies. 2019. № 6/2 (102). С. 28–51.

963. Висоцька В. А. Особливості моделювання синтаксису речення слов'янських та германських мов за допомогою породжувальних контекстно-вільних грамматик. Вісник НУ "Львівська політехніка" 2015. № 814. С. 246–276.

964. Кісь Я. П., Висоцька В. А., Чирун Л. Б., Фольтович В. М. Застосування контент-аналізу для опрацювання текстових масивів даних. Вісник НУ «Львівська політехніка» 2015. № 814. С. 282–292.

965. Висоцька В. А. Особливості рубрикації текстового комерційного контенту. Вісник Національного університету "Львівська політехніка". 2015. № 826. С. 359–367.

966. Чирун Л. Б., Кучковський В. В., Висоцька В. А. Особливості методів контент-аналізу текстових масивів даних web-ресурсів в межах регіону контенту. Вісник НУ "Львівська політехніка". 2015. № 829. С. 296–320.

967. Кучковський В. В., Висоцька В. А., Нитребич С. З., РОливко. М. Застосування методів Інтернет-маркетингу для аналізу Web-ресурсів в межах регіону. Вісник Національного університету «Львівська політехніка». Серія: Інформаційні системи та мережі : зб. наук. пр. 2015. № 832. С. 129–164.

968. Андруник В. А. Висоцька В. А., Чирун Л. В. Особливості формування електронних дайджестів. Вісник Національного університету "Львівська політехніка". Серія: Комп'ютерні науки та інформаційні технології : зб. наук. пр. 2016. № 843. С. 3–14.

969. Vysotska V., Chyrun L., Chyrun L. Online newspaper content analysis based on SEO technologies. Вісник Національного університету "Львівська політехніка". 2016. № 859. С. 3–16.

970. Фольтович В. М., Коробчинський М. В., Л Чирун. Б., Висоцька В. А. Метод контент-аналізу текстової інформації Інтернет газети. Вісник НУ «Львівська політехніка». 2017. № 864. С. 7–19.

971. Литвин В. В., В Висоцька. А., В Кучковський. В., Дупкевич С. Ю., Наум О. Метод інтеграції та управління контентом мережі інформаційних ресурсів туризму згідно з потребами користувача. Вісник Національного університету «Львівська політехніка» 2018. № 901. С. 22–36.

972. Chyrun L., Vysotska V. Features of the content-analysis method for text categorization of commercial content in processing online newspaper articles. Applied Computer Science. 2015. Vol. 11, № 1. P. 15–30. ISSN: 1895-3735, ISSN (Online): 2353-6977

973. Burov Y., Horodetska A., Bublyk M., Nashkerska M., Vysotska V. Intellectual Tourist Service with the Situation Context Processing. Advances in Social Science, Education and Humanities Research. 2021. Vol. 557. P. 233-243. ISSN (Online): 2352-5398

974. Vysotska V., Chyrun L., Chyrun L. The commercial content digest formation and distributional process. Computer science and information technologies : proc. of the XIth Intern. conf. CSIT'2016, 6–10 Sept., 2016, Lviv, Ukraine. Lviv: Publishing Lviv Polytechnic, 2016. P. 186–189.

975. Korobchinsky M., V Vysotska., Chyrun L., Chyrun L. Peculiarities of content forming and analysis in internet newspaper covering music news. Computer science and information technologies : proc. of the XIIth Intern. conf. CSIT'2017, 5–8 Sept., 2016, Lviv, Ukraine. Lviv: Publishing Lviv Polytechnic, 2017. P. 52–57.

976. Vysotska V., Lytvyn V., Hrendus M., O Brodyak., Kubinska S. Method of textual information authorship analysis based on stylometry. Комп'ютерні науки та інформаційні технології (CSIT-2018) : матеріали XIII-ої Міжнародної науково-технічної конференції, 11–14 вересня 2018 р., Львів. 2018. Т. 2. С. 9–16.

977. Vysotska V., Lytvyn V., Kovalchuk V., Kubinska S., Dilai M., Rusyn B., Pohreliuk L., Chyrun L., Chyrun S., Brodyak O. Method of similar textual content selection based on thematic information retrieval. Комп'ютерні науки та інформаційні технології: матеріали XIV-ої Міжнародної науково-технічної конференції CSIT-2019 (Львів, 17–20 вересня 2019 р.). 2019. Т. 3. С. 1–6.

978. Каніщева О., Главчева Ю., Висоцька В. Визначення стилю автора для виявлення плагіату в академічному середовищі. System analysis and information technology, SAIT 2017, May 22–25, 2017, Kyiv. P. 78–79.

979. Литвин В. В., Оборська О. В., Висоцька В. А., Бобик І. О. Метод аналізу авторства тексту на основі стилеметрії. Матеріали міжнародної наукової конференції «Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту» (ISDMCI'2018) 21–27 трав. 2018 р., Залізний Порт, Україна. С. 240–243.

980. Висоцька В. А., Литвин В. В., Олешек О. І. Автоматизований моніторинг змін у Web-ресурсах. Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту : збірка наукових праць Міжнародної наукової конференції (с. Залізний Порт, 21–25 травня 2019 р.). 2019. С. 30–32.

981. Демчук А. Б., Литвин В. В., Висоцька В. А. Технологія персоналізованого поширення комерційного контенту через Web-ресурс Е-комерції. Інтелектуальні системи прийняття рішень та проблеми обчислювального

інтелекту : збірка наукових праць Міжнародної наукової конференції (с. Залізний Порт, 21–25 травня 2019 р.). 2019. С. 49–51.

982. Tymoshenko K., Vysotska V. Algorithm of text recognizing in Ukrainian on the video mode. Computational Linguistics and Intelligent Systems. Proceedings of the 4th International conference "Computational linguistics and intelligent systems" COLINS, Lviv, Ukraine, June 23-24, 2020. Vol. II: Workshop. P. 81–89.
983. Висоцька В. А. Суб'єктивізм трактування академічної доброчесності в межах наукової діяльності видавництв. Академічна доброчесність: виклики сучасності : збірник наукових есе учасників дистанційного етапу наукового стажування для освітян. Польща, Варшава, 28.09 – 06.11.2020. С. 31-35.
984. Vysotska V., Demchuk A., Lytvyn V. Features of the architecture for Internet commercial content management system based on methods of Machine Learning, Web mining and SEO technologies. Радіоелектроніка. Інформатика. Управління. 2019. № 4. С. 121–135.
985. Zdebskyi P., Vysotska V., Peleshchak R., Peleshchak I., Demchuk A., Krylyshyn M. An application development for recognizing of view in order to control the mouse pointer. CEUR Workshop Proceedings. 2019. Vol. 2386. P. 55–74.
986. Lytvyn V., Vysotska V., Mykhailyshyn V., Rzheuskyi A., Semianchuk S. System development for video stream data analyzing. Advances in Intelligent Systems and Computing (AISC). – 2020. Vol. 1020. P. 315–331.
987. Peleshko D., Rak T., Noennig J. R., Lytvyn V., Vysotska V. Drone monitoring system DROMOS of urban environmental dynamics. CEUR Workshop Proceedings. 2020. Vol. 2565. P. 178–193.
988. Krislata I., Katrenko A., Lytvyn V., Vysotska V., Burov Y. Traffic flows system development for smart city. CEUR Workshop Proceedings. 2020. Vol. 2565. P. 280–294. E-ISSN: 1613-0073
989. Vysotska V., Lytvyn V., Danylyk V., Vyshemyrska S., Lurie I., Luchkevych M. Detecting items with the biggest weight based on neural network and machine learning methods. Communications in Computer and Information Science. 2020. Vol. 1158. P. 383-396. ISSN1865-0929, E-ISSN1865-0937
990. Kalinina I., Vysotska V., Bidyuk P., Gozhyj A. Methods for forecasting nonlinear non-stationary processes in machine learning. Communications in Computer and Information Science. 2020. Vol. 1158. P. 470–485.
991. Matseliukh Y., Bublyk M., Vysotska V. Development of intelligent system for visual passenger flows simulation of public transport in Smart City based on neural network. CEUR Workshop Proceedings. 2021. Vol. 2870. P. 1087–1138.
992. Lytvyn V., Pashchetnyk O., Klymovych O., Polishchuk L., Kolb I., Burov Y., Vysotska V. Assessment of the hydro-meteorological conditions impact on the combat troops operations preparation and conduct in the geo-information subsystem of the automated battlefield management system. CEUR Workshop Proceedings. 2021. Vol. 2870. P. 1063–1076. E-ISSN: 1613-0073.
993. Dokhnyak B., Vysotska V. Intelligent Smart Home System Using Amazon Alexa Tools. CEUR Workshop Proceedings. 2021. Vol. 2917. P. 441-464. E-ISSN: 1613-0073.
994. Zdorenko Y., Lavrut O., Lavrut T., Lytvyn V., Burov Y., Vysotska V. Route Selection Method in Military Information and Telecommunication Networks Based on ANFIS. CEUR Workshop Proceedings. 2021. Vol. 2917. P. 514-524.
995. Коробчинський М. В Чирун, Л. Б., Висоцька В. А., Нич М. О. Особливості прогнозування результатів матчів у кіберспорті. Радіоелектроніка. Інформатика. Управління. 2017. № 3 (42). С. 95–105.
996. Кравець П. О., Литвин В. В., Висоцька В. А. Моделювання ігрової задачі призначення персоналу для виконання іт-проектів на основі онтологій. Радіоелектроніка, інформатика, управління. 2022. № 1 (60). С. 130–145.

997. Литвин В. В., Бублик М. І., Висоцька В. А., Мацелюх Ю. Р. Технологія візуальної симуляції пасажиропотоків у сфері громадського транспорту Smart City. Радіоелектроніка, інформатика, управління. 2021 № 4 (59). С. 106-121.

998. Литвин В. В., Висоцька В. А., Кучковський В. В., Оливко Р. М. Архітектура інформаційної системи інтеграції та формування контенту про криптовалюти на основі аналізу діяльності бірж. Вісник Національного університету "Львівська політехніка". 2018. № 901. С. 43–60.

999. Литвин В. В., Наум О., Висоцька В. А., Дверій М. В. Архітектура системи онлайн-туризму для пошуку та планування подорожей із урахуванням потреб користувача. Вісник Національного університету "Львівська політехніка" 2019. Вип. 6. С. 13–29.

1000. Lytvyn V., Peleshchak I., Peleshchak R., Vysotska V. Satellite spectral information recognition based on the synthesis of modified dynamic neural networks and holographic data processing techniques. CSIT : матеріали XIII-ої Міжнародної науково-технічної конференції (Львів,11–14 вересня 2018 р.). 2018. Т. 1. С. 330–334.

1001. Lytvyn V., Kuchkovskiy V., Vysotska V., Markiv O., Pabyrivskyy V. Architecture of system for content integration and formation based on cryptographic consumer needs. Комп'ютерні науки та інформаційні технології (CSIT-2018) : матеріали XIII-ої Міжнародної науково-технічної конференції (Львів,11–14 вересня 2018 р.). С. 391–395.

1002. Lytvyn V., Peleshchak I., R Peleshchak., Vysotska V. Information encryption based on the synthesis of a neural network and AES algorithm. AICT–2019 (Lviv, Ukraine, July 2–6 2019). P. 447–450.

1003. Lytvyn V., Vysotska V., Mykhailyshyn V., Peleshchak I., Peleshchak R., Kohut I. Intelligent system of a smart house. Advanced information and communication technologies, AICT–2019 : proceedings of the 3rd International conference (Lviv, Ukraine, July 2–6 2019). 2019. P. 282–287.

1004. Kalinina I., Vysotska V., Bidyuk P., Gozhyj A., Vasilev M., Malets R. Forecasting nonlinear nonstationary processes in machine learning task. Data stream mining & processing (DSMP) : proceedings of the IEEE Third international conference, Lviv, Ukraine. 2020. P. 28–32.

1005. Lytvyn V., Vovnyanka R., Oborska O., Dosyn D., Vysotska V., Panasyuk V. Intelligent agent behavior simulation based on reinforcement learning. Комп'ютерні науки та інформаційні технології : матеріали XV Міжнародної науково-технічної конференції, Збараж, 23–26 вересня, 2020. P. 285–290.

1006. Peleshchak R., Lytvyn V., Peleshchak I., Vysotska V. Stochastic Pseudo-Spin Neural Network with Tridiagonal Synaptic Connections. IEEE International Conference on Smart Information Systems and Technologies (SIST), 28-30 April 2021, Nur-Sultan, Kazakhstan. Art. 9465998.

1007. Висоцька В. А., Чирун Л. Б., Чирун Л. В. Інтернет-портал аматорських спортивних ігор. Міжнародна наукова конференція «Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту», 22–26 трав. 2017, Залізний Порт. С. 45–47.

1008. Литвин В. В., Висоцька В. А., Михайлишин В. Ю., Сем'янчук С. О. Розроблення інформаційної системи аналізу даних відеопотоку. Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту : збірка наукових праць Міжнародної наукової конференції (с. Залізний Порт, 21–25 травня 2019 р.). С. 94–97.

1009. Bengfort B. The Age of the Data Product, 2015. URL: http://bit.ly/2GJBEEP.

1010. Arun Kumar, Robert McCann, Jeffrey Naughton, and Jignesh M. Patel, Model Selection Management Systems: The Next Frontier of Advanced Analytics, 2015. URL: http://bit.ly/2GOFa0G.

1011. Hadley Wickham, Dianne Cook, and Heike Hofmann, Visualizing Statistical Models: Removing the Blindfold, 2015. URL: http://bit.ly/2JHq92J.

1012. Neal Caren, Using Python to see how the Times writes about men and women, 2013. URL: http://bit. ly/2GJBGfV.

1013. Tararoys. Hunspell dictionary for Ogden's Basic English. URL: https://gist.github.com/tararoys/76417eda6af9e331b587.

1014. Sennrich R., Haddow B., Birch A. Neural machine translation of rare words with subword units. arXiv preprint 2015. arXiv: 1508.07909. URL: https://arxiv.org/abs/1508.07909.

1015. Gage P. A new algorithm for data compression. C Users Journal, 1994, Vol. 12(2), 23-38. https://www.derczynski.com/papers/archive/BPE_Gage.pdf.

1016. Висоцька В.А. Концептуальна модель процесу формування семантики речення природною мовою. Інформаційні системи та мережі. Вісник НУ "Львівська політехніка", 2014, № 805. Стор. 258-278.

1017. Newell A., Langer S., Hickey M. The rôle of natural language processing in alternative and augmentative communication. Natural Language Engineering, 1998, Vol. 4(1), 1-16.

1018. Levenshtein V. I. Binary codes capable of correcting deletions, insertions, and reversals. In Soviet physics doklady, 1966, February, Vol. 10, No. 8, pp. 707-710.

1019. Bellman R., Kalaba R. Dynamic programming and statistical communication theory. Proceedings of the National Academy of Sciences of the United States of America, 1957, Vol. 43(8), 749.

1020. Bellman R., Kalaba R. On the role of dynamic programming in statistical communication theory. IRE Transactions on Information Theory, 1957, Vol. 3(3), 197-203.

1021. Bellmam R. Dynamic programming. Princeton university press. Princeton. New Jersey, 1957.

1022. Bellman R. On the approximation of curves by line segments using dynamic programming. Communications of the ACM, 1961, Vol. 4(6), 284.

1023. Wagner R. A., Fischer M. J. The string-to-string correction problem. Journal of the ACM (JACM), 1974, Vol. 21(1), 168-173.

1024. Gusfield D. Algorithms on stings, trees, and sequences: Computer science and computational biology. Acm Sigact News, 1997, Vol. 28(4), 41-60.

1025. Forney G. D. The viterbi algorithm. Proceedings of the IEEE, 1973, Vol. 61(3), 268-278.

1026. Даревич Р. Р., Досин ,Д. Г., Литвин В. В., Назарчук З. Т. Оцінка подібності текстових документів на основі визначення інформаційної ваги елементів бази знань. Штучний інтелект, 2006, Vol. 3, 500-509.

1027. Литвин В. В., Черна Т. І., Ковалевич В. М. Метод квазіреферування текстових документів на основі онтології предметної області. Відбір і обробка інформації, 2014, Vol. 41, 100-108.

1028. Даревич Р. Р. Підвищення ефективності інтелектуального аналізу тексту шляхом зважування понять у моделі онтології. Искусственный интеллект, 2005, Vol. (3), 571-577.

1029. Даревич Р., Досин Д., Литвин В. Метод побудови інтелектуальних метапошукових систем на основі адаптації онтології, 2008. URL: http://vlp.com.ua/files/24_13.pdf

1030. Литвин В. Метод видобування знань з природомовних текстів для автоматизованої розбудови онтологій. Автоматизированные системы управления и приборы автоматики, 2013, Vol. 164, 67-72.

1031. Литвин В. В., Гопяк М. Я., Оборська О. В., Вовнянка Р. В. Метод побудови інтелектуальних агентів на основі адаптивних онтологій. Вісник Національного університету Львівська політехніка. Серія: Інформаційні системи та мережі, 2015, Vol. 829, 186-200.

1032. Литвин В. В., Мороз О. В. Метод контекстного пошуку на основі тезаурусу предметної області. Восточно-Европейский журнал передовых технологий, 2013, Vol. 6(2 (66)).

1033. Simons G. F., Fennig C. D. Language Status. Ethnologue: Languages of the World. 21st edn. Dallas, 2018, Texas: SIL International. Online version: https://www. ethnologue. com/about/language-status (23 March, 2018).
1034. Згуровський М.З., Панкратова Н.Д. Основи системного аналізу. К.: Вид. група BHV, 2007. – 544 с.
1035. Волкова В.Н., Денисов А.А. Теория систем и системный анализ. – М.: Юрайт, 2010. – 680 с.
1036. Лямец В. И., Тевяшев А. Д. Системный анализ. Харьков, ХТУРЭ, 1998, 252 с.
1037. Яковлев С.В. Теория систем и системный анализ. – М.: Гор. линия телеком, 2015. – 320 с.
1038. Сурмин Ю.П. Теория систем и системный анализ. – К.: МАУП, 2003. – 368 с.
1039. Keygeneratortext. URL: http://msurf.ru/tools/keygeneratortext/.
1040. Keygeneratorurl. URL: http://webmasta.org/tools/keygeneratorurl/.
1041. Keywordstext. URL: http://www.keywordstext.therealist.ru/.
1042. Keygeneratortext. URL: http://syn1.ru/tools/keygeneratortext/.
1043. Terminology extraction. URL: http://labs.translated.net/terminology-extraction/.
1044. Advego. URL: http://advego.ru/text/seo/.

ДОДАТОК А. ТАБЛИЦІ

Таблиця А.1

Порівняльні ознаки українських/англійських лінгвістичних ознак [716, 862-864]

Частина мови	Українська мова	Англійська мова
Іменник	Є граматичний рід.	Відсутній граматичний рід.
	Поділ на роди чоловічого, жіночого й середнього	Поділ на людей з одного боку за статтю, і на явища, інші живі істоти і предмети.
	Сім відмінків	Два відмінки – загальний і присвійний
	Зв’язки через відмінки	Зв’язки через прийменники.
Артикль	–	Дві форми – неозначений та означений
Інфінітив	Проста форма	Крім простої як в українській є ще 5 складних
Займенник	Поділ на 9 розрядів	Поділ на 7 розрядів
	2 форми 2ої особи: в однині *ти*, у множині *ви*.	Відсутній особовий займенник *ти* (його функцію виконує займенник *ви – you*)
	Особові: вона замінює всі іменники жіночого роду; він – чоловічого роду; воно – середнього роду.	Особові: he – живі істоти чоловічої статті; she – живі істоти жіночої статі; it – тварини або неживі предмети.
Дієслово	Вираз завершеності чи незавершеності дій, які не завжди залежать від тих чинників стосовно англійського дієслова.	Або відбувалася до якоїсь іншої дії в минулому тощо.чи вона відбувається в момент мовлення, чи протягом часу, який ще триває, або відбувається дія взагалі, завжди, постійно, повторно
	–	Часто застосовують з прислівниками без лексичного значення
Безособові дієслова	Наявні, наприклад, вечоріє.	–
Герундій	–	6 форм
Прикметник	Узгоджуються з іменником та змінюються за відмінками, числами, родами	Не змінюються і не узгоджуються за відмінками, числами, родами
Дієприкметник	Лише одна форма	2 форми дієприкметника: теперішнього й минулого часу, має деякі дієприслівникові властивості
Дієприслівник	2 форми	Відсутній у «чистому» вигляді
Числівник	Узгоджуються за відмінками, родами	Не узгоджуються за відмінками, родами
Службові слова	Прислівник, прийменник, сполучник та вигук не мають суттєвих відмінностей	
Речення	Порядок слів у реченні вільний.	Порядок: підмет – присудок – інші члени речення.

Таблиця А.2

Орфографічні/фонетичні особливості української/англійської мови [716, 862]

Одиниця	Українська мова	Англійська мова
Звуки	38	44
Літери	32	26
Голосні літери	10	6
Приголосні літери	22	20
Голосні звуки	6, поділу немає.	12, є довгі й короткі, заміна одного звука іншим призводить до зміни значення слова.
Приголосні звуки	32, є тверді і м’які, перед деякими голосними відбувається пом’якшення приголосних (звук [с] у *сіно* і *сірий*).	24, поділу немає; майже всі вимовляються твердо перед будь-яким голосним.
Дзвінки приголосні	В таких випадках оглушуються, наприклад, віз [віс], Бог [бох], дуб [дуп]	Завжди вимовляються дзвінко в кінці слова та перед глухими приголосними; їх оглушення часто приводить до зміни значення слова.
Дифтонги (звуки)	Немає	8, деякі голосні складаються з 2 елементів, що вимовляються в межах одного складу.

Таблиця А.3

Критерії графемного аналізу вхідного тексту [211-212]

№	Абревіатура	Розшифровка	Назва
1	*Grammar*	Grammar	Граматика графемного аналізу тексту
2	*Alphabet*	Alphabet	Алфавіт граматики ГА
3	*Terms*	Term	Термінальні символи граматики
4	*Symbol*	Initial character	Початковий символ граматки

№	Абревіатура	Розшифровка	Назва
5	*PrRules*	Production rules	Продукційні правила граматики
6	*Sb*	Symbol	Символ/знак
7	*Sp*	Space	Пробіл
8	*Dgt*	Digit	Цифра
9	*Ssb*	Special symbol	Спеціальний символ
10	*Ssg*	Syntactic sign	Синтаксичний знак
11	*Ltr*	Letter	Літера
12	*Lat*	Latin letter	Латинські літери
13	*Cyr*	Cyrillic letter	Літери кирилиці
14	*Eng*	English alphabet	Англійський алфавіт
15	*Ger*	German alphabet	Німецький алфавіт
16	*Pol*	Polish alphabet	Польський алфавіт
17	*Ukr*	Ukrainian alphabet	Український алфавіт
18	*Rus*	Russian alphabet	Російський алфавіт
19	*Osb*	Official symbol	Службовий символ
20	*Bsb*	Brackets	Дужки
21	*Msb*	Mathematical symbol	Математичний символ
22	*Cpl*	Capital letter	Заголовна літера
23	*Sml*	Small letter	Мала літера
24	*Lcp*	Latin capital letter	Латинська заголовна літера
25	*Lsm*	Latin small letter	Латинська мала літера
26	*Ccp*	Cyrillic capital letter	Заголовна літера кирилиці
27	*Csm*	Cyrillic small letter	Мала літера кирилиці
28	*Ecp*	English capital letter	Англійська заголовна літера
29	*Esm*	English small letter	Англійська мала літера
30	*Gcp*	German capital letter	Німецька заголовна літера
31	*Gsm*	German small letter	Німецька мала літера
32	*Pcp*	Polish letter	Польська заголовна літера
33	*Psm*	Polish small letter	Польська мала літера
34	*Ucp*	Ukrainian capital letter	Українська заголовна літера
35	*Usm*	Ukrainian small letter	Українська мала літера
36	*Rcp*	Russian capital letter	Російська заголовна літера
37	*Rsm*	Russian small letter	Російська мала літера
38	*Cnl*	Consonant letter	Приголосна літера
39	*Vwl*	Vowel letter	Голосна літера
40	*Lcc*	Latin capital consonant letter	Латинська заголовна приголосна літера
41	*Lsc*	Latin small consonant letter	Латинська мала приголосна літера
42	*Lcv*	Latin capital vowel letter	Латинська заголовна голосна літера
43	*Lsv*	Latin small vowel letter	Латинська мала голосна літера
44	*Ccc*	Cyrillic capital consonant letter	Заголовна приголосна літера кирилиці
45	*Csc*	Cyrillic small consonant letter	Мала приголосна літера кирилиці
46	*Ccv*	Cyrillic capital vowel letter	Заголовна голосна літера кирилиці
47	*Csv*	Cyrillic small vowel letter	Мала голосна літера кирилиці
48	*Ecc*	English capital consonant letter	Англійська заголовна приголосна літера
49	*Esc*	English small consonant letter	Англійська мала приголосна літера
50	*Ecv*	English capital vowel letter	Англійська заголовна голосна літера
51	*Esv*	English small vowel letter	Англійська мала голосна літера
52	*Gcc*	German capital consonant letter	Німецька заголовна приголосна літера
53	*Gsc*	German small consonant letter	Німецька мала приголосна літера
54	*Gcv*	German capital vowel letter	Німецька заголовна голосна літера
55	*Gsv*	German small vowel letter	Німецька мала голосна літера
56	*Pcc*	Polish capital consonant letter	Польська заголовна приголосна літера
57	*Psc*	Polish small consonant letter	Польська мала приголосна літера
58	*Pcv*	Polish capital vowel letter	Польська заголовна голосна літера
59	*Psv*	Polish small vowel letter	Польська мала голосна літера
60	*Ucc*	Ukrainian capital consonant letter	Українська заголовна приголосна літера
61	*Usc*	Ukrainian small consonant letter	Українська мала приголосна літера
62	*Ucv*	Ukrainian capital vowel letter	Українська заголовна голосна літера
63	*Usv*	Ukrainian small vowel letter	Українська мала голосна літера
64	*Rcc*	Russian capital consonant letter	Російська заголовна приголосна літера
65	*Rsc*	Russian small consonant letter	Російська мала приголосна літера
66	*Rcv*	Russian capital vowel letter	Російська заголовна голосна літера
67	*Rsv*	Russian small vowel letter	Російська мала голосна літера

Таблиця А.4

Правила класифікації графем у вигляді послідовності символів [211-212]

№	Назва з поясненням	Правило класифікації графеми
1	Вхідний текст як набір знаків та пробілів;	*Gr* : = *Sb* ∪ *Sp*
2	Пробіл як термінальний символ;	*Sp* = {_}
3	Кортеж літер, цифр, спецсимволів та синтаксичних знаків;	*Sb* : = *Ltr* ∪ *Dgt* ∪ *Ssb* ∪ *Ssg*
4	Кортеж латинських, кириличних та інших літер деяких європейських мов і апострофа;	*Ltr* : = *Lat* ∪ *Cyr* ∪ *Eng* ∪ *Ger* ∪ *Pol* ∪ *Ukr* ∪ ∪ *Rus* ∪ *Cpl* ∪ *Sml* ∪ *Cnl* ∪ *Vwl* \ ∪ {‘}
5	Множина цифр;	*Dgt* = { 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 }
6	Кортеж всіх службових символів;	*Ssb* : = *Osb* ∪ *Bsb* ∪ *Msb*
7	Множина синтаксичних знаків;	*Ssg* : = { «, », “, ”, :, -, ?, !} ∪ {;} ∪ {,}∪ {.}
8	Множина заголовних літер відповідних мов;	*Cpl* := *Lcp*∪*Ccp*∪*Ecp*∪*Gcp*∪*Pcp*∪*Ucp*∪*Rcp*
9	Множина малих літер відповідних мов;	*Sml* := *Lsm*∪*Csm*∪*Esm*∪*Gsm*∪*Psm*∪*Usm*∪*Rsm*
10	Множина всіх латинських літер;	*Lat* : = *Lcp* ∪ *Lsm*
11	Множина всіх кириличних;	*Cyr* : = *Ccp* ∪ *Csm*
12	Множина всіх англійських літер;	*Eng* : = *Ecp* ∪ *Esm*
13	Множина всіх німецьких літер;	*Ger* : = *Gcp* ∪ *Gsm*
14	Множина всіх польських літер;	*Pol* : = *Pcp* ∪ *Psm*
15	Множина всіх українських літер;	*Ukr* : = *Ucp* ∪ *Usm*
16	Множина всіх російських літер;	*Rus* : = *Rcp* ∪ *Rsm*
17	Множина службових символів;	*Osb* : = { №, %, /, @, #, $, &, *, \ }
18	Множина термінальних символів дужок;	*Bsb* : = { [,], {, }, (,) }
19	Множина математичних символів;	*Msb* : = { +, <, >, = }
20	Множина всіх приголосних літер (великих та малих) відповідних мов;	*Cnl* : = *Ecc*∪*Esc*∪*Gcc*∪*Gsc*∪*Pcc*∪*Psc*∪ ∪*Ucc*∪*Usc*∪*Rcc*∪*Rsc*
21	Множина всіх голосних літер (великих та малих) відповідних мов;	*Vwl* : = *Ecv*∪*Esv*∪*Gcv*∪*Gsv*∪*Pcv*∪*Psv*∪ ∪*Ucv*∪*Usv*∪*Rcv*∪*Rsv*
22	Множина латинських заголовних літер;	*Lcp* : = *Lcc* ∪ *Lcv* ∪ { Q, V, X }
23	Множина латинських малих літер;	*Lsm* : = *Lsc* ∪ *Lsv* ∪ { q, v, x }
24	Множина кириличних заголовних літер;	*Ccp* : = *Ccc* ∪ *Csv* ∪ { Ь, Й }
25	Множина кириличних малих літер;	*Csm* : = *Csc* ∪ *Csv* ∪ { ь, й }
26	Множина англійських заголовних літер;	*Ecp* : = *Lcc* ∪ *Lcv* ∪ {Q, V, X }
27	Множина англійських малих літери;	*Esm* : = *Lsc* ∪ *Lsv* ∪ { q, v, x }
28	Множина німецьких заголовних літер;	*Gcp* : = *Lcc* ∪ *Lcv* ∪ { Ä, Ö, Ü, Q, V, X }
29	Множина німецьких малих літер;	*Gsm* : = *Lsc* ∪ *Lsv* ∪ {ä, ö, ü, ß, q, v, x }
30	Множина польських заголовних літер;	*Pcp* : = *Lcc* ∪ *Lcv* ∪ { Ą, Ć, Ę, Ł, Ń, Ó, Ś, Ź, Ż }
31	Множина польських малих літер;	*Psm* : = *Lsc* ∪ *Lsv* ∪ { ą, ć, ę, ł, ń, ó, ś, ź, ż }
32	Множина українських заголовних літер;	*Ucp* : = *Ccc* ∪ *Ccv* ∪ { Є, І, Ї, Ґ }
33	Множина українських малих літер;	*Usm* : = *Csc* ∪ *Csv* ∪ { є, і, ї, ґ }
34	Множина російських заголовних літер;	*Rcp* : = *Ccc* ∪ *Ccv* ∪ { Ы, Э, Ъ }
35	Множина російських малих літер;	*Rsm* : = *Csc* ∪ *Csv* ∪ { ы, э, ъ }
36	Латинські заголовні приголосні літери;	*Lcc* :={B,C,D,F,G,H,J,K,L,M,N,P,R,S,T,W,Z}
37	Латинські заголовні голосні літери;	*Lcv* : = { A, E, I, O, U, Y}
38	Термінальні латинські малі приголосні літери;	*Lsc* := {b,c,d,f,g,h,j,k,l,m,n,p,r,s,t,w,x,z}
39	Термінальні латинські малі голосні літери;	*Lsv*: = { a, e, i, o, u, v, y }
40	Множина термінальних кириличних заголовних приголосних літер;	*Ccc* : = { Б, В, Г, Д, Ж, З, К, Л, М, Н, П, Р, С, Т, Ф, Х, Ц, Ч, Ш, Щ }
41	Кириличні заголовні голосні літери;	*Csv* : = { А, Е, И, О, У, Ю, Я }
42	Кириличні малі приголосні літери;	*Csc* := {б,в,г,д,ж,з,к,л,м,н,п,р,с,т,ф,х,ц,ч,ш,щ}
43	Термінальні кириличні малі голосні літери.	*Csv* : = { а, е, и, о, у, ю, я }

Таблиця А.5

Класифікація алгоритмів стемінгу лексем природної мови [256-259, 534-535]

Назва	Особливість	Недоліки	Переваги	Приклад
Лематизація	1. Ідентифікація частин мови у реченні (POST). 2. Застосування правила стемінгу згідно частини	Сильна залежність від коректності розпізнавання частин мови.	Висока якість і мінімальний відсоток помилок при коректному розпізнаванні частини мови слова.	Лексеми *вітальне* (прикметник) та *пальне* (іменник) проходять через різні ланцюжки правил: Rules = { Ending (*льне*) → Cut (*ьне*);

Назва	Особливість	Недоліки	Переваги	Приклад
	мови слова для приведення в нормальну форму. 3. Пошук в словнику відповідності.			Ending (*льне*) → Cut (*е*) }.
Відсікання суфіксів та флексій	Застосування правил скорочення слова до основи (з префіксом) Rules = { Ending (*льне*) → Cut (*ьне*); Ending (*ційним*) → Cut (*ійним*); Ending (*ційний*) → Cut (*ійний*); Ending (*ційне*) → Cut (*ійне*); Ending (*ційна*) → Cut (*ійна*); …….}.	Наявність хибних виведень і спотворень форм стемінгу (*пальне* стане *пал* замість *пальн*). Із-за особливості конкретної мови множина правил є різного рівня складності та кількості. Присутнє опрацювання винятків, наприклад, при чергуванні літер в основі слова (*бігом*, *біжу*). Необхідне ускладнення правил, де просте відсікання негативно впливає на якість стемінгу.	Продуктивний та компактний, так як число правил набагато менша за таблиці з усіма словоформами для всіх частин мови, осіб, відмінків, родів тощо.	Word={*національне*} → Stemming={*національ*}; Word={*кульмінаційний*} → Stemming={*кульмінац*}; Word={*приватизаційний*} → Stemming={*приватизац*}; Word={*цивілізаційний*} → Stemming={*цивілізац*}; Word={*інформаційний*} → Stemming={*інформац*}; ……
Відокремлення префіксів	Поряд із відсіканням закінчень та суфіксів лексеми, відокремлення при наявності префіксів.	Ймовірність утворення протилежних за змістом слів, тобто Word={*незалежний*} → Stemming={*залежн*}.	Суттєва важливість лише для деяких природніх мов.	Word={*проголошую*, *наголошувати*, *виголошував*} → Stemming={*голошу*}.
ІІП за таблицею	В словнику зібрані всі/ймовірні варіанти слів та їх форми після стемінгу.	Не працює з новими словами або з тими, форми яких не представленні словнику. Великі розміри таблиці для мов із складною морфологією (аглютинативні, слов'янські, в тому числі українська).	Простота, швидкість та зручність опрацювання винятків з правил. Для мов із простою морфологією (англійська) таблиці малі.	Stemming={*інформац*} → Word={*інформаційний*, *інформаційна*, *інформаційне*, *інформаційним*, *інформаційними*, *інформаційних*, *інформаційні*, *інформаційній*, *інформаційнім*, *інформаційного*, *безпритульної*, *інформаційному*, *інформаційною*, *інформаційну*}
ІІП відповідності	Застосвують базу знань лише з основами слів після стемінгу.	Ймовірність помилок стемінгу зростає при некоректному описі правил та формуванні таблиці закінчень/флексій	Через систему правил (довжина збігу слова та його основи) ІІП для найвідповіднішої форми з БЗ.	KnowledgeBase={*чорн*, *чорняв*} → Word={*чорнява*} → Count={4, 6}→ Stemming={*чорнява*}. Алгоритм обере довший варіант.
Стемінг різними мовами	Орієнтація на конкурентну мову.	Від особливостей мови залежить складність написання алгоритмів стемінгу.	Основні академічні/ практичні роботи присвячені лише англійській.	Якщо стемінг англійської є простою задачею, то стемінг для української - на декілька рівнів складніша.
Стемінг українською	Варіанти стемінгу для української мови як частина інших NLP-задач, але в більшості випадків є комерційним проектами	Мало досліджень в цьому напрямку та відсутня вільна загальнодупна з відкритим кодом реалізація подібних алгоритмів.	Певні кроки у цьому напрямку вже зроблені.	Детальний опис некомерційного алгоритму стемінгу для української є справою часу.

Назва	Особливість	Недоліки	Переваги	Приклад
Стохастичні алгоритми	Базуються на ймовірності визначення основи слова на основы БЗ. Лематизація має стохастичні властивості, коли частину мови визначають без урахування контексту, в якому це слово було вжито в реченні.	Після опрацювання слова може з'явитися декілька варіантів основи слова, з яких алгоритм обере найімовірніший варіант. Ймовірність помилок стемінгу зростає. Перевага віддається найвірогіднішій частині мови для цього слова.	Є лише одне логічне правило за яким від слова відсікаємо останні літери. Алгоритми мають здатність навчатися і чим краща та більша база навчання тим кращий результат їх роботи. База знань для цих алгоритмів - це набір логічних правил та таблиці ІІП.	Word={*особистість*}→ Stemming={*особист*}→ End={*ість*}; Word={*спогади*}→ Stemming={*спогад*}→ End={*и*}; Word={*дивними*}→ Stemming={*дивн*}→ End={*ими*}, де End – результат навчання алгоритму, тобто Word(*кияни*) → {End(*ість*) = FALSE, End(*и*) = TRUE, End(*ими*) = FALSE}→ Cut (*и*) або Word(*чуйними*) → {End (*ість*) = FALSE, End (*и*) = TRUE, End (*ими*) = TRUE}→ Cut (*и*) OR Cut (*ими*).
Гібридний підхід	Використовують комбінацію наведених вище алгоритмів.	Ймовірність помилок стемінгу зростає при некоректному описі правил та формуванні таблиці закінчень	Таблиця містить не всі словоформи, а винятки з правил, які невірно опрацьовуються алгоритмом відсікання.	Наприклад, алгоритм може використовувати метод відсікання закінчень та суфіксів, але на першому етапі виконувати ІІП по таблиці.

Таблиця А.6

Лінгвістичні характеристики деяких класів морфем основ дієслів [404, 716, 882]

Дієслово	Розбір	Основи	Дієприкметник
фарбувати(ся)	фарб-ува-ти(-ся)	фарб-($t, \bar{d}, I, atem, y, ся - \overline{ся}$)	фарб-ова-н-ий
усміхнутися	усміх-ну-ти-ся	усміх-($\bar{t}, d, I, atem,$ *н, ся*)	усміх-н-ен-ий
стогнати	стогн-а-ти	стогн-($\bar{t}, \bar{d}, I, \tilde{a}, \emptyset, \overline{ся}$)	стогн-уч-ий
спитати(ся)	спит-а-ти(-ся)	спит-($t, \bar{d}, I, a, \emptyset, ся - \overline{ся}$)	спит-а-юч-ий
сміятися	смій-а-ти-ся	смій-($\bar{t}, \bar{d}, I, \tilde{a}, \emptyset, ся$)	смій-уч-ий
розфарбувати(ся)	розфарб-ува-ти(-ся)	розфарб-($t, d, I, atem, y, ся - \overline{ся}$)	розфарб-ова-н-ий
привести(ся)	привес-ти(-ся)	привес-($t, d, I, atem, \emptyset, ся - \overline{ся}$)	приве-ен-ий
поділити(ся)	поділ-и-ти(-ся)	поділ-($t, d,$ II, $\tilde{\imath}, \emptyset, ся - \overline{ся}$)	поділ-ен-ий
побудувати(ся)	побуд-ува-ти(-ся)	побуд-($t, d, I, atem, y, ся - \overline{ся}$)	побуд-ова-н-ий
нести(ся)	нес-ти(-ся)	нес-($t, \bar{d}, I, atem, \emptyset, ся - \overline{ся}$)	нес-ен-ий
молоти(ся)	мол-о-ти(-ся)	мол-($t, d, I, o, \emptyset, ся - \overline{ся}$)	мол-о-т-ий і мел-ен-ий
малювати(ся)	мал-ювa-ти(-ся)	мал'-($t - \bar{t}, d - \bar{d}, I, atem, y, ся - \overline{ся}$)	малй-ова-н-ий
любити(ся)	люб-и-ти(-ся)	люб-($t, \bar{d},$ II, $\tilde{\imath}, \emptyset, ся - \overline{ся}$)	любл-ен-ий
кохати(ся)	кох-а-ти(-ся)	кох-($t, \bar{d}, I, a, \emptyset, ся - \overline{ся}$)	кох-а-юч-ий
змарніти	змарн-і-ти	змарн-($\bar{t}, d, I, i, \emptyset, \overline{ся}$)	змарн-і-л-ий
запізнюватися → запізнитися	запізн-юва-ти-ся → запізн-и-ти-ся	запізн-($\bar{t}, d, I, \tilde{\imath}, \tilde{y}, ся$)	запізн-юва-н-ий → запізн-ен-ий
досліджувати(ся) → дослідити(ся)	дослідж-ува-ти(-ся) → дослід-и-ти(-ся)	дослідж-($t, d - \bar{d}, I, \tilde{\imath}, \tilde{y}, ся - \overline{ся}$)	дослідж-ува-н-ий → дослідж-ен-ий
втручатися	втруч-а-ти-ся	втруч-($\bar{t}, \bar{d}, I, a, \emptyset, ся$)	втруч-ен-ий
втратити → втрачати	втрат-и-ти → втрач-а-ти	втрач-($t, d - \bar{d},$ II, $\tilde{\imath}, \emptyset, \overline{ся}$)	втрач-ен-ий
вести(ся)	вес-ти(-ся)	вес-($t - \bar{t}, \bar{d}, I, atem, \emptyset, ся - \overline{ся}$)	вед-ен-ий
будувати(ся)	буд-ува-ти(-ся)	буд-($t - \bar{t}, d - \bar{d}, I, atem, y, ся - \overline{ся}$)	буд-ова-н-ий
автоматизувати(ся)	автоматиз-ува-ти(-ся)	автоматиз-($t - \bar{t}, d - \bar{d}, I, atem, y, ся - \overline{ся}$)	автоматиз-ова-н-ий

Таблиця А.7

Основні правила формування українських дієприкметників [404, 716, 882]

Клас	Назва	Складові правила
I	Правило загальної будови	1. У словоформу повинно входити не більше 1 морфеми кожного класу. 2. Морфеми повинні застосовуватися в порядку нумерації класів. 3. Морфеми класів 1, 4, 5 (основа + суфікс + флексія) є обов'язковими.
II	Правило несумісності	Лексема одночасно не може містити: 1. Морфеми класів 2 і 3 (тематичний елемент і суфікс). 2. Основу з ознакою $\overline{ся}$ і закінчення *-ся*.

		3. Основу з ознакою $a/i/o$ і суфікс дієприкметника з ознакою act. 4. Основу з ознакою $\emptyset$ і суфікс для утворення форм дієслів. 5. Основу з ознакою d за відсутності дієслівного суфікса і дієприкметниковий суфікс з ознакою $pres$ (від доконаного виду дієслів неможливі теперішнього часу дієприкметники). 6. Основу з ознакою I і без ознаки $atem$ і суфікс дієприкметника з ознакою II (дієслова I дієвідмінювання не допускають суфіксів II дієвідмінювання). 7. Основу з ознакою II за відсутності суфікса для утворення форм дієслів і суфікс дієприкметника з ознакою I. 8. Суфікс для утворення форм дієслів і суфікс дієприкметника з ознакою II (суфікс для утворення форм дієслів переводить будь-яке дієслово в I дієвідміні). 9. Основу з ознакою $atem$ і суфікс дієприкметника з ознакою II, відмінний від *-ач-/-яч-* (не ТЕ-дієслова не мають суфіксів II дієвідміни, за винятком *-ач-/-яч-*). 10. Основу з ознакою $\tilde{\imath}$, $\tilde{a}$ або o, тематичний елемент і суфікс дієприкметника, що починається голосною (якщо при даній основі ТЕ не обов'язковий, то перед суфіксом дієприкметника, що починається голосною, він не використовується). 11. Основу з ознакою $atem$ (відповідно без ознаки $atem$) і суфікс -юч-/-яч- (відповідно -уч-/-ач-), наприклад, *зітхаю(ть) – зітхаючий*, *співаю(ть) – співаючий*, *квітну(ть) – квітнучий*, *лежа(ть) – лежачий*. Ці форми в сучасній українській мові мають обмежене вживання. 12. Суфікс дієприкметника з ознакою act і флексію з ознакою $\bar{f} = o$ (незмінна форма дієслова твориться від пасивних дієприкметників шляхом заміни закінчення на $\bar{f} = o$, наприклад, *зроблений – зроблено*, *забитий – забито*, *написаний – написано*, *розглянутий – розглянуто*). 13. Суфікс для утворення форм дієслів доконаного/недоконаного виду переважно іншомовного походження і суфікс дієприкметника з ознаками act, наприклад, *наслідувати – наслідуваний*, *гарантувати – гарантований*, *інтенсифікувати – інтенсифікований*, *засохнути – засохлий*, *телеграфувати – телеграфований*, *яровизувати – яровизований*, *організувати – організований*, *організовувати – організований*, *телефонувати – телефонований*, *воєнізувати – воєнізований*, *атакувати – атакований*, *промокнути – промоклий*. 14. Суфікс дієприкметника і закінчення -ся (дієприкметники не можуть мати *-ся*). 15. Основу з ознакою $\tilde{\imath}$ (відповідно з ознакою $\tilde{a}$) і ТЕ, відмінний від *-и(і,ї)-* (відповідно від *-а-/-я-*).
III	Правило невіддільності	Словоформа обов'язково повинна містити: 1. За наявності основи з ознакою i – ТЕ *-и(і,ї)-*. 2. За наявності основи з ознакою a – ТЕ *-а-/-я-*. 3. За наявності основи з ознакою $atem$ і суфікса дієприкметника з початком на приголосний – або ТЕ, або суфікс для утворення форм дієслів доконаного і недоконаного виду переважно іншомовного походження. 4. За наявності основи інфінітива з ознакою на *-а* (*-я*), *-ува-* (*-ювa-*), *-овува-* – до неї додається суфікс *-н-* (*-ий*, *-а*, *-е*, *-і*), наприклад, *посія-(ти) – посіяний*, *чита-(ти) – читаний*, *розпиля-(ти) – розпиляний*, *писа-(ти) – писаний*, *зігна-(ти) – зігнаний*; *загоювати – загоювaний*, *оспівувати – оспівуваний*, *застосовувати – застосовуваний*; суфікс *-ува-* (*-ювa-*), якщо наголос переходить на перший голосний, змінюється на *-ова-*, наприклад, *роздрукува(ти) – роздрукований*, *сформулюва(ти) – сформульований*, *реконструюва(ти) – реконструйований*, *запрограмува(ти) – запрограмований*. 5. За наявності основи з ознакою o – ТЕ *-ор(л)о-* та можливість утворення паралельних форм дієприкметників для дієслів інфінітива (*колоти – колотий* і *колений*, *пороти – поротий* і *порений*, *молоти – молотий* і *мелений*). 6. За наявності основи із ознакою ся – відсутність в дієприкметниках частинки *ся*.
IV	Морфонологічні і фонологічні правила	Морфонологічні правила відносяться до послідовностей графем, і обов'язково враховують їх морфологічну роль. Фонологічні правила мають справу просто із послідовностями фонем, незалежно від їхнього морфологічного статусу. 1. Якщо основа інфінітива закінчується на голосні *-и*, *-і* (*-ї*) або приголосні, то формотворчим є суфікс *-ен-* (*-єн-*); кінцеві голосні основи випадають, а приголосні здебільшого зазнають змін, наприклад, *втрат-и-ти → втрач-ен-ий*. 2. Усі дієслова на *-отіти* I дієвідміни, що мають відповідники на *-отати* (*цокотіти – цокотять*, а *цокотати – цокочуть*): *муркотіти*, *булькотіти*, *тріскотати* тощо. Деякі дієслова з основою на *-отати* I дієвідміни не мають відповідники на *-отіти*, наприклад, *бельк-ота-ти → бельк-оч-у*, *бельк-оч-уть*, *мурм-ота-ти → мурм-оч-у*, *мурм-оч-уть*. Щоб описати невраховані тут випадки типу *бельк-ота-ти* (чергування *і/а* неможливе) або *цокотіти – цокотати* (чергування *і/а* можливо, але не обов'язково), необхідно ввести ще одну ознаку основ: чергування *і/а* перед *-ти* можливе/неможливе/обов'язкове. 3. У словоформі, що містить суфікс *-ува-* (*-ювa-*) та наголос переходить на перший голосний, суфікс змінюється на *-ова-*, наприклад, *реконструюва(ти) – реконструйований*, *сформулюва(ти) – сформульований*, *роздрукува(ти) – роздрукований*, *запрограмува(ти) – запрограмований*. 4. У пасивних дієприкметниках *-н-* не подвоюється, наприклад, *намальований*, *зав'язаний*, *зроблений*, *натхнений* тощо. 5. В основах дієслів суфікс *-ну-* при зміні виду не зберігається, наприклад, *стукнути* (d, що зробити) – *стукати* ($\bar{d}$, що робити), *крикнути* (d, що зробити) – *кричати* ($\bar{d}$, що робити). При утворені дієприкметників, як правило, також випадає, наприклад, *засохну(ти) – засохлий*, *промокну(ти) – промоклий*. 6. Між двома сусідніми голосними, що належать до різних морфем, з'являється *j*, наприклад, *розділ′ + а + йуч + ий*, або *посij + а + н + ий*, *розпил' + а + н + ий*. 7. При словозміні та словотворенні у дієслівних формах *г-ж*, *к-ч*, *х-ш*, наприклад, *берегти – бережу – бережений*, *стерегти – стережу – стережений*. 8. При словозміні та словотворенні у коренях дієслів (Таблиця А.6). 9. При утворенні дієприкметників в деяких випадках відбувається чергування приголосних в особових формах (Таблиця А.6). 10. Якщо основа інфінітива закінчується на голосні *-и*, *-і* (*-ї*) або приголосні, та формотворчим є суфікс *-ен-* (*-єн-*), то кінцеві голосні основи випадають, а приголосні зазвичай зазнають змін, наприклад, *пекти – печений*, *заспокоїти – заспокоєний*, *вертіти – верчений*, *пустити – пущений*, *запрягти – запряжений*, *змусити – змушений*, *узгодити – узгоджений*, *вразити – вражений*, *загоїти – загоєний*. Перед цими

		суфіксами після губних з'являється *-л-*, наприклад, *купити – куплений, зробити – зроблений, вловити – вловлений, зломити – зломлений*. 11. Інколи утрачається *-ва-* в залежності від часу дієслова, наприклад, *вбивати – вбити, купувати – купити,* але *друкувати – надрукувати, співати – проспівати*. 12. В словоформі, яка має суфікс *-у(ю)ва-*, або основа має закінчення *и/а* коренева голосна *о* в деяких випадках замінюється на *а*. Щоби описувати невраховані тут випадки типу *заспокоїти – заспокоювати* та *заспокоєний – заспокоюваний* (чергування *о/а* неможливе) або *ломити – ламати* та *ломлений – ламаний* (чергування *о/а* можливе, але не обов'язкове), необхідно ввести ще одну ознаку основи – чергування *о/а* перед *-у(ю)ва-* можливе/неможливе/обов'язкове. Чергування у коренях дієслів відбувається для таких голосних (Таблиця А.6). 13. Існують правила застосування дієслівних суфіксів (Таблиця А.6). 14. Дієприкметникові суфікси не подвоюються, оскільки наголос у дієприкметниках падає на корінь (Таблиця А.6). 15. Перед суфіксами *-е(є)н-, -у(ю)ва-, -ова-, -овува-* тверді кінцеві приголосні атематичних основ пом'якшуються: *д-д′, с-с′* і т. д. 16. Перед суфіксами *-е(є)н-, -у(ю)ва-, -ова-, -овува-* кінцева приголосна основи *-с′-* замінюється на *-ш-*, а кінцева приголосна *-б′-* на *-бл′-* (аналогічно, *д′-ж, т′-ч, в′-вл* і т. д.; але в нашому списку немає основ на *-д′-, -т′-, -в′-*), наприклад, *любити – люблю – люблений, полюбляти – полюблений, висіти – вишу, вивішувати – вивішений*; *улюблений, робити – роблю, роблений, виробляю – вироблений*. 17. Незмінювана форма дієслова твориться від пасивних дієприкметників шляхом заміни закінчення на суфікс *-о*, наприклад, *зроблений – зроблено, забитий – забито, написаний – написано, розглянутий – розглянуто*. Використовувати форму на *-но*, *-то* треба замість пасивних дієприкметників, коли є потреба наголосити на дії, а не на ознаці, наприклад, *урок закінчено, книжки здано*. 18. Поєднання *ju* замінюється на *і*.
V	Графічно-орфографічні правила	1. Поєднання *ja, jy, je, ji* зображаються буквами *я, ю, є, ї* відповідно. 2. Поєднання *X′a, X′y, X′e, X′i, X′u* зображаються на листі як *Хя, Хю, Хє, Хї, Хі* відповідно (*X′* – будь-яка парна м'яка приголосна).

Таблиця А.8

Додаткові уточнення морфонологічних та фонологічних правил [534-535, 716]

№	Правило	Приклад
A	*Основні правила чергування приголосних в особових формах*	
1	Дієвідміна I – приголосні змінюються наприкінці основи, якщо є чергування в 1-й особі однини – *г-ж, з-ж, к-ч, х-ш, с-ш, т-ч, ст-щ, ск-щ*	*засвистати – засвищу, хотіти – хочу, чесати – чешу, колихати – колишу, мазати – мажу, могти – можу, полоскати – полощу, пекти –печу – печений*;
2	Дієвідміна II – звукові зміни маємо лише в 1-й особі однини – *д-дж, т-ч, з-ж, с-ш, зд-ждж, ст-щ*	*їздити – їжджу, просити – прошу, возити – вожу, тремтіти – тремчу, водити – воджу, мостити – мощу – мощений*. Виняток становить лише дієслово *бігти* (і похідні: *перебігти, забігти* тощо), у якому *г* чергується з *ж* в усіх особових формах, наприклад: *бігти – біжу, біжиш, біжить, біжать* (*вибігти – вибіжу, вибіжиш* тощо);
B	*Правила чергування приголосних у коренях дієслів*	
1	*б-бл*	*полюбляти – полюблений, любити – люблю – люблений, улюблений, робити – роблю, роблений, виробляю – вироблений*;
2	*в-вл*	*ловити – ловлю, виловлювати – виловлений*;
3	*д-дж*	*городити – огороджувати – огороджений*; *городити – загородити – загороджений*;
4	*зд-ждж*	*їздити – їжджу – їжджений*;
5	*з-ж*	*возити – вожу, вивожу – вивезений, лазити – лажу*;
6	*м-мл*	*громити – громлю – погромити – погромлений*;
7	*п-пл*	*терпіти – терплю – терплячий*;
8	*ст-щ*	*розмістити – розміщу, розмістити – розміщувати – розміщений, мастити – мащу, намащую – намащений, мостити – мощу, замощую – замащений*;
9	*с-ш*	*висіти – вишу, вивішувати – вивішений*;
10	*т-д*	*вести – водити, виводити –виведений*;
11	*т-ч*	*летіти – лечу, платити – плачу, сплачувати – сплачений, крутити – кручу – кручений, накручую – накручений*; *платити – сплатити – сплачений*;
12	*ф-фл*	*графити – графлю – графлений, розграфлювати*;
C	*Правила чергування у коренях дієслів для голосних о та а*	
1	з *а* – повторювана, багаторазова дія, недоконаний вид	*скакати – скакаючий*; *ламати – ламаючий*; *краяти – краючий*; *катати – катаючий*; *хапати – хапаючий*; *ганяти – ганяючий*; *кланятися*; *допомагати*; виняток – *вимовляти*; *прощати*; *заспокоювати*; *установлювати*.
2	з *о* – тривала, нерозчленована дія або одноразова, закінчена, доконаний вид	*гонити – гонений*; *схопити – схоплений*; *котити – кочений*; *клонити – клонений*; *кроїти*; *ломити*; *допомогти*; *скочити*; виняток – *вимовити*; *простити*; *заспокоїти*; *установити*;
D	*Правила чергування у коренях дієслів для голосних е (невипадний) та і*	
1	з *і* – у префіксальних дієсловах недоконаного виду	*викорінювати*; *зберігати*; *нарікати*; *випікати*; *замітати*; *вигрібати*; *причіпляти й зачіпати*;
2	з *е* – у префіксальних дієсловах доконаного виду	*викоренити*; *зберегти*; *наректи*; *випекти*; *замести*; *вигребти*; *причепити*
3	у дієсловах із суфіксом *-ува-* (*-ювa-*) з наголосом на кореневий і та в похідних від цих дієслів іменниках на *-ння*	*полоскати – виполіскувати – виполіскування, чекати – очікувати – очікування, завертіти – завірчувати – завірчування, брехати – набріхувати – набріхування*, але: *потребувати – потребування, вивершувати – вивершування, прищеплювати – прищеплювання*.

E	*Правила чергування у коренях дієслів для голосних е (випадний) та и перед л, р*		
1	з *и* – у дієслівних коренях		*стирати – стертий – стираючий, завмирати – завмираючий, вибирати – вибраний – вибираючий, умирати – умираючий*.
2	з *е* – у дієслівних коренях		*завмер – замру – завмираючий, беру – брати – вибраний – вибираючий, вистелю – вислати –висланий – вистеляючий, стер – стертий – стираючий, умерти – умру –умираючий*;
F	*Правила застосування дієприкметникових суфіксів*		
1	*-ян*(*ий*)		*порівняний*
2	*-ен*(*ий*)		*завішений, незлічений, нескінчений, неоцінений, куплений*
3	*-ан*(*ий*)		*сказаний, завішаний, вихований*
G	*Правила застосування дієслівних суфіксів*		
1	на перший суфіксальний голосний	*-овува-*	*завойовувати – завойовування – завойований*; *перемальовувати – перемальовування –перемальований*.
		-ова-	*підпорядкований,* але *підпорядкувати, підпорядкування*; *мальований,* але *малювати, малювання*; *друкований,* але *друкувати, друкування*; *риштований, риштування,* але *риштувати, риштування*;
2	на корені у похідних словах і формах (віддієслівних іменниках та дієприкметниках)	*-ювa-*	*підбілювати – підбілювання – підбілюваний*;
		-ува-	*маршірувати – маршірування, бомбувати – бомбування, маркувати – маркування, вивершувати – вивершування – вивершуваний, очікувати – очікування – очікуваний*;

Таблиця А.9

Аналіз граматичних/морфологічних ознак української/англійської мов [534-535, 716, 862]

Лінгвістична одиниця	Трактування означення	Мова	
		Українська	Англійська
Іменник	ім'я (*Robert*), особа або річ (*a teacher* – вчитель, *a table* – стіл), дія (*a conversation* – розмова).	Мають граматичний рід.	Не мають граматичного роду
Іменникове означення	іменник, який є означенням іншого *a stone* bridge (тобто камінний міст).	Немає	Існує
Займенник (Pronouns)	слово, що вживається замість іменника (A boy reads books – *He* reads books). Займенники за своїм лексичним значенням і морфологічними ознаками поділяються на кілька розрядів:	9: особові, зворотний, питальні, відносні, присвійні, вказівні, означальні, неозначені, заперечні	8: особисті (Personal), присвійні (Possessive), зворотні (Reflexive), питальні (Interrogative), вказівні (Demonstrative), відносні (Relative), невизначені (Indefinite), вВзаємні (Reciprocal)
Зворотні займенники	Від присвійних займенників **my, our, your** тощо шляхом додавання закінчень	*себе*	**The Reflexive Pronouns** - *myself, yourself, himself, herself, itself, oneself, ourselves, yourselves, themselves*.
Дієслово	окреме слово чи фраза, що описує стан або дію (He *loves* his children. Children *play* in the yard). Дієслова в англійській мові, як і в українській, означають дію (*to go, to build*), стан (*to sleep, to rest*), почуття (*to hear, to like*), процеси мислення (*to think, to realize*). В англійській мові є кілька складних дієслів, які мають дві основи: *to whitewash, to browbeat, to machine-gun*. Багато англійських дієслів збігаються за формою з Іменниками (рідше — з прикметниками):	В українській мові дієслово має 5 типових форм. Ці форми можна розпізнати за характерними закінченнями: 1) неозначена форма (інфінітив); 2) особова форми: (він) *пиш*-**е**, *писа*-**в-Ø**, *напиш*-**е**, *буде* + *писа*-**ти**, *писати*-**ме**, *писа*-**в-Ø** + *би, хай* + *пиш*-**е**; 3) дієприкметник: *пожовті*-**л-ий**, *посиві*-**л-ий**; *писа*-**н-ий**, *підписа*-**н-ий**; *залюбл*-**ен-ий**, *бач*-**ен-ий**; *вими*-**т-ий**, *коло*-**т-ий**; активні дієприкметники (пишучий) українській мові не притаманні, цю функцію виконують описові конструкції — що (або який) пише. 4) безособова на -но/-то: *написа*-**но**, *зробле*-**но**,*прожи*-**то**, *випи*-**то**; 5) дієприслівник: *пиш*-**учи**, *любл*-**ячи**, *підписа*-**вши**, *полюби*-**вши**.	Модальні, почуттєві, фразові, неправильні. Дієслова бувають ***прості***, ***похідні***, ***складні*** і ***складені***. ***Прості*** дієслова складаються з однієї непохідної основи: *to run, to speak, to go, to try* та ін. ***Похідні*** дієслова мають суфікси або префікси: *to organize, rewrite, to discover, to mispronounce*. ***Складені*** дієслова складаються з двох частин — дієслівної основи і відокремленого суфікса, які пишуть ся окремо і можуть роз'єднуватися іншими словами: *to stand up, to sit down, to go away, to put on*та ін. *Sit down, please! Put your cap on!* Складені дієслова дуже поширені в англійській мові. Всі закінчення у таких дієсловах приєднуються до основи. *He always wakes up at 7 o'clock. I'm writing down your address.*
Інфінітив	початкова форма дієслова	без частки: *писа*-**ти**, *говори*-**ти**, *літа*-**ти**, *гримі*-**ти**, *мерзну*-**ти**, *дивува*-**ти**;	як правило, з часткою to – *to write* (He likes to write letters).

Лінгвістична одиниця	Трактування означення	Мова	
		Українська	Англійська
Неправильні форми	ті, які змінюються за звичайними правилами Див. правильні форми.	Немає	(дієслова – be *was/wher been* write *wrote written*, ступенів порівняння прикметників/прислівників – good better best).
Дієприкметник	це форма дієслова, яка означає ознаку предмета за дією або станом і відповідає на питання який? яка? яке? які? (хмарою повіті, врятована планета, зачарований красою).	Активні дієприкметники виражають ознаку предмета за його ж дією (палаюче небо). Пасивні дієприкметники виражають ознаку предмета за дією, яка зумовлена дією іншого предмета над ним (посіяне жито (хтось посіяв)).	Неособова форма англійського дієслова, що має властивості дієслова, прислівника та прикметника. В українській мові англійський дієприкметник відповідає дієприслівнику та дієприкметнику.
Дієприкметник-1 теперішнього часу (Present Participle або Participle I).	Він має дві форми: Present Participle Simple, що відповідає українському дієприкметнику теперішнього часу. Present Participle Perfect, що відповідає укр. дієприкметнику теперішнього часу та дієприслівнику недоконаного виду.	Активні дієприкметники теперішнього часу утворюються від основи теперішнього часу перехідних і неперехідних дієслів недоконаного виду за допомогою суфіксів- ***уч****(ий),* ***-юч****(ий)* для дієслів 1-ої дієвідміни і - ***ач****(ий),* ***-яч****(ий)* для дієслів 2-ої дієвідміни (*реве→ревучий, працює→працюючий*).	-ing форма дієслова (reading the book he makes notes – читаючи книгу, він робить позначки). Present Participle Simple в активному стані утворюється за допомогою додавання закінчення **–ing** до 1 форми дієслова – так само, як і герундій. На укр. він перекладається дієприкметником в активному стані
Дієприкметник-2 минулого часу (Past Participle або Participle II).	Він відповідає дієприкметнику минулого часу в укр. мові. Пасивні дієприкметники в укр.. мові творяться від основи інфінітива перехідних дієслів доконаного і недоконаного виду за допомогою суфіксів ***-т****(ий),* ***-н****(ий),* ***-ен****(ий),* ***-єн****(ий): мити→митий, засіяти→ засіяний, везти→везений, засвоїти→засвоєний.*	Активні дієприкметники минулого часу утворюються від основи інфінітива лише неперехідних дієслів доконаного виду за допомогою суфікса - ***л****(ий)*:*замерзнути→ замерзлий, побіліти→побілілий.*	третя форма дієслова break-broke-*broken* (a *broken* cup – розбита чашка). Дієприкметник минулого часу має лише пасивну форму і перекладається як дієприкметник минулого стану на українську мову.
Прислівник	не змінюване за числом, відмінком слово, що вказує як, коли, куди, де тощо відбувається дія	Він говорить швидко	He speaks *slowly*
Кількісні прислівники	незмінна самостійна частина мови, що виражає ознаку дії, стан предмета або ознаку якості	відповідає на питання *як*? *де*? *звідки*? *наскільки*?*якою мірою*?	слово, що вказує на кількість чогось: *many, much, some, any*.
Прикметник	описує особу, річ, подію тощо (див. ступеневі порівняння прикметників)	змінюване за числом, родом та відмінком слово	не змінюване за числом, родом та відмінком слово (a *tall* boy, a *happy* end, a *long* holiday).
Присвійні займенники	Означає належність до когось або чогось	Мій, твій, його, її, наш, їхній.	*mine, yours, his, hers, ours, theirs*
Прийменники	вживаються перед іменником для позначення місця, часу, напряму	Існують, але не мають буквального перекладу з англійської	такі слова, як *at, in, on, to, under, near* (*in* the street, *on* Wednesday, *at* home).
Однина	одна річ або особа (*a girl, a man, a child, a room*).	Існує та вживається без артикля	Існує та вживається з артиклем
Особа	граматична особа займенника	(1-ша особа – *я, мене*; 2-а особа – ти; 3-я особа – *він, вона, воно, вони*).	(1-ша особа – *I, me*; 2-а особа – you; 3-я особа – *he, she, it, one, they*).
Сполучник	такі слова як *and* (і, а), *but* (але), *when* (коли), *because* (тому що, бо), що з'єднують речення.	Йому подобається тяжкий рок, але мені подобається класична музика	He likes hard rock *but* I like classical music.
Час	форма дієслова, що вказує на час	3 форми дієслова в залежності від часу	12 форм дієслова в залежності від часу (теперішній – *present*, минулий – *past*, майбутній – *future*).

Таблиця А.10

Аналіз синтаксичних/семанитичних ознак української/англійської мов [534-535, 716, 862]

Лінгвістична одиниця	Трактування означення	Мова	
		Українська	Англійська
Речення	наказове, окличне, питальне, розповідне, заперечне, стверджувальне.	Довільний порядок слів	Строгий порядок слів

Лінгвістична одиниця	Трактування означення	Мова	
		Українська	Англійська
Підмет	іменник, займенник або інша частина мови, що передує головному дієслову (присудку).	Машина має двоє дверцят. Ми щасливі.	*A car* has two doors. *We* are happy.
Додаток	частина мови (іменник, займенник, дієслово тощо), яка йде за головним дієсловом речення (присудком) і відповідає на питання що? кого?	Додаток виражається тими ж частинами мови, що й підмет.	Додаток може бути прямим, непрямим, прийменниковим. Додаток може бути виражений іменником, займенником, інфінітивом, герундієм, цілим підрядним додатковим реченням. (I can see *a bus*. They ask me to help.)
Розповідне речення	в якому щось стверджується або заперечується (He speaks English. He doesn't speak English.).	Довільний порядок слів	Строгий порядок слів
yes/no питання	Загальне питання, яке потребує відповіді *yes* (так)/*no* (ні).	Довільний порядок слів	Строгий порядок слів
Заперечне речення	речення з часткою not (не) (He *doesn't* speak English. We *don't* like classical music).	Довільний порядок слів	Строгий порядок слів
Стверджувальне речення	не заперечне і не питальне (He speaks English. We like classical music.).	Довільний порядок слів	Строгий порядок слів
Окличне речення	виражає здивування, гнів тощо (What a nice day! – Який чудовий день!)	Довільний порядок слів	Строгий порядок слів
Питання	альтернативне, загальне *yes/no* питання до підмета, роз'єднувальне, спеціальне *wh*-питання.	Довільний порядок слів	Строгий порядок слів
Коротка відповідь	відповідь, що містить підмет+дієслово-присудок (Who came? – *Mike did*).	Довільний порядок слів	Строгий порядок слів
Необчислювані іменники	іменники, що не вживаються у множині: *air* – повітря, *snow* - сніг, *milk* – молоко.	Існують як виключення. Мають або форму однини, або форму множини.	Існують як виключення. Мають або форму однини, або форму множини.
Активний стан	дію виконує підмет. Див. *пасивний стан*.	Довільний порядок слів (хлопчик розбив чашку або чашку розбив хлопчик).	Строгий порядок слів(A boy *broke* a cup).
Пасивний стан	дія спрямована на підмет	Немає	*be*+дієприкметник – A cup *was broken* (чашку розбито).
Альтернативне питання	питання, що пропонує вибір	Довільний порядок слів (Він говорить англійською чи іспанською? або Говорить він англійською чи іспанською?).	Строгий порядок слів (Does he speak English *or* Spanish?).
Питання до підмета	питання, що запитує про підмет (утворюється без допоміжного дієслова)	Довільний порядок слів	Строгий порядок слів (Who came late? – Jack did.)
Вищий ступінь порівняння	форма прикметника або прислівника, яка вживається для порівняння двох осіб, речей, понять тощо	Два ступені порівняння: вищий (смачніший) і найвищий (найсмачніший)	Три ступені порівняння: звичайний (the Positive Degree), вищий (the Comparative Degree) і найвищий (the Superlative degree): (*smaller than* – менший ніж, *more expensive* – дорожчий ніж).
Відносне підрядне речення	речення, що вводиться відносним займенником	Довільний порядок слів (Це книга, яку я купив вчора або Це книга, яку вчора я купив)	Строгий порядок слів (This is the book *which I bought yesterday*).
Відносні займенники	використовуються для зв'язку головного і підрядного речень.	що, хто, скільки, який, чий, котрий.	*who* (хто, той, що який, котрий); *whom* (кого кому); *that* (який); *which* (котрий, який, хто, що); *whose* (чий, чия, чиє, чиї).
Дійсний спосіб	Дійсний спосіб (індикатив) означає реальну дію; є найбільш уживаним. Дієслова в дійсному способі змінюються за часами.	Значення минулого і теперішнього часу є реальним, а майбутнього — гіпотетичним, тому він може мати відтінок значення недійсного способу.	показує, що дія розглядається як реальний факт у теперішньому, минулому або майбутньому часі.
Множина	більше одного (*girls, men, children, rooms*). Див. однина.	Додавання різних флексій взалежності від роду іменника.	Додавання закінчення *s* крім винятків (там чергування літер в слові або закінчення en)
Модальні дієслова	це функціонально-семантична категорія, яка виражає	Граматично модальність виражається поєднанням	*can, could, may, might, will, would, shall, should, must, ought to, need, needn't, used to*.

Лінгвістична одиниця	Трактування означення	Мова	
		Українська	Англійська
	відношення змісту висловлювання до дійсності і мовця до змісту висловлювання. Модальність є істотною ознакою речення.	дієслова (чи іншого предиката) з модальними частками, прислівниками, дієсловами, а також словосполученнями та реченнями. Порівняйте також різну модальність у прикладі: *піди-но, піди, ти б пішов, нехай би ти пішов, бодай би ти пішов*.	
Почуттєві дієслова	такі дієслова, як *feel* (почувати), *hear* (чути), *look* (дивитися), *smell* (відчувати запах, нюхати), *sound* (звучати).	Не існують синтаксичні особливості використання	Існують синтаксичні особливості використання
Найвищий ступінь	форма прикметника або прислівника, що виражає найвищу міру	Проста форма найвищого ступеня утворюється від форми вищого ступеня за допомогою префікса най- . Для підсилення можуть вживатися префікси як-і що-. Складена форма ступенів порівняння прислівників утворюється додаванням до звичайного прислівника: для вищого ступеня слів більш, менш; для найвищого ступеня слів найбільш, найменш.	*the largest*, *the most important*
Наказовий спосіб	дієслово у формі наказу, вказівки тощо	Довільний порядок слів	Строгий порядок слів(Open the book! Don't smoke!).
Умовний спосіб	показує, що мовець розглядає свою дію як реальний факт, як щось допустиме, бажане.	Не існує	Існує
Правильні форми	ті, які змінюються за звичайними правилами. Див. неправильні форми.	Всі дієслова є правильної форми	(дієслова – *play – played – played*, ступенів порівняння прикметників/прислівників – *small – smaller – the smallest*)
Присвійний відмінок	утворений за допомогою флексій	Або зміні слова (мама-мамин, тато - татовий)	апострофа й літери *s* ('*s*): 's додається іменникам (загальним чи власним), що указати на належність John's father (батько Джона).
Прислівники ступеня	такі слова, як *enough* (досить), *fairly* (досить, цілком), *hardly* (ледве, насилу), *quite* (цілком, зовсім, абсолютно, повністю), *rather* (швидше, краще, переважно).	Існують	Існують (The film was *quite* good.)
Простий час	(минулий, теперішній, майбутній) виражає нетривалу дію	Не існує	Існує (The *dance* well. He *lived* in Kyiv).
Роз'єднувальне питання	коротка питальна частина, що йде за розповідною	Немає	Існує. (He speaks English, *doesn't he*? He doesn't speak English, *does he*?).
Складний	(іменник, займенник тощо)з двох або більше частин	Хто-небудь, вухогорлоніс.	*schoolboy* (школяр), *somebody* (хтось).
Спеціальне wh-питання	питання, яке починається з питальних слів *who*(*m*), *what, when, which, why, where, whose*.	Довільний порядок слів	Строгий порядок слів
Тривалий час	форма дієслова, що утворюється з допоміжного дієслова *be* і смислового дієслова із закінченням *ing be + V-ing*	Немає	вказує, що дія відбувається, відбулась або відбуватиметься у розвитку. (She *is writing* (вона пише, тобто зараз); She *was writing* (вона писала); She *will be writing* (вона писатиме).
Форма -ing	дієслово, прикметник або іменник, що закінчується на -*ing*.	Немає	He is *reading* a *boring* book; I like *reading*.
Фразове дієслово	дієслово, що вживається з прийменниками	Немає	*Look at* the picture. *Come in*.
Часові позначення	прислівники, що вказують, коли відбувається дія	Аналогічно англійській	*last year*, *today*, *in* 1994, *on Sunday*

Таблиця А.11

Основні RE типу SFX для МА українських іменників на основі https://goroh.pp.ua/ [269-276]

№	Клас	F1	F2	RE	Відм.	Число	Група	Ознака	Приклад 1	Приклад 2	В-нок	№
1	I/a	а	и	[^жчшщ]а	1	одн.	тверда	-а	хата	хати	Р.	1
2			і	[^гґкх]а						хаті	Д.М.	2
3			у	а						хату	З.	3
4			ою	[^жчшщ]а						хатою	О.	4
5		а	ею	[жчшщ]а			мішана	-а	душа	душею	О.	5
6		га	зі	га			тверда	-а перед (г,ґ,к,х)	допомога	допомозі	Д.М.	6
7		га	зі	га					дзига	дзизі		7
8		ка	ці	[^к]ка					ріка	ріці		8
9		кка	цці	кка					мекка	мецці		9
10		ха	сі	ха					свекруха	свекрусі		10
11		я	і	[^іеиаоу'ь]я			м'яка	-я	вишня	вишні	Р.Д.М.	11
12			ю	я						вишню	З.	12
13			ею	[^іеиаоу'ь]я						вишнею	О.	13
14			ї	['ьіеиаоу]я			м'яка	-'я	сім'я	сім'ї	Р.Д.М.	14
15			єю	['ьіеиаоу]я						сім'єю	О.	15
16		ір	ору	[^л]ір	2		-	-ір з черг. -і/-о викл.: звір	вибір	вибору	Р.	16
17			орові							виборові	Д.	17
18			ором							вибором	О.	18
19			орі							виборі	М.	19
20			ьору	лір					колір	кольору	Р.	20
21			ьорові							кольорові	Д.	21
22			ьором							кольором	О.	22
23			ьорі							кольорі	М.	23
24		ін	ону	ін				-ін з черг. -і/-о	загін	загону	Д.Р.	24
25			онові							загонові	Д.	25
26			оном							загоном	О.	26
27			оні							загоні	М.	27
28		іг	огу	іг				-іг з черг. -і/-о викл.: оберіг	батіг	батогу	Д.Р.	28
29			огові							батогові	Д.М.	29
30			огом							батогом	О.	30
31			озі							батозі	М.	31
32		ід	оду	[^л]ід				-ід з черг. -і/-о	провід	проводу	Д.Р.	32
33			одові							проводові	Д.	33
34			одом							проводом	О.	34
35			оді							проводі	М.	35
36			ьоду	[^пг]лід					лід	льоду	Д.Р.	36
37			ьодові							льодові	Д.	37
38			ьодом							льодом	О.	38
39			ьоді							льоді	М.	39
40			оду	[пг]лід					плід	плоду	Д.Р.	40
41			одові							плодові	Д.	41
42			одом							плодом	О.	42
43			оді							плоді	М.	43
44		іб	обу	іб				-іб з черг. -і/-о	засіб	засобу	Д.Р.	44
45			обові							засобові	Д.	45
46			обом							засобом	О.	46
47			обі							засобі	М.	47
48		іп	опу	іп				-іп з черг. -і/-о викл.: чіп	піп	попу	Д.Р.	48
49			опові							попові	Д.	49
50			опом							попом	О.	50
51			опі							попі	М.	51
52		івш	овшу	івш				-івш з черг. -і/-о	ківш	ковшу	Д.Р.	52
53			овшеві							ковшеві	Д.	53
54			овшем							ковшем	О.	54
55			овші							ковші	М.	55
56		ізд	озду	ізд				-ізд з черг. -і/-о	дрізд	дрозду	Д.Р.	56
57			оздові							дроздові	Д.	57
58			оздом							дроздом	О.	58
59			озді							дрозді	М.	59
60		іл	олу	іл				-іл з черг. -і/-о	дозвіл	дозволу	Д.Р.	60
61			олові							дозволові	Д.	61
62			олом							дозволом	О.	62
63			олі							дозволі	М.	63
64		ів	ову	ів				-ів з черг. -і/-о	острів	острову	Д.Р.	64
65			овом							островом	О.	65
66			ові							острові	М.	66
67		їв	єву	їв				-їв з черг. -і/-о	Київ	Києву	Д.Р.	67
68			євом							Києвом	О.	68
69			єві							Києві	М.	69
70		ік	оку	ік				-ік з черг. -і/-о	рік	року	Д.Р.	70
71			окові							рокові	Д.	71
72			оком							роком	О.	72
73			оці							році	М.	73

№	Клас	F1	F2	RE	Відм.	Число	Група	Ознака	Приклад 1	Приклад 2	В-нок	№
74		іск	оску	іск				-іск з черг. -і/-о	віск	воску	Д.Р.	74
75			оскові							воскові	Д.	75
76			оском							воском	О.	76
77			осці							восці	М.	77
78		іст	осту	іст				-іст з черг. -і/-о	ріст	росту	Д.Р.	78
79			остові							ростові	Д.	79
80			остом							ростом	О.	80
81			ості							рості	М.	81
82		іс	осу	[кнч]іс				-іс з черг. -і/-о	ніс	носу	Д.Р.	82
83			осові							носові	Д.М.	83
84			осом							носом	О.	84
85			осі							носі	М.	85
86		іт	оту	[^л]іт				-іт з черг. -і/-о	гуркіт	гуркоту	Д.Р.	86
87			отові							гуркотові	Д.	87
88			отом							гуркотом	О.	88
89			оті							гуркоті	М.	89
90			ьоту	[^п]літ					політ	польоту	Д.Р.	90
91			ьотові							польотові	Д.	91
92			ьотом							польотом	О.	92
93			ьоті							польоті	М.	93
94			оту	[п]літ					пліт	плоту	Д.Р.	94
95			отові							плотові	Д.	95
96			отом							плотом	О.	96
97			оті							плоті	М.	97
98		із	озу	із				-із з черг. -і/-о	віз	возу	Д.Р.	98
99			озові							возові	Д.	99
100			озом							возом	О.	100
101			озі							возі	М.	101
102		іж	ожу	[^тбд]іж				-іж з черг. -і/-о	ніж	ножу	Д.Р.	102
103			ожеві							ножеві	Д.	103
104			ожем							ножем	О.	104
105			ожі							ножі	М.	105
106			ожу	е[тб]іж					небіж	небожу	Д.Р.	106
107			ожеві							небожеві	Д.	107
108			ожем							небожем	О.	108
109			ожі							небожі	М.	109
110			ежу	[^е][тбд]іж					рубіж	рубежу	Д.Р.	110
111			ежеві							рубежеві	Д.	111
112			ежем							рубежем	О.	112
113			ежі							рубежі	М.	113
114		ен	ну	ен				-ен з випад. 'е'	човен	човну	Д.Р.	114
115			нові							човнові	Д.	115
116			ном							човном	О.	116
117			ні							човні	М.	117
118		інь	оню	[к]інь				-кінь	кінь	коню	Д.	118
119			оневі							коневі	Д.М.	119
120			онем							конем	О.	120
121			оні							коні	М.	121
122			еню	[^ксв]інь				-інь з черг. -і -е	корінь	кореню	Д.	122
123			еневі							кореневі	Д.	123
124			енем							коренем	О.	124
125			ені							корені	М.	125
126			еню	[^о][св]інь					ревінь	ревеню	Р.Д.М.	126
127			еневі							ревеневі	Д.	127
128			енем							ревенем	О.	128
129			ені							ревені	М.	129
130		ень	ню	ень			м'яка	-ень, -онь	день	дню	Д.	130
131			неві							дневі	З.	131
132			ні							дні	М.	132
133			нем							днем	О.	133
134		онь	ню	онь					вогонь	вогню	Д.	134
135			неві							вогневі	З.	135
136			ні							вогні	М.	136
137			нем							вогнем	О.	137
138		оль	лю	оль				-оль	кухоль	кухлю	Д.	138
139			леві							кухлеві	З.	139
140			лі							кухлі	М.	140
141			лем							кухлем	О.	141
142		оть	тю	оть				-оть	ніготь	нігтю	Д.	142
143			теві							нігтеві	З.	143
144			ті							нігті	М.	144
145			тем							нігтем	О.	145
146		ій	ою	ій			-	-ій з черг. -і/-о	рій	рою	Д.Р.	146
147			оєві							роєві	Д.	147
148			оєм							роєм	О.	148
149			ої							рої	М.	149
150		ідь	едю	ідь				-ідь з черг. -і/-е	ведмідь	ведмедю	Д.	150
151			едеві							ведмедеві	Д.	151

№	Клас	F1	F2	RE	Відм.	Число	Група	Ознака	Приклад 1	Приклад 2	В-нок	№
152			едем							ведмедем	О.	152
153			еді							ведмеді	М.	153
154		іль	елю	іль				-іль з черг. -і/-е	важіль	важелю	Д.	154
155			елеві							важелеві	Д.	155
156			елем							важелем	О.	156
157			елі							важелі	М.	157
158		ість	остю	ість				-ість з черг. -і/-о	гість	гостю	Д.	158
159			остеві							гостеві	Д.	159
160			остем							гостем	О.	160
161			ості							гості	М.	161
162		ок	ку	ок				-ок з випад. -о	будиночок	будиночку	З.	162
163			кові							будиночкові	Д.	163
164			ком							будиночком	О.	164
165		ол	лу	ол				-ол з випад. -о	вузол	вузлу	З.	165
166			лові							вузлові	Д.	166
167			лі							вузлі	М.	167
168			лом							вузлом	О.	168
169		ор	ру	ор				-ор з випад. -о	свекор	свекру	Д.	169
170			ром							свекром	О.	170
171			рі							свекрі	М.	171
172			рові							свекрові	М.	172
173		ер	ру	ер				-ер з випад. -е	вітер	вітру	Д.	173
174			ром							вітром	О.	174
175			рі							вітрі	М.	175
176			рові							вітрові	М.	176
177		ел	лу	ел				-ел з випад. -е	осел	ослу	Д.	177
178			лом							ослом	О.	178
179			лі							ослі	М.	179
180			лові							ослові	М.	180
181		ет	ту	ет				-ет з випад. -е	оцет	оцту	Д.	181
182			том							оцтом	О.	182
183			ті							оцті	М.	183
184			тові							оцтові	М.	184
185		ес	су	ес				-ес з випад. -е	пес	псу	Д.	185
186			сом							псом	О.	186
187			сі							псі	М.	187
188			сові							псові	М.	188
189		ель	лю	ель				-ель з випад. -е	журавель	журавлю	Д.	189
190			лем							журавлем	О.	190
191			лі							журавлі	М.	191
192			леві							журавлеві	М.Д.	192
193		єць	йцю	[^о]єць			м'яка	-[^о]єць з випад. є	англієць	англійцю	Д.	193
194			йцеві							англійцеві	М.Д.	194
195			йці							англійці	М.	195
196			йцем							англійцем	О.	196
197		оєць	ійцю	оєць				оєць з випад. є	боєць	бійцю	Д.	197
198			ійцеві							бійцеві	З.	198
199			ійці							бійці	М.	199
200			ійцем							бійцем	О.	200
201		ець	цю	[^лрнв]ець				-ець з випад. е	німець	німцю	Д.	201
202			цеві							німцеві	З.	202
203			цем							німцем	О.	203
204			ці							німці	М.	204
205		лець	льцю	лець				-лець з випад. е	бразилець	бразильцю	Д.	205
206			льцеві							бразильцеві	З.	206
207			льцем							бразильцем	О.	207
208			льці							бразильці	М.	208
209		рець	цю	[аєєиіїоуюя]рець				-рець з випад. е	борець	борцю	Д.	209
210			цеві							борцеві	Д.М	210
211			ці							борці	М.	211
212			цем							борцем	О.	212
213			ерцю	[^аєєиіїоуюя]рець					жрець	жерцю	Д.	213
214			ерцеві							жерцеві	З.	214
215			ерці							жерці	М.	215
216			ерцем							жерцем	О.	216
217		ець	цю	[аєєиіїуюяго][нв]ець				-ець з випад. е	українець	українцю	Д.	217
218			цеві							українцеві	З.	218
219			цем							українцем	О.	219
220			ці							українці	М.	220
221		нець	енцю	[^аєєиіїо уюяг]нець				-нець з випад. е	жнець	женцю	Д.	221
222			енцеві							женцеві	З.	222
223			енцем							женцем	О.	223
224			енці							женці	М.	224
225			онцю	[о]нець					гонець	гонцю	Д.	225
226			онцеві							гонцеві	З.	226
227			онцем							гонцем	О.	227
228			інці							гонці	М.	228
229		вець	евцю	[^аєєиіїоуюя]				-вець з	швець	шевцю	Д.	229

№	Клас	F1	F2	RE	Відм.	Число	Група	Ознака	Приклад 1	Приклад 2	В-нок	№
230			евцеві	вець				випад. -е		шевцеві	З.	230
231			евцем							шевцем	О.	231
232			евці							шевці	М.	232
233			івцю	[о]вець					вдовець	вдівцю	Д.	233
234			івцеві							вдівцеві	З.	234
235			івцем							вдівцем	О.	235
236			івці							вдівці	М.	236
237		ь	и	ять	-	-	-	числівники -ять, -сят, -сто	двадцять	двадцяти	Р.	237
238			ьом							двадцятьом	Д.	238
239			ьма							двадцятьма	О.	239
240			ьох	сят						двадцятьох	М.	240
241		-	и						п'ятдесят	п'ятдесяти	Р.	241
242			ьом							п'ятдесятьом	Д.	242
243			ьма							п'ятдесятьма	О.	243
244			ьох	сто						п'ятдесятьох	М.	244
245		о	а						сто	ста	Р.	245
246			ам							стам	Д.	246
247			ами							стами	О.	247
248			ах							стах	М.	248
1	I/b	а	-	[^клн]а	1	мн.	тверда	-а, викл. сестра	хата	хат	Р.	249
2				[^вклнршч]а					рама	рам		250
3				[^ст]ла					щогла	щогл		251
4			ей	[шч]а					миша	мишей		252
5		ла	ел	[ст]ла					мітла	мітел		253
6		ва	-	[^к]ва					глава	глав		254
7		ква		[^р]ква					буква	букв		255
8		рква	рков	рква					церква	церков		256
9		на	-	[^сжзвм]на					частина	частин		257
10			ен	[сжвм]на					сосна	сосен		258
11		изна	н	изна					тризна	тризн		259
12		на	ен	озна					борозна	борозен		260
13		ка	-	[аєєоуиіїяю]ка					техніка	технік		261
14			ок	[^аєєоуиіїяю]ка					відпустка	відпусток		262
15		шка	шок	шка					дошка	дошок		263
16		я	ь	[^іеуиаон'ьлтдр]я			м'яка	на -я	Вася	Вась		264
17				еря					вечеря	вечерь		265
18		ря	-	[^ео]ря					буря	бур		266
19		оря	ір	оря					зоря	зір		267
20		я	ь	[іеиаоюуїяє]ня					богиня	богинь		268
21		ня	онь	[кх]ня					кухня	кухонь		269
22		йня	єнь	йня					бойня	боєнь		270
23		ьня	ень	ьня					вітальня	віталень		271
24		ня	ень	[^іеиаоюуїяєнкхйь]ня					вишня	вишень		272
25			сль	сля					тесля	тесль		273
26		ля	ель	[^лоеуаієюяїиск]ля					будівля	будівель		274
27		я	ь	[лоеуаієюяїиск]ля					Валя	Валь		275
28			ей	адя					попадя	попадей		276
29			ів	[^да]дя					дядя	дядів		277
30			ь	[^тс]тя					Катя	Кать		278
31		ля	ей	лля					рілля	рілей		279
32		я	ів	уддя					суддя	суддів		280
33		дя	ей	аддя					баддя	бадей		281
34		тя	ей	ття					стаття	статей		282
35		я	ь	[^о]стя					причастя	причасть		283
36			ей	остя					гостя	гостей		284
37		'я	ей	[^р]'я					сім'я	сімей		285
38		я	й	[іеиаоу]я					мрія	мрій		286
39		-	м	[ая]					хата	хатам	Д.	287
40			ми	[ая]						хатами	О.	288
41			х	[ая]						хатах	М.	289
42		ір	ори	[^л]ір	2			-ір із черг. -і/-о викл.: звір	вибір	вибори	Н.	290
43			орів							виборів	Р.	291
44			орам							виборам	Д.	292
45			орами							виборами	О.	293
46			орах							виборах	М.	294
47			ьори	лір					колір	кольори	Н.	295
48			ьорів							кольорів	Р.	296
49			ьорам							кольорам	Д.	297
50			ьорами							кольорами	О.	298
51			ьорах							кольорах	М.	299
52		ін	они	ін				-ін із черг.	загін	загони	Н.	300
53			онів							загонів	Р.	301

№	Клас	F1	F2	RE	Відм.	Число	Група	Ознака	Приклад 1	Приклад 2	В-нок	№
54			онам					-і/-о		загонам	Д.	302
55			онами							загонами	О.	303
56			онах							загонах	М.	304
57		іп	опи	іп				-іп із черг. -і/-о	піп	попи	Н.	305
58			опів							попів	Р.	306
59			опам							попам	Д.	307
60			опами							попами	О.	308
61			опах							попах	М.	309
62		івш	овші	івш				-івш із черг. -і/-о	ківш	ковші	Н.	310
63			овшів							ковшів	Р.	311
64			овшам							ковшам	Д.	312
65			овшами							ковшами	О.	313
66			овшах							ковшах	М.	314
67		ізд	озди	ізд				-ізд із черг. -і/-о	дрізд	дрозди	Н.	315
68			оздів							дроздів	З.	316
69			оздам							дроздам	Д.	317
70			оздами							дроздами	О.	318
71			оздах							дроздах	М.	319
72		іг	оги	іг				-іг із черг. -і/-о	батіг	батоги	Н.	320
73			огів							батогів	З.	321
74			огам							батогам	Д.	322
75			огами							батогами	О.	323
76			огах							батогах	М.	324
77			еги					-ріг із черг. -і/-е як викл.	оберіг	обереги	Н.	325
78			егів							оберегів	З.	326
79			егам							оберегам	Д.	327
80			егами							оберегами	О.	328
81			егах							оберегах	М.	329
82		ід	оди	[^л]ід				-ід із черг. -і/-о	провід	проводи	Н.	330
83			одів							проводів	З.	331
84			одам							проводам	Д.	332
85			одами							проводами	О.	333
86			одах							проводах	М.	334
87			оди	[пг]лід					плід	плоди	Н.	335
88			одів							плодів	З.	336
89			одам							плодам	Д.	337
90			одами							плодами	О.	338
91			одах							плодах	М.	339
92			ьоди	[^пг]лід					лід	льоди	Н.	340
93			ьодів							льодів	З.	341
94			ьодам							льодам	Д.	342
95			ьодами							льодами	О.	343
96			ьодах							льодах	М.	344
97		іб	оби	іб				-іб із черг. -і/-о	спосіб	способи	Н.	345
98			обів							способів	З.	346
99			обам							способам	Д.	347
100			обами							способами	О.	348
101			обах							способах	М.	349
102		іл	оли	іл				-іл із черг. -і/-о	дозвіл	дозволи	Н.	350
103			олів							дозволів	З.	351
104			олам							дозволам	Д.	352
105			олами							дозволами	О.	353
106			олах							дозволах	М.	354
107		ів	ови	ів				-ів із черг. -і/-о	острів	острови	Н.	355
108			овів							островів	Р.	356
109			овам							островам	Д.	357
110			овами							островами	О.	358
111			овах							островах	М.	359
112		їв	єви	їв				-їв із черг. -і/-о	Київ	Києви	Н.	360
113			євів							Києвів	Р.	361
114			євам							Києвам	Д.	362
115			євами							Києвами	О.	363
116			євах							Києвах	М.	364
117		ік	оки	ік				-ік із черг. -і/-о	рік	роки	Н.	365
118			оків							років	З.	366
119			окам							рокам	Д.	367
120			оками							роками	О.	368
121			оках							роках	М.	369
122		іск	оски	іск				-іск із черг. -і/-о	віск	воски	Н.	370
123			осків							восків	З.	371
124			оскам							воскам	Д.	372
125			осками							восками	О.	373
126			осках							восках	М.	374
127		іст	ости	іст				-іст із черг. -і/-о	наріст	нарости	Н.	375
128			остів							наростів	З.	376
129			остам							наростам	Д.	378
130			остами							наростами	О.	379
131			остах							наростах	М.	380

№	Клас	F1	F2	RE	Відм.	Число	Група	Ознака	Приклад 1	Приклад 2	В-нок	№
132		іс	оси	[кнч]іс				-іс із черг. -і/-о	ніс	носи	Н.	381
133			осів							носів	Р.	382
134			осам							носам	Д.	383
135			осами							носами	О.	384
136			осах							носах	М.	385
137		іт	оти	[^л]іт				-іт із черг. -і/-о	гніт	гноти	Н.	386
138			отів							гнотів	З.	387
139			отам							гнотам	Д.	388
140			отами							гнотами	О.	389
141			отах							гнотах	М.	390
142			оти	[п]літ					пліт	плоти	Н.	391
143			отів							плотів	З.	392
144			отам							плотам	Д.	393
145			отами							плотами	О.	394
146			отах							плотах	М.	395
147			ьоти	[^п]літ					політ	польоти	Н.	396
148			ьотів							польотів	З.	397
149			ьотам							польотам	Д.	398
150			ьотами							польотами	О.	399
151			ьотах							польотах	М.	400
152		із	ози	із				-іс із черг. -і/-о	віз	вози	Н.	401
153			озів							возів	З.	402
154			озам							возам	Д.	403
155			озами							возами	О.	404
156			озах							возах	М.	405
157		іж	ожі	[^тбд]іж				-іж із черг. -і/-о	ніж	ножі	Н.	406
158			ожів							ножів	З.	407
159			ожам							ножам	Д.	408
160			ожами							ножами	О.	409
161			ожах							ножах	М.	410
162			ожі	е[тб]іж					небіж	небожі	Н.	411
163			ожів							небожів	З.	412
164			ожам							небожам	Д.	413
165			ожами							небожами	О.	414
166			ожах							небожах	М.	415
167			ежі	[^е][тбд]іж					рубіж	рубежі	Н.	416
168			ежів							рубежів	З.	417
169			ежам							рубежам	Д.	418
170			ежами							рубежами	О.	419
171			ежах							рубежах	М.	420
172		ен	ни	ен				-ен із вип. -е	човен	човни	Н.	421
173			нів							човнів	Р.	422
174			нам							човнам	Д.	423
175			нами							човнами	О.	424
176			нах							човнах	М.	425
177		ет	ти	ет				-ет із вип. -е	оцет	оцти	мн.	426
178			тів							оцтів	Р.З.	427
179			там							оцтам	Д.	428
180			тами							оцтами	О.	429
181			тах							оцтах	М.	430
182		ес	си	ес				-ес із вип. -е	пес	пси	мн.	431
183			сів							псів	Р.З.	432
184			сам							псам	Д.	433
185			сами							псами	О.	434
186			сах							псах	М.	435
187		інь	оней	[к]інь				-кінь	кінь	коней	Р.	436
188			оням							коням	Д.	437
189			онями							конями	О.	438
190			ми							кіньми	О.	439
191			онях							конях	М.	440
192			енів	[^ксв]інь				-інь із черг. -і/-е	корінь	коренів	Р.	441
193			еням							кореням	Д.	442
194			енями							кореняами	О.	443
195			енях							кореняах	М.	444
196			енів	[^о][св]інь					ревінь	ревенів	Р.	445
197			еням							ревеням	Д.	446
198			енями							ревенями	О.	447
199			енях							ревенях	М.	448
200		ій	ої	ій				-ій	рій	рої	Н.	449
201			оїв							роїв	З.	450
202			оям							роям	Д.	451
203			оями							роями	О.	452
204			оях							роях	М.	453
205		ідь	еді	ідь				-ідь	ведмідь	ведмеді	Н.	454
206			едів							ведмедів	Р.	455
207			едям							ведмедям	Д.	456
208			едями							ведмедями	О.	457
209			едях							ведмедях	М.	458

№	Клас	F1	F2	RE	Відм.	Число	Група	Ознака	Приклад 1	Приклад 2	В-нок	№
210			елі					-іль із черг. -і/-е	важіль	важелі	Н.	459
211			елів							важелів	Р.	460
212			елям							важелям	Д.	461
213			елями							важелями	О.	462
214			елях							важелях	М.	463
215		ість	ості	ість				-ість із черг. -і/-е	гість	гості	Н.	464
216			остів							гостів	Р.	465
217			остям							гостям	Д.	466
218			остями							гостями	О.	467
219			остях							гостях	М.	468
220		ок	ки	ок				-ок із вип. о	будиночок	будиночки	Н.	469
221			ків							будиночків	Р.	470
222			кам							будиночкам	Д.	471
223			ками							будиночками	О.	472
224			ках							будиночках	М.	473
225		ол	ли	ол				-ол із вип. о	вузол	вузли	Н.	474
226			лів							вузлів	Р.	475
227			лам							вузлам	Д.	476
228			лами							вузлами	О.	477
229			лах							вузлах	М.	478
230		ор	ри	ор				-ор із вип. о	свекор	свекри	Н.	479
231			рів							свекрів	Р.	480
232			рам							свекрам	Д.	481
233			рами							свекрами	О.	482
234			рах							свекрах	М.	483
235		ер	ри	ер				-ер із вип. е	вітер	вітри	Н.	484
236			рів							вітрів	Р.	485
237			рам							вітрам	Д.	486
238			рами							вітрами	О.	487
239			рах							вітрах	М.	488
240		ел	ли	ел				-ел із вип. е	осел	осли	Н.	489
241			лів							ослів	Р.	490
242			лам							ослам	Д.	491
243			лами							ослами	О.	492
244			лах							ослах	М.	493
245		ель	лів	ель				-ель із вип. е	журавель	журавлів	Р.З.	494
246			лям							журавлям	Д.	495
247			лями							журавлями	О.	496
248			лях							журавлях	М.	497
249		єць	йці	[^о]єць			м'яка	-[^о]єць	англієць	англійці	Н.	498
250			йців							англійців	Р.	499
251			йцям							англійцям	Д.	500
252			йцями							англійцями	О.	501
253			йцях							англійцях	М.	502
254		оєць	ійці	оєць				-[о]єць	боєць	бійці	Н.	503
255			ійців							бійців	Р.	504
256			ійцям							бійцям	Д.	505
257			ійцями							бійцями	О.	506
258			ійцях							бійцях	М.	507
259		ець	ці	[^лрнв]ець				-[^л]ець	німець	німці	Н.	508
260			ців							німців	Р.	509
261			цям							німцям	Д.	510
262			цями							німцями	О.	511
263			цях							німцях	М.	512
264		лець	льці	лець				-лець -льця	бразилець	бразильці	Н.	513
265			льців							бразильців	Р.	514
266			льцям							бразильцям	Д.	515
267			льцями							бразильцями	О.	516
268			льцях							бразильцях	М.	517
269		рець	ерці	[^аеєиіїоуюя]рець				-рець	жрець	жерці	Мн.	518
270			ерців							жерців	Р.	519
271			ерцям							жерцям	Д.	520
272			ерцями							жерцями	О.	521
273			ерцях							жерцях	М.	522
274			рці	[аеєиіїоуюя]рець					борець	борці	Мн.	523
275			рців							борців	Р.	524
276			рцям							борцям	Д.	525
277			рцями							борцями	О.	526
280			рцях							борцях	М.	527
281		ець	ці	[аеєиіїуюяго][нв]ець				-[нв]ець	українець	українці	Мн.	528
282			ців							українців	Р.	529
283			цям							українцям	Д.	530
284			цями							українцями	О.	531
285			цях							українцях	М.	532
286		нець	енці	[^аеєиіїоуюяг]нець				-нець	жнець	женці	Мн.	533
287			енців							женців	Р.	534
288			енцям							женцям	Д.	535
289			енцями							женцями	О.	536

№	Клас	F1	F2	RE	Відм.	Число	Група	Ознака	Приклад 1	Приклад 2	В-нок	№
290			енцях							женцях	М.	537
291			інці	[о]нець				-[о]нець	гонець	гінці	Мн.	538
292			інців							гінців	Р.	539
293			інцям							гінцям	Д.	540
294			інцями							гінцями	О.	541
295			інцях							гінцях	М.	542
296		вець	евці	[^аеєиіїоуюя]вець				-вець	швець	шевці	Мн.	543
297			евців							шевців	Р.	544
298			евцям							шевцям	Д.	545
299			евцями							шевцями	О.	546
300			евцях							шевцях	М.	547
301			івці	[о]вець				-[о]вець	вдовець	вдівці	Мн.	548
302			івців							вдівців	Р.	549
303			івцям							вдівцям	Д.	550
304			івцями							вдівцями	О.	551
305			івцях							вдівцях	М.	552
306		ень	ні	-[ое]нь				-[ое]нь	день	дні	Мн.	553
307			нів							днів	Р.	554
308			ням							дням	Д.	555
309			нями							днями	О.	556
310			нях							днях	М.	557
311		оль	лі	оль				-оль	кухоль	кухлі	Мн.	558
312			лів							кухлів	Р.	559
313			лям							кухлям	Д.	560
314			лями							кухлями	О.	561
315			лях							кухлях	М.	562
316		оть	ті	оть				-оть	ніготь	нігті	Мн.	563
317			тів							нігтів	Р.	564
318			тям							нігтям	Д.	565
319			тями							нігтями	О.	566
320			тях							нігтях	М.	567
321		и	ів	[^кнд]и				-и	шаровари	шароварів	Р.	568
322			ів	[^ую]ди				-і	народи	народів		569
323			ей	[ую]ди					люди	людей		570
324			ів	[^ияо]ни					човни	човнів		571
325			-	[ия]ни					громадяни	громадян		572
326			-	[лс]они					панталони	панталон		573
327			ів	сони					кальсони	кальсонів		574
328				[гр]они					макарони	макаронів		575
329				[^бвджлнпртчь]ки					вершки	вершків	Р.З.	576
330				стки					підмостки	підмостків		577
331		ки	ок	[^с]тки					колготки	колготок		578
332			ів	[^і]вки					висновки	висновків		579
333			ок	івки					висівки	висівок		580
334				[^у]нки					поминки	поминок		581
335			ів	унки					лаштунки	лаштунків		582
336			ок	[бджлпрчь]ки					лапки	лапок		583
337		и	ям	сани					сани	саням	Д.	584
338				сіни					сіни	сіням		585
339			ам	[^с]ани					кайдани	кайданам		586
340				[^с]іни					Афіни	Афінам		587
341				[^аі]ни					човни	човнам		588
342				[^нд]и					шаровари	шароварам		589
343				[^ую]ди					мандри	мандрам		590
344			ям	[ую]ди					люди	людям		591
345			ьми	сани					сани	саньми	О.	592
346				сіни					сіни	сіньми		593
347			ами	[^с]ани					кайдани	кайданами		594
348				[^с]іни					Афіни	Афінами		595
349				[^аі]ни					човни	човнами		596
350				[^дн]и					шаровари	шароварами		597
351				[^ую]ди					мандри	мандрами		598
352			ьми	[ую]ди					люди	людьми		599
353			ях	сани					сани	санях	М.	600
354				сіни					сіни	сінях		601
355			ах	[^с]ани					кайдани	кайданах		602
356				[^с]іни					Афіни	Афінах		603
357				[^аі]ни					човни	човнах		604
358				[^нд]и					шаровари	шароварах		605
359				[^ую]ди					кеди	кедах		606
360			ях	[ую]ди					люди	людях		607
361		і	ів	[^лзнцрш]і					хвастощі	хвастощів	Р.	608
362			ей	ші					грошІ	грошей		609
363			ам	[^лзнцр]і					хвастощі	хвастощам	Д.	610
364			ами							хвастощами	О.	611
365			ах							хвастощах	М.	612
366		-	в	[лзн]і				-і перед	штанці	штанців	Р.	613

№	Клас	F1	F2	RE	Відм.	Число	Група	Ознака	Приклад 1	Приклад 2	В-нок	№
367		і	ь	иці				л,з,н,р,ц	вечорниці	вечорниць		614
368		-	в	[^ри]ці					рубці	рубців		615
369				[^е]рці					майорці	майорців		617
370				[^в]ерці					мерці	мерців		618
371		ці	ець	верці					дверці	дверець		619
372		і	ей	[аяєєиії]рі					двері	дверей		620
373		-	в	[^аяєєиії]рі					нетрі	нетрів		621
374		і	ям	[лзнцр]і					штанці	штанцям	Д.	622
375			ями	[лзнц]і					мазі	мазям		623
376				[^е]рі					нетрі	нетрями	О.	624
377			ми	ері					двері	дверми		625
378			има	єрі						дверима		626
379			ях	[лзнцр]і					штанці	штанцях	М.	627
380		ї	й	[^ея]ї				-ї	геніталії	геніталій	Р.	628
381			їв	[ея]ї					хазяї	хазяїв		629
382			ям	ї					геніталії	геніталіям	Д.	630
383			ями	ях						геніталіями	О.	631
384										геніталіях	М.	632
1	І/с	ір	ора	[^л]ір	2	одн	мішана	в Р. відм. на -а	вечір	вечора	З.	633
2			ьора	лір					колір	кольора	Р.	634
3		ін	она	ін					загін	загона		635
4		іг	ога	іг					батіг	батога		636
5			ега						оберіг	оберега		637
6		ід	ода	[^л]ід					провід	провода		638
7				[^пг]лід					поріг	порога		639
8				[пг]лід					плід	плода		640
9		іп	опа	іп					піп	попа		641
10		івш	овша	івш					ківш	ковша		642
11		ізд	озда	ізд					дрізд	дрозда		643
12		іб	обу	іб					засіб	засобу		644
13		іл	олу	іл					дозвіл	дозволу		645
14		ів	ову	ів					острів	острова		646
15		їв	єва	їв					Київ	Києва		647
16		ік	оку	ік					сік	соку		648
17		іск	оску	іск					віск	воску		649
18		іст	осту	іст					піст	посту		650
19		іс	оса	[кнч]іс					ніс	носа		651
20		іт	іту	[^л]іт					гніт	гніту		652
21			ьоту	[^п]літ					політ	польоту		653
22			от[ау]	[п]літ					пліт	плот[ау]		654
23		із	оза	із					віз	воза		655
24		іж	ожа	[^тб]іж					ніж	ножа		656
25				е[тб]іж					небіж	небожа		657
26			ежу	[^е][тб]іж					рубіж	рубежу		658
27		ій	оя	ій					рій	роя		659
28		ен	на	ен					рожен	рожна		660
29		ет	ту	ет					оцет	оцту		661
30		ес	са	ес					пес	пса		662
31		інь	оня	[к]інь					кінь	коня		663
32		інь	еня	[^к]інь				-інь з черг. і/е	корінь	кореня		664
33		ідь	едя	ідь					ведмідь	ведмедя		665
34		ість	остя	ість					гість	гостя		666
35		ок	ка	ок				-о[крл] із вип. -о	будиночок	будиночка		667
36		ол	ла	ол					вузол	вузла		668
37		ор	ра	ор					вугор	вугра		669
38		ер	ру	ер				-[еєо] [рлцн]ь* із вип. -[еєо]	вітер	вітру		670
39		ел	ла	ел					орел	орла		671
40		ель	ля	ель					журавель	журавля		672
41		єць	йця	[^о]єць					англієць	англійця		673
42		оєць	ійця	оєць					боєць	бійця		674
43		ець	ця	[^лрвн] ець					німець	німця		675
44		лець	льця	лець					бразилець	бразильця		676
45		рець	ерця	[^аєєиіїоуюя] рець					жрець	жреця		677
46		рець	рця	[аєєиіїоуюя] рець					борець	борця		678
47		ець	ця	[аєєиіїуюяго] [нв]ець					кореєць	корейця		679
48		нець	енця	[^аєєиіїоу юяг]нець					жнець	женці		680
49		вець	евця	[^аєєиіїо уюя]вець					швець	шевці		681
50			івця	вець					вдівець	вдівця		682
51		ень	ня	ень					день	дня		683
52		онь	ню	онь					вогонь	вогню		684
53		оль	ля	оль					кухоль	кухля		685
54		оть	тя	оть					ніготь	нігтя		686
1	І/о	а	и	а	1	мн	тверда	-а з	школа	школи	Н.	687

№	Клас	F1	F2	RE	Відм.	Число	Група	Ознака	Приклад 1	Приклад 2	В-нок	№
2		ола	іл	ола				черг. і/о та появою о(е) в Р.в. є викл -она		шкіл	Р.	688
3		оба	іб	оба					доба	діб		689
4		ода	ід	ода					борода	борід		690
5		ога	іг	ога					нога	ніг		691
6		га		рга					кочерга	кочерг		692
7		оха	іх	оха					блоха	бліх		693
8		ока	ік	ока					щока	щік		694
9		она	ін	она					борона	борін		695
10			он						корона	корон		696
11		опа	іп	опа					копа	кіп		697
12		ора	ір	ора					гора	гір		698
13		ра	ер	тра					сестра	сестер		699
14			ор	[^от]ра					іскра	іскор		700
15		оса	іс	оса					коса	кіс		701
16		ота	іт	ота					сирота	сиріт		702
17		ва	ов	[^о]ва					молитва	молитов		703
18		ова	ів	[о]ва					удова	удів		704
19		оза	із	[^ь]оза					коза	кіз		705
20		ьоза		ьоза					сльоза	сліз		706
21		еза	ів	еза					береза	беріз		707
22		а		жа					магараджа	магараджів		708
23		ійня	оєнь	ійня					бійня	боєнь		709
24		а	м	[ая]					школа	школам	Д.	710
25			ми							школами	О.	711
26			х							школах	М.	712
27		о	а	о	2			-о сер. роду з черг. і/о	слово	слова	Н.З.	713
28		ово	ів	ово						слів	Р.	714
29		ото	іт	ото					болото	боліт		715
30		ето		ето					решето	решіт		716
31		оло	іл	оло					коло	кіл		717
32		ело		ело					село	сіл		718
33		есо	іс	есо					колесо	коліс		719
34		осо	-	осо					просо	-		720
35		но	он	[кг]но					вікно	вікон		721
36		ло	ол	[кз]ло					ікло	ікол		722
37		ко	ок	[ь]ко					серденько	серденьок		723
38		ло	ел	[пбдт]ло					житло	жител		724
39		но	ен	[дтрв]но					зерно	зерен		725
40		ро	ер	[дб]ро					ребро	ребер		726
41		мо	ем	рмо					ярмо	ярем		727
42		мо	ом	смо					пасмо	пасом		728
43		о	ам	о					слово	словам	Д.	729
44			ами							словами	О.	730
45			ах							словах	М.	731
46		г	зі	г				черед. г/з викл. варяг	протяг	протязі		732
47		г		г					луг	лузі		733
48		к	ці	к				черед. к/ц	крок	кроці		734
49		х	сі	х				черед. с/х	реп'ях	реп'ясі		735
50		и	-	и				викл. Карпати	Карпати	Карпат	Р.	736
51			ам							Карпатам	Д.	737
52			ами							Карпатами	О.	738
53			ах							Карпатами	М.	739
1	I/d	а	о	[^жчшщ]а	1	одн	тверда	кличний на [ая]	хата	хато	К.	740
2			е	[жчщ]а					вежа	веже		741
3		я		[лнц]я					вишня	вишне		742
4			є	['ьіеиаоу]я					сім'я	сім'є		743
5			ю	[нср]я					Таня	Таню	К.З.	744
6		ір	оре	[^л]ір	2			кличний на [рнгдблвк] з черг. і/о	вибір	виборе	К.	745
7			ьоре	лір					колір	кольоре		746
8		ін	оне	ін					загін	загоне		747
9		іг	оже	[^р]іг					батіг	батоже		748
10			еже	ріг					оберіг	обереже		749
11		ід	оде	[^л]ід					провід	проводе		750
12			ьоде	[^пг]лід					лід	льоде		751
13			оде	[пг]лід					плід	плоде		752
14			іде	[ос]лід					слід	сліде		753
15		іб	обе	іб					засіб	засобе		754
16		іп	опе	іп					піп	попе		755
17		івш	овше	івш					ківш	ковше		756
18		ізд	озде	ізд					дрізд	дрозде		757
19		іл	оле	іл					дозвіл	дозвзоле		758
20		ів	ове	ів					острів	острове		759
21		їв	єве	їв					Київ	Києве		760
22		ік	оче	ік					рік	роче		761
23		іск	оску	іск					обеліск	обеліску		762
24		іст	осте	іст					піст	посте		763
25		іт	оте	[^л]іт					гніт	гноте		764
26			ьоте	[^п]літ					політ	польоте		765

№	Клас	F1	F2	RE	Відм.	Число	Група	Ознака	Приклад 1	Приклад 2	В-нок	№
27			оте	[п]літ					поте	поте		766
28		із	озе	із					віз	возе		767
29		іж	оже	[^тб]іж					ніж	ноже		768
30				е[тб]іж					небіж	небоже		769
31			ежу	[^е][тб]іж					рубіж	рубежу		770
32				іж					падіж	падежу		771
33		ет	те	ет					Єгипет	Єгипте		772
34		інь	оне	[к]інь				-кінь	кінь	коне		773
35		ідь	едю	ідь				з черг. і/е	ведмідь	ведмедю		774
36		іль	елю	іль					важіль	важелю		775
37		ість	остю	ість					гість	гостю		776
38		інь	ене	[^ксв]інь					корінь	корене		777
39				[о][св]інь					осінь	осене		778
40			еню	[^о][св]інь					ревінь	ревеню		779
41		ор	ре	ор				з вип. -о-	свекор	свекре		780
42		ол	ле	ол					вузол	вузле		781
43		ел	ле	ел				з вип. -е-	орел	орле		782
44		ер	ре	ер					вітер	вітре		783
1	II/e	о	а	о			тверда	ч.р. на -о	батько	батька	Р.	784
2			у							батьку	Д.З.К.	785
3			ові							батькові	Д.	786
4			ом							батьком	О.	787
5		-	у	[^аєиійоуьюя]				ч.р. на 0 крім ов, овов	завод	заводу	Р.	788
6			ові	[^аєиійоуь юяжчшщв]						заводові	Д.М.	789
7				[^о]в					спів	співові		790
8			ом	[^аєиійоуь юяжчшщв]					завод	заводом	О.	791
9			і	[^аєиійоу ьюяггкх]						заводі	М.	792
10			еві	[жчшщ]			мішана	ч.р. на [жчшщ]	товариш	товаришеві	О.	793
11			ем							товаришем	М.	794
12		й	ю	й			м'яка	ч.р. на -й - ій	край	краю	Р.	795
13			єві							краєві	ДМ.	796
14			єм							краєм	О.	797
15			ї							краї	М.	798
16		ь	ю	ь					кріль	кролю	Д.	799
17			еві							кролеві	З.	800
18			ем							кролем	О.	801
19			і							кролі	М.	802
1	II/f	-	и	[^аєиійоуь юяжчшщ]	2	мн	тверда	ч.р. на 0	завод	заводи	Н.	803
2			ів	[^аєиійоуьюя]						заводів	Р.	804
3			ам							заводам	Д.	805
4			ами							заводами	О.	806
5			ах							заводах	М.	807
6		о	и	о				ч.р. на о	батько	батьки	Н.	808
7			ів							батьків	Р.	809
8			ам							батькам	Д.	810
9			ами							батьками	О.	811
10			ах							батьках	М.	812
11		й	ї	й				ч.р. на -й -ій	край	краї	Н.	813
12			їв							країв	Р.	814
13			ям							краям	Д.	815
14			ями							краями	О.	816
15			ях							краях	М.	817
16		-	і	[жчшщ]			мішана	[жчшщ]	товариш	товариші	Н.	818
17		ь	і	ь			м'яка	ч.р. на -ь крім -[ое][нц]ь з вип. [ое]	кріль	кролі	Н.	819
18			ів							кролів	Р.	820
19			ям							кролям	Д.	821
20			ями							кролями	О.	822
21			ях							королях	М.	823
22		я	ів	ття				ття - ттів, викл. з гр. II/і	життя	життів	Р.	824
23			ям							життям	Д.	825
24			ями							життями	О.	826
25			ях							життях	М.	827
1	II/g	-	а	[^аєиійоуьюя]		одн	мішана	ч.р. на -а Р.в.	ягуар	ягуара	Р.	828
2		й	я	й					багатій	багатія		829
3		ь		ь					король	короля		830
1	II/h	-	е	[^аєиійоуь юякгжчшщ]				ч.р. на 0 К.в.	завод	заводе	К.	831
2		к	че	[уая]к					козак	козаче		832
3		к	у	[^уая]				ч.р. на 0 К.в. з Д. співпадає	братик	братику		833
4		г	же	[уо]г					друг	друже		834
5		о	е	[рл]о					Петро	Петре		835
1	III/i	ь	і	ь	3			ж.р. на 0 крім -[ії]сть	міць	міці	Р.Д.М.	836
2		-		[^і][вф]					кров	крові		837
3		ь	ю	[^аоуеи іяюєї].ь					смерть боязнь	смертю боязню	О.	838

№	Клас	F1	F2	RE	Відм.	Число	Група	Ознака	Приклад 1	Приклад 2	В-нок	№
4		-	'ю	[^і][вф]					кров	кров'ю		839
5				[ауі]р					глазур	глазур'ю		840
6			тю	ть				ж.р. на 0 крім -[ії]сть після [аоуеиі яюєї]	благодать	благодаттю		841
7			ню	нь					тінь	тінню		842
8			дю	дь					мідь	міддю		843
9			лю	ль					сіль	сіллю		844
10			зю	зь					галузь	галуззю		845
11			сю	сь					Русь	Руссю		846
12			і	[жчшщр]				ж.р. на [жчшщр] крім -ь	зустріч	зустрічі	Р.	847
13			ю	[^аоуеиіяюєї] [жчшщ]					фальш	фальшю	О.	848
14			чю	[аоуеиіяюєї]ч					зустріч	зустріччю		849
15			жю	[аоуеиіяюєї]ж					подорож	подорожжю		850
16			шю	[аоуеиіяюєї]ш					розкіш	розкішшю		851
17		о	а	о	2		тверда	сер.р. на -о	озеро	озера	Р.	852
18			у							озеру	Д.	853
19			ом							озером	О.	854
20			і	[^кх]о						озері	М.	855
21		хо	сі	хо					вухо	вусі		856
22		ко	ці	око					молоко	молоці		857
23		е	я	[^жчшщ]е			м'яка	сер.р. на -е крім [жчшщ]	море	моря	Р.	858
24			ю							морю	Д.	859
25			ем							морем	О.	860
26			і							морі	М.	861
27			а	[жчшщ]е			мішана	сер.р. на -е після [жчшщ]	прізвище	прізвища	Р.	862
28			у							прізвищу	Д.	863
29			ем							прізвищем	О.	864
30			і							прізвищі	М.	865
31		я	ю	я				сер.р. на -я	завдання	завданню	Д.	866
32			ям							завданням	О.	867
33			і	[^'ь]я						завданні	М.	868
34			ї	['ь]я					бездощів'я	бездощів'ї		869
35		ий	ого	ий				прикметн. чол. р. без -ций/їй	хорунжий	хорунжого	Р.З.	870
36			ому							хорунжому	Д.М.	871
37			им							хорунжим	О.	872
38		ій	ього	ій					Вишній	Вишнього	Р.З.	873
39			ьому							Вишньому	Д.М.	874
40			ім							Вишнім	О.	875
41		а	ої	[тнк]а				прикметн. жін. р.	чебуречна	чебуречної	Р.	876
42			ій							чебуречній	Д.М.	877
43			у							чебуречну	З.	878
44			ою							чебуречною	О.	879
45		е	ого	[нк]е				прикметн. сер. р.	подільне	подільного	Р.З.	880
46			ому							подільному	Д.М.	881
47			им							подільним	О.	882
1	III/j	ь	ей	ь	3	мн.		ж.р. на 0 крім -[ії]сть	тінь	тіней	Р.	883
2			ям							тіням	Д.	884
3			ями							тінями	О.	885
4			ях							тінях	М.	886
5		-	ей	[жчшщ]				ж.р. на [жчшщр] крім -ь	зустріч	зустрічей	Р.	887
6			ам							зустрічам	Д.	888
7			ами							зустрічами	О.	889
8			ах							зустрічах	М.	890
9			ей	[ау]р					глазур	глазурей	Р.	891
10			ям							глазурям	Д.	892
11			ями							глазурями	О.	893
12			ях							глазурях	М.	894
13			ей	ф					верф	верфей	Р.	895
14			ям							верфям	Д.	896
15			ями							верфями	О.	897
16			ях							верфях	М.	898
17		о	а	о	2		тверда	сер.р. на -о	гасло	гасла	Н.	899
18		ло	ел	сло					гасло	гасел	Р.	900
19		о	-	[^с]ло					тіло	тіл		901
20		ко	ок	[^ьоаеиіу]ко					коліщатко	коліщаток		902
21		о	-	[ьоаеиіу]ко					лико	лик		903
22				[^лк]о					озеро	озер		904
23			ам	о						озерам	Д.	905
24			ами							озерами	О.	906
25			ах							озерах	М.	907
26		е	ів	[^жчшщц]е			м'яка	сер.р. на -е крім [жчшщ]	море	морів	Р.	908
27			ям							морям	Д.	909
28			ями							морями	О.	910
29			ях							морях	М.	911
30		це	дець	серце			мішана	сер.р. на -е після [жчшщ]	серце	сердець	Р.	912
31			ець	[^с]ерце					озерце	озерець		913
32		е	ь	ісце					місце	місць		914
33		йце	єць	йце					яйце	яєць		915

№	Клас	F1	F2	RE	Відм.	Число	Група	Ознака	Приклад 1	Приклад 2	В-нок	№
34		ьце	ець	ьце					бильце	билець		916
35		це		[^йрсь]це					віконце	віконець		917
36		е	-	[жшщ]е					прізвище	прізвищ		918
37		-	й	че					плече	плечей		919
38		е	ам	[жчшщ]е					прізвище	прізвищам	Д.	920
39			ами							прізвищами	О.	921
40			ах							прізвищах	М.	922
41		я	їв	'я				сер.р.	бездощів'я	бездощів'їв	Р.	923
42		ня	ь	ння				на -я	завдання	завдань		924
43		я	ів	[у]жжя					подружжя	подружжів		925
44		жя	-	[^у]жжя					бемежжя	безмеж		926
45		тя	ь	ття					заняття	занять		927
46		я	т	[^т]тя					дитя	дитят		928
47		дя	ь	ддя					безвладдя	безвладь		929
48		я		[^д]дя					безглуздя	безглуздь		930
49		ля		лля					зусилля	зусиль		931
50		чя	-	ччя					сторіччя	сторіч		932
51		шя		шшя					затишшя	затиш		933
52		ся	ь	сся					Полісся	Полісь		934
53		я		[^лджтншчь' аоуеіис] я					повітря	повітрь		935
54			ями	я					завдання	завданнями	О.	936
55			ях	я						завданнях	М.	937
56		ий	і	ий				прикметн.	хорунжий	хорунжі	Н.	938
57			их					чол. р.		хорунжих	Р.	939
58			ими					без -ций/їй		хорунжими	О.	940
59		і	их	і					морські	морських	Р.З.М.	941
60			им							морським	Д.	942
61			ими							морськими	О.	943
62		н	-	ин				кияни,	львів'янин	львів'яни	Н.	944
63								львів'янин,		львів'ян	Р.	945
64			ам					бояри....		львів'янам	Д.	946
65			ами							львів'янами	О.	947
66			ах							львів'янах	М.	948
1	III/k	ь	е	ь	3	одн		кличний	смерть	смерте	К.	949
2		-		[чшж]				ж.р. на 0	зустріч	зустріче		950
3				[^і]в					кров	крове		951
4		ий	а	ий				ж.р.	вожатий	вожата	Н.	952
5			ій					прикметн		вожатій	Д.	953
6			ої							вожатої	Р.	954
7			у							вожату	З.	955
8			ою							вожатою	О.	956
1	IV/l	інь	ені	о[св]інь				з черг. і/е	осінь	осені	Р.Д.М.	957
2		ь	ню							осінню	О.	958
3		іль	олі	іль				з черг. і/о	сіль	солі	Р.Д.М.	959
4		ь	лю	іль						сіллю	О.	960
5		іць	оці	іць					міць	моці	Р.Д.М.	961
6		ь	цю	іць						міццю	О.	962
7		іч	ечі	[^н]іч				на шипл.	піч	печі	Р.	963
8			очі	ніч				крім -ь	ніч	ночі		964
9		іш	оші	іш					розкіш	розкоші		965
10		-	чю	іч					піч	піччю	О.	966
11			шю	іш					розкіш	розкішшю		967
12		ість	ості	ість				-ість -їсть	ніжність	ніжності	Р.Д.М	968
13		їсть	йості	їсть					безкрайсть	безкрайості		969
14		ь	і	[^ії]сть					водорість	водорості		970
15			ю	сть					ніжність	ніжністю	О.	971
16		-	ти	[^'][ая]	4			на -а -я	цуценя	цуценяти	Р.	972
17			ті					викл.		цуценяті	Д.М.	973
18			м	[ая]				хлоп'я		цуценям	О.	974
19			ти	п'я				хлоп'я	хлоп'я	хлоп'яти	Р.	975
20			ті						ім'я	хлоп'яті	Д.М.	976
21		'я	ені	[^п]'я				на -'я		імені	Р.	977
22			енем							іменем	О.	978
23		одець	ідця	одець	2			ч. р. на	виходець	вихідця	Р.З	979
24			ідцю					-о[дв]ець		вихідцю	Д.	980
25			ідцеві					з вип. е і		вихідцеві		981
26			ідцем					черг. о-і,		вихідцем	О.	982
27			ідці					в Р. в. на -а		вихідці	М.	983
28		овець	івця	овець					вдовець	вдівця	Р.З.	984
29			івцю							вдівцю	Д.	985
30			івцеві							вдівцеві		986
31			івцем							вдівцем	О.	987
32			івці							вдівці	М.	988
33		-	у	яр			мішана	на -яр	газетяр	газетяру	Р.	989
34			ем							газетярем	О.	990
35			еві							газетяреві	М.Д.	991
36			і							газетярі	М.	992

№	Клас	F1	F2	RE	Відм.	Число	Група	Ознака	Приклад 1	Приклад 2	В-нок	№
37			ю	[аиу]р				ч. р. на -ар -ир наголош.	кобзар	кобзарю	Д.М.	993
38			еві							кобзареві		994
39			ем							кобзарем	О.	995
40			і							кобзарі	М.	996
1	IV/m	інь	еней	о[св]інь	3			з черг. і е	осінь	осеней	Р.	997
2			еням							осеням	Д.	998
3			енями							осенями	О.	999
4			енях							осенях	М.	1000
5		іль	олей	іль				з черг. і о	сіль	солей	Р.	1001
6			олям							солям	Д.	1002
7			олями							солями	О.	1003
8			олях							солях	М.	1004
9		іч	ечей	[^н]іч		мн		з черг. і е	піч	печей	Р.	1005
10			ечам							печам	Д.	1006
11			ечами							печами	О.	1007
12			ечах							печах	М.	1008
13			очей	ніч				з черг. і о	ніч	ночей	Р.	1009
14			очам							ночам	Д.	1010
15			очами							ночами	О.	1011
16			очах							ночах	М.	1012
17		іш	ошей	іш					розкіш	розкошей	Р.	1013
18			ошам							розкошам	Д.	1014
19			ошами							розкошами	О.	1015
20			ошах							розкошах	М.	1016
21		ість	остей	ість				на -ість	ніжність	ніжностей	Р.	1017
22			остям							ніжностям	Д.	1018
23			остями							ніжностями	О.	1019
24			остях							ніжностях	М.	1020
25		їсть	йостей	їсть				на -їсть	безкрайсть	безкрайостей	Р.	1021
26			йостям							безкрайостям	Д.	1022
27			йостями							безкрайостями	О.	1023
28			йостях							безкрайостях	М.	1024
29		ь	ей	[^їі]сть				на -[їі]сть	водорість	водоростей	Р.	1025
30			ям							водоростям	Д.	1026
31			ями							водоростями	О.	1027
32			ях							водоростях	М.	1028
33		-	та	[^'][яа]	4			с. р. на -а -я викл. хлоп'я	цуценя	цуценята	Н.	1029
34			т							цуценят	Р.	1030
35			там							цуценятам	Д.	1031
36			тами							цуценятами	О.	1032
37			тах							цуценятах	Д.	1033
38			та	п'я				хлоп'я	хлоп'я	хлоп'ята	Н.	1034
39			т							хлоп'ят	Р.	1035
40			там							хлоп'ятам	Д.	1036
41			тами							хлоп'ятами	О.	1037
42			тах							хлоп'ятах	Д.	1038
43		'я	ена	[^п]'я				на -'я	ім'я	імена	Н.	1039
44			ен							імен	Р.	1040
45			енам							іменам	Д.	1041
46			енами							іменами	О.	1042
47			енах							іменах	М.	1043
48		одець	ідці	одець	2			ч. р. на -о[дв]ець з вип. е і черг. о-і в Р. в. на -а	виходець	вихідці	Мн.	1044
49			ідців							вихідця	Р.З.	1045
50			ідцям							вихідцю	Д.	1046
51			ідцями							вихідцеві	О.	1047
52			ідцях							вихідцем	М.	1048
53		овець	івці	овець					вдовець	вдівці	Мн.	1049
54			івців							вдівців	Р.З.	1050
55			івцям							вдівцям	Д.	1051
56			івцями							вдівцями	О.	1052
57			івцях							вдівцях	М.	1053
58		-	і	яр				на -яр	газетяр	газетярі	Н.	1054
59			ів							газетярів	Р.	1055
60			ам							газетярам	Д.	1056
61			ами							газетярами	О.	1057
62			ах							газетярах	М.	1058
63			ів	[аиу]р			м'яка	ч. р. на -ар -ир наголошені	кобзар	кобзарів	Д.	1059
64			ям							кобзарям	Р.	1060
65			ями							кобзарями	О.	1061
66			ях							кобзарях	М.	1062
1	IV/n	інь	ене	о[св]інь		од		з черг. і е	осінь	осене	К.	1063
2		іль	оле	іль				з черг. і о	сіль	соле		1064
3		іч	ече	[^н]іч				на шиплячий крім -ь	піч	пече		1065
4			оче	ніч					ніч	ноче		1066
5		іш	оше	іш					розкіш	розкоше		1067
6		ість	осте	ість	3			ж. р. на 0	ніжність	ніжносте		1068
7		їсть	йосте	їсть					безкрайсть	безкрайосте		1069
8		ь	те	[^їі]сть					водорість	водоросте		1070

№	Клас	F1	F2	RE	Відм.	Число	Група	Ознака	Приклад 1	Приклад 2	В-нок	№
9		-	е	[аиуя]р	2		-	на -яр – ар -ир	газетяр	газетяре		1071
1	IV/q		а	яр			мішана	на -яр	газетяр	газетяра	Р.	1072
2			я	[аиу]р			м'яка	на -ар -ир	кобзар	кобзаря	Р.З.	1073
1	V/p	я	йович	ія	-	од/мн	-	-ія	Єремія Єремійович Єреміївна	Єремійович	Н.ч.о.	1074
2			ївна							Єреміївна	Н.ж.о.	1075
3			йовича							Єремійовича	Р.Д.ч.о.	1076
4			ївни							Єреміївни	Р.ж.о.Н.ж.м	1077
5			йовичу							Єремійовичу	Д.М.К.ч.о.	1078
6			йовичеві							Єремійовичеві	Д.М.ч.о.	1079
7			ївні							Єреміївні	Д.М.ж.о.	1080
8			ївну							Єреміївну	З.ж.о.	1081
9			йовичем							Єремійовичем	О.Д.ч.о.	1082
10			ївною							Єреміївною	О.ж.о.	1083
11			йовичі							Єремійовичі	М.ч.о.Н.ч.м	1084
12			ївно							Єреміївно	К.ж.о.	1085
13			йовичів							Єремійовичів	Р.З.ч.м.	1086
14			ївн							Єреміївн	Р.З.ж.м.	1087
15			йовичам							Єремійовичам	Д.ч.м.	1088
16			ївнам							Єреміївнам	Д.ж.м.	1089
17			йовичами							Єремійовичами	О.ч.м.	1090
18			ївнами							Єреміївнами	О.ж.м.	1091
19			йовичах							Єремійовичах	М.ч.м.	1092
20			ївнах							Єреміївнах	М.ж.м.	1093
21			іч	[^і]я				патроніми на -[^і]я	Ілля Ілліч Іллівна	Ілліч	Н.ч.о.	1094
22			івна							Іллівна	Н.ж.о.	1095
23			іча							Ілліча	Р.Д.ч.о.	1096
24			івни							Іллівни	Р.ж.о.Н.ж.м	1097
25			ічу							Іллічу	Д.М.К.ч.о.	1098
26			ічеві							Іллічеві	Д.М.ч.о.	1099
27			івні							Іллівні	Д.М.ж.о.	1100
28			івну							Іллівну	З.ж.о.	1101
29			ічем							Іллічем	О.Д.ч.о.	1102
30			івною							Іллівною	О.ж.о.	1103
31			ічі							Іллічі	М.ч.о.Н.ч.м	1104
32			івно							Іллівно	К.ж.о.	1105
33			ічів							Іллічів	Р.З.ч.м.	1106
34			івн							Іллівн	Р.З.ж.м.	1107
35			ічам							Іллічам	Д.ч.м.	1108
36			івнам							Іллівнам	Д.ж.м.	1109
37			ічами							Іллічами	О.ч.м.	1110
38			івнами							Іллівнами	О.ж.м.	1111
39			ічах							Іллічах	М.ч.м.	1112
40			івнах							Іллівнах	М.ж.м.	1113
41		а	ич	[^тс]а				патроніми на -а	Кузьма Кузьмич Кузьмівна	Кузьмич	Н.ч.о.	1114
42			ича							Кузьмича	Р.Д.ч.о.	1115
43			ичу							Кузьмичу	Д.М.К.ч.о.	1116
44			ичеві							Кузьмичеві	Д.М.ч.о.	1117
45			ичем							Кузьмичем	О.Д.ч.о.	1118
46			ичі							Кузьмичі	М.ч.о.Н.ч.м	1119
47			ович	[тс]а					Олекса	Олексович	Н.ч.о.	1120
48			овича							Олексовича	Р.Д.ч.о.	1121
49			овичу							Олексовичу	Д.М.К.ч.о.	1122
50			овичеві							Олексовичеві	Д.М.ч.о.	1123
51			овичем							Олексовичем	О.Д.ч.о.	1124
52			овичі							Олексовичі	М.ч.о.Н.ч.м	1125
53			івна	а					Кузьма	Кузьмівна	Н.ж.о.	1126
54			івни							Кузьмівни	Р.ж.о.Н.ж.м	1127
55			івні							Кузьмівні	Д.М.ж.о.	1128
56			івну							Кузьмівну	З.ж.о.	1129
57			івною							Кузьмівною	О.ж.о.	1130
58			івно							Кузьмівно	К.ж.о.	1131
59			ичів	[^тс]а						Кузьмичів	Р.З.ч.м.	1132
60			ичам							Кузьмичам	Д.ч.м.	1133
61			ичами							Кузьмичами	О.ч.м.	1134
62			ичах							Кузьмичах	М.ч.м.	1135
63			овичів	[тс]а					Олекса	Олексовичів	Р.З.ч.м.	1136
64			овичам							Олексовичам	Д.ч.м.	1137
65			овичами							Олексовичами	О.ч.м.	1138
66			овичах							Олексовичах	М.ч.м.	1139
67			івн						Кузьма	Кузьмівн	Р.З.ж.м	1140
68			івнам							Кузьмівнам	Д.ж.м.	1141
69			івнами							Кузьмівнами	О.ж.м.	1142
70			івнах							Кузьмівнах	М.ж.м.	1143
71		ір	орович	ір				патроніми на -ір	Федор Федорович Федорівна	Федорович	Н.ч.о.	1144
72			орівна							Федорівна	Н.ж.о.	1145
73			оровича							Федоровича	Р.Д.ч.о.	1146
74			орівни							Федорівни	Р.ж.о.Н.ж.м	1147

№	Клас	F1	F2	RE	Відм.	Число	Група	Ознака	Приклад 1	Приклад 2	В-нок	№
75			оровичу							Федоровичу	Д.М.К.ч.о.	1148
76			оровичеві							Федоровичеві	Д.М.ч.о.	1149
77			орівні							Федорівні	Д.М.ж.о.	1150
78			орівну							Федорівну	З.ж.о.	1151
79			оровичем							Федоровичем	О.Д.ч.о.	1152
80			орівною							Федорівною	О.ж.о.	1153
81			оровичі							Федоровичі	М.ч.о.Н.ч.м	1154
82			орівно							Федорівно	К.ж.о.	1155
83			оровичів							Федоровичів	Р.З.ч.м.	1156
84			орівн							Федорівн	Р.З.ж.м.	1157
85			оровичам							Федоровичам	Д.ч.м.	1158
86			орівнам							Федорівнам	Д.ж.м.	1159
87			оровичами							Федоровичами	О.ч.м.	1160
88			орівнами							Федорівнами	О.ж.м.	1161
89			оровичах							Федоровичах	М.ч.м.	1162
90			орівнах							Федорівнах	М.ж.м.	1163
91		ін	онович	и[мх]ін				патроніми на -ін	Пимін Пимонович Пимонівна	Пимонович	Н.ч.о.	1164
92			онівна							Пимонівна	Н.ж.о.	1165
93			оновича							Пимоновича	Р.Д.ч.о.	1166
94			онівни							Пимонівни	Р.ж.о.Н.ж.м	1167
95			оновичу							Пимоновичу	Д.М.К.ч.о.	1168
96			оновичеві							Пимоновичеві	Д.М.ч.о.	1169
97			онівні							Пимонівні	Д.М.ж.о.	1170
98			онівну							Пимонівну	З.ж.о.	1171
99			оновичем							Пимоновичем	О.Д.ч.о.	1172
100			онівною							Пимонівною	О.ж.о.	1173
101			оновичі							Пимоновичі	М.ч.о.Н.ч.м	1174
102			онівно							Пимонівно	К.ж.о.	1175
103			оновичів							Пимоновичів	Р.З.ч.м.	1176
104			онівн							Пимонівн	Р.З.ж.м.	1177
105			оновичам							Пимоновичам	Д.ч.м.	1178
106			онівнам							Пимонівнам	Д.ж.м.	1179
107			оновичами							Пимоновичами	О.ч.м.	1180
108			онівнами							Пимонівнами	О.ж.м.	1181
109			оновичах							Пимоновичах	М.ч.м.	1182
110			онівнах							Пимонівнах	М.ж.м.	1183
111		ів	ович	ів					Яків Якович Яківна	Якович	Н.ч.о.	1184
112			івна							Яківна	Н.ж.о.	1185
113			овича							Яковича	Р.Д.ч.о.	1186
114			івни							Яківни	Р.ж.о.Н.ж.м	1187
115			овичу							Яковичу	Д.М.К.ч.о.	1188
116			овичеві							Яковичеві	Д.М.ч.о	1189
117			івні							Яківні	Д.М.ж.о.	1190
118			івну							Яківну	З.ж.о.	1191
119			овичем							Яковичем	О.Д.ч.о.	1192
120			івною							Яківною	О.ж.о.	1193
121			овичі							Яковичі	М.ч.о.Н.ч.м	1194
122			івно							Яківно	К.ж.о.	1195
123			овичів							Яковичів	Р.З.ч.м.	1196
124			івн							Яківн	Р.З.ж.м.	1197
125			овичам							Яковичам	Д.ч.м.	1198
126			івнам							Яківнам	Д.ж.м.	1199
127			овичами							Яковичами	О.ч.м.	1200
128			івнами							Яківнами	О.ж.м.	1201
129			овичах							Яковичах	М.ч.м.	1202
130			івнах							Яківнах	М.ж.м.	1203
131		о	ович	о				патроніми на -о	Павло Павлович Павлівна	Павлович	Н.ч.о.	1204
132			івна							Павлівна	Н.ж.о.	1205
133			овича							Павловича	Р.Д.ч.о.	1206
134			івни							Павлівни	Р.ж.о.Н.ж.м	1207
135			овичу							Павловичу	Д.М.К.ч.о.	1208
136			овичеві							Павловичеві	Д.М.ч.о.	1209
137			івні							Павлівні	Д.М.ж.о.	1210
138			івну							Павлівну	З.ж.о.	1211
139			овичем							Павловичем	О.Д.ч.о.	1212
140			івною							Павлівною	О.ж.о.	1213
141			овичі							Павловичі	М.ч.о.Н.ч.м	1214
142			івно							Павлівно	К.ж.о.	1215
143			овичів							Павловичів	Р.З.ч.м.	1216
144			івн							Павлівн	Р.З.ж.м.	1217
145			овичам							Павловичам	Д.ч.м.	1218
146			івнам							Павлівнам	Д.ж.м.	1219
147			овичами							Павловичами	О.ч.м.	1220
148			івнами							Павлівнами	О.ж.м.	1221
149			овичах							Павловичах	М.ч.м.	1222
150			івнах							Павлівнах	М.ж.м.	1223
151		-	ович	[^врнуеа оійїєяю]				на приголосні	Антон Антонович	Антонович	Н.ч.о.	1224
152			овича							Антоновича	Р.Д.ч.о.	1225

№	Клас	F1	F2	RE	Відм.	Число	Група	Ознака	Приклад 1	Приклад 2	В-нок	№
153			овичу					крім [^врм]	Антонівна	Антоновичу	Д.М.К.ч.о.	1226
154			овичем							Антоновичем	О.Д.ч.о.	1227
155			овичеві							Антоновичеві	Д.М.ч.о.	1228
156			овичі							Антоновичі	М.ч.о.Н.ч.м	1229
157			івна	[^врнйуеа оіиїєяюь]						Антонівна	Н.ж.о.	1230
158			івни							Антонівни	Р.ж.о.Н.ж.м	1231
159			івні							Антонівні	Д.М.ж.о.	1232
160			івну							Антонівну	З.ж.о.	1233
161			івною							Антонівною	О.ж.о.	1234
162			івно							Антонівно	К.ж.о.	1235
163		й	ївна	й					Анатолій Анато-лійович Анато-ліївна	Анатоліївна	Н.ж.о.	1236
164			ївни							Анатоліївни	Р.ж.о.Н.ж.м	1237
165			ївні							Анатоліївні	Д.М.ж.о.	1238
166			ївну							Анатоліївну	З.ж.о.	1239
167			ївною							Анатоліївною	О.ж.о.	1240
168			ївно							Анатоліївно	К.ж.о.	1241
169		-	овичів	[^врнуеа оіиїєяю]						Антоновичів	Р.З.ч.м.	1242
170			івн							Антонівн	Р.З.ж.м.	1243
171			овичам							Антоновичам	Д.ч.м.	1244
172			івнам							Антонівнам	Д.ж.м.	1245
173			овичами							Антоновичами	О.ч.м.	1246
174			івнами							Антонівнами	О.ж.м.	1247
175			овичах							Антоновичах	М.ч.м.	1248
176			івнах							Антонівнах	М.ж.м.	1249
177		ь	івна	ь				ж. р. на -ь	Василь	Василівна	Н.ж.о.	1250
178			івни							Василівни	Р.ж.о.Н.ж.м	1251
179			івні							Василівні	Д.М.ж.о.	1252
180			івну							Василівну	З.ж.о.	1253
181			івною							Василівною	О.ж.о.	1254
182			івно							Василівно	К.ж.о.	1255
183		-	ович	[^і][врн]				[^і][врн]	Єгор Єгорович Єгорівна	Єгорович	Н.ч.о.	1256
184			івна							Єгорівна	Н.ж.о.	1257
185			овича							Єгоровича	Р.Д.ч.о.	1258
186			івни							Єгорівни	Р.ж.о.Н.ж.м	1259
187			овичу							Єгоровичу	Д.М.К.ч.о.	1260
188			овичеві							Єгоровичеві	Д.М.ч.о.	1261
189			івні							Єгорівні	Д.М.ж.о.	1262
190			івну							Єгорівну	З.ж.о.	1263
191			овичем							Єгоровичем	О.Д.ч.о.	1264
192			івною							Єгорівною	О.ж.о.	1265
193			овичі							Єгоровичі	М.ч.о.Н.ч.м	1266
194			івно							Єгорівно	К.ж.о.	1267
195			овичів							Єгоровичів	Р.З.ч.м.	1268
196			івн							Єгорівн	Р.З.ж.м.	1269
197			овичам							Єгоровичам	Д.ч.м.	1270
198			івнам							Єгорівнам	Д.ж.м.	1271
199			овичами							Єгоровичами	О.ч.м.	1272
200			івнами							Єгорівнами	О.ж.м.	1273
201			овичах							Єгоровичах	М.ч.м.	1274
202			івнах							Єгорівнах	М.ж.м.	1275
203			ович	[^и]мін				[^и]мін	Фомін Фомінович Фомінівна	Фомінович	Н.ч.о.	1276
204			івна							Фомінівна	Н.ж.о.	1277
205			овича							Фоміновича	Р.Д.ч.о.	1278
206			івни							Фомінівни	Р.ж.о.Н.ж.м	1279
207			овичу							Фоміновичу	Д.М.К.ч.о.	1280
208			овичеві							Фоміновичеві	Д.М.ч.о.	1281
209			івні							Фомінівні	Д.М.ж.о.	1282
210			івну							Фомінівну	З.ж.о.	1283
211			овичем							Фоміновичем	О.Д.ч.о.	1284
212			івною							Фомінівною	О.ж.о.	1285
213			овичі							Фоміновичі	М.ч.о.Н.ч.м	1286
214			івно							Фомінівно	К.ж.о.	1287
215			овичів							Фоміновичів	Р.З.ч.м.	1288
216			івн							Фомінівн	Р.З.ж.м.	1289
217			овичам							Фоміновичам	Д.ч.м.	1290
218			івнам							Фомінівнам	Д.ж.м.	1291
219			овичами							Фоміновичами	О.ч.м.	1292
220			івнами							Фомінівнами	О.ж.м.	1293
221			овичах							Фоміновичах	М.ч.м.	1294
222			івнах							Фомінівнах	М.ж.м.	1295

Таблиця А.12

Основні RE типу SFX для МА українських дієслів на основі https://goroh.pp.ua/ [269-276]

№	Клас	F1	F2	RE	Час	НФ	Ознака	Приклад 1	Приклад 2	Займенник	№
1	А	ти	ла	[^сй]ти	МН	-	(окрім Я, Ти, Він) для всіх, крім -йти/-сти	клонувати	клонувала	Вона	1
2			ло						клонувало	Воно	2

№	Клас	F1	F2	RE	Час	НФ	Ознака	Приклад 1	Приклад 2	Займенник	№
3			ли						клонували	Ми,Ви,Вони	3
4			в	[аеиіїоуя]ти					клонував	Я, Ти, Він	4
5		вати	ю	[ауюя]вати	Т		-вати (Т недоконаної форми, МБ доконаної)		клоную	Я	5
6			єш						клонуєш	Ти	6
7			є						клонує	Він	7
8			ємо						клонуємо	Ми	8
9			єте						клонуєте	Ви	9
10			ють						клонують	Вони	10
11		вати	й	[ую]вати	МН	-ати	-ати(Т недоконаної форми, МБ доконаної)		клонуй	Ти	11
12			ймо						клонуймо	Ми	12
13			йте						клонуйте	Ви	13
14		ти	й	[ая]вати				вставати	вставай	Ти	14
15			ймо						вставаймо	Ми	15
16			йте						вставайте	Ви	16
17		ати	у	[рз]вати	Т		-рвати, -звати (Т недоконаної форми, МБ доконаної)	рвати	рву	Я	17
18			еш						рвеш	Ти	18
19			е						рве	Він	19
20			емо						рвемо	Ми	20
21			ете						рвете	Ви	21
22			уть						рвуть	Вони	22
23			и						рви	Ти	23
24			імо						рвімо	Ми	24
25			іть						рвіть	Ви	25
26		зати	жу	зати			-зати з черг. з/ж -зати (Т недоконаної форми, МБ доконаної)	казати	кажу	Я	26
27			жеш						кажеш	Ти	27
28			же						каже	Він	28
29			жемо						кажемо	Ми	29
30			жете						кажете	Ви	30
31			жуть						кажуть	Вони	31
32			ж	ізати	МН	-зати	МН -зати: жи (д.ф. зв'яжи), -ж (різати, зарізати - ріж), і в гр. /І -зай (вирізати - вирізай)	різати	ріж	Ти	32
33				мазати				мазати	маж		33
34			жи	казати				казати	кажи		34
35				[еия]зати				лизати	лижи		35
36			жмо	ізати				відрізати	відріжмо	Ми	36
37				мазати				мазати	мажмо		37
38			жімо	казати				казати	кажімо		38
39				[еия]зати				лизати	лижімо		39
40			жте	ізати				відрізати	відріжте	Ви	40
41				мазати				мазати	мажте		41
42			жіть	казати				казати	кажіть		42
43				[еия]зати				лизати	лижіть		43
44		ати	у	[днжщ]ати	Т	-	Т недоконаної форми, МБ доконаної викл: жати (І) - жму (гр /К), жати (ІІ) - жну викл: (за)(і)ржати в гр. М, слати"	блищати	блищу	Я	44
45				[^ао]чати				деренчати	деренчу		45
46			ну	[ао]чати				зачати	зачну		46
47		тати	чу	[^с]тати				шептати	шепчу		47
48		кати		[^с]кати				плакати	плачу		48
49		сати	шу	сати				писати	пишу		49
50		хати		хати				брехати	брешу		50
51		стати		стати				свистати	свищу		51
52		скати		скати				плескати	плещу		52
53		слати	шлю	слати				послати	пошлю		53
54		ати	лю	пати				сипати	сиплю		54
55			ю	орати				орати	орю		55
56		рати	еру	[бдп]рати				брати	беру		56
57		ати	еш	[дн]ати				стогнати	стогнеш	Ти	57
58			иш	[жщ]ати				блищати	блищиш		58
59				[^оа]чати				бряжчати	бряжчиш		59
60			неш	[ао]чати				зачати	зачнеш		60
61		тати	чеш	[^с]тати				шептати	шепчеш		61
62		кати		[^с]кати				плакати	плачеш		62
63		сати	шеш	сати				писати	пишеш		63
64		хати		хати				брехати	брешеш		64
65		стати	щеш	стати				свистати	свищеш		65
66		скати		скати				плескати	плещеш		66
67		слати	шлеш	слати				послати	пошлеш		67
68		ати	леш	ипати				сипати	сиплеш		68
69			иш	спати				спати	спиш		69
70			еш	орати				орати	ореш		70
71		рати	ереш	[бдп]рати				брати	береш		71
72		ати	е	[дн]ати				стогнати	стогне	Він	72
73			ить	[жщ]ати				блищати	блищить		73
74				[^оа]чати				бряжчати	бряжчить		74
75			не	[ао]чати				зачати	зачне		75
76		тати	че	[^с]тати				шептати	шепче		76
77		кати		[^с]кати				плакати	плаче		77
78		сати	ше	сати				писати	пише		78
79		хати		хати				брехати	бреше		79
80		стати	ще	стати				свистати	свище		80
81		скати		скати				плескати	плеще		81
82		слати	шле	слати				послати	пошле		82
83		ати	ле	ипати				сипати	сипле		83
84			ить	спати				спати	спить		84
85			е	орати				орати	оре		85

№	Клас	F1	F2	RE	Час	НФ	Ознака	Приклад 1	Приклад 2	Займенник	№
86		рати	ере	[бдп]рати				брати	бере		86
87		ати	емо	[дн]ати				стогнати	стогнемо	Ми	87
88			имо	[жщ]ати				блищати	блищимо		88
89			немо	[ао]чати				зачати	зачнемо		89
90		чати	чимо	[^оа]чати				бряжчати	бряжчимо		90
91		тати	чемо	[^с]тати				шептати	шепчемо		91
92		кати		[^с]кати				плакати	плачемо		92
93		сати	шемо	сати				писати	пишемо		93
94		хати		хати				брехати	брешемо		94
95		стати	щемо	стати				свистати	свищемо		95
96		скати		скати				плескати	плещемо		96
97		слати	шлемо	слати				послати	пошлемо		97
98		ати	лемо	ипати				сипати	сиплемо		98
99			имо	спати				спати	спимо		99
100			емо	орати				орати	оремо		100
101		рати	еремо	[бдп]рати				брати	беремо		101
102		ати	ете	[дн]ати				стогнати	стогнете	Ви	102
103			ите	[жщ]ати				блищати	блищиш		103
104			нете	[ао]чати				зачати	зачнете		104
105		чати	чите	[^оа]чати				бряжчати	бряжчиш		105
106		тати	чете	[^с]тати				шептати	шепчете		106
107		кати		[^с]кати				плакати	плачете		107
108		сати	шете	сати				писати	пишете		108
109		хати		хати				брехати	брешете		109
110		стати	щете	стати				свистати	свищете		110
111		скати		скати				плескати	плещете		111
112		слати	шлете	слати				послати	пошлете		112
113		ати	лете	ипати				сипати	сиплете		113
114			ите	спати				спати	спите		114
115			ете	орати				орати	орете		115
116		рати	ерете	[бдп]рати				брати	берете		116
117		ати	уть	[дн]ати				стогнати	стогнуть	Вони	117
118			ать	[жщ]ати				блищати	блищать		118
119			нуть	[ао]чати				зачати	зачнуть		119
120			ать	[^ао]чати				бряжчати	бряжчать		120
121		тати	чуть	[^с]тати				шептати	шепчуть		121
122		кати		[^с]кати				плакати	плачуть		122
123		сати	шуть	сати				писати	пишуть		123
124		хати		хати				брехати	брешуть		124
125		стати	щуть	стати				свистати	свищуть		125
126		скати		скати				плескати	плещуть		126
127		слати	шлють	слати				послати	пошлють		127
128		ати	лють	ипати				сипати	сиплють		128
129			лять	спати				спати	сплять		129
130			ють	орати				орати	орють		130
131		рати	еруть	[бдп]рати				брати	беруть		131
132		ати	и	[днжщ]ати	МН	-ати	-нати	блищати	блищи	Ти	132
133			ни	[ао]чати				зачати	зачни		133
134			и	[^ао]чати				деренчати	деренчи		134
135		тати	чи	[^с]тати				шептати	шепчи		135
136		сати	ши	сати				писати	пиши		136
137		хати		хати				брехати	бреши		137
138		кати	ч	лакати				плакати	плач		138
139			чи	какати				скакати	скачи		139
140				ткати				ткати	тчи		140
141			ч	икати				кликати	клич		141
142		скати	щи	скати				плескати	плещи		142
143		стати		стати				свистати	свищи		143
144		слати	шли	слати				послати	пошли		144
145		пати	пи	спати				сипати	спи		145
146		пати	п	ипати				сипати	сип		146
147		ати	и	орати				орати	ори		147
148		рати	ери	[бдп]рати				брати	бери		148
149		ати	імо	[днжщ]ати				блищати	блищімо	Ми	149
150			німо	[ао]чати				зачати	зачнімо		150
151			імо	[^оа]чати				бряжчати	бряжчімо		151
152		тати	чімо	[^с]тати				шептати	шепчімо		152
153		сати	шімо	сати				писати	пишімо		153
154		хати	шімо	хати				брехати	брешімо		154
155		кати	чмо	лакати				плакати	плачмо		155
156			чімо	какати				скакати	скачімо		156
157				ткати				ткати	тчімо		157
158			чмо	икати				кликати	кличмо		158
159		скати	щімо	скати				плескати	плещімо		159
160		стати		стати				свистати	свищімо		160
161		слати	шлімо	слати				послати	пошлімо		161
162		ати	мо	ипати				сипати	сипмо		162
163			імо	спати				спати	спімо		163
164				орати				орати	орімо		164
165		рати	ерімо	[бдп]рати				брати	берімо		165
166		ати	іть	[джцн]ати				блищати	блищіть	Ви	166
167			ніть	[ао]чати				зачати	зачніть		167
168		чати	чіть	[^оа]чати				бряжчати	бряжчіть		168

№	Клас	F1	F2	RE	Час	НФ	Ознака	Приклад 1	Приклад 2	Займенник	№
169		тати		[^с]тати				шептати	шепчіть		169
170		сати	шіть	сати				писати	пишіть		170
171		хати		хати				брехати	брешіть		171
172		кати	чте	лакати				плакати	плачте		172
173			чіть	какати				скакати	скачіть		173
174				ткати				ткати	тчіть		174
175		стати	щіть	стати				свистати	свищіть		175
176		скати	щіть	скати				плескати	плещіть		176
177		слати	шліть	слати				послати	пошліть		177
178		кати	чте	[^ста]кати				плакати	плачте		178
179		ати	те	ипати				сипати	сипте		179
180			іть	спати				спати	спіть		180
181				орати				орати	оріть		181
182		рати	еріть	[бдп]рати				брати	беріть		182
183		ити	жу	[^з]дити	Т	-	-ити не повинно бути в корені - інші в гр. /І - пити, бити Т недоконаної форми, МБ доконаної	входити	входжу	Я	183
184		здити	жджу	здити				їздити	їжджу		184
185		зити	жу	зити				возити	вожу		185
186		ити	у	[жчшщ]ити				бентежити	бентежу		186
187		сити	шу	сити				місити	мішу		187
188		тити	чу	[^с]тити				тратити	трачу		188
189		стити	щу	стити				мостити	мощу		189
190		ити	лю	[бвмпф]ити				вимовити	вимовлю		190
191			ю	[лнр]ити				творити	творю		191
192		ти	ш	ити				бентежити	бентежиш	Ти	192
193		и	ь						бентежить	Він	193
194		ти	мо						бентежимо	Ми	194
195		ити	ите						бентежите	Ви	195
196			ать	[жчшщ]ити					бентежать	Вони	196
197			лять	[бвмпф]ити				вимовити	вимовлять		197
198			ять	[дзлнрст]ити				входити	входять		198
199			ив	ити	МН			бентежити	бентежив	Я, Ти, Він	199
200		іти	жу	діти	Т		-іти Т недоконаної форми, МБ доконаної	смердіти	смерджу	Я	200
201			у	[шж]іти				кишіти	кишу		201
202		сіти	шу	сіти				висіти	вишу		202
203		тіти	чу	[^с]тіти				летіти	лечу		203
204		стіти	щу	стіти				шелестіти	шелещу		204
205		іти	лю	[бвмп]іти				шуміти	шумлю		205
206			ю	[нлр]іти					велю		206
207			иш	іти					велиш	Ти	207
208				отіти				шепотіти	шепотиш		208
209				[^о]тіти				шелестіти	шелестиш		209
210				[^т]іти				смердіти	смердиш		210
211			ить	іти				веліти	велить	Він	211
212				отіти				шепотіти	шепотить		212
213				[^о]тіти				шелестіти	шелестить		213
214				[^т]іти				смердіти	смердить		214
215			имо	іти				веліти	велимо	Ми	215
216				отіти				шепотіти	шепотимо		216
217				[^о]тіти				шелестіти	шелестимо		217
218				[^т]іти				смердіти	смердимо		218
219			ите	іти				веліти	велите	Ви	219
220				отіти				шепотіти	шепотите		220
221				[^о]тіти				шелестіти	шелестите		221
222				[^т]іти				смердіти	смердите		222
223				отіти				шепотіти	шепотять	Вони	223
224				[^о]тіти				шелестіти	шелестять		224
225			ать	шіти				кишіти	кишать		225
226			лять	[бвмп]іти				шуміти	шумлять		226
227			ять	[^бвмпш]іти				входити	входять		227
228			в	ти	МН			смердіти	смердів	Я, Ти, Він	228
229			и	[^д]іти		-іти	-діти визначається в гр C/D та E/F	скрипіти	скрипи	Ти	229
230			імо						скрипімо	Ми	230
231			іть						скрипіть	Ви	231
232		ути	у	нути	Т	-	-нути Т недоконаної форми, МБ доконаної	тягнути	тягну	Я	232
233			еш						тягнеш	Ти	233
234			е						тягне	Він	234
235			емо						тягнемо	Ми	235
236			ете						тягнете	Ви	236
237			уть						тягнуть	Вони	237
238			в		МН				тягнув	Я, Ти, Він	238
239		ти	ду	бути	Т		-бути Т недоконаної форми, МБ доконаної	забути	забуду	Я	239
240			деш						забудеш	Ти	240
241			де						забуде	Він	241
242			демо						забудемо	Ми	242
243			дете						забудете	Ви	243
244			дуть						забудуть	Вони	244
245			в		МН				забув	Я, Ти, Він	245
246			дь			-бути			забудь	Ти	246
247			дьмо						забудьмо	Ми	247
248			дьте						забудьте	Ви	248
249		оти	ю	оти	Т	-	-оти Т недоконаної форми, МБ доконаної	бороти	борю	Я	249
250			еш						бореш	Ти	250
251			е						боре	Він	251

№	Клас	F1	F2	RE	Час	НФ	Ознака	Приклад 1	Приклад 2	Займенник	№
252			емо						боремо	Ми	252
253			ете						борете	Ви	253
254			ють						борють	Вони	254
255			в		МН				боров	Я, Ти, Він	255
256			и			-оти			бори	Ти	256
257			імо						борімо	Ми	257
258			іть						боріть	Ви	258
259		їти	ю	їти	Т	-	-їти Т недоконаної форми, МБ доконаної	клеїти	клею	Я	259
260		ти	ш						клеїш	Ти	260
261			ть						клеїть	Він	261
262			мо						клеїмо	Ми	262
263			те						клеїте	Ви	263
264		їти	ять						клеять	Вони	264
265			в		МН				клеїв	Я, Ти, Він	265
266		ти	у	[збв]ти	Т		зти, -бти, -вти Т недоконаної форми, МБ доконаної	везти	везу	Я	266
267			еш						везеш	Ти	267
268			е						везе	Він	268
269			емо						веземо	Ми	269
270			ете						везете	Ви	270
271			уть						везуть	Вони	271
272		ебти	іб	ебти	МН			гребти	гріб	Я, Ти, Він	272
273		езти	із	езти				везти	віз		273
274		зти	з	[^е]зти				гризти	гриз		274
275		ти	ів	евти				ревти	ревів		275
276		вти	в	[^е]вти				пливти	плив		276
277		бти	б	убти				скубти	скуб		277
278		ти	ь	ізти	-	-ти		лізти	лізь	Ти	278
279			и	[^і]зти				везти	вези		279
280				[бв]ти				гребти	греби		280
281			ьмо	ізти				лізти	лізьмо	Ми	281
282			імо	[^і]зти				везти	везімо		282
283				[бв]ти				гребти	гребімо		283
284			ьте	ізти				лізти	лізьте	Ви	284
285			іть	[^і]зти				везти	везіть		285
286				[бв]ти				гребти	гребіть		286
287		сти	ла	[^о]сти	МН	-	-сти (також в гр. І,К, М) викл.: пасти-пасла, нести-несла, трясти-трясла - в гр. /І	плести	плела	Вона	287
288			ло						плело	Воно	288
289			ли						плели	Вони	289
290			сла	ости				рости	росла	Вона	290
291			сло						росло	Воно	291
292			сли						росли	Вони	292
293		ти	ту		Т		-ости Т недоконаної форми, МБ доконаної		росту	Я	293
294			теш						ростеш	Ти	294
295			те						росте	Він	295
296			темо						ростемо	Ми	296
297			тете						ростете	Ви	297
298			туть						ростуть	Вони	298
299		сти	ту	[еі]сти			-цвісти, -[лм]ести	цвісти	цвіту	Я	299
300			теш						цвітеш	Ти	300
301			те						цвіте	Він	301
302			темо						цвітемо	Ми	302
303			тете						цвітете	Ви	303
304			туть						цвітуть	Вони	304
305			ну	ляcти			-ляcти	кляcти	кляну	Я	305
306			неш						клянеш	Ти	306
307			не						кляне	Він	307
308			немо						клянемо	Ми	308
309			нете						клянете	Ви	309
310			нуть						клянуть	Вони	310
311			в	[ія]сти	МН		-вати	прясти	пряв	Я, Ти, Він	311
312		есть	ів	есть				плести	плів		312
313		ости	іс	ости				рости	ріс		313
314		сти	ти	[еі]сти		-вати		плести	плети	Ти	314
315			тімо						плетімо	Ми	315
316			тіть						плетіть	Ви	316
317			ни	ляcти				кляcти	кляни	Ти	317
318			німо						клянімо	Ми	318
319			ніть						кляніть	Ви	319
320		ости	ости	ости				рости	рости	Ти	320
321		и	імо						ростімо	Ми	321
322			іть						ростіть	Ви	322
323		кти	чу	кти	Т	-	-кти Т недоконаної форми, МБ доконаної	текти	течу	Я	323
324			чеш						течеш	Ти	324
325			че						тече	Він	325
326			чемо						течемо	Ми	326
327			чете						течете	Ви	327
328			чуть						течуть	Вони	328
329		екти	ік	екти	МН				тік	Я, Ти, Він	329
330		ікти		ікти				одсікти	одсік		330
331		окти		окти				волокти	волік		331
332		вкти		вкти				товкти	товк		332
333		кти	чи	кти		-вати		текти	течи	Ти	333
334			чімо						течімо	Ми	334

№	Клас	F1	F2	RE	Час	НФ	Ознака	Приклад 1	Приклад 2	Займенник	№
335			чіть						течіть	Ви	335
336		гти	жу	[еоиія]гти	Т	-	-гти з черг. г/ж (без - в гр. М) Т недоконаної форми, МБ доконаної	допомогти	допоможу	Я	336
337			жиш	ігти				бігти	біжиш	Ти	337
338			жеш	[еоия]гти				допомогти	допоможеш		338
339			жить	ігти				бігти	біжить	Він	339
340			же	[еоия]гти				допомогти	допоможе		340
341			жимо	ігти				бігти	біжимо	Ми	341
342			жемо	[еоия]гти				допомогти	допоможемо		342
343			жите	ігти				бігти	біжите	Ви	343
344			жете	[еоия]гти				допомогти	допоможете		344
345			жать	ігти				бігти	біжать	Вони	345
346			жуть	[еоия]гти				допомогти	допоможуть		346
347		егти	іг	егти	МН			зберегти	зберіг	Я, Ти, Він	347
348		огти		огти				допомогти	допоміг		348
349		ягти		ягти				лягти	ліг		349
350		гти	г	[иі]гти				стригти	стриг		350
351			ж	лягти		-гти		лягти	ляж	Ти	351
352			жи	рягти				запрягти	запряжи		352
353				[еоіи]гти				допомогти	допоможи		353
354			жмо	лягти				лягти	ляжмо	Ми	354
355			жімо	рягти				запрягти	запряжімо		355
356				[еоіи]гти				допомогти	допоможімо		356
357			жте	лягти				лягти	ляжте	Ви	357
358			жіть	рягти				запрягти	запряжіть		358
359				[еоіи]гти				допомогти	допоможіть		359
360		ерти	ру	[^дж]ерти	Т	-	-ерти Т недоконаної форми, МБ доконаної обтерти, розтерти, стерти - обітри... (поява "і" між приставкою і коренем)	терти	тру	Я	360
361			реш						треш	Ти	361
362			ре						тре	Він	362
363			ремо						тремо	Ми	363
364			рете						трете	Ви	364
365			руть						труть	Вони	365
366		рти	р	рти	МН		-рти		тер	Я, Ти, Він	366
367		ерти	ри	[^дж]ерти		-ерти			три	Ти	367
368			рімо						трімо	Ми	368
369			ріть						тріть	Ви	369
370		рти	ру	[дж]ерти	Т	-	-ерти	жерти	жеру	Я	370
371			реш						жереш	Ти	371
372			ре						жере	Він	372
373			ремо						жеремо	Ми	373
374			рете						жерете	Ви	374
375			руть						жеруть	Вони	375
376		ти	и		МН	-ерти	-рти		жери	Ти	376
378			імо						жерімо	Ми	378
379			іть						жеріть	Ви	379
380			ю	[аі]яти	Т	-	-ти викл. з групи /І: паяти, сіяти-сіяю (ІІ)	паяти	паяю	Я	380
381			єш						паяєш	Ти	381
382			є						паяє	Він	382
383			ємо						паяємо	Ми	383
384			єте						паяєте	Ви	384
385			ють						паяють	Вони	385
386			в		МН		-ивати		паяв	Я, Ти, Він	386
387			й			ивати			паяй	Ти	387
388			ймо						паяймо	Ми	388
389			йте						паяйте	Ви	389
390		зяти	ізьму	взяти	Т	-	доконаної форми з чергуванням узяти - візьму, *взяти - *візьму	взяти	візьму	Я	390
391			ізьмеш						візьмеш	Ти	391
392			ізьме						візьме	Він	392
393			ізьмемо						візьмемо	Ми	393
394			ізьмете						візьмете	Ви	394
395			ізьмуть						візьмуть	Вони	395
396			в		МН		стяти, відтяти – тільки минулий час		взяв	Я, Ти, Він	396
397			ізьми			зяти			візьми	Ти	397
398			ізьмімо						візьмімо	Ми	398
399			ізьміть						візьміть	Ви	399
400		няти	му	йняти	Т	-	-няти, -'яти зняти, підняти, заняти (-няти), зім'яти, нам'яти (-'яти)...	зайняти	займу	Я	400
401			меш						займеш	Ти	401
402			ме						займе	Він	402
403			мемо						займемо	Ми	403
404			мете						займете	Ви	404
405			муть						займуть	Вони	405
406			ми		МН	ивати			займи	Ти	406
407			мімо						займімо	Ми	407
408			міть						займіть	Ви	408
409			йму	[аеиіїоу юя]няти	Т	-			займу	Я	409
410			ймеш						займеш	Ти	410
411			йме						займе	Він	411
412			ймемо						займемо	Ми	412
413			ймете						займете	Ви	413
414			ймуть						займуть	Вони	414
415			йми		МН	няти			займи	Ти	415
416			ймімо						займімо	Ми	416
417			йміть						займіть	Ви	417
418		яти	іму	[здб]няти	Т	-	чергування я/і в корені		підніму	Я	418

№	Клас	F1	F2	RE	Час	НФ	Ознака	Приклад 1	Приклад 2	Займенник	№
419			імеш						піднімеш	Ти	419
420			іме						підніме	Він	420
421			імемо						піднімемо	Ми	421
422			імете						піднімете	Ви	422
423			імуть						піднімуть	Вони	423
424			іми		МН	яти			підніми	Ти	424
425			імімо						піднімімо	Ми	425
426			іміть						підніміть	Ви	426
427		'яти	ну	'яти	Т	-		зім'яти	зімну	Я	427
428			неш						зімнеш	Ти	428
429			не						зімне	Він	429
430			немо						зімнемо	Ми	430
431			нете						зімнете	Ви	431
432			нуть						зімнуть	Вони	432
433			ни		МН	'яти			зімни	Ти	433
434			німо						зімнімо	Ми	434
435			ніть						зімніть	Ви	435
1	С	ити	-	[вжчшщб мпр]ити		-, -ь, -мо, -те	НФ на у "графити", "проштрафити" НФ немає, тому без -фити	бентежити	бентеж	Ти	436
2			мо						бентежмо	Ми	437
3			те						бентежте	Ви	438
4		ити	ь	[дтзснл]ити				заходити	заходь	Ти	439
5			ьмо						заходьмо	Ми	440
6			ьте					проводити	проводьте	Ви	441
7		іти	ь	діти				посидіти	посидь	Ти	442
8			ьмо						посидьмо	Ми	443
9			ьте						посидьте	Ви	444
10		ути	ь	нути				кинути	кинь	Ти	445
11			ьмо						киньмо	Ми	446
12			ьте						киньте		447
13		їти	й	їти				клеїти	клей	Ти	448
14			ймо						клеймо	Ми	449
15			йте						клейте	Ви	450
1	Е	ити	и	ити		-и, -імо, -іть	НФ на -и, -імо, -іть	учити	учи	Ти	451
2			імо						учімо	Ми	452
3			іть						учіть	Ви	453
4		іти	и	діти				сидіти	сиди	Ти	454
5			імо						сидімо	Ми	455
6			іть						сидіть	Ви	456
7		ути	и	нути				кашлянути	кашляни	Ти	457
8			імо						кашлянімо	Ми	458
9			іть						кашляніть	Ви	459
10		їти	ї	їти				напоїти	напої	Ти	460
11			їмо						напоїмо	Ми	461
12			їть						напоїть	Ви	462
1	G	-	му	ти	МБ	-	майбутній час для недоконаної форми -вати (МБ недоконаної, відсутня в доконаній)	абонувати	абонуватиму	Я	463
2			меш						абонуватимеш	Ти	464
3			ме						абонуватиме	Він	465
4			мемо						абонуватимемо	Ми	466
5			мете						абонуватимете	Ви	467
6			муть						абонуватимуть	Вони	468
1	I	ти	ла	ти	МН		(окрім Я, Ти, Він) для -ти зворотня форма - група /J доконана форма = /IG та доконана зворотня = /JH	вбивати	вбивала	Вона	469
2			ло						вбивало	Воно	470
3			ли						вбивали	Вони	471
4			в	[аиіуя]ти					вбивав	Я, Ти, Він	472
5		яти	ю	[аяіо]яти	Т		-ати, -яти, -іти (як -аю, аєш...-яю,-яєш...) Т недоконаної форми, МБ доконаної + З на -ути: окути, дути, чути, слова: лаяти, сіяти (II) - сіяю, на -няти, -'яти, які мають тільки Минулий Час - у /А	віяти	вію	Я	473
6		ти		[илнрц]яти				ганяти	ганяю		474
7				[аіу]ти				вбивати	вбиваю		475
8		яти	їш	ояти				стояти	стоїш	Ти	476
9			єш	[аяі]яти				віяти	вієш		477
10		ти		[илнрц]яти				ганяти	ганяєш		478
11				[аіу]ти				вбивати	вбиваєш		479
12		яти	їть	ояти				стояти	стоїть	Він	480
13			є	[аяі]яти				віяти	віє		481
14		ти		[илнрц]яти				ганяти	ганяє		482
15				[аіу]ти				вбивати	вбиває		483
16		яти	їмо	ояти				стояти	стоїмо	Ми	484
17			ємо	[аяі]яти				віяти	віємо		485
18		ти		[илнрц]яти				ганяти	ганяємо		486
19				[аіу]ти				вбивати	вбиваємо		487
20		яти	їте	ояти				стояти	стоїте	Ви	488
21			єте	[аяі]яти				віяти	вієте		489
22		ти		[илнрц]яти				ганяти	ганяєте		490
23				[аіу]ти				вбивати	вбиваєте		491
24		яти	ять	ояти				стояти	стоять	Вони	492
25			ють	[аяі]яти				віяти	віють		493
26		ти		[илнрц]яти				ганяти	ганяють		494
27				[аіу]ти				вбивати	вбивають		495
28		ояти	ій	ояти	МН	ивати	-ивати	стояти	стій	Ти	496
29		яти	й	[аяі]яти				віяти	вій		497
30		ти		[илнрц]яти				ганяти	ганяй		498
31				[аіу]ти				вбивати	вбивай		499
32		ояти	іймо	ояти				стояти	стіймо	Ми	500
33		яти	ймо	[аяі]яти				віяти	віймо		501

№	Клас	F1	F2	RE	Час	НФ	Ознака	Приклад 1	Приклад 2	Займенник	№
34		ти		[илнрц]яти				ганяти	ганяймо		502
35				[аіу]ти				вбивати	вбиваймо		503
36		ояти	ійте	ояти				стояти	стійте	Ви	504
37		яти	йте	[аяі]яти				віяти	війте		505
38		ти	йте	[илнрц]яти				ганяти	ганяйте		506
39				[аіу]ти				вбивати	вбивайте		507
40		ити	'ю	[бвп]ити	Т	-	короткі слова на -ити як корень слова (бити, пити, вити-в'ю (І), вити-вию (ІІ), жити, рити, ...)	бити	б'ю	Я	508
41			'єш						б'єш	Ти	509
42			'є						б'є	Він	510
43			'ємо						б'ємо	Ми	511
44			'єте						б'єте	Ви	512
45			'ють						б'ють	Вони	513
46			ию	[врмнш]ити				рити	рию	Я	514
47			иєш						риєш	Ти	515
48			иє						риє	Він	516
49			иємо						риємо	Ми	517
50			иєте						риєте	Ви	518
51			иють						риють	Вони	519
52			лю	лити				лити	ллю	Я	520
53			лєш						ллєш	Ти	521
54			лє						ллє	Він	522
55			лємо						ллємо	Ми	523
56			лєте						ллєте	Ви	524
57			лють						ллють	Вони	525
58		ти	ву	жити				жити	живу	Я	526
59			веш						живеш	Ти	527
60			ве						живе	Він	528
61			вемо						живемо	Ми	529
62			вете						живете	Ви	530
63			вуть						живуть	Вони	531
64			в		МН			бити	бив	Я, Ти, Він	532
65			й	[^ж]ити		-ити			бий	Ти	533
66			ймо						биймо	Ми	534
67			йте						бийте	Ви	535
68			ви					жити	живи	Ти	536
69			вімо						живімо	Ми	537
70			вість						живіть	Ви	538
71			у	сти	Т	-	нести-несла, пасти-пасла, трясти-трясла (впасти-впала - у гр. /А)	пасти	пасу	Я	539
72			еш						пасеш	Ти	540
73			е						пасе	Він	541
74			емо						пасемо	Ми	542
75			ете						пасете	Ви	543
76			уть						пасуть	Вони	544
77		сти	с	[ая]сти	МН				пас	Я, Ти, Він	545
78		єсти	іс	єсти				нести	ніс		546
79		ти	и	сти		ти		пасти	паси	Ти	547
80			імо						пасімо	Ми	548
81			іть						пасіте	Ви	549
1	К		ла	[^ус]ти		-	для всіх закінчень викл.: з чергув., тощо зворотня L, доконана KG, доконана зворотня LH	гнати	гнала	Вона	550
2			ло						гнало	Воно	551
3			ли						гнали	Вони	552
4			в	[аеиіоя]ти				зігнати	зігнав	Я, Ти, Він	553
5		нути	ла	нути			укорочений минулий	змерзнути	змерзла	Вона	554
6			ло						змерзло	Воно	555
7			ли						змерзли	Вони	556
8			-						змерз	Я, Ти, Він	557
9			нула						змерзнула	Вона	558
10			нуло						змерзнуло	Воно	559
11			нули						змерзнули	Вони	560
12			нув						змерзнув	Я, Ти, Він	561
13		ути	у		МБ		укорочений майбутній		змерзну	Я	562
14			еш						змерзнеш	Ти	563
15			е						змерзне	Він	564
16			емо						змерзнемо	Ми	565
17			ете						змерзнете	Ви	566
18			уть						змерзнуть	Вони	567
19		ти	ну	чити				відпочити	відпочину	Я	568
20			неш						відпочинеш	Ти	569
21			не						відпочине	Він	570
22			немо						відпочинемо	Ми	571
23			нете						відпочинете	Ви	572
24			нуть						відпочинуть	Вони	573
25		ігнати	жену	ігнати	Т		зігнати – зжену увігнати, ввігнати – аналоги вгнати, угнати	відігнати	віджену	Я	574
26			женеш						відженеш	Ти	575
27			жене						віджене	Він	576
28			женемо						відженемо	Ми	577
29			женете						відженете	Ви	578
30			женуть						відженуть	Вони	579
31			в		МН			зігнати	зігнав	Я, Ти, Він	580
32			жени			ігнати		відігнати	віджени	Ти	581
33			женімо						відженімо	Ми	582
34			женіть						відженіть	Ви	583
35		іпрати	перу	іпрати	Т	-	*іпрати - *перу	відіпрати	відперу	Я	584

№	Клас	F1	F2	RE	Час	НФ	Ознака	Приклад 1	Приклад 2	Займенник	№
36			переш						відпереш	Ти	585
37			пере						відпере	Він	586
38			перемо						відперемо	Ми	587
39			перете						відперете	Ви	588
40			перуть						відперуть	Вони	589
41			в		МН				відправ	Я, Ти, Він	590
42			пери			іпрати			відпери	Ти	591
43			перімо						відперімо	Ми	592
44			періть						відперіть	Ви	593
45		ібрати	беру	ібрати	Т	-	*ібрати - *беру	відібрати	відберу	Я	594
46			береш						відбереш	Ти	595
47			бере						відбере	Він	596
48			беремо						відберемо	Ми	597
49			берете						відберете	Ви	598
50			беруть						відберуть	Вони	599
51			в		МН			брати	брав	Я, Ти, Він	600
52			бери			ібрати		відібрати	відбери	Ти	601
53			берімо						відберімо	Ми	602
54			беріть						відберіть	Ви	603
55		ідрати	деру	ідрати	Т	-	*ідрати - *деру	відідрати	віддеру	Я	604
56			дереш						віддереш	Ти	605
57			дере						віддере	Він	606
58			деремо						віддеремо	Ми	607
59			дерете						віддерете	Ви	608
60			деруть						віддеруть	Вони	609
61			в		МН			брати	драв	Я, Ти, Він	610
62			дери			ідрати		відідрати	віддери	Ти	611
63			дерімо						віддерімо	Ми	612
64			деріть						віддеріть	Ви	613
65		бити	іб'ю	[бдз]бити	Т	-	короткі слова на -ити як корень слова (бити, пити, вити-в'ю (I), вити-вию (II), лити, жити, рити, ...) з'являється "і" у приставці	надбити	надіб'ю	Я	614
66			іб'єш						надіб'єш	Ти	615
67			іб'є						надіб'є	Він	617
68			іб'ємо						надіб'ємо	Ми	618
69			іб'єте						надіб'єте	Ви	619
70			іб'ють						надіб'ють	Вони	620
71		вити	ів'ю	[бдз]вити				обвити	обів'ю	Я	621
72			ів'єш						обів'єш	Ти	622
73			ів'є						обів'є	Він	623
74			ів'ємо						обів'ємо	Ми	624
75			ів'єте						обів'єте	Ви	625
76			ів'ють						обів'ють	Вони	626
77		пити	іп'ю	[бдз]пити				надпити	надіп'ю	Я	627
78			іп'єш						надіп'єш	Ти	628
79			іп'є						надіп'є	Він	629
80			іп'ємо						надіп'ємо	Ми	630
81			іп'єте						надіп'єте	Ви	631
82			іп'ють						надіп'ють	Вони	632
83		лити	іллю	[бвдз]лити				надлити	наділлю	Я	633
84			іллєш						наділлєш	Ти	634
85			іллє						наділлє	Він	635
86			іллємо						наділлємо	Ми	636
87			іллєте						наділлєте	Ви	637
88			іллють						наділлють	Вони	638
89			в		МН				надлив	Я, Ти, Він	639
90		ти	й	[бвлп]ити		-ити		надбити	надбий	Ти	640
91			ймо						надбиймо	Ми	641
92			йте						надбийте	Ви	642
93			м	дати	Т	-	*дати - *дам	дати	дам	Я	643
94			си						даси	Ти	644
95			сть						дасть	Він	645
96			мо						дамо	Ми	646
97			сте						дасте	Ви	647
98			дуть						дадуть	Вони	648
99			в		МН				дав	Я, Ти, Він	649
100			й			ти			дай	Ти	650
101			ймо						даймо	Ми	651
102			йте						дайте	Ви	652
103			в		Т	-	брати - драв	брати	драв	Я, Ти, Він	653
104		вати	ю	[тд]авати			*давати - *даю	давати	даю	Я	654
105			єш						даєш	Ти	655
106			є						дає	Він	656
107			ємо						даємо	Ми	657
108			єте						даєте	Ви	658
109			ють						дають	Вони	659
110			в		МН				давав	Я, Ти, Він	660
111			вай			вати			давай	Ти	661
112			ваймо						давайме	Ми	662
113			вайте						давайте	Ви	663
114		ати	му	жати	Т	-	жати - жму (викл: жати-жну), -звати Т недоконаної форми, МБ доконаної + "зову, зовеш.."	жати	жму	Я	664
115			меш						жмеш	Ти	665
116			ме						жме	Він	666
117			мемо						жмемо	Ми	667
118			мете						жмете	Ви	668

№	Клас	F1	F2	RE	Час	НФ	Ознака	Приклад 1	Приклад 2	Займенник	№
119			муть				до гр. А - "зву, звеш..."		жмуть	Вони	669
120			в		МН		-нути		жав	Я, Ти, Він	670
121			ми			нути			жми	Ти	671
122			мімо						жмімо	Ми	672
123			міть						жміть	Ви	673
124		ти	ну	діти	Т	-	*стати - *стану	подіти	подіну	Я	674
125			неш						подінеш	Ти	675
126			не						подіне	Він	676
127			немо						подінемо	Ми	677
128			нете						подінете	Ви	678
129			нуть						подінуть	Вони	679
130			в		МН				подів	Я, Ти, Він	680
131			нь			ти		одіти	одінь	Ти	681
132			ньмо						одіньмо	Ми	682
133			ньте						одіньте	Ви	683
134			ну	стати	Т	-	*діти - *діну	постати	постану	Я	684
135			неш						постанеш	Ти	685
136			не						постане	Він	686
137			немо						постанемо	Ми	687
138			нете						постанете	Ви	688
139			нуть						постануть	Вони	689
140			в		МН				постав	Я, Ти, Він	690
141			нь			ти			постань	Ти	691
142			ньмо						постаньмо	Ми	692
143			ньте						постаньте	Ви	693
144		олоти	елю	олоти	Т	-	*молоти - *мелю	молоти	мелю	Я	694
145			елеш						мелеш	Ти	695
146			еле						меле	Він	696
147			елемо						мелемо	Ми	697
148			елете						мелете	Ви	698
149			елють						мелють	Вони	699
150			в		МН				молов	Я, Ти, Він	700
151			ели			олоти			мели	Ти	701
152			елімо						мелімо	Ми	702
153			еліть						меліть	Ви	703
154		істи	яду	істи	Т	-	-сісти	сісти	сяду	Я	704
155			ядеш						сядеш	Ти	705
156			яде						сяде	Він	706
157			ядемо						сядемо	Ми	707
158			ядете						сядете	Ви	708
159			ядуть						сядуть	Вони	709
160		сти	ла		МН				сіла	Вона	710
161			ло						сіло	Воно	711
162			ли						сіли	Вони	712
163			в						сів	Я, Ти, Він	713
164		істи	ядь			істи			сядь	Ти	714
165			ядьмо						сядьмо	Ми	715
166			ядьте						сядьте	Ви	716
167		ти	в	тріти		-	стріти - *стрів	стріти	стрів	Я, Ти, Він	717
1	М		ла	[^йіс]ти			складні слова –	ржати	ржала	Вона	718
2			ло				виключення		ржало	Воно	719
3			ли				окрім Я, Ти, Він		ржали	Вони	720
4			в				для всіх закінчень		ржав	Я, Ти, Він	721
5		ати	у	ржати	Т		тільки "(за)(і)ржати"		ржу	Я	722
6			еш				інші в гр. А		ржеш	Ти	723
7			е						рже	Він	724
8			емо						ржемо	Ми	725
9			ете						ржете	Ви	726
10			уть						ржуть	Вони	727
11			в		МН				ржав	Я, Ти, Він	728
12			и			ати			ржи	Ти	729
13			імо						ржімо	Ми	730
14			іть						ржіть	Ви	731
15		жати	іжму	[^р]жати	Т	-	-жати (від)жати	жати	жму	Я	732
16			іжмеш				- (від)іжму (жати - в гр. К)		жмеш	Ти	733
17			іжме						жме	Він	734
18			іжмемо						жмемо	Ми	735
19			іжмете						жмете	Ви	736
20			іжмуть						жмуть	Вони	737
21			іжми			жати			жми	Ти	738
22			іжмімо						жмімо	Ми	739
23			іжміть						жміть	Ви	740
24		ти	ду	[йі]ти		-	-йти	йти	йду	Я	741
25			деш				Т недоконаної форми,		йдеш	Ти	742
26			де				МБ доконаної		йде	Він	743
27			демо						йдемо	Ми	744
28			дете						йдете	Ви	745
29			дуть						йдуть	Вони	746
30			шов		МН				йшов	Я, Ти, Він	747
31			шла						йшла	Вона	748
32			шло						йшло	Воно	749
33			шли						йшли	Вони	750
34			ди			-йти			йди	Ти	751

№	Клас	F1	F2	RE	Час	НФ	Ознака	Приклад 1	Приклад 2	Займенник	№
35			дімо						йдімо	Ми	752
36			діть						йдіть	Ви	753
37		хати	ду	хати	Т	-	-їхати Т недоконаної форми, МБ доконаної	їхати	їду	Я	754
38			деш						їдеш	Ти	755
39			де						їде	Він	756
40			демо						їдемо	Ми	757
41			дете						їдете	Ви	758
42			дуть						їдуть	Вони	759
43			дь			їхати			їдь	Ти	760
44			дьмо						їдьмо	Ми	761
45			дьте						їдьте	Ви	762
46		гнати	жену	гнати		-	- [^і]гнати - *жену не працює для "гнати" де зникає "і" у приставці - у К	гнати	жену	Я	763
47			женеш						женеш	Ти	764
48			жене						жене	Він	765
49			женемо						женемо	Ми	766
50			женете						женете	Ви	767
51			женуть						женуть	Вони	768
52			в		МН				гнав	Я, Ти, Він	769
53			жени			гнати			жени	Ти	770
54			женімо						женімо	Ми	771
55			женіть						женіть	Ви	772
56		чити	в	чити		-	-почити, які тільки МН	відпочити	відпочив	Я, Ти, Він	773
57		ти	ну	[ия]гти	Т		-гти без черг. г/ж з ним в А Т недоконаної форми, МБ доконаної	одягти	одягну	Я	774
58			неш						одягнеш	Ти	775
59			не						одягне	Він	776
60			немо						одягнемо	Ми	777
61			нете						одягнете	Ви	778
62			нуть						одягнуть	Вони	779
63		гти	г		МН				одяг	Я, Ти, Він	780
64		ти	ни	ягти		-гти	тільки для -ягти, для "встигти" відсутня		одягни	Ти	781
65			німо						одягнімо	Ми	782
66			ніть						одягніть	Ви	783
67		сти	ду	[еая]сти	Т	-	-[вр]ести, -асти, -прясти	вести	веду	Я	784
68			деш						ведеш	Ти	785
69			де						веде	Він	786
70			демо						ведемо	Ми	787
71			дете						ведете	Ви	788
72			дуть						ведуть	Вони	789
73			ла	сти	МН		-їсти, -овісти	їсти	їла	Вона	790
74			ло						їло	Воно	791
75			ли						їли	Вони	792
76			в	[^е]сти					їв	Він	793
77			ів	ести					вів	Я, Ти, Він	794
78			ди	[еая]сти		сти	-[вр]ести, -асти, -прясти	вести	веди	Ти	795
79			дімо						ведімо	Ми	796
80			діть						ведіть	Ви	797
81			м	[ії]сти	Т	-	-їсти, -овісти	їсти	їм	Я	798
82		ти	и						їси	Ти	799
83			ть						їсть	Він	800
84			те						їсте	Ви	801
85		сти	мо						їмо	Ми	802
86			дять	їсти					їдять	Вони	803
87			ж						їж	Ти	804
88			жмо						їжмо	Ми	805
89			жте						їжте	Ви	806
90			нь	овісти			не плутати з відповідай	відповісти	-	Ти	807
91			ньмо							Ми	808
92			ньте							Ви	809
93			дять	[і]сти			відсутня форма 3ї особи мн			Вони	810

ДОДАТОК Б. РИСУНКИ

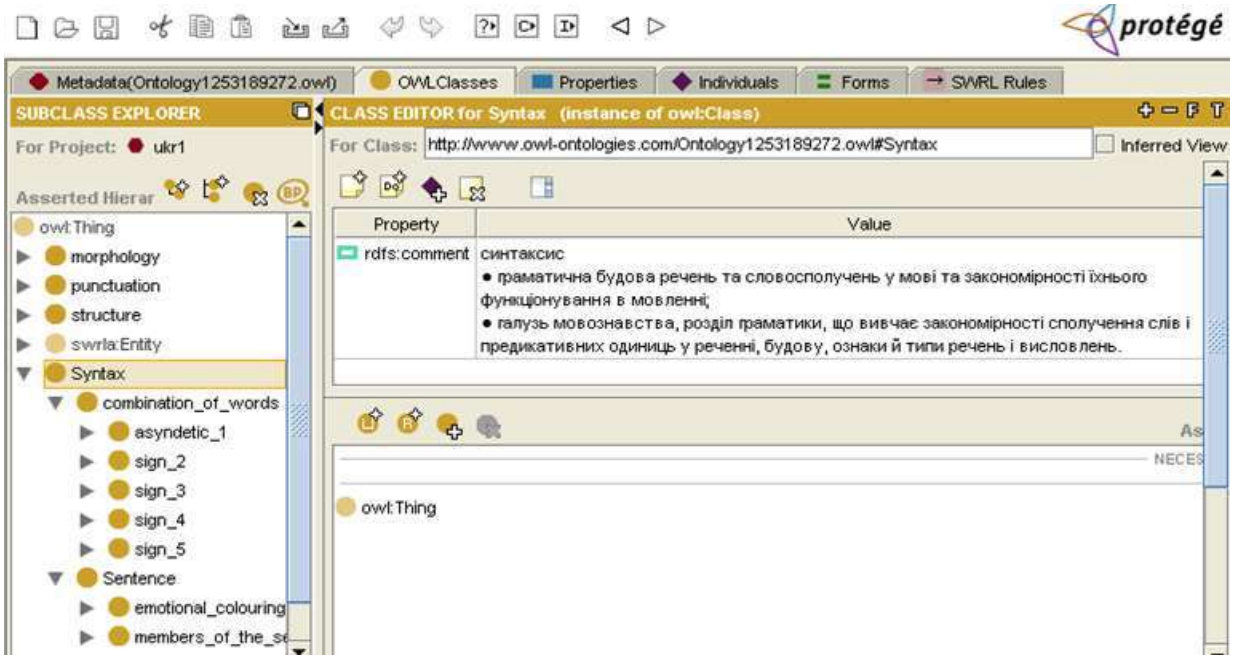


Рис. Б.1. Ієрархії класів в OWLClasses в Protégé 3.4.7

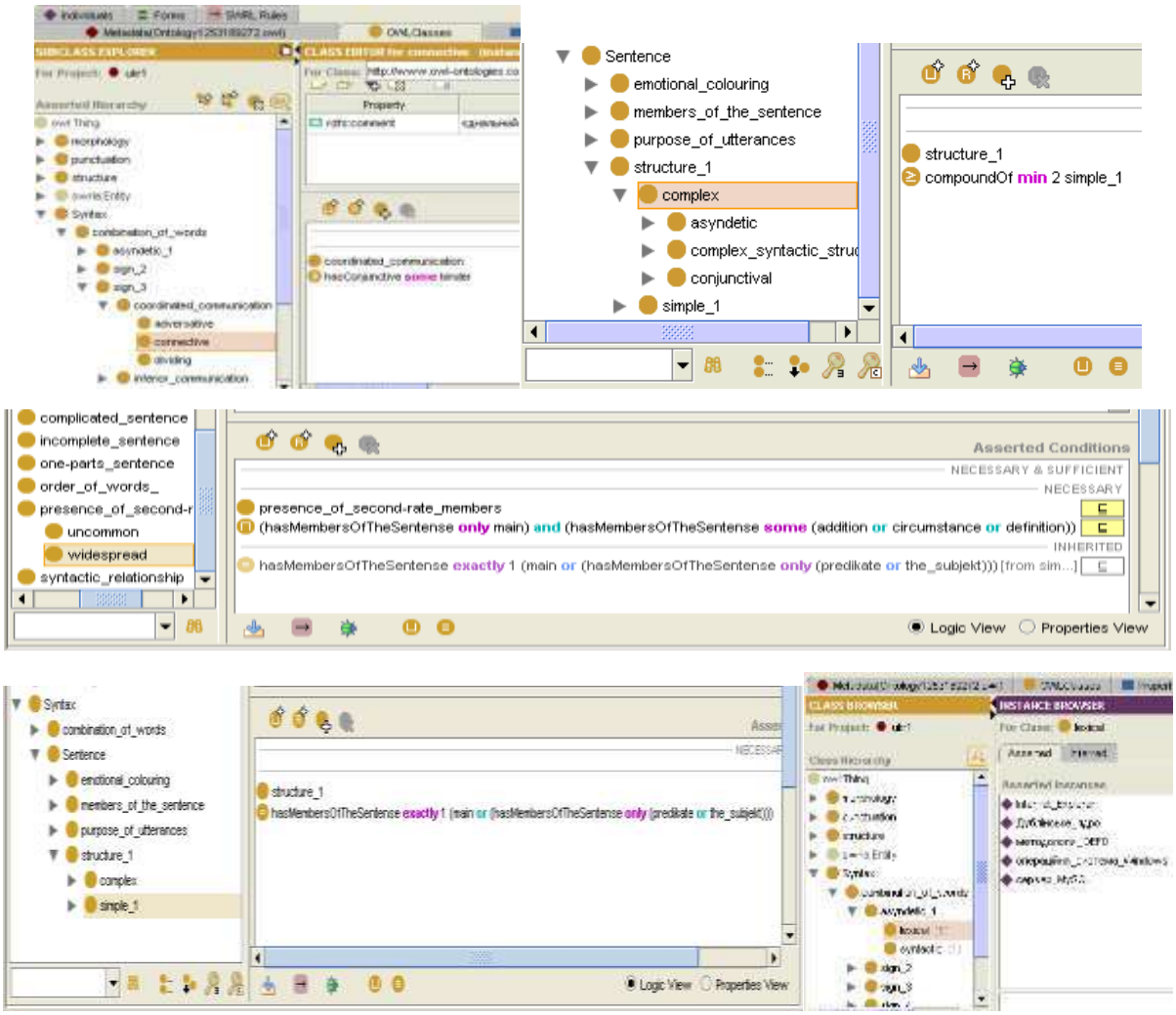


Рис. Б.2. Класи Connective, Complex, Simple_1, Widespread та Lexical

SWRL Rules

Enabled	Name	Expression
☑	Rule-1	main_word(?x) ∧ noun(?x) → name(?x)
☑	Rule-10	Sentence(?x) ∧ hasPunctuation(?x, ?y) ∧ question_mark(?y) → interrogative(?x)
☑	Rule-11	hasConjunctive(?x, ?y) ∧ binder(?x) → with_opposition_conjuctions(?x)
☑	Rule-12	hasConjunctive(?x, ?y) ∧ contrasting(?x) → with_confrontation_conjuctions(?x)
☑	Rule-13	hasConjunctive(?x, ?y) ∧ dividing_1(?x) → with_disjunctive_coordination(?x)
☑	Rule-2	main_word(?x) ∧ adjective_1(?x) → adjective(?x)
☑	Rule-3	main_word(?x) ∧ verb(?x) → verbal(?x)
☑	Rule-4	main_word(?x) ∧ dependent_word(?x) → inferior_communication(?x)
☑	Rule-5	main_word(?x) ∧ hasConjunctive(?x, ?y) ∧ main_word(?x) ∧ contrasting(?y) → adversative(?x
☑	Rule-6	main_word(?x) ∧ hasConjunctive(?x, ?y) ∧ main_word(?x) ∧ binder(?y) → connective(?x)
☑	Rule-7	main_word(?x) ∧ hasConjunctive(?x, ?y) ∧ main_word(?x) ∧ dividing_1(?y) → dividing(?x)
☑	Rule-8	Sentence(?x) ∧ hasPunctuation(?x, ?y) ∧ exclamation_point(?y) → imperative(?x)
☑	Rule-9	Sentence(?x) ∧ hasPunctuation(?x, ?y) ∧ point(?y) → narrative(?x)

Рис. Б.3. Основні правила БЗ онтології синтаксичного аналізу

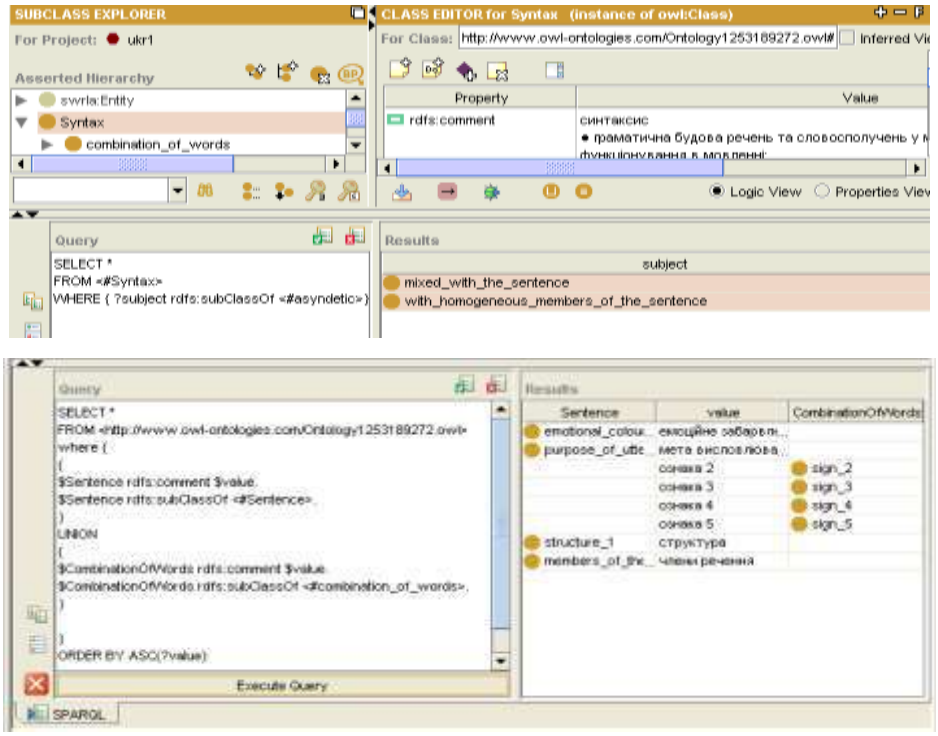


Рис. Б.4. Основні запити до БЗ онтології синтаксичного аналізу

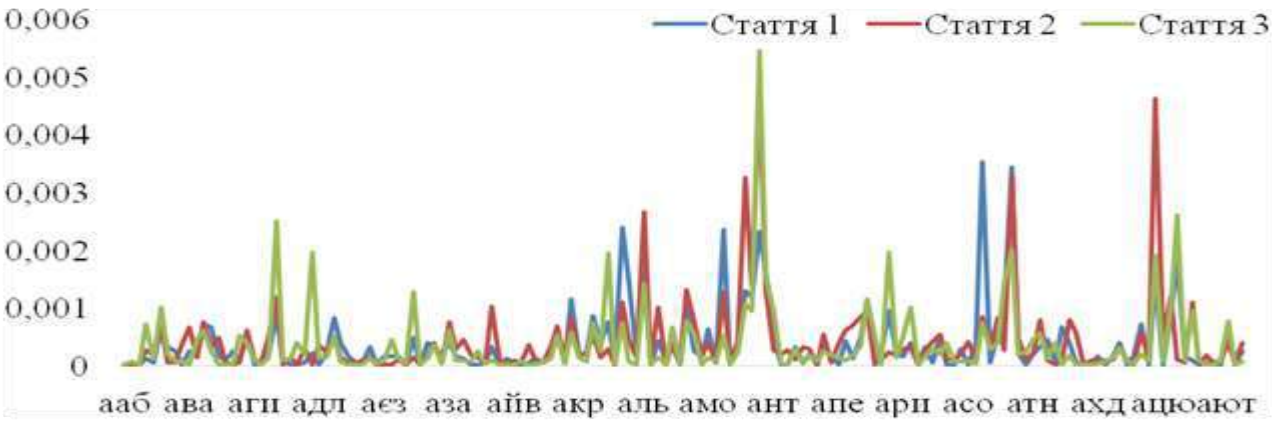


Рис. Б.5. Графік вживання 3-грам, які починаються з літери а

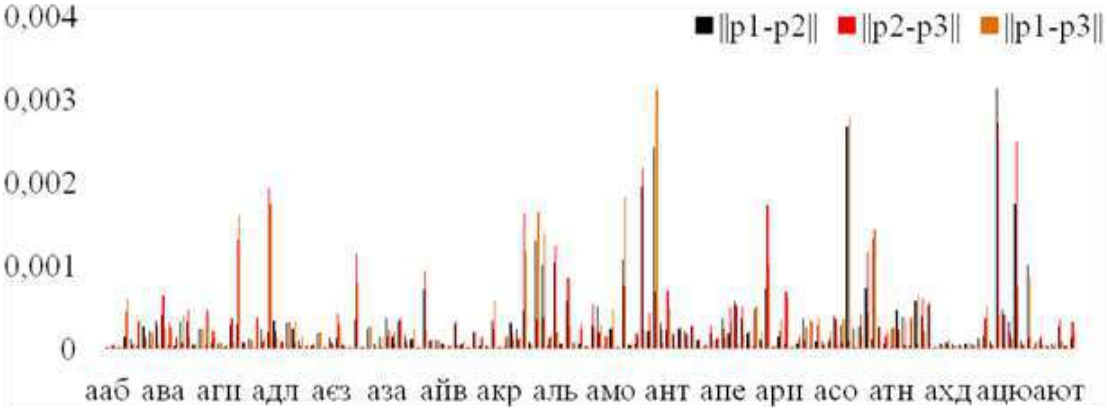


Рис. Б.6. Графік різниці вживання 3-грам, які починаються з літери а

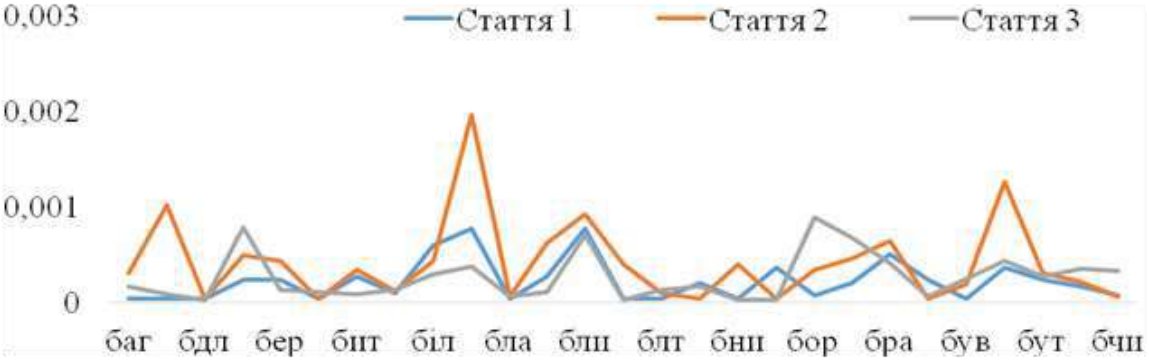


Рис. Б.7. Графік вживання 3-грам, які починаються з літери б

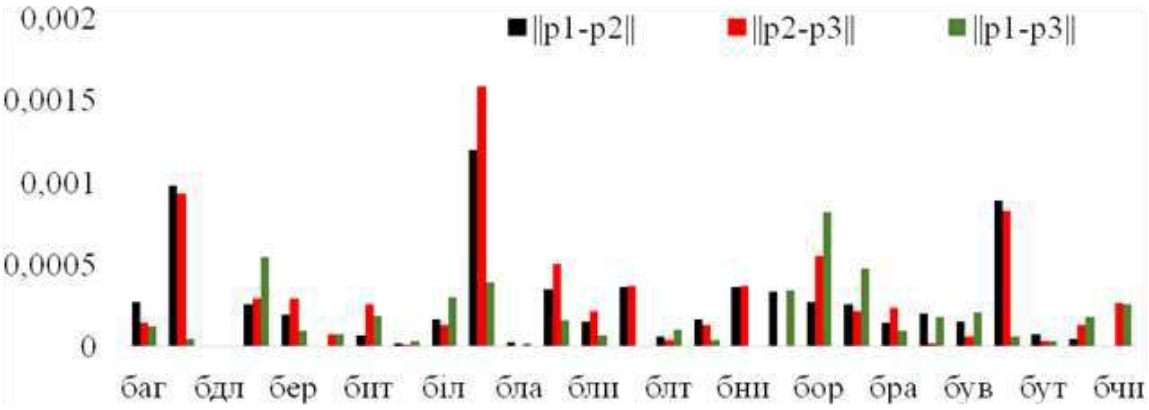


Рис. Б.8. Графік різниці вживання 3-грам, які починаються з літери б

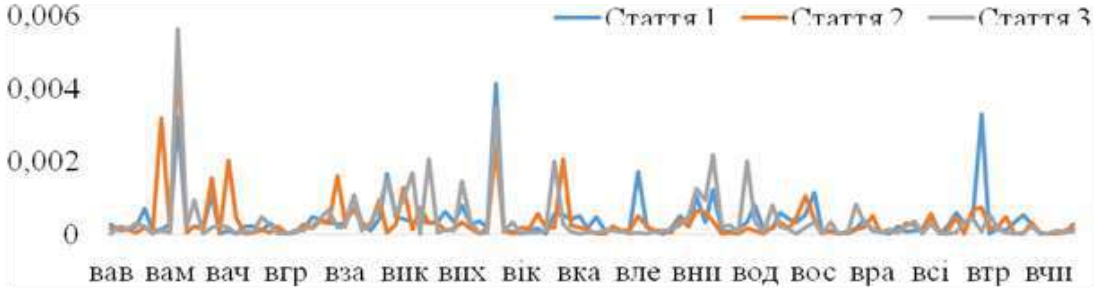


Рис. Б.9. Графік вживання 3-грам, які починаються з літери в

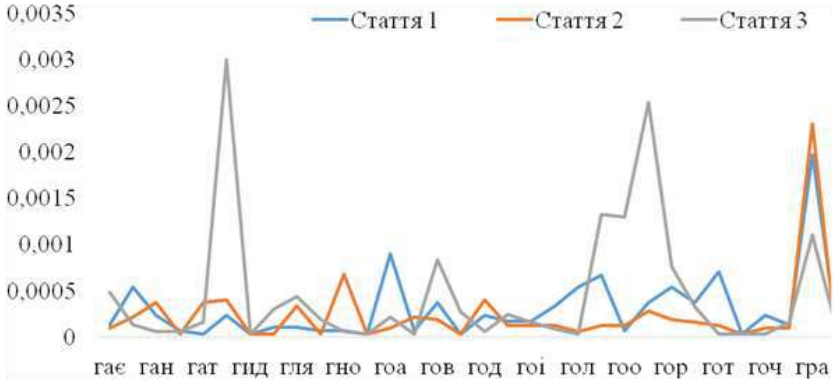


Рис. Б.10. Графік вживання 3-грам, які починаються з літери г

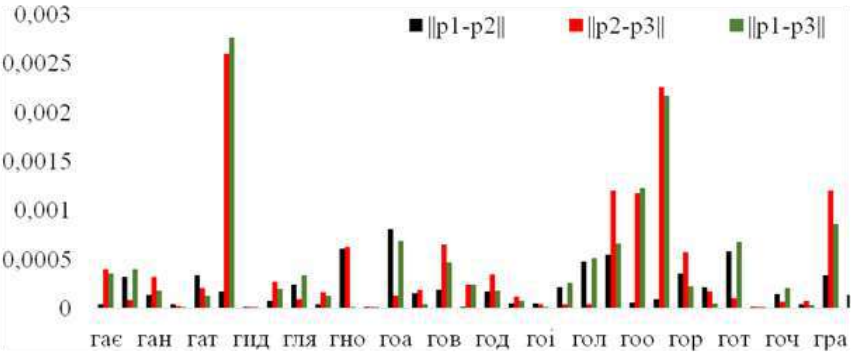


Рис. Б.11. Графік різниці вживання 3-грам, які починаються з літери г

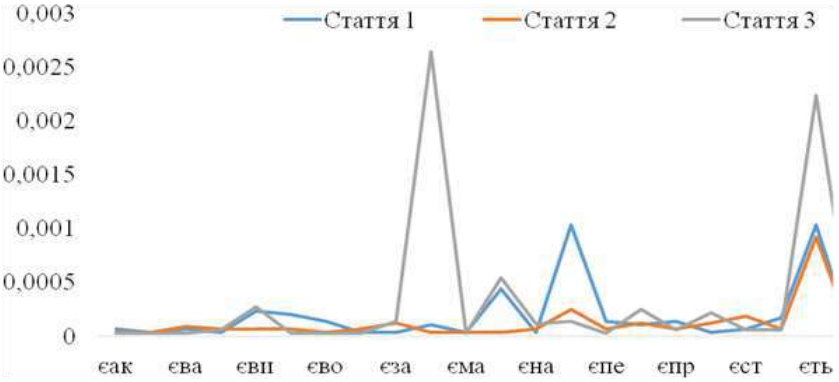


Рис. Б.12. Графік вживання 3-грам, які починаються з літери є

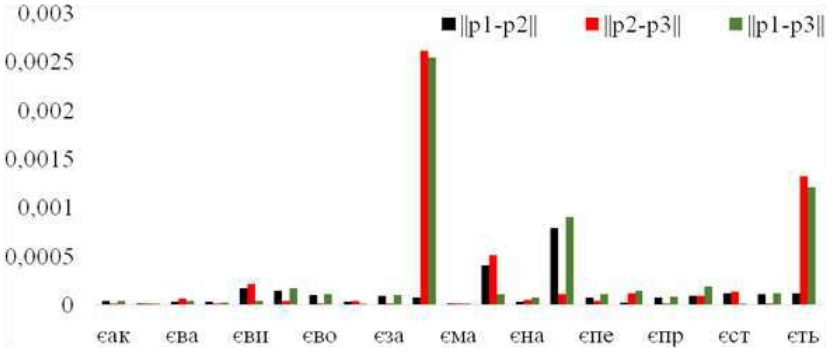


Рис. Б.13. Графік різниці вживання 3-грам, які починаються з літери є

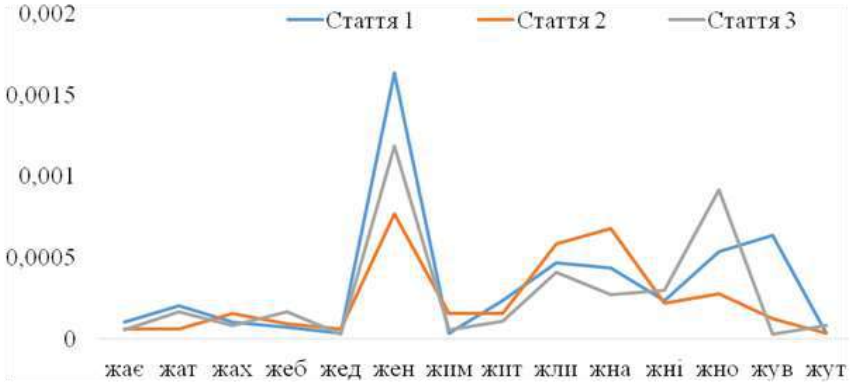


Рис. Б.14. Графік вживання 3-грам, які починаються з літери ж

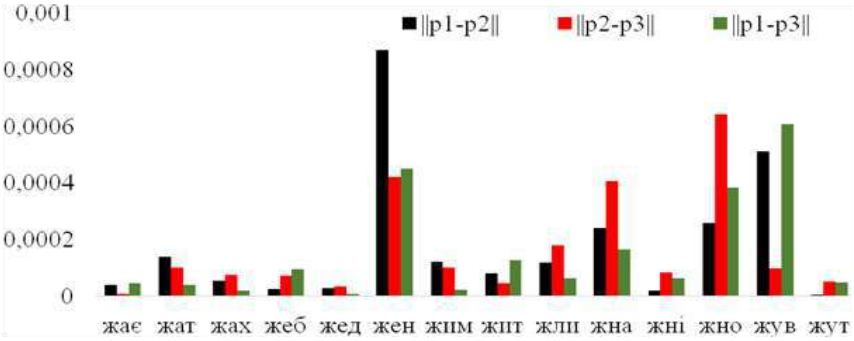


Рис. Б.15. Графік різниці вживання 3-грам, які починаються з літери ж

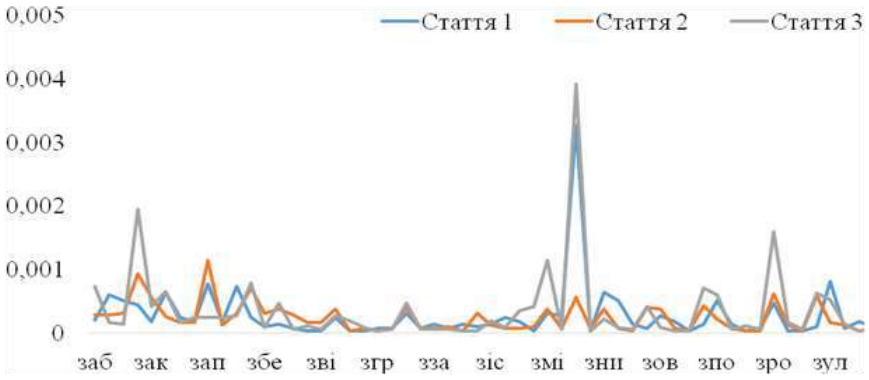


Рис. Б.16. Графік вживання 3-грам, які починаються з літери з

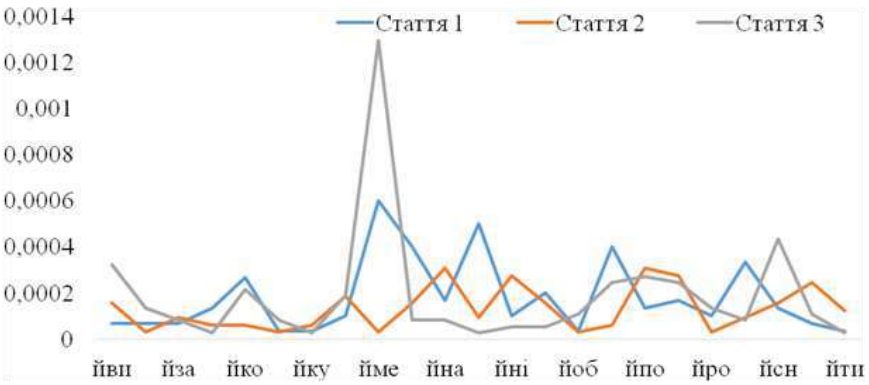


Рис. Б.17. Графік вживання 3-грам, які починаються з літери й

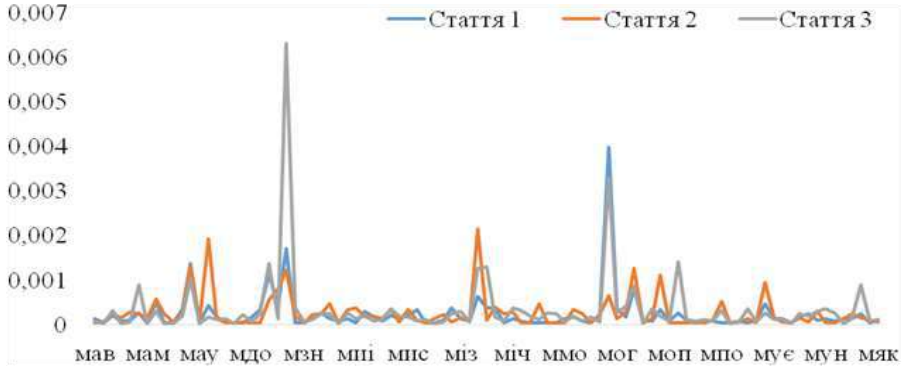


Рис. Б.18. Графік вживання 3-грам, які починаються з літери м

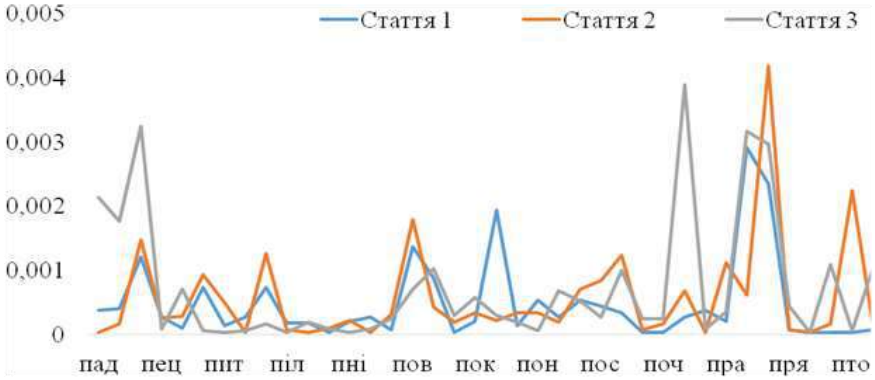


Рис. Б.19. Графік вживання 3-грам, які починаються з літери п

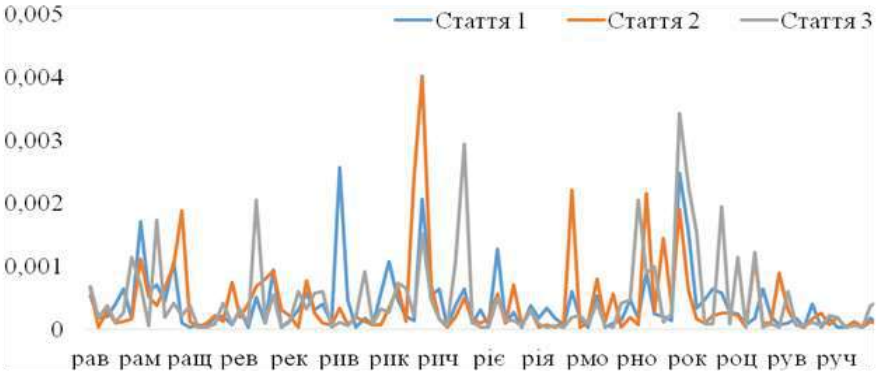


Рис. Б.20. Графік вживання 3-грам, які починаються з літери р

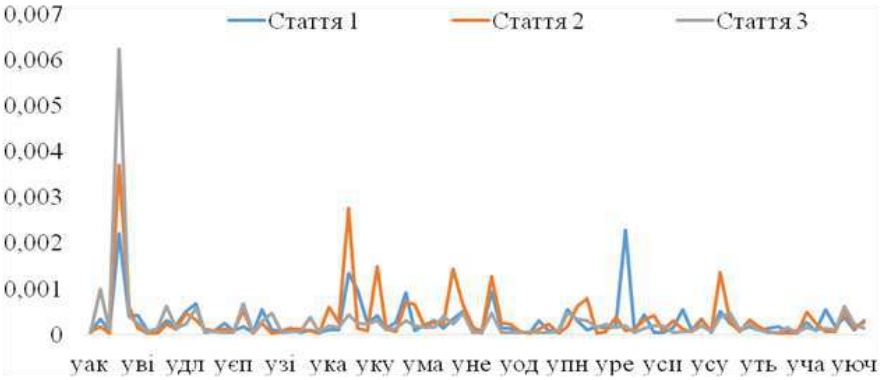


Рис. Б.21. Графік вживання 3-грам, які починаються з літери у

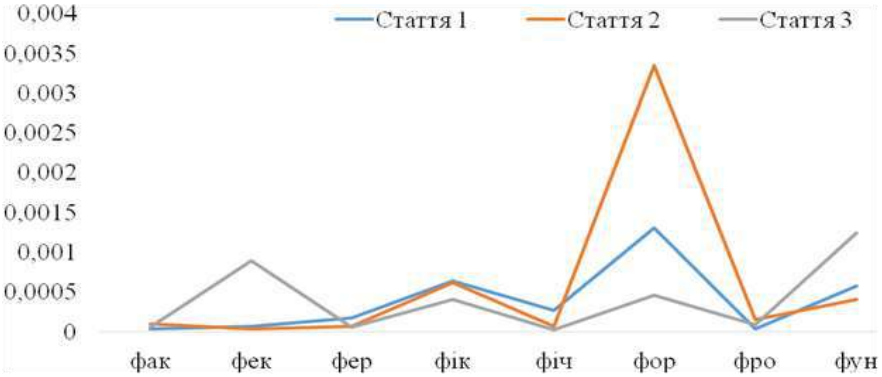


Рис. Б.22. Графік вживання 3-грам, які починаються з літери ф

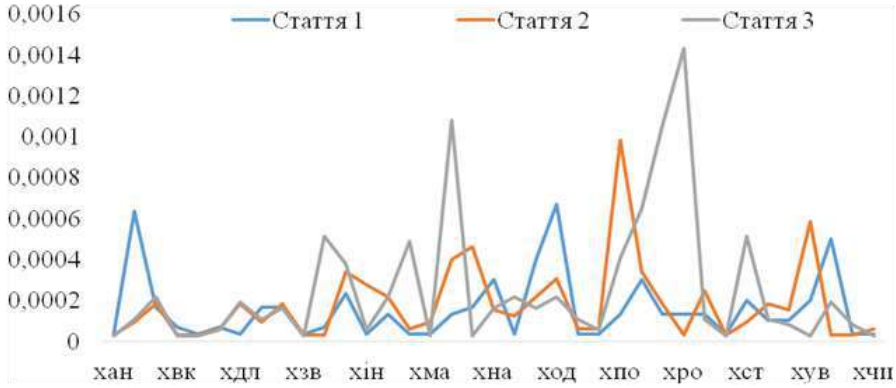


Рис. Б.23. Графік вживання 3-грам, які починаються з літери х

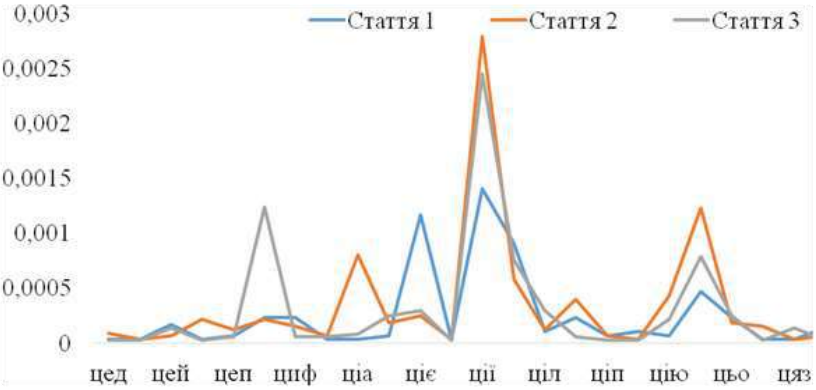


Рис. Б.24. Графік вживання 3-грам, які починаються з літери ц

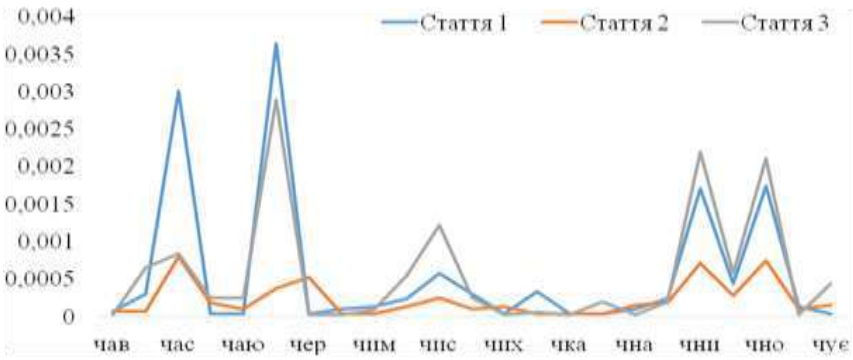


Рис. Б.25. Графік вживання 3-грам, які починаються з літери ч

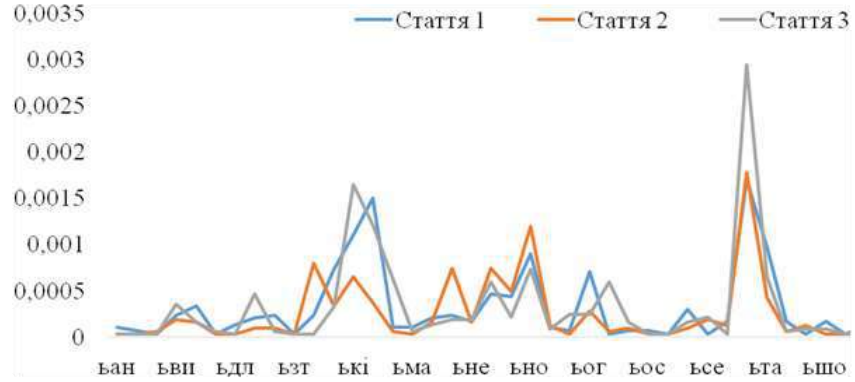


Рис. Б.26. Графік вживання 3-грам, які починаються з літери ь

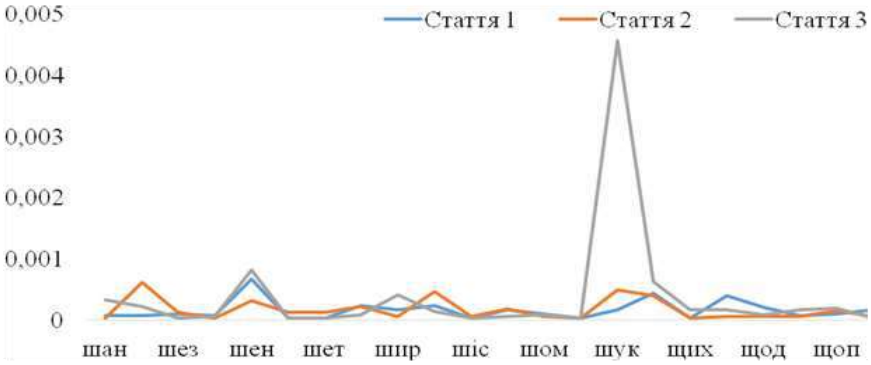


Рис. Б.27. Графік вживання 3-грам, які починаються з літер ш та щ

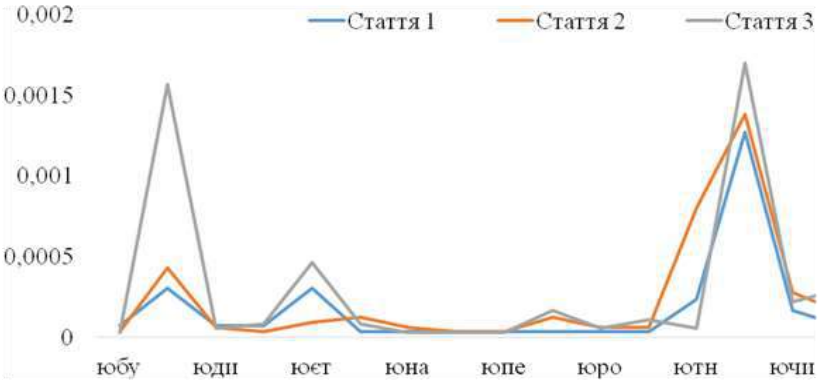


Рис. Б.28. Графік вживання 3-грам, які починаються з літери ю

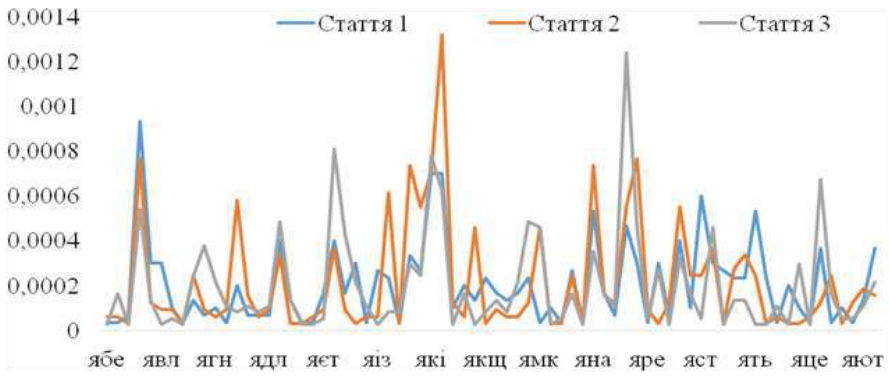
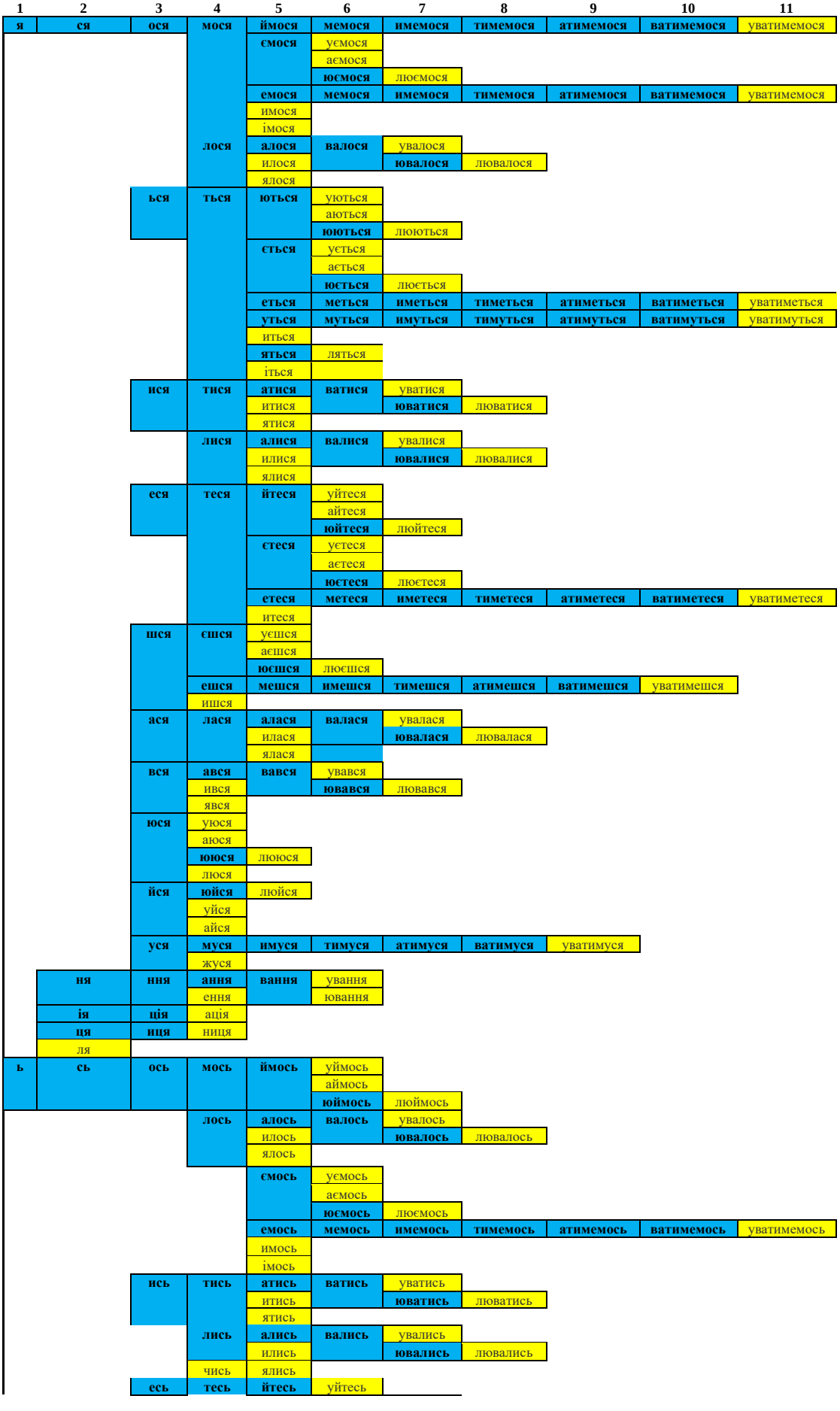
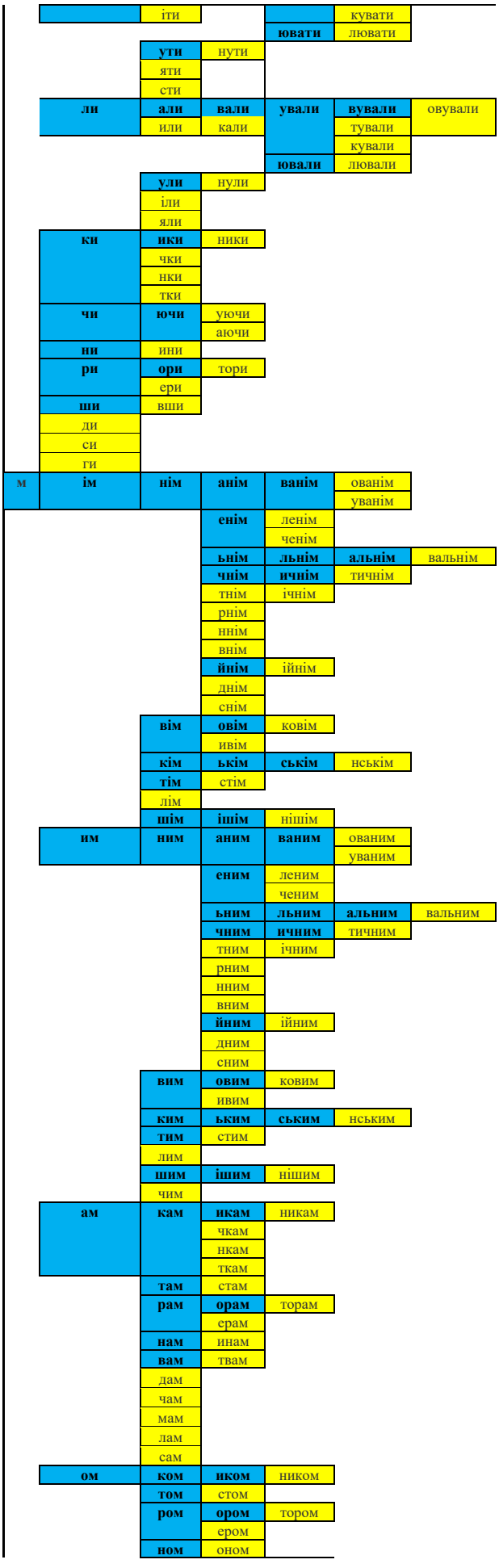


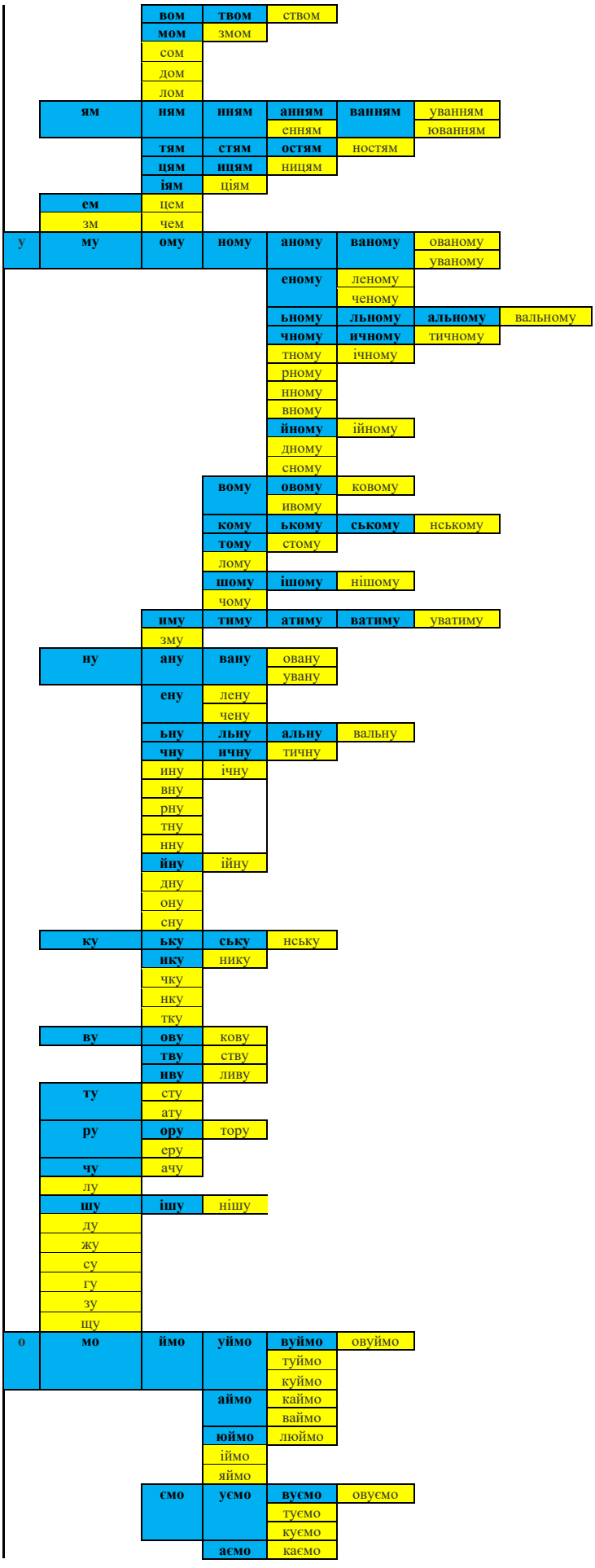
Рис. Б.29. Графік вживання 3-грам, які починаються з літери я

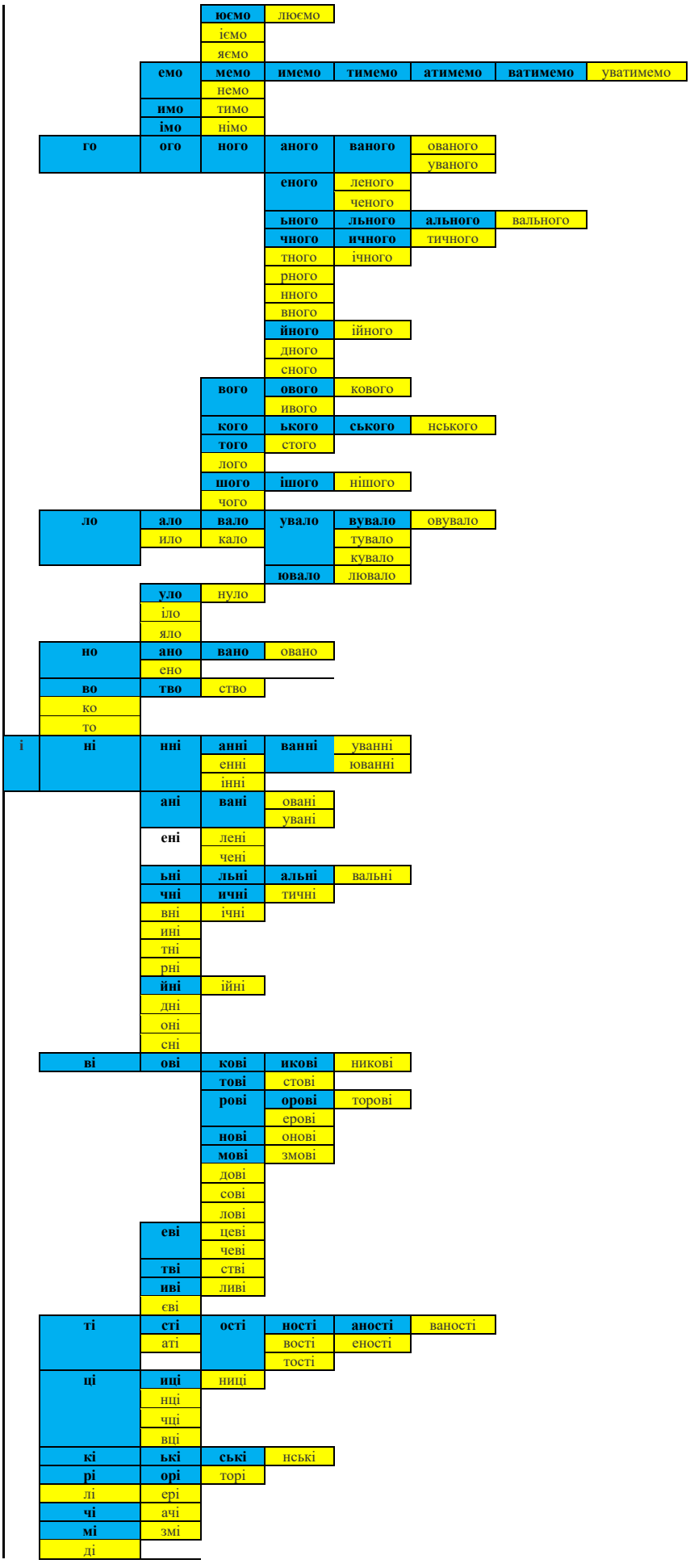
ДОДАТОК В. ДЕРЕВО ЗАКІНЧЕНЬ СЛІВ В УКРАЇНСЬКІЙ МОВІ



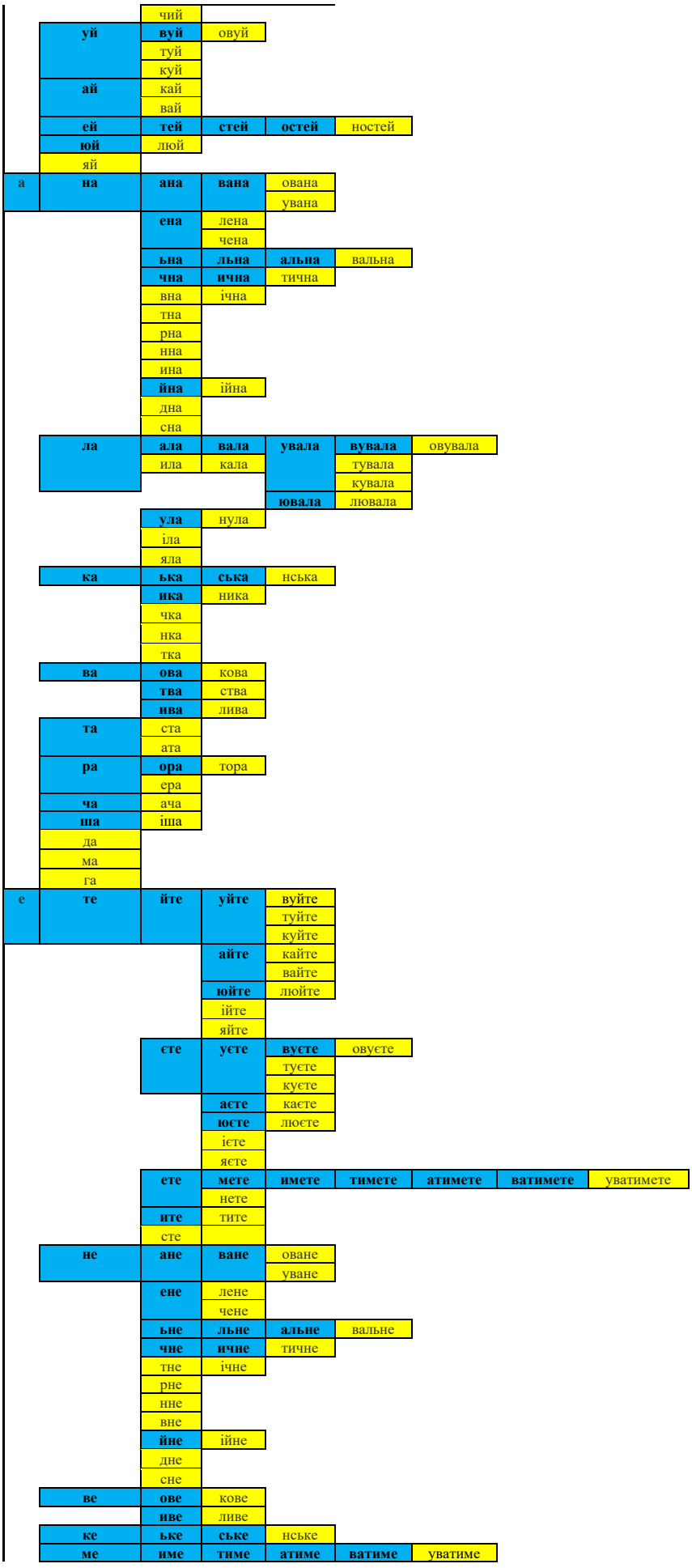

айтесь
юйтесь
люйтесь
стесь
уєтесь
аєтесь
юстесь
люєтесь
стесь
метесь
иметесь
тиметесь
атиметесь
ватиметесь
уватиметесь
итесь
ась
лась
алась
валась
увалась
илась
ювалась
ювалась
лювалась
ялась
всь
авсь
вавсь
увавсь
ивсь
ювавсь
ювавсь
лювавсь
явсь
юсь
уюсь
аюсь
ююсь
люються
люсь
йсь
уйсь
айсь
юйсь
люйсь
усь
мусь
имусь
тимусь
атимусь
ватимусь
уватимусь
жусь
ть
ють
ують
вують
овують
тують
кують
ають
кають
ються
люють
іють
яють
сть
ість
ність
аність
ваність
вість
еність
тість
уть
муть
имуть
тимуть
атимуть
ватимуть
уватимуть
нуть
ить
тить
ять
лять
тять
іть
ніть
ать
нь
ань
вань
увань
ень
ць
ець
ниць
ль
ниць
и
ми
ими
ними
аними
ваними
ованими
уваними
еними
леними
ченими
ьними
льними
альними
вальними
чними
ичними
тичними
тними
ічними
рними
ннми
вними
йними
ійними
дними
сними
вими
овими
ковими
ивими
кими
ькими
ськими
ньскими
тими
стими
лими
шими
ішими
нішими
ами
ками
иками
никами
чками
нками
тками
тами
стами
рами
орами
торами
ерами
нами
инами
вами
твами
дами
чами
мами
лами
сами
ями
нями
ннями
аннями
ваннями
уваннями
еннями
тями
стями
остями
ностями
ціями
іциями
ницями
іями
ціями
ти
ати
вати
увати
вувати
овувати
ити
кати
тувати

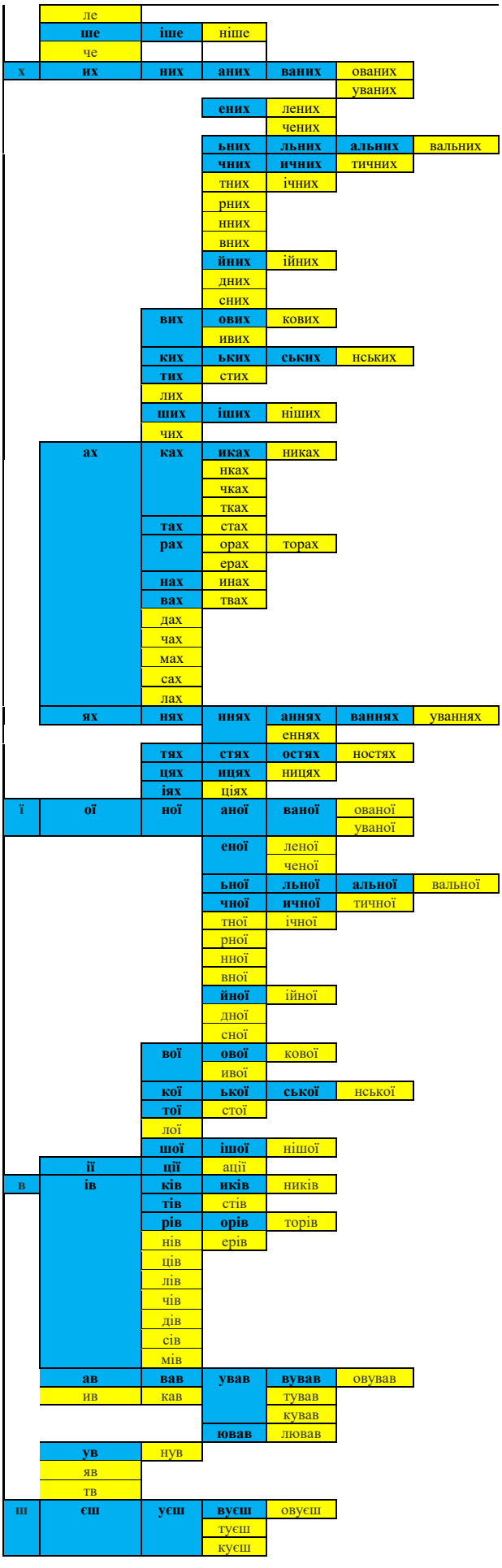

іти
кувати
ювати
лювати
ути
нути
яти
сти
ли
али
вали
ували
вувати
овували
или
кали
тувати
кувати
ювали
лювали
ули
нули
іли
яли
ки
ики
ники
чки
нки
тки
чи
ючи
уючи
аючи
ни
ини
ри
ори
тори
ери
ши
вши
ди
си
ги
м
ім
нім
анім
ванім
ованім
уванім
енім
ленім
ченім
ьнім
льнім
альнім
вальнім
чнім
ичнім
тичнім
тнім
ічнім
рнім
ннім
внім
йнім
ійнім
днім
снім
вім
овім
ковім
ивім
кім
ькім
ським
ньскім
тім
стім
лім
шім
ішім
нішім
им
ним
аним
ваним
ованим
уваним
еним
леним
ченим
ьним
льним
альним
вальним
чним
ичним
тичним
тним
ічним
рним
ннім
вним
йним
ійним
днім
сним
вим
овим
ковим
ивим
ким
ьким
ським
ньким
тим
стим
лим
шим
ішим
нішим
чим
ам
кам
икам
никам
чкам
нкам
ткам
там
стам
рам
орам
торам
ерам
нам
инам
вам
твам
дам
чам
мам
лам
сам
ом
ком
иком
ником
том
стом
ром
ором
тором
еромном
оном

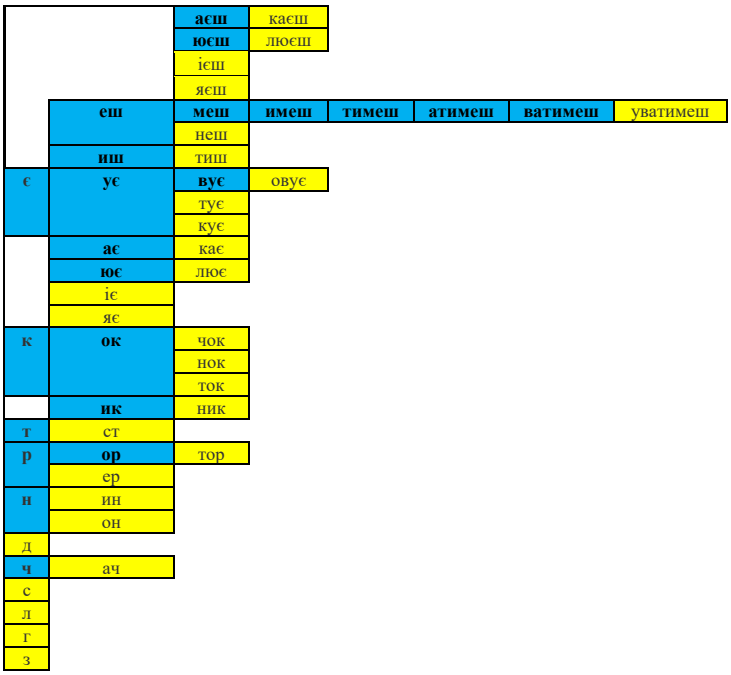

вом твом ством
мом змом
сом
дом
лом
ям ням нням анням ванням уванням
енням юванням
тям стям остям ностям
цям ицям ницям
іям ціям
ем цем
зм чем
у му ому ному аному ваному ованому уваному
еному леному ченому
ьному льному альному вальному
чному ичному тичному
тному ічному
рному
нному
вному
йному ійному
дному
сному
вому овому ковому
ивому
кому ькому ському нському
тому стому
лому
шому ішому нішому
чому
иму тиму атиму ватиму уватиму
зму
ну ану вану овану
увану
ену лену
чену
ьну льну альну вальну
чну ичну тичну
ину ічну
вну
рну
тну
нну
йну ійну
дну
ону
сну
ку ьку ську нську
ику нику
чку
нку
тку
ву ову кову
тву ству
иву ливу
ту сту
ату
ру ору тору
еру
чу ачу
лу
шу ішу нішу
ду
жу
су
гу
зу
щу
о мо йомо уйомо вуйомо овуйомо
туйомо
куйомо
аймо каймо
ваймо
юймо люймо
іймо
яймо
смо усмо вуємо овуємо
туємо
куємо
аємо каємо


юємо
люємо
іємо
яємо
емо
емо
имемо
тимемо
атимемо
ватимемо
уватимемо
немо
имо
тимо
імо
німо
го
ого
ного
аного
ваного
ованого
уваного
еного
леного
ченого
ьного
льного
ального
вального
чного
ичного
тичного
тного
ічного
рного
нного
вного
йного
ійного
дного
сного
вого
ового
кового
ивого
кого
ького
ського
нського
того
стого
лого
шого
ішого
нішого
чого
ло
ало
вало
увало
вувало
овувало
ило
кало
тувало
кувало
ювало
лювало
уло
нуло
ілo
яло
но
ано
вано
овано
ено
во
тво
ство
ко
то
і
ні
нні
анні
ванні
уванні
енні
ювanні
інні
ані
вані
овані
увані
ені
лені
чені
ьні
льні
альні
вальні
чні
ичні
тичні
вні
ічні
ини
тні
рні
йні
ійні
дні
они
сні
ві
ові
кові
икові
никові
тові
стові
рові
орові
торові
ерові
нові
онові
мові
змові
дові
сові
лові
еві
цеві
чеві
тві
ствi
иві
ливі
єві
ті
сті
ості
ності
аності
ваності
аті
вості
еності
тості
ці
иці
ниці
нці
чці
вці
кі
ькі
ські
нські
рі
орі
торі
лі
ері
чі
ачі
мі
змі
ді

	сі					
	щі	**іші**	ніші			
	зі					
ю	**ою**	**ною**	**аною**	**ваною**	ованою	
					уваною	
			еною	леною		
				ченою		
			ьною	**льною**	**альною**	вальною
			чною	**ичною**	тичною	
			вною	ічною		
			тною			
			рною			
			ннею			
			йною	ійною		
			иною			
			дною			
			сною			
		кою	**ькою**	**ською**	ньською	
			нкою			
			чкою			
			ткою			
		вою	**овою**	ковою		
			ивою	ливою		
		тою	стою			
		лою				
		шою	**ішою**	нішою		
		рою				
	ню	**нню**	**анню**	**ванню**	уванню	
			енню		ованню	
			інню			
	тю	**стю**	**істю**	**ністю**	**аністю**	ваністю
				вістю	еністю	
				тістю		
	ую	**вую**	овую			
		тую				
		кую				
	ію	**цію**	ацію			
	аю	каю				
	сю	**ісю**	**цісю**	ацією		
	цю	**ицю**	ницю			
	ею	**цею**	**ицею**	ницею		
	лю					
	юю	люю				
	яю					
й	**ій**	**ній**	**аній**	**ваній**	ованій	
					уваній	
			еній	леній		
				ченій		
			ьній	**льній**	**альній**	вальній
			чній	**ичній**	тичній	
			тній	ічній		
			рній			
			ннній			
			вній			
			йній	ійній		
			дній			
			сній			
		вій	**овій**	ковій		
			ивій			
		кій	**ькій**	**ській**	ньській	
		тій	стій			
		лій				
		шій	**ішій**	нішій		
		цій				
	ий	**ний**	**аний**	**ваний**	ований	
					уваний	
			ений	лений		
				чений		
			ьний	**льний**	**альний**	вальний
			чний	**ичний**	тичний	
			тний	ічний		
			рний			
			нний			
			вний			
			йний	ійний		
			дний			
			сний			
		вий	**овий**	ковий		
			ивий			
		кий	**ький**	**ський**	ньський	
		тий	стий			
		лий				
		ший	**ішний**	нішний		


чий
уй вуй овуй
туй
куй
ай кай
вай
ей тей стей остей ностей
юй люй
яй
а на ана ванаована
увана
ена лена
чена
ьна льна альна вальна
чна ична тична
вна ічна
тна
рна
нна
ина
йна ійна
дна
сна
ла ала вала увала вувала овувала
ила кала тувала
кувала
ювала лювала
ула нула
іла
яла
ка ька ська нська
ика ника
чка
нка
тка
ва ова кова
тва ства
ива лива
та ста
ата
ра ора тора
ера
ча ача
ша іша
да
ма
га
е те йте уйте вуйте
туйте
куйте
айте кайте
вайте
юйте люйте
ійте
яйте
сте усте вусте овусте
тусте
кусте
асте каєте
юсте люєте
ієте
яєте
ете мете имете тимете атимете ватимете уватимете
нете
ите тите
сте
не ане ване ование
уване
ене лене
чене
ьне льне альне вальне
чне ичне тичне
тне ічне
рне
нне
вне
йне ійне
дне
сне
ве ове ковe
иве ливе
ке ьке ське нське
ме име тиме атиме ватиме уватиме


ле
ше іше ніше
че
х их них аних ваних ованих
уваних
ених ленних
ченних
ьних льних альних вальних
чних ичних тичних
тних ічних
рних
нних
вних
йних ійних
дних
сних
вих ових кових
ивих
ких ьких ських ноських
тих стих
лих
ших іших ніших
чих
ах ках иках никах
нках
чках
тках
тах стах
рах орах торах
ерах
нах инах
вах твах
дах
чах
мах
сах
лах
ях ніях ніях аніях ваніях уваннях
еннях
тях стях остях ностях
цях ицях ницях
іях ціях
ї ої ної аної ваної ованої
уваної
еної ленної
ченної
ьної льної альної вальної
чної ичної тичної
тної ічної
рної
ннної
вної
йної ійної
дної
сної
вої ової кової
ивої
кої ької ської носької
тої стої
лої
шої ішої нішої
ії ції ації
в ів ків иків ників
тів стів
рів орів торів
нів ерів
ців
лів
чів
дів
сів
мів
ав вав ував вував овував
ив кав тував
кував
ював лював
ув нув
яв
тв
ш єш уєш вуєш овуєш
туєш
куєш


аєш
каєш
юєш
люєш
ієш
яєш
еш
меш
имеш
тимеш
атимеш
ватимеш
уватимеш
неш
иш
тиш
є
ує
вує
овує
тує
кує
ає
кає
ює
лює
іє
яє
к
ок
чок
нок
ток
ик
ник
т
ст
р
ор
тор
ер
н
ин
он
д
ч
ач
с
л
г
з

ДОДАТОК Д. СТАТИСТИЧНІ ДАНІ

Таблиця Д.1

Список за рейтингом частоти появи стійких словосполучень для 3 рандомних статей

№	Авторські	Victana.lviv.ua (згідно закону Зіпфа)	FREG, t-тест	LR	χ^2
Q	A	B	C,D	F	G
У роботі[1] українською мовою					
1	**Стиль автора**	Стоп-слово	Відносна частота	*Коефіцієнт кореляції*	*Коефіцієнт кореляції*
2	Статистичний аналіз	Метод визначення	*Коефіцієнт кореляції*	Відносна частота	Відносна частота
3	Лінгвістичний аналіз	**Визначення стилю**	**Стиль автора**	*Частота появи*	*Частота появи*
4	Квантитативна лінгвістика	**Стиль автора**	**Визначення стилю**	Стопове слово	Авторська атрибуція
5	Авторська атрибуція	Аналіз уривку	Стопове слово	Україномовний текст	**Стиль автора**
6	**Визначення стилю**	*Частота появи*	Україномовний текст	**Стиль автора**	Україномовний текст
7	Україномовні тексти	*Автор тексту*	*Частота появи*	Поява слова	Стопове слово
8	Технологія лінгвометрії	Уривок тексту	Авторська атрибуція	Авторська атрибуція	**Визначення стилю**
9	Технологія стилеметрії	*Коефіцієнт кореляції*	Поява слова	**Визначення стилю**	Поява слова
10	Технологія глоттохронології	Дослідження тексту	*Автор тексту*	Слова уривку	Слова уривку
У роботі[2] українською мовою					
1	**Web Mining**	**Ключове слово**	**Ключове слово**	*Текстовий контент*	*Текстовий контент*
2	Контент-моніторінг	**Контент-аналіз**	*Текстовий контент*	**Ключове слово**	Тематичний словник
3	**Ключові слова**	Визначена системою	**Web Mining**	Тематичний словник	**Ключове слово**
4	**Контент-аналіз**	*Формування системою*	Тематичний словник	*Слова контенту*	*Слова контенту*
5	Стеммер Портера	**Web Mining**	Визначення слів	Ключове словосполучення	*Множина слів*
6	Лінгвістичний аналіз	*Слова контенту*	Ключове словосполучення	Визначення слів	*Формування системою*
7	Метод визначення	*Текстовий контент*	*Слова контенту*	*Формування системою*	**Web Mining**
8	Визначення слів	Аналіз статистики	*Множина слів*	**Web Mining**	Визначення слів
9	Слов'янськомовні тексти	Ключове словосполучення	*Формування системою*	*Слова контенту*	*Слова контенту*
10	Технологія NLP	*Множина слів*	**Контент-аналіз**	Контент-моніторінг	Контент-моніторінг
У роботі[3] українською мовою					
1	Інформаційний ресурс	**Контент-аналіз**	Психологічний стан	Психологічна особистість	Психологічна особистість
2	**Контент-аналіз**	Стоп- слово	Психологічна особистість	Психологічний стан	Психологічний стан
3	Лінгвістичний аналіз	Тематичний словник	**Контент-аналіз**	Формування зрізу	Формування зрізу
4	Морфологічний аналіз	Пости користувача	Марковане слово	*Стан особистості*	Зріз стану
5	**Соціальна мережа**	Повідомлення користувача	Психологічний зріз	Марковане слово	Марковане слово
6	Формування зрізу	Користувач мережі	*Стан особистості*	Психологічний зріз	**Контент-аналіз**
7	Зріз розуміння	*Стан особистості*	Формування зрізу	**Контент-аналіз**	Психологічний зріз
8	Розуміння особистості	Аналізована особистість	Зріз стану	Зріз стану	*Стан особистості*
9	Україномовні тексти	**Соціальна мережа**	Зріз особистості	Аналізована особистість	**Соціальна мережа**
10	Big-Five	Диспозиції особистості	**Соціальна мережа**	**Соціальна мережа**	Аналізована особистість
У роботі[1] англійською мовою					
1	**Style of the author**	*Reference fragment*	*Reference fragment*	Words fragment	Words fragment
2	Statistical analysis	**Author's style**	Words fragment	*Reference fragment*	*Reference fragment*
3	Linguistic analysis	*Author's text*	*Syntactic words*	*Stop words*	Recognition author
4	Quantitative linguistics	*Syntactic words*	Frequency fragment	Swadesh list	*Stop words*
5	Author's attribution	*Stop words*	Swadesh list	Recognition author	Swadesh list
6	Recognition of style	Formatted fragments	*Stop words*	*Syntactic words*	*Syntactic words*
7	Ukrainian texts	*Anchor words*	**Author style**	Frequency fragment	Frequency fragment
8	Linguometry technology	Author's language	Recognition author	*Author's text*	*Author's text*
9	Stylemetry technology	Method of anchor	*Author's text*	*Anchor words*	**Author style**
10	Glottochronology technology	Frequency dictionary	*Anchor words*	**Author style**	*Anchor words*
У роботі[2] англійською мовою					
1	Web Mining	*Text content*	*Text content*	Web mining	Web mining
2	**Content monitoring**	**Content analysis**	Web mining	*Text content*	*Text content*
3	**Content analysis**	Analysis of statistics	Keywords text	*Keywords content*	*Keywords content*
4	Porter stemmer	Defined systematically	Keywords defined	Keywords text	Analysis text
5	Linguistic analysis	*Stop word*	Analysis text	Keywords defined	Keywords text

6	Determining the keywords	Potential keywords	*Keywords content*	*Stop word*	Keywords defined
7	Slavic language	**Content monitoring**	**Content monitoring**	Analysis text	*Stop word*
8	Slavic texts	*Author's keywords*	**Content analysis**	*Author's keywords*	**Content monitoring**
9	Method for determining	*Keywords content*	*Stop word*	**Content monitoring**	**Content analysis**
10	Web technology	Direct word	*Author's keywords*	**Content analysis**	*Author's keywords*
У роботі[3] англійською мовою					
1	Information resource	**Content analysis**	**Content analysis**	Psychological personality	**Content analysis**
2	**Content analysis**	*Psychological state*	Psychological personality	*Psychological state*	Psychological personality
3	Linguistic analysis	Personality analysis	*Psychological state*	**Content analysis**	*Psychological state*
4	Morphological analysis	Personality disposition	Social networks	Based analysis	Based analysis
5	Social network	Psychological analysis	Marked words	State personality	Psychological base
6	Status of personality	Personality model	State personality	Psychological base	State personality
7	Personality understanding	Stop words	Based analysis	Social networks	Social networks
8	Formation of the status	Psychological disposition	Psychological base	Marked words	Psychological base
9	Stop words	Content monitoring	State based	State based	Marked words
10	Method of formation	Social network	Based content	Psychological base	State based

Таблиця Д.2

Відмінності методів за рейтинговим списком із 100 стійких словосполечень

Q	A	B	C	D	F	G	A	B	C	D	F	G	A	B	C	D	F	G
Для україномовних статтей [1-3]																		
A	1	0,23	0,47	0,35	0,27	0,21	1	0,27	0,51	0,39	0,31	0,25	1	0,25	0,49	0,36	0,29	0,23
B	0,23	1	0,63	0,61	0,52	0,43	0,27	1	0,65	0,63	0,57	0,47	0,25	1	0,64	0,62	0,55	0,45
C	0,47	0,63	1	0,93	0,17	0,71	0,51	0,65	1	0,94	0,25	0,73	0,49	0,64	1	0,93	0,21	0,72
D	0,35	0,61	0,93	1	0,19	0,75	0,39	0,63	0,94	1	0,26	0,77	0,36	0,62	0,93	1	0,22	0,76
F	0,27	0,52	0,17	0,19	1	0,26	0,31	0,57	0,25	0,26	1	0,39	0,29	0,55	0,21	0,22	1	0,33
G	0,21	0,43	0,71	0,75	0,26	1	0,25	0,47	0,73	0,77	0,39	1	0,23	0,45	0,72	0,76	0,33	1
Для англомовних статтей [1-3]																		
A	1	0,27	0,51	0,47	0,31	0,27	1	0,31	0,55	0,51	0,35	0,31	1	0,29	0,53	0,49	0,33	0,29
B	0,27	1	0,66	0,64	0,55	0,47	0,31	1	0,69	0,67	0,59	0,49	0,29	1	0,68	0,65	0,57	0,48
C	0,51	0,66	1	0,95	0,23	0,76	0,55	0,69	1	0,96	0,27	0,77	0,53	0,68	1	0,95	0,24	0,75
D	0,47	0,64	0,95	1	0,21	0,79	0,51	0,67	0,96	1	0,29	0,81	0,49	0,65	0,95	1	0,25	0,78
F	0,31	0,55	0,23	0,21	1	0,31	0,35	0,59	0,27	0,29	1	0,41	0,33	0,57	0,24	0,25	1	0,37
G	0,27	0,47	0,76	0,79	0,31	1	0,31	0,49	0,77	0,81	0,41	1	0,29	0,48	0,75	0,78	0,37	1

Таблиця Д.3

Відмінності інших методів за рейтингом частоти появи стійких словосполучень

Метод	Мова	Робота [1]	Робота [2]	Робота [3]
A_1	UA	('контент_моніторінгу', 13)	('тематичного_словника', 11) ('слов_янськомовних', 10)	('психологічного_стану', 16) ('формування_зрізу', 12) ('sfx_a', 12) ('структурну_схему', 7) ('відкритість_досвіду', 6) ('зрізу_психологічного', 2)
	ENG	('swadesh_list', 18) ('based_on', 15)	('based_on', 20) ('slavic_language', 15) ('author_s', 13)	('based_on', 35) ('psychological_state', 26) ('social_networks', 22) ('his_her', 11) ('following_structural', 8) ('big_five', 7) ('let_us', 7) ('structural_scheme', 4)
A_2	UA	(('службових', 'слів'), 32) (('стопових', 'слів'), 24) (('визначення', 'визначення'), 23) (('стилю', 'стилю'), 22) (('слів', 'слів'), 22) (('списку', 'сводеша'), 20) (('в', 'уривку'), 19) (('опорних', 'слів'), 18) (('стилю', 'автора'), 17) (('автора', 'автора'), 17)	(('ключових', 'слів'), 72) (('текстового', 'контенту'), 21) (('на', 'етапі'), 17) (('визначення', 'ключових'), 16) (('крок', '1'), 16) (('крок', '2'), 16) (('web', 'mining'), 15) (('слів', 'в'), 14) (('тематичного', 'словника'), 11) (('для', 'визначення'), 10)	(('на', 'основі'), 21) (('психологічного', 'стану'), 18) (('контент', 'аналізу'), 16) (('маркованих', 'слів'), 15) (('зрізу', 'психологічного'), 14) (('стану', 'особистості'), 14) (('формування', 'зрізу'), 12) (('особистості', 'на'), 12) (('sfx', 'a'), 12) (('основі', 'контент'), 11)
	ENG	(('of', 'the'), 107) (('author', 's'), 52) (('of', 'a'), 51) (('in', 'the'), 46) (('the', 'author'), 45)	(('of', 'the'), 134) (('in', 'the'), 61) (('by', 'the'), 45) (('analysis', 'of'), 39) (('of', 'a'), 31)	(('of', 'the'), 134) (('is', 'the'), 117) (('the', 'content'), 45) (('of', 'a'), 43) (('analysis', 'of'), 37)

		(('reference', 'fragment'), 31) (('analysis', 'of'), 24) (('words', 'in'), 22) (('to', 'the'), 21) (('the', 'method'), 21)	(('the', 'text'), 30) (('the', 'system'), 30) (('to', 'the'), 29) (('of', 'keywords'), 28) (('text', 'content'), 27)	(('based', 'on'), 35) (('on', 'the'), 34) (('in', 'the'), 33) (('content', 'analysis'), 30) (('the', 'process'), 27)
A_3	UA	(('слів', 'слів'), 88) (('стилю', 'автора'), 68) (('службових', 'слів'), 63) (('визначення', 'стилю'), 61) (('списку', 'сводеша'), 56) (('стопових', 'слів'), 48) (('визначення', 'автора'), 45) (('авторського', 'мовлення'), 33) (('опорних', 'слів'), 31) (('стилю', 'стилю'), 30)	(('ключових', 'слів'), 74) (('слів', 'в'), 24) (('web', 'mining'), 22) (('текстового', 'контенту'), 21) (('на', '2'), 20) (('визначення', 'ключових'), 19) (('ключових', 'в'), 19) (('визначення', 'слів'), 18) (('слів', 'для'), 18) (('на', 'крок'), 18)	(('на', 'основі'), 21) (('психологічного', 'стану'), 18) (('психологічного', 'особистості'), 17) (('контент', 'аналізу'), 16) (('стану', 'особистості'), 15) (('маркованих', 'слів'), 15) (('зрізу', 'психологічного'), 14) (('зрізу', 'стану'), 14) (('зрізу', 'особистості'), 14) (('особистості', 'на'), 14)
	ENG	(('of', 'the'), 186) (('the', 'of'), 169) (('of', 'of'), 152) (('of', 'a'), 81) (('the', 'the'), 75) (('the', 'author'), 66) (('and', 'of'), 63) (('in', 'the'), 57) (('of', 'author'), 57) (('of', 'words'), 55)	(('of', 'the'), 258) (('the', 'of'), 235) (('of', 'of'), 137) (('the', 'the'), 122) (('of', 'keywords'), 72) (('in', 'the'), 71) (('a', 'of'), 70) (('and', 'of'), 69) (('by', 'the'), 64) (('of', 'content'), 63)	(('the', 'of'), 304) (('of', 'the'), 243) (('the', 'the'), 168) (('of', 'of'), 162) (('is', 'the'), 154) (('of', 'a'), 91) (('the', 'is'), 76) (('the', 'content'), 71) (('is', 'of'), 61) (('and', 'the'), 57)
A_4		(('слів', 'слів'), 88) (('стилю', 'автора'), 68) (('службових', 'слів'), 63) (('визначення', 'стилю'), 61) (('списку', 'сводеша'), 56) (('стопових', 'слів'), 48) (('визначення', 'автора'), 45) (('авторського', 'мовлення'), 33) (('опорних', 'слів'), 31) (('стилю', 'стилю'), 30)	(('text', 'content'), 30) (('web', 'mining'), 24) (('keywords', 'text'), 23) (('keywords', 'defined'), 22) (('stage', '1'), 20) (('analysis', 'text'), 18) (('step', '2'), 18) (('keywords', 'content'), 17) (('content', 'monitoring'), 17) (('step', '1'), 17)	(('на', 'основі'), 21) (('психологічного', 'стану'), 18) (('психологічного', 'особистості'), 17) (('контент', 'аналізу'), 16) (('стану', 'особистості'), 15) (('маркованих', 'слів'), 15) (('зрізу', 'психологічного'), 14) (('зрізу', 'стану'), 14) (('зрізу', 'особистості'), 14) (('особистості', 'на'), 14)
		(('fragment', 'fragment'), 37) (('reference', 'fragment'), 35) (('words', 'fragment'), 25) (('syntactic', 'words'), 21) (('frequency', 'fragment'), 19) (('swadesh', 'list'), 19) (('stop', 'words'), 18) (('author', 'style'), 17) (('fragment', '3'), 17) (('recognition', 'author'), 16)	(('ключових', 'слів'), 74) (('слів', 'в'), 24) (('web', 'mining'), 22) (('текстового', 'контенту'), 21) (('на', '2'), 20) (('визначення', 'ключових'), 19) (('ключових', 'в'), 19) (('визначення', 'слів'), 18) (('слів', 'для'), 18) (('на', 'крок'), 18)	(('content', 'analysis'), 40) (('psychological', 'personality'), 27) (('psychological', 'state'), 26) (('social', 'networks'), 22) (('marked', 'words'), 21) (('state', 'personality'), 20) (('based', 'analysis'), 19) (('psychological', 'based'), 18) (('state', 'based'), 18) (('based', 'content'), 18)

Таблиця Д.4

Абсолютні та відносні частоти появи стопових слів в Уривку та еталоні

Уривок	Стоп-слово	АЧ	ВЧ	Частина мови	ВЧ в еталоні
1 (107 слів)	але	1	0,0093	Сполучник	0,0074
	в	2	0,0187	Прийменник	0,0140
	для	3	0,0280	Прийменник	0,0024
	до	1	0,0093	Прийменник	0,0113
	з	1	0,0093	Прийменник	0,0129
	і	14	0,1308	Сполучник	0,0300
	й	1	0,0093	Сполучник	0,0038
	мов	1	0,0093	Частка	0,0022
	не	2	0,0187	Частка	0,0237
	про	2	0,0187	Прийменник	0,0040
	та	2	0,0187	Сполучник	0,0047
	що	1	0,0093	Сполучник	0,0206
2 (117 слів)	а	2	0,0171	Сполучник	0,0116
	в	3	0,0256	Прийменник	0,0140
	від	1	0,0085	Прийменник	0,0034
	до	1	0,0085	Прийменник	0,0113
	ж	1	0,0085	Сполучник	0,0033
	з	2	0,0171	Прийменник	0,0129
	за	1	0,0085	Прийменник	0,0053
	і	2	0,0171	Сполучник	0,0300
	й	2	0,0171	Сполучник	0,0038
	на	1	0,0085	Прийменник	0,0159
	над	1	0,0085	Прийменник	0,0005

	не	2	0,0171	Частка	0,0237
	ні	1	0,0085	Частка	0,0024
	ось	1	0,0085	Частка	0,0012
	от	1	0,0085	Частка	0,0005
	се	1	0,0085	Частка	0,0074
	хіба	1	0,0085	Частка	0,0006
	хоч	1	0,0085	Частка	0,0010
	що	2	0,0171	Сполучник	0,0206
	як	1	0,0085	Сполучник	0,0060
3 (162 слів)	а	4	0,0247	Сполучник	0,0116
	але	2	0,0123	Сполучник	0,0074
	без	1	0,0062	Прийменник	0,0008
	бо	1	0,0062	Сполучник	0,0012
	в	1	0,0062	Прийменник	0,0140
	від	1	0,0062	Прийменник	0,0034
	ж	1	0,0062	Сполучник	0,0033
	з	4	0,0247	Прийменник	0,0129
	за	2	0,0123	Прийменник	0,0053
	і	1	0,0062	Сполучник	0,0300
	й	4	0,0247	Сполучник	0,0038
	на	6	0,0370	Сполучник	0,0159
	навіть	2	0,0123	Частка	0,0011
	не	3	0,0185	Частка	0,0237
	під	4	0,0247	Прийменник	0,0011
	таки	1	0,0062	Частка	0,0004
	тож	1	0,0062	Сполучник	0,0001
	у	4	0,0247	Прийменник	0,0088
	що	3	0,0185	Сполучник	0,0206
	щоб	1	0,0062	Сполучник	0,0028
	як	1	0,0062	Сполучник	0,0060
4 (149 слів)	адже	1	0,00671	Частка	0,0011
	але	2	0,01342	Сполучник	0,0074
	би	1	0,00671	Частка	0,0033
	в	1	0,00671	Прийменник	0,0140
	ж	1	0,00671	Сполучник	0,0033
	з	3	0,02013	Прийменник	0,0129
	за	1	0,00671	Прийменник	0,0053
	і	4	0,02685	Прийменник	0,0300
	мов	1	0,00671	Частка	0,0022
	на	7	0,04698	Прийменник	0,0159
	не	4	0,02685	Частка	0,0237
	отсе	1	0,00671	Частка	0,0003
	при	1	0,00671	Прийменник	0,0018
	про	2	0,01342	Прийменник	0,0040
	се	1	0,00671	Частка	0,0074
	у	2	0,01342	Прийменник	0,0088
	чи	2	0,01342	Сполучник	0,0027
	що	7	0,04698	Сполучник	0,0206
	щоб	1	0,00671	Сполучник	0,0028
	як	1	0,00671	Сполучник	0,0060

Таблиця Д.5

Результат роботи алгоритму аналізу стилю автора публікації

№	N	W	W_1	W_{10}	P	Z	S	K_l	K_s	K_z	I_{wt}	I_{kt}
1	671,3	395,6	299	6	44,2	57,1	41,1	0,59	0,89	0,76	0,76	0,015
2	662,5	410,3	303	5	37,8	39,8	34,8	0,61	0,9	0,67	0,74	0,012
3	668,8	418,3	325,8	6,8	29,8	56	57	0,63	0,93	1,28	0,78	0,016
4	708	419	309	8	36	64	28	0,59	0,91	0,85	0,74	0,019
5	661,1	402,7	299,7	4,7	44,7	54,7	24,8	0,61	0,89	0,6	0,74	0,012
6	694,5	417,4	313,1	6,4	54,3	58,5	38,1	0,6	0,87	0,62	0,75	0,015
7	691,8	403,4	301,6	7,8	47,8	60	47,8	0,58	0,88	0,79	0,75	0,019
8	682,5	394,2	291	5	49	61	39,7	0,58	0,88	0,74	0,74	0,013
9	733,5	486,5	392	5	50	65	45	0,66	0,9	0,76	0,8	0,01
10	729	380	261	7	62	75	32	0,52	0,84	0,58	0,69	0,018
11	686,5	414,5	312,6	5,9	41,1	56,9	45	0,6	0,9	0,86	0,75	0,012
12	665,5	399	299	6	35,5	72	43	0,6	0,91	1,09	0,75	0,015
13	724,2	394,2	278,8	5,8	59,6	68,4	36,8	0,55	0,85	0,61	0,71	0,015
14	691	396,7	289	7	39	55,3	42,3	0,57	0,9	0,85	0,73	0,018
15	745	439	319	6	45	59	61	0,59	0,9	0,89	0,73	0,014
16	768	452,5	323	5,5	51,5	58	47	0,59	0,89	0,68	0,71	0,012
17	647	422	308	3	62	50	32	0,65	0,85	0,44	0,73	0,007
18	677,5	373,5	255	6,5	64,5	72	36	0,55	0,86	0,57	0,68	0,018

19	680	379	251	5	42	55	33	0,56	0,89	0,7	0,66	0,013
20	642	337,5	230,3	7,8	44,8	52,3	56,8	0,52	0,87	0,81	0,68	0,023
21	665	376	275,7	7,7	41,7	65	32,3	0,57	0,89	0,79	0,73	0,02
22	731	420	301	7	49	71	54	0,57	0,88	0,85	0,72	0,017
23	691,7	425,7	331,3	6,5	41,8	58,2	50	0,62	0,9	0,88	0,78	0,015
24	668,8	368,3	262,5	6,8	44	55,8	34,5	0,55	0,88	0,73	0,71	0,018
25	691	421	311	4	47	65	40	0,6	0,89	0,74	0,74	0,01
26	708,5	434	323,5	6,5	42	57,5	47,5	0,61	0,9	0,84	0,75	0,015
27	665	406	309	5	41	42	28	0,61	0,9	0,57	0,76	0,012
28	700	418,5	320,5	6	40	68,5	35	0,6	0,9	0,88	0,77	0,014
29	704,5	412	303,5	5,5	59	47,5	38	0,58	0,86	0,49	0,74	0,013
30	688,8	416,8	321,9	6	49,7	49,3	41,3	0,6	0,88	0,67	0,77	0,016
31	711	396	268	6	60	67	19	0,56	0,85	0,48	0,68	0,015
32	691	436,7	336,7	5,7	40	51	44,7	0,63	0,91	0,82	0,77	0,013
33	695	422,5	318,3	7,5	38,5	61,3	41	0,6	0,91	0,89	0,75	0,018
34	699	427	314	6	49,5	60	41	0,61	0,88	0,69	0,74	0,014
35	683	438	339	4	38	52	42	0,64	0,91	0,82	0,77	0,009
36	730	440	323	6	42	62	39	0,6	0,9	0,8	0,73	0,014
37	714,5	418,5	304,5	6,5	46	65	48,5	0,59	0,89	0,86	0,73	0,016
38	717,5	433,5	321,5	6,5	56	57,5	26,5	0,6	0,87	0,5	0,74	0,015
39	728	430	313	6	49	59	51	0,59	0,89	0,75	0,73	0,014
40	666	401,5	305	6,5	40	63	35,5	0,6	0,9	0,82	0,76	0,016
41	715,5	352	223,5	8,5	45	58	34	0,49	0,87	0,68	0,63	0,024
42	699	401	302	6	46	68	32	0,57	0,89	0,72	0,75	0,015
43	620	411	323	2	36	55	40	0,66	0,91	0,88	0,79	0,005
44	645	403	302,3	4,3	39,3	58,7	37,7	0,62	0,9	0,84	0,74	0,011
45	708	475	392	5	49	83	46	0,67	0,9	0,88	0,83	0,011
46	708	442,5	336,5	5,5	43,5	62	56,5	0,63	0,9	0,91	0,76	0,012
47	689	458	369	7	44	65	36	0,66	0,9	0,77	0,81	0,015
48	1602	442	245	30	100	3	1	0,28	0,77	0,01	0,55	0,068
49	644	400	310	8	28	66	37	0,62	0,93	1,23	0,78	0,02
50	661,5	402,5	302	5	32	49,5	31	0,6	0,92	0,84	0,75	0,012
51	705	474	369	1	31	50	49	0,67	0,93	1,06	0,78	0,002
52	656	422,5	341,5	4,5	50	57,5	46	0,64	0,88	0,69	0,81	0,011
53	704,8	458,8	360	6	54,8	60	45,8	0,65	0,88	0,66	0,78	0,013
54	716	413,5	293	5,5	47	74,5	27,5	0,58	0,89	0,73	0,71	0,013
55	652	389	287	4	55	46	36	0,6	0,86	0,5	0,74	0,01
56	666	412	318	7	44	55	49	0,62	0,89	0,79	0,77	0,017
57	732	402	290	6	53	63	45	0,55	0,87	0,68	0,72	0,015
58	670	449	356	3	38	55	30	0,67	0,92	0,75	0,79	0,007
59	693	366	242	8	45	44	60	0,53	0,88	0,77	0,66	0,022
60	761	440	315,8	5,3	39,3	48,5	28,3	0,58	0,91	0,65	0,71	0,012
61	717	422	310	6	45	53	46	0,59	0,89	0,73	0,73	0,014
62	673,5	419	329	7,5	39	58	33	0,62	0,91	0,78	0,79	0,018
63	679	381	280	5	64	50	32	0,56	0,83	0,43	0,73	0,013
64	682,6	416,2	318	6,2	60	47,8	45	0,6	0,86	0,59	0,76	0,015
65	658	399	277	3	41	48	47	0,6	0,9	0,78	0,69	0,008
66	683	446	357	5,5	48,5	63	43,5	0,65	0,89	0,74	0,8	0,012
67	689,5	407,5	296	5,5	47,5	57	28	0,59	0,88	0,61	0,73	0,014
68	726	493	399	4	42	56	46	0,68	0,91	0,81	0,81	0,008
69	1325	538	360	19	66	9	2	0,4	0,88	0,06	0,67	0,035
70	697	450	361,5	5	56	59,5	46	0,65	0,88	0,63	0,8	0,011
71	652	405	296	2	34	45	28	0,62	0,92	0,72	0,73	0,005
72	598	386	309	4	83	40	0	0,65	0,78	0,16	0,8	0,01
73	726,3	441,3	332,3	6,7	51	61,3	39	0,6	0,88	0,68	0,75	0,015
74	846	440	299	10	54	57	26	0,52	0,88	0,51	0,68	0,023
75	712,5	442,5	331,5	4	51	48	33	0,62	0,88	0,53	0,75	0,009
76	706	374	275	8,5	45	68,5	31	0,53	0,88	0,74	0,73	0,023
77	682,3	398,7	296,3	4,7	39	50,3	37,7	0,58	0,9	0,75	0,74	0,012
78	654	361	240	5	39	35	28	0,55	0,89	0,54	0,66	0,014
79	631	350	249	7	34	45	31	0,55	0,9	0,75	0,71	0,02
80	661	391	275	4	63	53	24	0,59	0,84	0,41	0,7	0,01
81	709,5	399	292,5	9,5	48	58	49,5	0,56	0,88	0,75	0,73	0,024
82	695	436	332	3	53	51	39	0,63	0,88	0,57	0,76	0,007
83	700	485	406	6	50	46	42	0,69	0,9	0,59	0,84	0,012
84	674	404	316	7	39	63	35	0,9	0,9	0,84	0,78	0,017
85	685	432	333	5	42	53	39	0,63	0,9	0,73	0,77	0,012
86	780	479	366	6	41	43	34	0,61	0,91	0,63	0,76	0,013
87	723	401	280	6	41	54	35	0,55	0,9	0,72	0,7	0,015
88	665	425	324	4	40	46	33	0,64	0,91	0,66	0,76	0,009
89	730	433	317	7	41	70	51	0,59	0,91	0,98	0,73	0,016
90	734	381	273	7	30	26	29	0,52	0,92	0,61	0,72	0,018
91	749	478	375	7	46	73	49	0,64	0,9	0,88	0,78	0,015
92	732	429	329	6	55	59	67	0,59	0,87	0,76	0,77	0,014

93	709	398	285	6	52	46	35	0,56	0,87	0,52	0,72	0,015
94	680	414	314	4	55	62	34	0,6	0,87	0,58	0,76	0,01
95	622	397	305	5	37	42	48	0,64	0,91	0,81	0,77	0,013
96	614	391	287	4	46	69	32	0,64	0,88	0,73	0,73	0,01
97	658	345	241	8	31	59	42	0,52	0,91	1,07	0,7	0,023
98	631,3	377,7	277,7	5,7	38	56,7	40,7	0,6	0,9	0,88	0,73	0,015

Таблиця Д.6

Частотності появи літер в еталоні та досліджуваних уривках

Літера	Частотність вживання літер української мови (еталон)	Абсолютна частота літер в Уривку 1	Відносна частота вживання літер в Уривку 1	Абсолютна частота літер в Уривку 2	Відносна частота вживання літер в Уривку 2
« »	0,133	80	0,14	82	0,15
о	0,082	37	0,07	41	0,08
а	0,074	43	0,08	31	0,06
н	0,068	33	0,06	30	0,06
и	0,054	27	0,05	27	0,05
в	0,047	29	0,05	19	0,04
т	0,046	25	0,04	20	0,04
е	0,038	26	0,05	45	0,08
р	0,036	15	0,03	16	0,03
с	0,033	22	0,04	27	0,05
м	0,031	10	0,02	13	0,02
к	0,031	22	0,04	20	0,04
л	0,028	17	0,03	30	0,06
д	0,028	16	0,03	4	0,01
у	0,025	19	0,03	14	0,03
п	0,025	11	0,02	21	0,04
я	0,024	15	0,03	6	0,01
з	0,018	9	0,02	8	0,01
б	0,016	7	0,01	5	0,01
ч	0,015	5	0,01	11	0,02
г	0,012	4	0,01	6	0,01
ю	0,012	2	0,00	2	0,00
б	0,011	7	0,01	5	0,01
х	0,01	4	0,01	7	0,01
ц	0,009	7	0,01	1	0,00
ж	0,007	3	0,01	7	0,01
й	0,007	4	0,01	6	0,01
ш	0,005	3	0,01	2	0,00
щ	0,004	3	0,01	1	0,00
ф	0,003	1	0,00	0	0,00
інші	0,0605	51	0,09	34	0,06

ДОДАТОК Е. СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

Статті у періодичних виданнях, індексованих у Scopus та Web of Science

1. Lytvyn V., Pukach P., Vysotska V., Vovk M., Kholodna N. Identification and correction of grammatical errors in Ukrainian texts based on machine learning technology. Mathematics. 2023. Vol. 11. 904. ISSN 2227-7390. (квартиль Q2 відповідно до SCImago Journal).
2. Bisikalo O., Danylchuk O., Kovtun V., Kovtun O., Nikitenko O., Vysotska V. Modeling of operation of information system for critical use in the conditions of influence of a complex certain negative factor. International Journal of Control, Automation and Systems. 2022. Vol. 20. P. 904–1913. Print ISSN 1598-6446. (квартиль Q2 відповідно до SCImago Journal).
3. Bublyk M., Kowalska-Styczeń A., Lytvyn V., Vysotska V. The Ukrainian economy transformation into the circular based on fuzzy-logic cluster analysis. Energies. 2021. Vol. 14(18). Art. 5951. ISSN:1996-1073. (квартиль Q2 відповідно до SCImago Journal).
4. Lytvyn V., Vysotska V., Peleshchak I., Rishnyak I., Peleshchak R. Time dependence of the output signal morphology for nonlinear oscillator neuron based on Van der Pol model. International Journal of Intelligent Systems and Applications. 2018.Vol. 10(4). P. 8–17. ISSN: 2074-904X. (квартиль Q2 відповідно до SCImago Journal).
5. Висоцька В. Метод авторифікації тексту науково-технічних публікацій на основі лінгвістичного аналізу коефіцієнтів мовної різноманітності. Радіоелектроніка. Інформатика. Управління. 2020. № 1(52). С. 108–124.
6. Висоцька В. Інформаційна технологія просування інтернет-ресурсів в пошукових системах на основі контент-аналізу ключових слів web-сторінок. Радіоелектроніка, інформатика, управління. 2021 № 3 (58). С. 133-151.
7. Алєксєєва К. А., Берко А. Ю., Висоцька В. А. Технологія управління комерційним web-ресурсом на основі нечіткої логіки. Радіоелектроніка. Інформатика. Управління. 2015. № 3 (34). С. 71–79.
8. Бісікало О. В., Висоцька В. А. Виявлення ключових слів на основі методу контент-моніторингу україномовних текстів. Радіоелектроніка. Інформатика. Управління. 2016. № 1 (36). С. 74–83.
9. Бісікало О. В., Висоцька В. А. Застосування методу синтаксичного аналізу речень для визначення ключових слів україномовного тексту. Радіоелектроніка. Інформатика. Управління. 2016. № 3 (38). С. 54–65.
10. Lytvyn V., Pukach P., Bobyk I., Vysotska V. The method of formation of the status of personality understanding based on the content analysis. Eastern-European Journal of Enterprise Technologies. 2016. Vol. 5. P. 4–12.
11. Литвин В. В., Бобик І. О., Висоцька В. А. Застосування системи алгоритмічних алгебр для граматичного аналізу символьних обчислень виразів логіки висловлювань. Радіоелектроніка. Інформатика. Управління. 2016. № 4 (39). С. 77–89.
12. Lytvyn V., Vysotska V., Pukach P., Bobyk I., Pakholok B. A method for constructing recruitment rules based on the analysis of a specialist's competences. Eastern-European Journal of Enterprise Technologies. 2016. Vol. 6/2 (84). P. 4–14.
13. Lytvyn V., Vysotska V., Pukach P., Brodyak O., Ugryn D. Development of a method for determining the keywords in the Slavic language texts based on the technology of web mining. Eastern-European Journal of Enterprise Technologies. 2017. Vol. 2/2 (86). P. 14–23.
14. Lytvyn V., Vysotska V., Pukach P., Vovk M., Ugryn D. Method of functioning of intelligent agents, designed to solve action planning problems based on ontological approach. Eastern-European Journal of Enterprise Technologies. 2017. Vol. 3/2 (87). P. 11–17.

15. Lytvyn V., Vysotska V., Pukach P., Bobyk I., Uhryn D. Development of a method for the recognition of author's style in the Ukrainian language texts based on linguometry, stylemetry. Eastern-European Journal of Enterprise Technologies. 2017. Vol. 4/2 (88). P. 10–18.
16. Коробчинський М. В., Чирун Л. Б., Висоцька В. А., Нич М. О. Особливості прогнозування результатів матчів у кіберспорті. Радіоелектроніка. Інформатика. Управління. 2017. № 3. С. 95–105.
17. Коробчинський М. В., Чирун Л. Б., Висоцька В. А., Кондратьєв Є. О. Особливості формування та аналізу контенту інтернет-газети музичних новин. Радіоелектроніка. Інформатика. Управління. 2017. № 4. С. 139–150.
18. Lytvyn V., Vysotska V., Uhryn D., Hrendus M., Naum O. Analysis of statistical methods for stable combinations determination of keywords identification. Eastern-European Journal of Enterprise Technologies. 2018. Vol. 2/2 (92). P. 23–37.
19. Lytvyn V., Vysotska V., Maria H. Method of data expression from the Ukrainian content based on the ontological approach. Радіоелектроніка. Інформатика. Управління. 2018. № 3 (46). P. 144–157.
20. Lytvyn V., Vysotska V., Pukach P., Nytrebych Z., Demkiv I., Kovalchuk R., Huzyk N. Development of the linguometric method for automatic identification of the author of text content based on statistical analysis of language diversity coefficients. Eastern-European Journal of Enterprise Technologies. 2018. Vol. 5. P. 16–28.
21. Lytvyn V., Vysotska V., Kuchkovskiy V., Pelekh I., Bobyk I., Malanchuk O., Ryshkovets Y., Brodyak O., Bobrivetc V., Panasyuk V. Development of the system to integrate and generate content considering the cryptocurrent needs of users. Eastern-European Journal of Enterprise Technologies. 2019. Vol. 1/2. P. 18–39.
22. Lytvyn V., Vysotska V., Pukach P., Nytrebych Z., Demkiv I., Senyk A., Malanchuk O., Sachenko S., Kovalchuk R., Huzyk N. Analysis of the developed quantitative method for automatic attribution of scientific and technical text content written in Ukrainian. Eastern-European Journal of Enterprise Technologies. 2018. Vol. 6/2 (96). P. 19–31.
23. Berko A., Vysotska V., Lytvyn V., Naum O. Planning the activities of intellectual agents in the electronic commerce systems. Радіоелектроніка. Інформатика. Управління. 2018. № 4. С. 143–158.
24. Lytvyn V., Vysotska V., Demchuk A., Demkiv I., Ukhans'ka O., Hladun V., Kovalchuk R., Petruchenko O., Dzyubyk L., Sokulska N. Design of the architecture of an intelligent system for distributing commercial content in the internet space based on SEO-technologies, neural networks, and machine learning. Eastern-European Journal of Enterprise Technologies. 2019. Vol. 2/2(98). P. 15–34.
25. Lytvyn V., Vysotska V., Shatskykh V., Kohut I., Petruchenko O., Dzyubyk L., Bobrivetc V., Panasyuk V., Sachenko S., Komar M. Design of a recommendation system based on collaborative filtering and machine learning considering personal needs of the user. Eastern-European Journal of Enterprise Technologies. 2019. Vol. 4/2 (100). P. 6–28.
26. Vysotska V., Demchuk A., Lytvyn V. Features of the architecture for Internet commercial content management system based on methods of Machine Learning, Web mining and SEO technologies. Радіоелектроніка. Інформатика. Управління. 2019. № 4. С. 121–135.
27. Lytvyn V., Vysotska V., Budz I., Pelekh Y., Sokulska N., Kovalchuk R., Dzyubyk L., Tereshchuk O., Komar M. Development of the quantitative method for automated text content authorship attribution based on the statistical analysis of N-grams distribution. Eastern-European Journal of Enterprise Technologies. 2019. Vol. 6/2 (102). P. 28–51.
28. Кравець П., Литвин В., Висоцька В. Ігрова модель онтологічної підтримки проектів. Радіоелектроніка, інформатика, управління. 2021. № 1(56). С. 172–183.
29. Литвин В. В., Бублик М. І., Висоцька В. А., Мацелюх Ю. Р. Технологія візуальної симуляції пасажиропотоків у сфері громадського транспорту smart city. Радіоелектроніка, інформатика, управління. 2021 № 4 (59). С. 106-121.

30. Кравець П. О., Литвин В. В., Висоцька В. А. Моделювання ігрової задачі призначення персоналу для виконання ІТ-проектів на основі онтологій. Радіоелектроніка, інформатика, управління. 2022. № 1 (60). С. 130–145.
31. Lytvyn V., Vysotska V., Veres O., Rishnyak I., Rishnyak H. Classification methods of text documents using ontology based approach. Advances in Intelligent Systems and Computing. 2017. Vol. 512. P. 229–240.
32. Shakhovska N., Vysotska V., Chyrun L. Intelligent systems design of distance learning realization for modern youth promotion and involvement in independent scientific researches. Advances in Intelligent Systems and Computing. 2017. Vol. 512. P. 175–198. ISSN 2194-5357.
33. Lytvyn V., Vysotska V., Burov Y., Veres O., Rishnyak I. The contextual search method based on domain thesaurus. Advances in Intelligent Systems and Computing. 2018. Vol. 689. P. 310–319.ISSN 2194-5357.
34. Kanishcheva O., Vysotska V., Chyrun L., Gozhyj A. Method of integration and content management of the information resources network. Advances in Intelligent Systems and Computing. 2018. Vol. 689. P. 204–216.
35. Vysotska V., Fernandes B. V., Emmerich M. Web content support method in electronic business systems. CEUR Workshop Proceedings. 2018. Vol. 2136. P. 20–41. E-ISSN: 1613-0073.
36. Lytvyn V., Vysotska V., Dosyn D., Y Burov. Method for ontology content and structure optimization, provided by a weighted conceptual graph. Webology. 2018. Vol. 15(2). P. 66–85. E-ISSN: 1735-188X.
37. Lytvyn V., Vysotska V., Osypov M., Slyusarchuk O., Y Slyusarchuk. Development of intellectual system for data de-duplication and distribution in cloud storage. Webology. 2019. Vol. 16(2). P. 1-42.
38. Vysotska V., Lytvyn V., Burov Y., Gozhyj A., Makara S. The consolidated information web-resource about pharmacy networks in city. CEUR Workshop Proceedings. 2018. Vol. 2255. P. 239–255.
39. Lytvyn V., Sharonova N., Hamon T., Vysotska V., Grabar N., Kowalska-Styczen A. Computational linguistics and intelligent systems. CEUR Workshop Proceedings. 2018. Vol. 2136. 390 p.
40. Rusyn B., Lytvyn V., Vysotska V., Emmerich M., Pohreliuk L. The virtual library system design and development. Advances in Intelligent Systems and Computing (AISC). 2019. Vol. 871. P. 328–349.
41. Vysotska V., Fernandes B. V., Lytvyn V., Emmerich M., Hirnyak M. Method for determining linguometric coefficient dynamics of Ukrainian text content authorship. Advances in Intelligent Systems and Computing (AISC). 2019. Vol. 871. P. 132–151. ISSN 2194-5357.
42. Gozhyj A., V Vysotska., Yevseyeva I., Kalinina I., Gozhyj V. Web resources management method based on intelligent technologies. Advances in Intelligent Systems and Computing. 2019. Vol. 871. P. 206–221.
43. Vysotska V., Lytvyn V., Burov Y., Berezin P., Emmerich M., Fernandes B. V. Development of information system for textual content categorizing based on ontology. CEUR Workshop Proceedings. 2019. V. 2362. P. 53–70.
44. Burov Y., Vysotska V., Kravets P. Ontological approach to plot analysis and modeling. CEUR Workshop Proceedings. 2019. Vol. 2362. P. 22–31. E-ISSN: 1613-0073.
45. Zdebskyi P., Vysotska V., Peleshchak R., Peleshchak I., Demchuk A., Krylyshyn M. An application development for recognizing of view in order to control the mouse pointer. CEUR Workshop Proceedings. 2019. Vol. 2386. P. 55–74. E-ISSN: 1613-0073.
46. Lytvyn V., Vysotska V., Rusyn B., Pohreliuk L., Berezin P., Naum O. Textual content categorizing technology development based on ontology. CEUR Workshop Proceedings. 2019. Vol. 2386. P. 234–254.
47. Lytvyn V., Vysotska V., Rzheuskyi A. Technology for the psychological portraits formation of social networks users for the IT specialists recruitment based on Big Five, NLP and Big Data. CEUR Workshop Proceedings. 2019. M2392. P. 147–171. E-ISSN: 1613-0073.

48. Vysotska V., Burov Y., Lytvyn V., Oleshek O. Automated monitoring of changes in web resources. Advances in Intelligent Systems and Computing. 2020. Vol. 1020. P. 348–363. ISSN 2194-5357, E-ISSN: 2194-5365.
49. Demchuk A., Lytvyn V., Vysotska V., Dilai M. Methods and means of web content personalization for commercial information products distribution. Advances in Intelligent Systems and Computing. 2020. Vol. 1020. P. 332–347. ISSN 2194-5357, E-ISSN: 2194-5365.
50. Lytvyn V., Vysotska V. Mykhailyshyn V. Rzheuskyi A. Semianchuk S. System development for video stream data analyzing. Advances in Intelligent Systems and Computing. 2020. Vol. 1020. P. 315–331.
51. Kravets P., Burov Y., Lytvyn V., Vysotska V. Gaming method of ontology clusterization. Webology. 2019. Vol. 16(1). P. 55–76. ISSN: 1735-188X.
52. Chyrun L., Leshchynskyy E., Lytvyn V., Rzheuskyi A., Vysotska V., Borzov Y. Intellectual analysis of making decisions tree in information systems of screening observation. CEUR Workshop Proceedings. 2019. Vol. 2488. P. 281–296. E-ISSN: 1613-0073.
53. Lytvyn V., Kowalska A., Peleshko D., Rak T., Voloshyn V., Noennig J., Vysotska V., Nykolyshyn L., Pryshchepa H. Aviation aircraft planning system project development. AISC. 2020. Vol. 1080. P. 315–348.
54. Lytvyn V., Burov Y., Kravets P., Vysotska V., Demchuk A., Berko A., Ryshkovets Y., Shcherbak S., Naum O. Methods and models of intellectual processing of texts for building ontologies of software for medical terms identification in content classification. CEUR Workshop Proceedings. 2019. Vol. 2488. P. 354–368.
55. Lytvyn V., Vysotska V., Shakhovska N., Mykhailyshyn V., Medykovskyy M., Peleshchak I., Fernandes V.B., Peleshchak R., Shcherbak S. A smart home system development. Advances in Intelligent Systems and Computing. 2020. Vol. 1080. P. 804–830. ISSN 2194-5357, E-ISSN: 2194-5365.
56. Lytvyn V., Gozhyj A., I Kalinina., Vysotska V., Shatskykh V., Chyrun L., Borzov Y. An intelligent system of the content relevance at the example of films according to user needs. CEUR Workshop Proceedings. 2019. Vol. 2516. P. 1–23. E-ISSN: 1613-0073.
57. Peleshko D., Rak T., Noennig J. R., Lytvyn V., Vysotska V. Drone monitoring system DROMOS of urban environmental dynamics. CEUR Workshop Proceedings. 2020. Vol. 2565. P. 178–193. E-ISSN: 1613-0073.
58. Krislata I., Katrenko A., Lytvyn V., Vysotska V., Burov Y. Traffic flows system development for smart city. CEUR Workshop Proceedings. 2020. Vol. 2565. P. 280–294. E-ISSN: 1613-0073.
59. Bisikalo O., Vysotska V. Linguistic analysis method of Ukrainian commercial textual content. CEUR Workshop Proceedings. 2020. Vol. 2608. P. 224–244. E-ISSN: 1613-0073.
60. Bisikalo O., Vysotska V., Kravets Y., Burov P. Conceptual model of process formation for the semantics of sentence in natural language. CEUR Workshop Proceedings. 2020. Vol. 2604. P. 151–177.
61. Vysotska V. Ukrainian participles formation by the generative grammars use. CEUR Workshop Proceedings. 2020. Vol. 2604. P. 407–427. E-ISSN: 1613-0073.
62. Batiuk T., Vysotska V., Lytvyn V. Intelligent system for socialization by personal interests on the basis of SEO technologies and methods of machine learning. CEUR Workshop Proceedings. 2020. V. 2604. P. 1237–1250.
63. Oliinyk V., Vysotska V., Burov Y., Mykich K., Basto-Fernandes V. Propaganda detection in text data based on NLP and machine learning. CEUR Workshop Proceedings. 2020. Vol. 2631. P. 132–144.
64. Kalinina I., Vysotska V., Sachenko S., Kovalchuk R., Gozhyj A. Qualitative and quantitative characteristics analysis for information security risk assessment in e-commerce systems. CEUR Workshop Proceedings. 2020. Vol. 2762. P. 177–190. E-ISSN: 1613-0073.

65. Lytvyn V., Hryhorovych A., Hryhorovych V., Vysotska V., Bublyk M., Chyrun L. Medical content processing in intelligent system of district therapist. CEUR Workshop Proceedings. 2020. V. 2753. P. 415–429.
66. Bublyk M., Lytvyn V., Vysotska V., Sokulska N., Chyrun L. Matseliukh Y. The decision tree usage for the results analysis of the psychophysiological testing. CEUR Workshop Proceedings. 2020. Vol. 2753. P. 458–472.
67. Vysotska V., Bublyk M., Korolenko O., Matseliukh Y., Kopach T. Network modelling of resource consumption intensities in human capital management in digital business enterprises. CEUR Workshop Proceedings. 2021. Vol. 2851. P. 366–380. E-ISSN: 1613-0073.
68. Kravets P., Lytvyn V., Vysotska V., Burov Y., Andrusyak I. Game task of ontological project coverage /. CEUR Workshop Proceedings. 2021. Vol. 2851. P. 344–355. E-ISSN: 1613-0073.
69. Kravets P., Burov Y., Oborska O., Vysotska V., Dzyubyk L., Lytvyn V. Stochastic Game Model of Data Clustering. CEUR Workshop Proceedings. 2021. Vol. 2853. P. 198–213. E-ISSN: 1613-0073.
70. Bublyk M., Vysotska V., Panasyuk V., Brodyak O., Chyrun L. Assessing security risks method in e-commerce system for IT portfolio management. CEUR Workshop Proceedings. 2021. Vol. 2853. P. 462–479.
71. Vysotska V., Lytvyn V., Danylyk V. Vyshemyrska S. Lurie I. Luchkevych M. Detecting items with the biggest weight based on neural network and machine learning methods. Communications in Computer and Information Science.2020.V.1158.P.383-396.ISSN1865-0929.
72. Kalinina I., Vysotska V., Bidyuk P., Gozhyj A. Methods for forecasting nonlinear non-stationary processes in machine learning. Communications in Computer and Information Science. 2020. Vol. 1158. P. 470–485.
73. Matseliukh Y., Bublyk M., Vysotska V. Development of intelligent system for visual passenger flows simulation of public transport in Smart City based on neural network. CEUR Workshop Proceedings. 2021. Vol. 2870. P. 1087–1138. E-ISSN: 1613-0073.
74. Pashchetnyk O., Lytvyn V., Zhyvchuk V., Polishchuk L., Vysotska V., Rybchak Z., Pukach Y. The ontological decision support system composition and structure determination for commanders of Land Forces formations and units in Ukrainian Armed Force. CEUR Workshop Proceedings. 2021. Vol. 2870. P. 1077–1086.
75. Lytvyn V., Pashchetnyk O., Klymovych O., Polishchuk L., Kolb I., Burov Y., Vysotska V. Assessment of the hydro-meteorological conditions impact on the combat troops operations preparation and conduct in the geo-information subsystem of the automated battlefield system. CEUR Workshop Proceedings. 2021. Vol. 2870. P. 1063–1076. E-ISSN: 1613-0073.
76. Tymoshenko K., Vysotska V., Kovtun O., Holoshchuk R., Holoshchuk S. Real-time Ukrainian text recognition and voicing. CEUR Workshop Proceedings. 2021. Vol. 2870. P. 357–387. E-ISSN: 1613-0073.
77. Vysotska V., Holoshchuk S., Holoshchuk R. A comparative analysis for English and Ukrainian texts processing based on semantics and syntax approach. CEUR Workshop Proceedings. 2021. Vol. 2870. P. 311–356.
78. Dokhnyak B., Vysotska V. Intelligent Smart Home System Using Amazon Alexa Tools. CEUR Workshop Proceedings. 2021. Vol. 2917. P. 441-464. E-ISSN: 1613-0073.
79. Zdorenko Y., Lavrut O., T Lavrut., V Lytvyn., Burov Y., Vysotska V. Route Selection Method in Military Information and Telecommunication Networks Based on ANFIS. CEUR Workshop Proceedings. 2021. Vol. 2917. P. 514-524. E-ISSN: 1613-0073.
80. Balush I., Vysotska V., Albota S. Recommendation System Development Based on Intelligent Search, NLP and Machine Learning Methods. CEUR Workshop Proceedings. 2021. Vol. 2917. P. 584-617.

81. Kholodna N., Vysotska V., Albota S. A Machine Learning Model for Automatic Emotion Detection from Speech. CEUR Workshop Proceedings. 2021. Vol. 2917. P. 699-713. E-ISSN: 1613-0073.
82. Kravets P., Lytvyn V., Dobrotvor I., Sachenko O., Vysotska V., Sachenko A. Matrix Stochastic Game with Q-learning for Multi-agent Systems. Lecture Notes on Data Engineering and Communications Technologies. 2021. Vol. 83. P. 304–314. ISSN 23674512.
83. Kravets P., Burov Y., Lytvyn V., Vysotska V., Ryshkovets Y., Brodyak O., Vyshemyrska S. Markovian Learning Methods in Decision-Making Systems. Lecture Notes on Data Engineering and Communications Technologies. 2022. Vol. 77. P. 423-437. ISSN 23674512.
84. Vysotska V., Berko A., Lytvyn V., Kravets P., Dzyubyk L., Bardachov Y., Vyshemyrska S. Information resource management technology based on fuzzy logic. AISC. 2020. Vol. 1246. P. 164–182.
85. Kravets P., Lytvyn V., Vysotska V., Ryshkovets Y., Vyshemyrska S., Smailova S. Dynamic coordination of strategies for multi-agent systems. Advances in Intelligent Systems and Computing. 2020. Vol. 1246. P. 653–670.

Статті у наукових фахових виданнях України

86. Алєксєєва К. А., Берко А. Ю., Висоцька В. А. Управління Web-ресурсами за умов невизначеності. Технологічний аудит та резерви виробництва. 2015. № 2 (2). С. 4–7.
87. Висоцька В. А. Гопяк М. В., Козлов П. Ю. Особливості технології управління web-ресурсом. Інженерія програмного забезпечення. 2015. № 1 (21). С. 25–35.
88. Висоцька В. А., Чирун Л. В. Формальна модель опрацювання інформаційних ресурсів в системах електронної контент-комерції. Вісник НУ «Львівська політехніка». 2015. № 814. С. 44–54.
89. Берко А. Ю., Висоцька В. А., Чирун Л. В. Лінгвістичний аналіз текстового комерційного контенту. Вісник НУ «Львівська політехніка». 2015. № 814. С. 203–227.
90. Висоцька В. А. Особливості моделювання синтаксису речення слов'янських та германських мов за допомогою породжувальних контекстно-вільних грамматик. Вісник НУ «Львівська політехніка». 2015. № 814. С. 246–276.
91. Кісь Я. П., Висоцька В. А., Чирун Л. Б., Фольтович В. М. Застосування контент-аналізу для опрацювання текстових масивів даних. Вісник НУ «Львівська політехніка». 2015. № 814. С. 282–292.
92. Шестакевич Т. В., Висоцька В. А., Чирун Л. В., Чирун Л. Б. Моделювання семантики речення природною мовою за допомогою породжувальних граматик. Вісник НУ «Львівська політехніка». 2015. № 814. С. 335–352.
93. Висоцька В. А. Нога А. Ю., Козлов П. Ю. Управління Web-проектами електронного бізнесу для реалізації комерційного контенту. Вісник НУ «Львівська політехніка». 2015. № 814. С. 421–434.
94. Висоцька В. А., Чирун Л. В. Концептуальна модель опрацювання інформаційних ресурсів в системах електронної контент-комерції. Математичні машини і системи. 2015. № 3. С. 179–190.
95. Висоцька В. А., Чирун Л. В. Опрацювання інформаційних ресурсів у системах електронної контент-комерції. Відбір і обробка інформації. 2015. Вип. 42 (118). С. 84–92.
96. Алєксєєва К. А., Берко А. Ю., Висоцька В. А. Особливості процесу управління web-ресурсом комерційного контенту на основі нечіткої логіки. Вісник НУ «Львівська політехніка». 2015. № 826. С. 201–211.
97. Висоцька В. А. Особливості рубрикації текстового комерційного контенту. Вісник НУ «Львівська політехніка». 2015. № 826. С. 359–367.
98. Алєксєєва К. А., Берко А. Ю., Висоцька В. А. Інформаційна технологія управління Web-ресурсом на основі нечіткої логіки. Вісник НУ «Львівська політехніка». 2015. № 829. С. 7–28.

99. Висоцька В. А. Аналітичні методи опрацювання інформаційних ресурсів в системах електронної контент-комерції. Вісник НУ «Львівська політехніка». 2015. № 829. С. 76–101.
100. Гасько Р. В., Висоцька В. А., Чирун Л. Б. Інформаційна система аналізу психологічного стану особистості. Вісник НУ «Львівська політехніка». 2015. № 829. С. 102–128.
101. Бісікало О. В., Висоцька В. А. Експериментальне дослідження пошуку значущих ключових слів україномовного контенту. Вісник НУ «Львівська політехніка». 2015. № 829. С. 255–272.
102. Чирун Л. Б., Кучковський В. В., Висоцька В. А. Особливості методів контент-аналізу текстових масивів даних web-ресурсів в межах регіону контенту. Вісник НУ «Львівська політехніка». 2015. № 829. С. 296–320.
103. Андруник В. А., Висоцька В. А., Чирун Л. Б. Проект розроблення та впровадження системи електронної контент-комерції. Вісник НУ «Львівська політехніка». 2015. № 829. С. 321–348.
104. Козлов П. Ю., Висоцька В. А., Чирун Л. Б. Сучасні технології управління Web-ресурсами в інформаційній системі аналізу сервісу цифрової дистрибуції. Вісник НУ «Львівська політехніка». 015. № 832. С. 103–128.
105. Кучковський В. В., Висоцька В. А., Нитребич С. З., Оливко Р. М. Застосування методів Інтернет-маркетингу для аналізу Web-ресурсів в межах регіону. Вісник НУ «Львівська політехніка». 2015. № 832. С. 129–164.
106. Шаховська Н. Б., Висоцька В. А., Чирун Л. Б. Методи та засоби дистанційної освіти для заохочення і залучення сучасної молоді до проведення самостійних наукових досліджень. Вісник НУ «Львівська політехніка». 2015. № 832. С. 254–284.
107. Литвин В. В., Висоцька В. А., Досин Д. Г., Гірняк М. Г. Розроблення методів та засобів побудови інтелектуальних систем опрацювання інформаційних ресурсів з використанням онтологічного підходу. Вісник НУ «Львівська політехніка». 2015. № 832. С. 295–314.
108. Алєксєєва К. А., Берко А. Ю., Висоцька В. А. Аналіз процесу опрацювання web-ресурсу інформаційного продукту на основі нечіткої логіки та контент-аналізу. Вісник НУ «Львівська політехніка». 2016. № 843. С. 122–134.
109. Андруник В. А. Висоцька В. А., Чирун Л. В. Особливості формування електронних дайджестів. Вісник НУ «Львівська політехніка». 2016. № 843. С. 3–14.
110. Бісікало О. В., Висоцька В. А. Метод лінгвістичного аналізу україномовного комерційного контенту. Вісник НУ «Львівська політехніка». 2016. № 854. С. 185–204.
111. Вінтоняк С. М., Коробчинський М. В., Чирун Л. Б., Висоцька В. А. Аналіз особливостей Інтернет-порталу аматорських спортивних ігор. Вісник НУ «Львівська політехніка». 2016. № 854. С. 21–41.
112. Vysotska V., Chyrun L., Chyrun L. Online newspaper content analysis based on SEO technologies. Вісник НУ «Львівська політехніка». 2016. № 859. С. 3–16.
113. Литвин В. В., Ремешило-Рибчинська О. І., Висоцька В. А. Побудова онтології архітектурних термінів. Відбір і обробка інформації. 2017. Вип. 44 (120). С. 90–96.
114. Фольтович В. М., Коробчинський М. В., Чирун Л. Б., Висоцька В. А. Метод контент-аналізу текстової інформації Інтернет газети. Вісник НУ «Львівська політехніка». 2017. № 864. С. 7–19.
115. Гасько Р. В., Чирун Л. В., Висоцька В. А. Особливості контент-аналізу користувацької Інтернет-діяльності для формування зрізу психологічного стану особистості. Вісник НУ «Львівська політехніка». 2017. № 864. С. 221–238.
116. Lytvyn V., Vysotska V., Veres O., Brodyak O., Oryshchyn O. Big Data analytics ontology. Технологічний аудит та резерви виробництва. 2018. Vol. 1, № 2(39). С. 16–27.

117. Висоцька В. А., Наум О. М. Порівняння складності автоматичного опрацювання англійських та українських текстів з врахуванням семантики та синтаксису природних мов. Вісник НУ «Львівська політехніка». 2017. № 872. С. 149–162.
118. Шаховська Н. Б. Висоцька В. А., Скотар О. О. Розроблення архітектури інтелектуальної системи на основі інноваційних методів навчання студентів. Вісник НУ «Львівська політехніка». 2017. № 872. С. 220–229.
119. Русин Б., Висоцька В., Погрелюк Л. Особливості проектування та розроблення інформаційної системи Virtual Library. Оптико-електронні інформаційно-енергетичні технології. 2017. Т. 34, № 2. С. 18–33.
120. Литвин В. В., Висоцька В. А., Кучковський В. В., Дуткевич С. Ю., Наум О. Метод інтеграції та управління контентом мережі інформаційних ресурсів туризму згідно з потребами користувача. Вісник НУ «Львівська політехніка». 2018. № 901. С. 22–36.
121. Литвин В. В., Висоцька В. А., Кучковський В. В., Оливко Р. М. Архітектура інформаційної системи інтеграції та формування контенту про криптовалюти на основі аналізу бірж. Вісник НУ «Львівська політехніка». 2018. № 901. С. 43–60.
122. Русин Б. П., Погрелюк Л. В., Висоцька В. А., Осипов М. М., Варецький Я. Ю., Капшій О. В. Архітектура системи дедублікації та розподілу даних у хмарних сховищах під час резервного копіювання. Інформаційні технології та комп'ютерна інженерія. 2019. Т. 2, № 45. С. 40–63.
123. Литвин В. В., Наум О., Висоцька В. А., Дверій М. В. Архітектура системи онлайн-туризму для пошуку та планування подорожей із урахуванням потреб користувача. Вісник НУ «Львівська політехніка». 2019. Вип. 6. С. 13–29.
124. Русин Б., Погрелюк Л., Висоцька В., Осипов М. Метод дедублікації та розподілу даних у хмарних сховищах під час резервного копіювання даних Вісник НУ «Львівська політехніка». 2019. № 6. С. 1–12.
125. Пелещак Р. М., Литвин В. В., Пелещак І. Р., Висоцька В. А. Розробка штучної нейронної мережі з осциляторними нейронами для розпізнавання спектральних образів. Вісник НУ «Львівська політехніка». 2020. Вип. 7. С. 16–23.
126. Пелещак Р. М., Литвин В. В., Пелещак І. Р., Висоцька В. А., Черняк О. І. Побудова оптимізованої багатошарової нейронної мережі в межах нелінійної моделі узагальненої похибки. Вісник НУ «Львівська політехніка». 2021. Вип. 9. С. 53–60.
127. Батюк Т.М., Висоцька В. А. Розробка інтелектуальної системи підтримки соціалізації користувача за подібністю інтересівю Сучасний стан наукових досліджень та технологій в промисловості. 2022. Вип. 1(19). С. 13-26.
128. Batiuk T., Vysotska V. Decision-making support system to support of social networks users based similar common interests and preferences. Computer systems and information technologies. 2022. Vol. 1. P. 11-22.
129. Батюк Т.М., Висоцька В. А. Інформаційна підтримка процесів соціалізації особистості на основі інтересів. Вісник НУ «Львівська політехніка». 2022. №11. С. 56 – 86.
130. Олексів Н., Висоцька В. Мобільна інформаційна система контролю раціону харчування людини. Вісник НУ «Львівська політехніка». 2022. Вип. 11. С. 145 – 172.

Монографії

131. Lytvyn V., Vysotska V., Chyrun L., Dosyn D. Methods based on ontologies for information resources processing. Saarbrücken: LAP Lambert Academic Publishing, 2016. 324 p.

132. Литвин В. В., Висоцька В. А., Досин Д. Г. Методи та засоби опрацювання інформаційних ресурсів на основі онтологій. Львів: Піраміда, 2016. 404 с.
133. Висоцька В. А. Технології електронної комерції та Інтернет-маркетингу. Saarbrücken: LAP Lambert Academic Publishing, 2018. 285 с.
134. Vysotska V., Lytvyn V. Web resources processing based on ontologies. Saarbrücken: LAP Lambert Academic Publishing, 2018. 232 p.
135. Vysotska V., Shakhovska N. Information technologies of gamification for training and recruitment. Saarbrücken: LAP Lambert Academic Publishing, 2018. 248 p.
136. Vysotska V Internet systems design and development based on Web Mining and NLP. Saarbrücken: LAP Lambert Academic Publishing, 2018. 316 p.
137. Vysotska V. Computer linguistics for online marketing in information technology. Saarbrücken: LAP Lambert Academic Publishing, 2018. 396 p.
138. Висоцька В. А., Досин Д. Г., Микіч Х. І., Завущак І. І., Рибчак З. Л. Методи та засоби функціонування систем підтримки прийняття рішень на основі онтологій. Львів: Новий світ, 2019. 334 с.
139. Peleshchak R., Peleshchak I., Vysotska V. Methods for recognizing multispectral images based on neural networks. Beau Bassin: Lambert Academic Publishing, 2020. 153 p.

Статті у міжнародних виданнях

140. Vysotska V., Chyrun L. Linguistic analysis and modelling semantics of textual content for digest formation. MEST Journal. 2015. Vol. 3, № 1. P. 127–148. ISSN: 2334-7171.
141. Chyrun L., Vysotska V. Features of the content-analysis method for text categorization of commercial content in processing online newspaper articles. Applied Computer Science. 2015. Vol. 11, № 1. P. 15–30.
142. Chyrun L., Andrunyk V., Vysotska V. Electronic content commerce system development. MEST Journal. 2015. Vol. 3, № 2. P. 10–33. ISSN: 2334-7171.
143. Vysotska V., Chyrun L. The means structure of information resources processing in electronic content commerce systems. Journal of Information Sciences and Computing Technologies. 2015. Vol 3, №3. P. 241–248.
144. Vysotska V., Chyrun L. Methods and means of processing information resources in electronic content commerce systems. Applied Computer Science. 2015. Vol. 11(2). P. 68–85. ISSN: 2353-6977.
145. Chyrun L. Vysotska V., Laba R. Information resources analysis in electronic content commerce systems. Applied Computer Science. 2016. Vol. 12(1). P. 48–66. ISSN: 2353-6977.
146. Vysotska V. Chyrun L., Kozlov P. Analysis of business processes in electronic content-commerce systems. Econtechmod. 2016. Vol. 5, № 1. P. 111–125. ISSN: 2084-5715.
147. Vysotska V., Chyrun L., Kozlov P. Design and analysis features of generalized electronic content-commerce systems architecture. Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska. 2016. №6. P. 48–59.
148. Chyrun L., Vysotska V., Kozak I. Informational resources processing intellectual systems with textual commercial content linguistic analysis usage constructional means and tools development. Econtechmod. 2016. Vol. 5, № 2. P. 85–94. ISSN: 2084-5715.
149. Chyrun L., Andrunyk V., Vysotska V. Content analysis peculiarities of user internet activities for personality psychological state slice formation. MEST Journal. 2017. Vol. 6, № 2. P. 26–46. ISSN: 2334-7171.
150. Lytvyn V., Vysotska V., Veres O. Ontology of big data analytics. MEST Journal. 2018. Vol. 6(1). P. 41–60.

151. Lytvyn V., Vysotska V., Bublyk M., Nanivskyi R., Grudowski P., Matseliukh Y. Developing methods for building intelligent systems of information resources processing using an ontological approach. AISC. 2021. Vol. 1293. P. 345–370. ISSN 2194-5357.
152. Bisikalo O., V Vysotska., Lytvyn V., Brodyak O., Vyshemyrska S., Rozov Y. Experimental investigation of significant keywords search in Ukrainian content. AISC. 2021. Vol. 1293. P. 3–29.
153. Burov Y., Horodetska A., Bublyk M., Nashkerska M., Vysotska V. Intellectual Tourist Service with the Situation Context Processing. Advances in Social Science, Education and Humanities Research. 2021. Vol. 557. P. 233-243. ISSN (Online): 2352-5398.

Статті у міжнародних конференціях, які індексуються у Scopus та Web of Science

154. Alieksieieva K., Berko A., Vysotska V. Technology of commercial web-resource processing. CADSM : XIII Міжнар. наук.-техн. конф., 24–27 лют., Львів, Поляна, 2015. С. 340–344.
155. Andrunyk V., Chyrun L., Vysotska L. Electronic content commerce system development. CADSM : матер. XIII Міжнар. наук.-техн. конф., 24–27 лют., Львів, Поляна, 2015. С. 434–438.
156. Vysotska V. Chyrun L. Methods of information resources processing in electronic content commerce systems. CADSM: XIII Міжн. наук.-техн. конф., 24–27 лют., Львів, Поляна, 2015. С. 328–332.
157. Vysotska V., Chyrun L. Analysis features of information resources processing. CSIT : proc. of the Xth Intern. conf., 14–17 Sept., Lviv, Ukraine, 2015. P. 124–128.
158. Lytvyn V., Vysotska V. Designing architecture of electronic content commerce system. CSIT : proc. of the Xth Intern. conf., 14–17 Sept., Lviv, Ukraine, 2015. P. 115–119.
159. Vysotska V., Hasko R., Kuchkovskiy V. Process analysis in electronic content commerce system. CSIT: proc. of the X Intern. conf., 14–17 Sept., Lviv, Ukraine. 2015. P. 120–123.
160. Vysotska V. Linguistic analysis of textual commercial content for information resources processing. TCSET : proc. of the XIII Intern. conf, Feb. 23–26, Lviv, Slavske, Ukraine, 2016. P. 709–713.
161. Lytvyn V., Vysotska V., Veres O., Rishnyak I., Rishnyak H. Content linguistic analysis methods for textual documents classification. CSIT: proc. of the XIth Intern. conf., 6–10 Sept., Lviv. P. 190–192.
162. Vysotska V., Chyrun L., Chyrun L. The commercial content digest formation and distributional process. CSIT: proc. of the XIth Intern. conf. CSIT, 6–10 Sept., Lviv, Ukraine. 2016. P. 186–189.
163. Vysotska V., Chyrun L., Chyrun L. Information technology of processing information resources in electronic content commerce systems. Computer science and information technologies: proc. of the XIth Intern. conf., 6–10 Sept., Lviv, Ukraine. 2016. P. 212–222.
164. Shakhovska N., Vysotska V., Chyrun L. Features of e-learning realization using virtual research laboratory. CSIT: proc. of the XIth Intern. conf., 6–10 Sept., Ukraine. 2016. P. 143–148.
165. Lytvyn V., Vysotska V., Chyrun L., Chyrun L. Distance learning method for modern youth promotion and involvement in independent scientific researches. 1st IEEE International conference on data stream mining and processing, DSMP : proc. Aug. 23–27, Lviv, Ukraine. 2016. P. 269–274.
166. Lytvyn V., Vysotska V., Veres O., Rishnyak I., Rishnyak H. The risk management modelling in multi project environment. CSIT: proc. of the Intern. conf. 5–8 Sept., Lviv, Ukraine. 2017. P. 32–35.
167. Korobchinsky M., Vysotska V., Chyrun L., Chyrun L. Peculiarities of content forming and analysis in internet newspaper covering music news. Computer science and information technologies : proc. of the XIIth Intern. conf., 5–8 Sept., Lviv, Ukraine. 2017. P. 52–57.

168. Naum O., Chyrun L., Kanishcheva O., Vysotska V. Intellectual system design for content formation. CSIT: proc. of the XII Intern. conf., 5–8 Sept., Ukraine. 2016. P. 131–138.
169. Lytvyn V., Vysotska V., Dosyn D., Holoschuk R., Rybchak Z. Application of sentence parsing for determining keywords in Ukrainian texts. Computer science and information technologies : proc. of the XIIth Intern. conf., 5–8 Sept., Lviv, Ukraine. 2017. P. 326–331.
170. Su J., Vysotska V., Sachenko A., Lytvyn V., Burov Ye. Information resources processing using linguistic analysis of textual content. IDAACS : proc. conf., Bucharest, Sept. 21–23. 2017. P. 573–578.
171. Lytvyn V., Vysotska V., Burov Y., Demchuk A. Defining author's style for plagiarism detection in academic environment. DSMP : proc. of Intern. conf., Aug. 21–25, Lviv, Ukraine. 2018. P. 128–133.
172. Lytvyn V., Vysotska V., Lozynska O., Oborska O., Dosyn D. Methods of building intelligent decision support systems based on adaptive ontology. Data stream mining and processing : proc. of Intern. conf., Aug. 21–25, Lviv, Ukraine. 2018. P. 145–150.
173. Chyrun L., Vysotska V., Kis I., Chyrun L. Content analysis method for cut formation of human psychological stat. DSMP: proc. of Intern. conf., August 21–25, Lviv, Ukraine. 2018. P. 139–144.
174. Lytvyn V., Vysotska V., Burov Y., Bobyk I., Ohirko O. The linguometric approach for co-authoring author's style definition. IEEE IDAACS-SWS : proc., Lviv, 20–21 September 2018. P. 29–34.
175. Vysotska V., Lytvyn V., Hrendus M., Brodyak O., Kubinska S. Method of textual information authorship analysis based on stylometry. CSIT: proc. of Intern. conf., 11–14 вер., Львів. 2018. С. 9–16.
176. Vysotska V., Kanishcheva O., Hlavcheva Y. Authorship identification of the scientific text in Ukrainian with using the lingvometry methods. CSIT: proc. of the Intern. conf., 11 вересня 2018 р., Львів. 2018. С. 34–38.
177. Chyrun L., Kis I., Vysotska V., Chyrun L. Content monitoring method for cut formation of person psychological state in social scoring. CSIT: proc. of the Intern. conf., Львів, 11–14 вер. 2018 р., С. 106–112.
178. Su J., Lytvyn V., Vysotska V., Sachenko A., Dosyn D. Model of touristic information resources integration according to user needs. CSIT: proc. of Intern. conf., Львів, 11–14 вер., С. 113–116.
179. Rusyn B., Vysotska V., Pohreliuk L. Model and architecture for virtual library information system. CSIT: proc. of Intern. conf., Львів, 11–14 вер. 2018 р., С. 34–41.
180. Lytvyn V., Peleshchak I., Peleshchak R., Vysotska V. Satellite spectral information recognition based on the synthesis of modified dynamic neural networks and holographic data processing techniques. CSIT: proc. of Intern. conf., Львів,11–14 вер. 2018. Т. 1. С. 330–334.
181. Lytvyn V., Vysotska V., Burov Y., Demchuk A. Architectural ontology designed for intellectual analysis of e-tourism resources. CSIT: proc. of Intern. conf., Львів,11–14 вер. 2018. С. 335–338.
182. Gozhyj A., Kalinina I., Vysotska V., Gozhyj V. The method of web-resources management under conditions of uncertainty based on fuzzy logic. CSIT: proc. of the Intern. conf., Львів,11–14 вер. 2018. P. 343–346.
183. Lytvyn V., Kuchkovskiy V., Vysotska V., Markiv O., Pabyrivskyy V. Architecture of system for content integration and formation based on cryptographic consumer needs. Computer Sciences and Information Technologies (CSIT) : proc. of the IEEE Intern. conf., Львів,11–14 вер. 2018. С. 391–395.
184. Lytvyn V., Peleshchak I., Peleshchak R., Vysotska V. Information encryption based on the synthesis of a neural network and AES algorithm. Advanced information and communication technologies, AICT : proc. of the 3rd Intern. conf., Lviv, Ukraine, July 2–6 2019. P. 447–450.

185. Lytvyn V., Vysotska V., Mykhailyshyn V., Peleshchak I., Peleshchak R., Kohut I. Intelligent system of a smart house. AICT: proc. of the 3rd Inter. conf. (Lviv, Ukraine, July 2–6 2019. P. 282–287.

186. Vysotsky A., V Vysotska., Lytvyn V., Burov Y., Demchuk A., I Lyudkevych. Consolidated information web resource for online tourism based on data integration and geolocation. Computer Sciences and Information Technologies (CSIT): proc. of the IEEE Intern. conf., Львів, 17–20 вересня 2019. С. 15–20.

187. Lytvyn V., Vysotska V., Peleshchak I., Basyuk T., Kovalchuk V., Kubinska S., Chyrun L., Rusyn B., Pohreliuk L., Salo T. Identifying textual content based on thematic analysis of similar texts in big data . CSIT: proc. of Intern. conf., Львів, 17–20 вер. 2019. Т. 2. С. 84–91.

188. Vysotsky A., Lytvyn V., Vysotska V., Dosyn D., Lyudkevych I., Antonyuk N., Naum O., Vysotskyi A., Chyrun L., Slyusarchuk O. Online tourism system for proposals formation to user based on data integration from various sources. Computer Sciences and Information Technologies (CSIT) : proc. of the IEEE Intern. conf., Львів, 17–20 вер. 2019. С. 92–97.

189. Vysotska V., Lytvyn V., Kovalchuk V., Kubinska S., Dilai M., Rusyn B., Pohreliuk L., Chyrun L., Chyrun S., Brodyak O. Method of similar textual content selection based on thematic information retrieval. CSIT: proc. of Intern. conf., Львів, 17–20 вересня. 2019. Т. 3. С. 1–6.

190. Rzheuskyi A., Kutyuk O., Vysotska V., Burov Y., Lytvyn V., Chyrun L. The architecture of distant competencies analyzing system for IT recruitment. Computer Sciences and Information Technologies (CSIT): proc. of the IEEE Intern. conf., Львів, 17–20 вересня. 2019. Т. 3. С. 254–261.

191. Gozhyj A., Kalinina I., Gozhyj V., Vysotska V. Web service interaction modeling with colored petri nets. 10th IEEE IDAACS : proc., September 18–21, Metz, France. 2019. P. 319–323.

192. Shu C., Dosyn D., Lytvyn V., Vysotska V., Sachenko A., Jun S. Building of the predicate recognition system for the NLP ontology learning module. IDAACS : proc., September 18–21, Metz, France. 2019. P. 802–808.

193. Kalinina I., Vysotska V., Bidyuk P., Gozhyj A., Vasilev M., Malets R. Forecasting nonlinear nonstationary processes in machine learning task. DSMP: proc. of the 3rd inter. conf., Lviv, Ukraine. 2020. P. 28–32.

194. Lytvyn V., Vysotska V., Burov Y., Hryhorovych V. Knowledge novelty assessment during the automatic development of ontologies. DSMP: proc. of the Intern. conf., Lviv, Ukraine. 2020. P. 372–377.

195. Lytvyn V., Dosyn D., Vysotska V., Hryhorovych A. Method of ontology use in OODA. DSMP: proc. of the IEEE 3rd Intern. conf., Lviv, Ukraine. 2020. P. 409–413.

196. Vysotska V., Lytvyn V., Bublyk M., Demchuk A., Demkiv L., Shpak Y. Method of ontology quality assessment for knowledge base in intellectual systems based on ISO\IEC 25012. CSIT: proc. of Intern. conf., Збараж, 23–26 вересня, 2020. P. 109–113.

197. Vysotska V., Berko A., Bublyk M., Chyrun L., Vysotsky A., Doroshkevych K. Methods and tools for web resources processing in e-commercial content systems. Computer Sciences and Information Technologies (CSIT): proc. of the IEEE Intern. conf., Збараж, 23–26 вересня, 2020. P. 114–118.

198. Lytvyn V., Dosyn D., Vysotska V., Demchuk A., Demkiv L., Lytvyn I. Intellectual agent construction method based on the subject field ontology. Computer Sciences and Information Technologies (CSIT): proc. of the IEEE Intern. conf., Збараж, 23–26 вересня, 2020. P. 40-46.

199. Lytvyn V., Vysotska V., Burov Y., Brodyak O. Approach to automatic construction of interpretation functions during ontology learling. CSIT: proc. of Int. conf., Збараж, 23–26 вер., 2020. P. 267–271.

200. Burov Y., Lytvyn V., Vysotska V., I Shakleina. The basic ontology development process automation based on text resources analysis. CSIT: proc. of Int. conf., Збараж, 23–26 вер, 2020. P. 280–284.
201. Lytvyn V., Vovnyanka R., Oborska O., Dosyn D., Vysotska V., Panasyuk V. Intelligent agent behavior simulation based on reinforcement learning. Computer Sciences and Information Technologies (CSIT): proc. of the IEEE Intern. conf., Збараж, 23–26 вересня, 2020. P. 285–290.
202. Peleshchak R., Lytvyn V., Peleshchak I., Vysotska V. Stochastic Pseudo-Spin Neural Network with Tridiagonal Synaptic Connections. SIST, 28-30 April 2021, Nur-Sultan, Kazakhstan. Art. 9465998.
203. Lytvyn V., Bublyk M., Vysotska V., Panasyuk V., Brodyak O., Luchkevych M. Modelling of the Intelligent Agent's Behavior Scheduler Based on Petri Nets and Ontological Approach. SIST, 28-30 Apr. 2021, Nur-Sultan, Kazakhstan. Art. 9465994.
204. Lytvyn V., Y Burov., Vysotska V., Pukach Y., Tereshchuk O., Shakleina I. Abstracting Text Content Based on Weighing the TF-IDF Measure by the Subject Area Ontology. SIST, 28-30 April 2021, Nur-Sultan, Kazakhstan. Art. 9465978.
205. Tchynetskyi S., Peleshchak R., Peleshchak I., Vysotska V. A Neural Network Development for Multispectral Images Recognition. CSIT : proc. of the Intern. conf., 22-25 Sept., Lviv, Ukraine. 2021. P. 278–284.
206. Dmytriv A., Vysotska V., Bublyk M. The Speech Parts Identification for Ukrainian Words Based on VESUM and Horokh Using. Computer Sciences and Information Technologies (CSIT) : proc. of the IEEE 16th Intern. conf., 22-25 Sept., Lviv, Ukraine. 2021. Vol. 2. P. 21–33.
207. Kubinska S., Vysotska V., Matseliukh Y. User Mood Recognition and Further Dialog Support. CSIT: proc. of Intern. conf., 22-25 Sept., Lviv, Ukraine. 2021. Vol. 2. P. 34–39.
208. Ivanchyshyn D., Vysotska V., Albota S. The Film Script Generation Analysis Based on the Fiction Book Text Using Machine Learning. Computer Sciences and Information Technologies (CSIT) : proc. of the IEEE 16th Intern. conf., 22-25 Sept., Lviv, Ukraine. 2021. Vol. 2. P. 68–80.
209. Sartiukova A., Peleshchak R., Peleshchak I., Vysotska V. The Multiclass Classification of Objects Based on Multispectral Images Recognition. Computer Sciences and Information Technologies (CSIT): proc. of the IEEE 16th Intern. conf., 22-25 Sept., Lviv, Ukraine. 2021.Vol. 1. P. 52–60.
210. Voloshynskyi O., Vysotska V., Bublyk M. Cardiovascular Disease Prediction Based on Machine Learning Technology. CSIT: proc. of Intern. conf., 22-25 Sept., Lviv, Ukraine. 2021. Vol. 1. P. 69–75.
211. Aksonov D., Gozhyj A., Kalinina I., Vysotska V. Question-Answering Systems Development Based on Big Data Analysis. CSIT: proc. of Intern. conf., 22-25 Sept., Lviv, Ukraine. 2021. P. 113–118.
212. Mykytiuk A., Vysotska V., Albota S. Spam Filtration System with the Use of Machine Learning Technology. CSIT: proc. of Intern. conf., 22-25 Sept., Lviv, Ukraine. 2021. Vol. 1. P. 124–130.
213. Zanchak M., Vysotska V., Albota S. The Sarcasm Detection in News Headlines Based on Machine Learning Technology. CSIT: proc. of Intern. conf., 22-25 Sept., Lviv, Ukraine. 2021. P. 131–137.
214. Voloshyn S., Peleshchak R., Peleshchak I., Vysotska V. Big Data Analysis for Multispectral Images Recognition Based on Deep Learning. Computer Sciences and Information Technologies (CSIT): proc. of the IEEE 16th Intern. conf., 22-25 Sept., Lviv, Ukraine. 2021. Vol. 1. P. 160–170.
215. Lytvyn V., Vysotska V., Bublyk M., Gozhyj A., Schuchmann V. Solving Scheduling Issues Methods Analysis in Computational Grid. CSIT, 22-25 Sept., Lviv, Ukraine. 2021. Vol. 1. P. 267–273.

216. Kravets P., Lytvyn V., Burov Y., Vysotska V., Chyrun L., Panasyuk V. Making Optimal Decisions with Learning Method Based on Fuzzy Logic. Advanced Information and Communication Technologies (AICT): proc. of the IEEE 4th Intern. conf., 21-25 Sept., Lviv, Ukraine. 2021. P. 183–188.
217. Gozhyj A., Kalinina I., Nechakhin V., Gozhyj V., Vysotska V. Modeling an Intelligent Solar Power Plant Control System Using Colored Petri Nets. IDAACS, 22-25 Sept., Cracow, Poland. 2021. P. 626–631.
218. Peleshchak R., Lytvyn V., Kholodna N., Peleshchak I., Vysotska V. Two-Stage AES Encryption Method Based on Stochastic Error of a Neural Network. TCSET, Lviv-Slavske, Ukraine, Feb. 22 - 26, 2022.

Статті та тези доповідей у збірниках праць конференцій

219. Козлов П., Висоцька В. Особливості технології управління web-ресурсом. V Міжн. наук.-практ. конф. «Обробка сигналів і негаусівських процесів», 20-22 травня, 2015, Черкаси. С. 38–40.
220. Козлов П., Висоцька В. Аналіз процесу управління комерційним контентом. Міжн. наук. конф. ISDMCI, Залізний Порт, Україна, 25–28 трав. 2015. С. 36–38.
221. Висоцька В. А., Чирун Л. Б., Чирун Л. В. Контент-моніторинг текстової інформації Web-ресурсів. Міжнар. наук. конф. ISDMCI, Залізний Порт, Україна, 25–28 трав. 2015. С. 36–38.
222. Козлов П., Висоцька В. Технологія управління комерційними контентом в системах електронного бізнесу. Міжнар. наук. конф. ІКС, 20–23 трав. 2015, Львів, Славське. С. 48–49.
223. Кондратєв Є., Висоцька В Контент-аналіз текстових масивів даних. 4 Міжн. наукова конференція ІКС, 20–23 трав. 2015, Україна, Львів, Славське. С. 170–171.
224. Литвин В. В. Висоцька В.А., Оливко Р. М. Метод визначення семантичної метрики на основі тезаурусу предметної області. ІСПЛ, Харків, 14 квіт. 2016 р. С. 10–12.
225. Chyrun L., Vysotska V., Lytvyn V. Specifics informational resources processing for textual content linguistic analysis. Proceeding of MEMSTECH 2016, 20–24 Apr., 2016, Polyana, 2016. P. 214–219.
226. Литвин В. В., Висоцька В. А., Оливко Р. М., Черна Т. М. Особливості рубрикації текстових документів з використанням онтології. ISDMIT, Україна, 25–28 трав. 2016. С. 292–295.
227. Висоцька В. А. Чирун Л. Б., Чирун Л. В. Аналіз процесу супроводу текстового комерційного контенту. Міжнар. наук. конф. ISDMIT, Залізний Порт, Україна, 25–28 трав. 2016. С. 42–44.
228. А Берко. Ю., Висоцька В. А., Чирун Л. В., Чирун Л. Б. Аналітичний метод супроводу текстового контенту інформаційних ресурсів. Математика. Інформаційні технології. Освіта. Східноєвропейський НУ ім. Лесі Українки. Луцьк, 2016. С. 11–20.
229. Висоцька В. А., Козлов П. Ю. Управління Web-ресурсом.. Математика. Інформаційні технології. Освіта. V Міжн. наук.-практ. конф., 5–7 черв. 2016 р., Луцьк. С. 62–63.
230. Берко А. Ю., Висоцька В. А., Чирун Л. В., Чирун Л. Б. Особливості формування критеріїв оцінювання знань студентів згідно їх компетентності у ІТ-сфері. Математика. Інформаційні технології. Освіта. V Міжн. наук.-практ. конф., 5–7 черв. 2016 р., Луцьк. С. 117–118.
231. Каніщева О., Главчева Ю., Висоцька В. Визначення стилю автора для виявлення плагіату в академічному середовищі. SAIT 2017, May 22–25, 2017, Kyiv. P. 78–79.
232. Lytvyn V., Vysotska V., Chyrun L., Smolarz A., Naum O. Intelligent system structure for web resources processing and analysis. 1st Intern. conf., COLINS, 21 Apr. 2017, Kharkiv. P. 56–74.
233. Lytvyn V., Vysotska V., Wojcik W., Dosyn D. A method of construction of automated basic ontology. 1st Intern. conf., COLINS, 21 Apr. 2017, Kharkiv. P. 75–83.

234. Висоцька В. А. Методика аналізу компетентностей для рекрутингу. International scientific and practical conf. on Scientific Research Priorities., 22–23 June 2017, Nowy Sanz, Poland. P. 60–62.
235. Литвин В. В., Наум О. М., Висоцька В. А. Метод інтеграції та управління контентом мережі інформаційних ресурсів туризму згідно потреб користувача. Міжнар. наук. конф. ISDMCI, 22–26 трав. 2017, Залізний Порт. С. 78–80.
236. Висоцька В. А., Чирун Л. Б., Чирун Л. В. Інтернет-портал аматорських спортивних ігор. Міжнар. наук. конф. ISDMCI, 22–26 трав. 2017 Залізний Порт. С. 45–47.
237. Литвин В. В., Оборська О. В., Висоцька В. А., Бобик І. О. Метод аналізу авторства тексту на основі стилеметрії. Міжнар. наук. конф. ISDMCI, 21–27 трав. 2018 р., Залізний Порт. С. 240–243.
238. Чирун Л. Б., Чирун Л. В., Висоцька В. А. Метод визначення авторства текстового україномовного контенту. ISDMCI, 21–27 трав. 2018 р., Залізний Порт, Україна. С. 287–289.
239. Русин Б. П., Висоцька В. А., Погрелюк Л. В. Модель інформаційної системи Virtual Library. Міжнар. наук. конф. ISDMCI, 21 трав. 2018 р., Залізний Порт, Україна. С. 100–102.
240. Kovalchuk V., Lytvyn V., Vysotska V., Hrendus M., Naum O. The information system for identification of content set based on analysis of similar texts. Computational Linguistics and Intelligent Systems. Proceedings. Vol. 2: Proc. of the 2nd Intern. Conf. COLINS 2018. P. 122–127. ISSN 2523-4013.
241. Lytvyn V. Vysotska V., Chyrun L., Hrendus M., Naum O. Content analysis of text-based information in E-commerce systems. COLINS. Vol. 2: Proc. of the 2nd Intern. Conf. 2018. P. 81–94.
242. Rusyn B., Vysotska V., Pohreliuk L. Methods of information resources processing in virtual library. COLINS. Vol. 2 : Proc. of the 2nd Int. Conf., 2018. P. 28–39. ISSN 2523-4013.
243. Lytvyn V., Oborska O., Vysotska V., Dosyn D., Demchuk A. Ontology using for decision making in a competitive environment . COLINS. Vol. 2 : Proc. of the 2nd Intern. Conf. 2018. P. 17–27.
244. Chyrun L., Vysotska V., Chyrun L., Gozhyj A., Kalinina I. SEO technology for web resource processing. COLINS. Proceedings. Vol. 2 : Proc. of the 2nd Intern. Conf., 2018. P. 40–52.
245. Досин Д. Г., Висоцька В. А., Литвин В. В. Побудова системи підтримки прийняття рішень на базі адаптивної онтології. Обчислювальні методи і системи перетворення інформації: зб. пр. V-ї наук.-техн. конф., Львів, 4–5 жовт. 2018. С. 135–138.
246. Висоцька В. А., Литвин В. В., Олешек О. І. Автоматизований моніторинг змін у Web-ресурсах. Міжнар. наук. конф. ISDMCI, Залізний Порт, 21–25 трав. 2019. С. 30–32.
247. Литвин В. В., Висоцька В. А., Михайлишин В. Ю., Сем'янчук С. О. Розроблення інформаційної системи аналізу даних відеопотоку. ISDMCI, Україна, 21–25 трав. 2019. С. 94–97.
248. Демчук А. Б., Литвин В. В., Висоцька В. А. Технологія персоналізованого поширення комерційного контенту через Web-ресурс Е-комерції. ISDMCI, Україна, 21 трав. 2019. С. 49–51.
249. Lytvyn V., Oborska O., Vysotska V., Dosyn D., Demchuk A., Burov Y., Kravets P., Oleksiv N. Problems of ontology structure and meaning optimization and theirs solution methods. COLINS. Proceedings of the 4th Intern. Conf., Lviv, Ukraine; June 23-24, 2020. Vol. II. P. 21–40.
250. Kutyuk O., Lytvyn V., Oborska O., Vysotska V., Dosyn D., Demchuk A., Burov Y., Kravets P. Intelligent system development of distant matrix analysis for recruitment in the IT sector. COLINS. Proceedings of the 4th Intern. Conf., Lviv, Ukraine; June 23-24, 2020. Vol. II. P. 41–78.

251. Tymoshenko K., Vysotska V. Algorithm of text recognizing in Ukrainian on the video mode. COLINS. Proceedings of the 4th Intern. Conf., Lviv, Ukraine; June 23-24, 2020. Vol. II. P. 81–89.
252. Висоцька В. Суб'єктивізм трактування академічної доброчесності в межах наукової діяльності видавництв. Академічна доброчесність: виклики сучасності. Варшава, 06.11.2020. С. 31-35.
253. Bublyk M., Zahreva Y., Vysotska V., Matseliukh Y., Chyrun L., Korolenko O. Information system development for recording offenses in smart city based on cloud technologies and social networks. Webology. 2022. Vol. 19, No. 2. P. 1870–1898.
254. Bublyk M., Kalynii T., Varava L., Vysotska V., Chyrun L., Matseliukh Y. Decision support system design for low voice emergency medical calls at smart city based on chatbot management in social networks. Webology. 2022. Vol. 19, No. 2. P. 2135–2178.

ДОДАТОК Ж. ІНФОРМАЦІЯ ПРО АПРОБАЦІЮ РЕЗУЛЬТАТІВ ДИСЕРТАЦІЙНОЇ РОБОТИ ТА ВПРОВАДЖЕННЯ

"ЗАТВЕРДЖУЮ"
Проректор з наукової роботи
Національного університету
«Львівська політехніка»
_____ І.В. Демидов
«08» 07 2022 р.

АКТ

використання наукових результатів дисертаційної роботи
Висоцької Вікторії Анатоліївни
«Аналіз та синтез комп'ютерних лінгвістичних систем опрацювання україномовного текстового контенту», представленої на здобуття наукового ступеня
доктора технічних наук за спеціальністю
10.02.21 – структурна, прикладна і математична лінгвістика

Комісія в складі: голови комісії – начальника науково-дослідної частини, д.т.н., Небесного Р.В. та членів комісії – завідувача кафедри інформаційних систем та мереж, д.т.н., професора Литвина В.В., завідувача відділу науково-організаційного супроводу наукових досліджень, к.т.н. Лазько Г.В. і заступника начальника планово-фінансового відділу Чулой Т.М., цим актом підтверджують, що результати дисертаційної роботи Висоцької В.А., використовувалися при виконанні науково-дослідної роботи кафедри інформаційних систем та мереж, зокрема в рамках держбюджетних НДР:

«Розроблення інтелектуальних розподілених систем на основі онтологічного підходу з метою інтеграції інформаційних ресурсів» (номер державної реєстрації 0115U004228), зокрема, В.А. Висоцькою було проаналізовано можливості та особливості методів інтеграції інформаційних ресурсів для реалізації модулів опрацювання україномовного текстового контенту систем підтримки прийняття рішень для розв'язку відповідних NLP-задач;

«Методи та засоби функціонування систем підтримки прийняття рішень на основі онтологій» (номер державної реєстрації U0118U000269), запропоновано і вдосконалено моделі, методи та засобів лінгвістичного аналізу україномовного текстового контенту інформаційних ресурсів систем підтримки прийняття рішень для підвищення ефективності розв'язку відповідних NLP-задач;

«Система підтримки прийняття рішень розпізнавання мультиспектральних образів на основі технологій машинного навчання та онтологічного підходу» (номер державної реєстрації 0120U102203), в якій В.А. Висоцькою було проведено дослідження методів та засобів інтелектуального аналізу україномовного текстового потоку інформаційних ресурсів систем підтримки прийняття рішень для підвищення ефективності розв'язку відповідних NLP-задач.

Голова комісії:

начальник науково-дослідної частини, д.т.н. _____ Небесний Р.В.

Члени комісії:

зав. відділу науково-організаційного супроводу наукових досліджень, к.т.н. _____ Лазько Г.В.

заст. нач. планово-фінансового відділу _____ Чулой Т.М.

зав. каф. інформаційних систем та мереж, д.т.н., проф. _____ Литвин В.В.

«ЗАТВЕРДЖУЮ»
Проректор з науково-педагогічної роботи
Національного університету
«Львівська політехніка»
_____ **О.Р. Давидчак**
"___" _____ 2022 р.

А К Т

про впровадження в навчальний процес результатів
докторської дисертаційної роботи
Висоцької Вікторії Анатоліївни

Цей акт складено про те, що результати докторської дисертаційної роботи Висоцької Вікторії Анатоліївни на тему «Аналіз та синтез комп'ютерних лінгвістичних систем опрацювання україномовного текстового контенту», представленої на здобуття наукового ступеня доктора технічних наук, використовуються у навчальному процесі кафедри «Інформаційні системи та мережі» Національного університету «Львівська політехніка». Матеріали дисертаційного дослідження використовуються під час написання студентами курсових робіт, розрахункових робіт, кваліфікаційних бакалаврських та магістерських робіт, а також під час викладання дисциплін «Комп'ютерна лінгвістика», «Розпізнавання мови» та «Методи обчислень та візуалізація даних».

Зокрема, у навчальному процесі використовуються запропоновані В.А. Висоцькою:

- методи та моделі лінгвістичного опрацювання текстового контенту на основі графемного/фонологічного, морфологічного, лексичного, синтаксичного, семантичного, структурного, онтологічного та прагматичного аналізів для розв'язку конкретної NLP-задачі (дисципліна «Комп'ютерна лінгвістика» для студентів освітньо-кваліфікаційного рівня «бакалавр», що навчаються за спеціальністю 124 «Системний аналіз», тема 2 «Регулярні вирази, нормалізація тексту, редагування відстані. Моделювання мови за допомогою N-грам. Основи статистичного опрацювання природних мов. Фільтрація спаму. Пошукові системи»);
- методи та засоби онтологічного моделювання предметної області (дисципліна «Розпізнавання мови» для студентів освітньо-кваліфікаційного рівня «бакалавр», що навчаються за спеціальністю 124 «Системний аналіз», тема 8 «Методи квантитативної лінгвістики для розпізнавання мови»);
- метод визначення ключових слів в україномовних текстах та метод визначення стійких словосполучень при ідентифікації ключових слів україномовного тексту (дисципліна «Комп'ютерна лінгвістика» для студентів освітньо-кваліфікаційного рівня «бакалавр», що навчаються за спеціальністю 124 «Системний аналіз», тема 6 «Методика та парсинг безпосередніх складників. Статистичний парсинг безпосередніх складників та залежностей. Стемінг. Алгоритм Мартина Портера. Перетворення промову в текст , текст у промову»);
- метод визначення автора в україномовних текстах та метод визначення стилю автора тексту (дисципліна «Методи обчислень та візуалізація даних» для студентів освітньо-кваліфікаційного рівня «бакалавр», що навчаються за спеціальністю 124 «Системний аналіз», тема 10 «Аналіз статистичних даних.»).

Директор ІКІІ,
д.т.н., професор — М.О. Медиковський

Завідувач кафедри ІСМ,
д.т.н., професор — В.В. Литвин

Професор кафедри ІСМ, д.т.н. — Д.Г. Досин

ЗАТВЕРДЖУЮ:
Проректор НТУ «ХПІ»
Руслан МИГУЩЕНКО
"20" 06 2023 р.

АКТ

про впровадження докторського дисертаційного дослідження Висоцької Вікторії Анатоліївни «Аналіз та синтез комп'ютерних лінгвістичних систем опрацювання україномовного текстового контенту» у навчальний процес кафедри Інтелектуальних комп'ютерних систем Національного технічного університету «Харківський політехнічний інститут»

Матеріали дисертаційної роботи Висоцької Вікторії Анатоліївни, в яких розроблено нову методику побудови комп'ютерних лінгвістичних систем опрацювання україномовного текстового контенту для розв'язку різних NLP-задач на основі застосування інтелектуального аналізу текстового потоку з інформаційних ресурсів, використовуються у навчальному процесі Національного технічного університету «Харківський політехнічний інститут» на кафедрі інтелектуальних комп'ютерних систем навчально наукового інституту соціально-гуманітарних технологій при підготовці спеціалістів з освітньої програми «Філологія» за спеціальністю 035.10 «Прикладна та комп'ютерна лінгвістика».

При викладанні навчальної дисципліни «Інформаційно-ресурсне забезпечення лінгвістичної діяльності» загальним обсягом 169 годин використано спеціальний корпус академічних текстів, розроблений в дисертаційній роботі; у розділі «Принципи побудови інформаційно-пошукових та експертних систем» створено математичну модель ідентифікації семантичних зв'язків, яка базується на запропонованій методиці побудови комп'ютерних лінгвістичних систем опрацювання україномовного текстового контенту на основі застосування інтелектуального аналізу текстового потоку з інформаційних ресурсів.

При викладанні навчальної дисципліни «Автоматизована обробка природної мови» загальним обсягом 162 години у розділі «Екстракція знань та пошук фактографічної інформації» використано запропоновану в дисертаційній роботі інформаційну технологію визначення авторського стилю наукових документів українською мовою.

Результати роботи впроваджено у навчальний процес кафедри інтелектуальних комп'ютерних систем для студентів спеціальності «Прикладна та комп'ютерна лінгвістика», а саме за рахунок використання навчальних посібників: Математична лінгвістика. Книга 2. Комбінаторна лінгвістика: навчальний посібник / В. В. Пасічник, Ю. М. Щербина, В. А. Висоцька, Т. В. Шестакевич. Львів: Вид-во Львів. політехніки, 2019,

250 с.; Литвин В. В. Глибинне навчання: навч. посіб. / В. В. Литвин, Р. М. Пелещак, В. А. Висоцька. – Львів: Видавництво Львівської політехніки, 2021. – 264 с.; Чисельні методи в комп'ютерних науках / В. А. Андруник, В. А. Висоцька, В. В. Пасічник, Л. Б. Чирун, Л. В. Чирун. Львів: Новий Світ – 2000, 2017. Т. 1. 470 с.; Чисельні методи в комп'ютерних науках / В. А. Андруник, В. А. Висоцька, В. В. Пасічник, Л. Б. Чирун, Л. В. Чирун. Львів: Новий Світ – 2000, 2017. Т. 2. 536 с.; Ришковець Ю. В., Висоцька В. А. Алгоритмізація та програмування. Ч. 1: навч. посіб. Львів: Новий Світ – 200, 2018. 337 с.; Ришковець Ю. В., Висоцька В. А. Алгоритмізація та програмування. Ч. 2: навч. посіб. Львів: Новий Світ – 2000, 2018. 316 с.; Висоцька В. А., Литвин В. В., Лозинська О. В. Дискретна математика: практикум: навч. посіб. Львів: Новий Світ – 2000, 2019. 575 с.; Висоцька В. А., Оборська О. В. Python: алгоритмізація та програмування: навчальний посібник. Львів: Новий Світ – 2000, 2020. 516 с.

Результати дисертаційної роботи Висоцької В.А. «Аналіз та синтез комп'ютерних лінгвістичних систем опрацювання україномовного текстового контенту», які опубліковано у багатьох статтях, використовуються у роботі зі студентами при підготовці наукових публікацій, дипломних магістерських робіт.

Директор навчально-наукового
інституту соціально-гуманітарних технологій,
проф., д.т.н. Андрій КІПЕНСЬКИЙ

Завідувач кафедри ІКС,
д.т.н., проф. Наталія ШАРОНОВА

Доц. каф. ІКС НТУ "ХПІ", к.філ.н. Євген КУПРІЯНОВ

Доц. каф. ІКС НТУ "ХПІ", к.т.н. Надія БАБКОВА