УДК 004.89:81’322.2 **DOI:** https://doi.org/10.30837/bi.2026.1(104).02

О. М. Юрченко[1], Ю. Ю. Повесьма[2]

[1]НТУ «ХПІ», м. Харків, Україна, Olena.Yurchenko@khpi.edu.ua,
ORCID iD: 0000-0002-6074-0241

[2]НТУ «ХПІ», м. Харків, Україна, Україна, Yurii.Povesma@sgt.khpi.edu.ua,
ORCID iD: 0009-0008-9283-0517

СЕМАНТИЧНИЙ АНАЛІЗ КОНТЕКСТІВ РІЗНОЇ ДОВЖИНИ ЗА МЕТРИКОЮ ЛЕКСИЧНОЇ РІЗНОМАНІТНОСТІ MTLD

Семантичне розгортання сенсу речення в текст здійснюється головним чином шляхом збереження та варіації лексичних елементів, що утворюють спільне семантичне поле значимих концептів різних за розміром контекстів. Це дослідження на прикладі змісту коротких текстів (питання завдань) та довгих текстів (відповіді студентів у вигляді есе) з корпусу EFCAMDAT за допомоги міри лексичної текстової різноманітності (MTLD) розглядає лексичні методи представлення сенсу, здатні правильно відтворювати семантичну інформацію в контекстах різної довжини. Воно спрямоване на подолання нестачі даних для навчання великих мовних моделей (LLM) та сприяння професійній інтеграції та міжкультурній співпраці.

СЕМАНТИКА, СЕМАНТИЧНИЙ АНАЛІЗ ТЕКСТУ, КЛАСТЕРИЗАЦІЯ, МОДЕЛЮВАННЯ ТЕМ, ТЕМАТИЧНИЙ АНАЛІЗ, BERTopic, МІРА ЛЕКСИЧНОЇ РІЗНОМАНІТНОСТІ ТЕКСТУ (MTLD)

O. M. Yurchenko, Yu. Yu. Povesma. Semantic analysis of contexts of varying lengths using the MTLD lexical diversity metric. The semantic deployment of sentence meaning in text is achieved mainly by preserving and varying lexical elements that form a common semantic field of meaningful concepts in contexts of varying size. This study, using the example of content between short texts (task questions) and long texts (student answers in the form of essays) from the EFCAMDAT corpus, with the help of the Measure of Textual Lexical Diversity (MTLD), examines lexical methods of representing meaning that are capable of correctly reproducing semantic information in contexts of varying lengths. It aimed to overcome the lack of data for training large language models (LLMs) and to promote professional integration and intercultural cooperation.

SEMANTICS, SEMANTIC TEXT ANALYSIS, CLUSTERING, TOPIC MODELLING, TOPIC ANALYSIS, BERTopic, MEASURE OF TEXT LEXICAL DIVERSITY (MTLD)

Вступ

У сучасному світі професійні сфери не обмежуються місцевим рівнем, а є міжнародними та міждисциплінарними. Це передбачає інтеграцію іноземців в іншу мовну, культурну та іноді в нову професійну сферу, що, як правило, не відбувається швидко і створює чимало труднощів для людини з певним багажем знань та іншомовною професійною підготовкою. У цьому контексті мета нашого дослідження – знайти метод представлення сенсу, здатний правильно відтворювати семантичну інформацію в текстах різної довжини – є дуже актуальною, оскільки це може вирішити проблему нестачі існуючих даних для навчання великих мовних моделей (LLM), які полегшують вирішення завдань обробки природної мови (NLP).

Оскільки семантичний перенос сенсу речення в текст відбувається головним чином шляхом збереження та варіації лексичних елементів, що утворюють спільне лексико-семантичне поле, вважається, що методи лексико-тематичного аналізу, такі як, наприклад, тематичне моделювання, можуть ефективно виявляти семантичну схожість між реченням і текстом [1]. Представлення сенсу тексту у вигляді «торб слів» (bag-of-words, BoW), своєю чергою, робить актуальним застосування методів лексичного аналізу.

Діапазон лексики, що використовується у живому людському мовленні, відображає розмаїття словникового запасу, а також рівень володіння мовою, тому матеріалом дослідження було обрано корпус EFCAMDAT, який є відкритим корпусом робіт тих, хто вивчає англійську мову як іноземну [2]. Тож у цьому дослідженні ми вивчаємо, чи може лексична різноманітність (англ. Lexical Diversity (LD)) застосовуватися як для вимірювання розміру лексико-семантичних полів у текстах, написаних на одну ту саму тему, чи у відповідь на одне й те саме питання/завдання, так і для визначення ступеню семантичної близькості двох різних за довжиною контекстів: речення та тексту.

Ми виходимо з гіпотези, що подібні лексико-семантичні елементи у відповідях на одне й те саме питання в межах однієї мови можуть відрізнятися залежно від рівня володіння мовою, але завжди мають відповідати на поставлене питання і бути семантично співвідносними із заданою темою. Іншими словами, якщо 100 осіб відповідають на одне й те саме запитання, їхні відповіді можна звести до суті запитання, тобто до єдиного семантичного елемента, який може бути представленим у значимих поняттях/концептах, у ключових лексемах, а саме запитання можна вважати темою завдання, сформульованою у вигляді фрази.

Таким чином, метою нашої роботи є вирішення проблеми семантичної подібності чи семантичної спорідненості між контекстами різної довжини,

а саме завданням (питанням), сформульованим у вигляді речення, та відповідями на це запитання, у вигляді тексту.

Для досягнення зазначеної мети ми використовуємо два підходи: (1) результати тематичного аналізу та кластеризації та (2) порівняння лексичного представлення отриманих кластерів до кожного з питань корпусу за допомогою міри текстуальної лексичної різноманітності (англ. Measure of Textual Lexical Diversity (MTLD) [3].

Основними науковими питання дослідження є:

– Як проявляється процес семантичної схожості/подібності між реченнями та текстом з точки зору лексичних і семантичних відносин?

– Як методи тематичного аналізу дозволяють виявити спільне семантичне поле між реченням і текстом?

– Які критерії можна використовувати для підтвердження семантичної подібності між коротким і довгим текстом?

– Які обмеження існують у текстах щодо переносу семантичного ядра лексико-семантичного поля?

– Чи існує межа, за якою «перенесення сенсу» перестає бути лексичним і стає прагматичним або дискурсивним явищем?

Робота структурована наступним чином: у Вступі представлено гіпотезу та цілі дослідження, у розділі 1 наведено огляд сучасних рішень з лексико-семантичного аналізу текстів, у розділі 2 описано методи та матеріали дослідження, розділ 3 присвячено результатам поточного проекту. У висновках ми робимо підсумки і пропонуємо майбутні перспективи дослідження.

1. Огляд існуючих рішень з лексико-семантичного аналізу текстів

Семантична схожість текстів або документів широко вивчається в різних сферах, включаючи обробку природної мови (NLP), порівняння документів, штучний інтелект, семантичну мережу тощо. Сучасні підходи до лексико-семантичного аналізу текстів базуються, з одного боку, на розширенні поняття семантичної подібності, з іншого, на використанні великих мовних моделей, які дають змогу спиратися на контекст.

Головною метою визначення семантичної подібності є вимірювання відстані між семантичними значеннями пари слів, фраз, речень або документів [4]. Семантична подібність (англ. Semantic similarity) – це метрика, де поняття відстані між елементами базується на схожості їхнього значення або семантичного змісту, на відміну від лексикографічної подібності [5]. Але оскільки термін «семантична подібність» зазвичай включає лише відношення «є», в нашому дослідженні ми також використовуємо поняття «семантична спорідненість» (англ. Semantic relatednesse), як синонімічне, але більш широке поняття, оскільки визначення семантичної спорідненості також включає до себе лексичну ієрархію [6], використання поняття лексико-семантичного поля [7, 8] та методів його вимірювання [9, 10, 11].

Сучасні методи тематичного моделювання текстів за допомогою контекстних LLM [1, 12] мають великий потенціал щодо семантичного аналізу та подальшого вдосконалення міжмовних трансферних моделей [13], що зазвичай використовують лексику для перенесення сенсу. Наразі моделі сімейства BERT вважаються одними з найбільш контекстуалізованих [14] завдяки методу TF-IDF, і порівняно з попередніми методами тематичного моделювання, такими як латентний розподіл Діріхле (LDA), які виявляються недостатніми для вирішення проблеми кореляції між текстом і темою (питанням), оскільки базуються лише на принципі ключових слів [15, 16]. Однак щоб підтвердити ступінь семантичної схожості текстів, необхідно враховувати розмір контексту, щоб ваги були пропорційними [9].

Новизна нашого підходу полягає у тому, що наше завдання є зворотним до анотування/реферування тексту, а запропоноване рішення базується не на суто лексичному аналізі, а на контекстуалізованих векторних представленнях сенсу у вигляді набору лексем, що репрезентують текст. Вимірювання лексичної різноманітності тексту (MTLD), яке детально описано в роботах [17, 18, 19], не обмежується набором слів, відомих автору чи читачеві, а також включає способи використання цих слів у текстах [3, 20, 21].

Таким чином, наш підхід базується на припущенні, що семантичний перенос сенсу речення в текст відбувається головним чином шляхом збереження та варіації лексичних елементів, які утворюють спільне семантичне поле між реченнями та текстами. Отже, методи тематичного аналізу (тематичне моделювання та кластеризація) допомагають виявити семантичну подібність між питанням/завданням і текстом відповіді на нього, а лексичний аналіз допомагає вимірювати міру її близькості.

2. Матеріали та методи дослідження

Представлене дослідження[1] є першим кроком у низці досліджень та було проведено на прикладі англійської мови для текстів з корпусу EFCAMDAT[2], що складається з есе, написаних студентами різних рівнів володіння мовою у відповідь на задані викладачами завдання.

Корпус EFCAMDAT поділений відповідно до загальноприйнятих стандартів Загальноєвропейської системи оцінки мовних знань (CEFR)[3] на 6 рівнів: A1, A2, B1, B2, C1, C2. Він складається з відповідей сту-

[1] https://gitlab.univ-lille.fr/olena.yurchenko.etu/mtld
[2] https://github.com/amichw/EFCAMDAT
[3] https ://rm.coe.int/1680459f97

дентів на поставлені запитання: коротких запитань та довших текстів відповідей. На кожному рівні є 24 запитання (теми завдань), за винятком рівня C2, який містить 8 запитань (тем завдань). Загалом корпус містить 128 запитань, які не повторюються з рівня на рівень. Це створює певні труднощі для проведення досліджень на однакові теми на різних мовних рівнях.

Цей корпус текстів також містить інформацію про помилки учнів, частини мови та граматичні відносини; усі завдання оцінені та прокоментовані викладачами [22]. Дані корпусу були нами відфільтровані за позитивною оцінкою від 60 до 100 балів. Загальна кількість даних становить 1 169 298 текстів, розподілених за рівнями: A1 – 621 231, A2 – 304 983, B1 – 166 161, B2 – 60 538, C1 – 14 499, C2 – 1 886.

Порівняння питань (теми завдань у формі речень) та відповідей (відповіді студентів у формі тексту) було здійснено за допомогою наступних бібліотек Python: UMAP, numpy та K-Means у Google Colab для всіх рівнів мови A1–C2 корпусу EFCAMDAT. В ході лексико-семантичного аналізу текстів було використано наступні методи:

1) Тематичний аналіз моделлю BERTopic [14], з виділенням ключових концептів методами KeyBERT, [23], TF-IDF [24, 25] та порівнянням їх за вагою.

2) Кластерізація за допомогою моделі all-MiniLM-L6-v2 (Yin і Zhang, 2024) з алгоритмом HDBSCAN (англ. Hierarchical Density-Based Spatial Clustering of Applications with Noise) [26].

3) Аналіз схожості текстів за методом Жаккара [27].

4) Аналіз лексичного складу текстів за MTLD [19].

Згідно з Treffers-Daller et al. (2018), MTLD обчислюється як середня довжина послідовностей тексту, що мають стабільне співвідношення типу-токену (TTR). Формально:

$$MTLD = \frac{N}{\sum_{i=1}^{k} S_i} \qquad (1)$$

де N – загальна кількість слів у тексті, а S_i – довжина кожної послідовності, в якій TTR перевищує заданий поріг (зазвичай 0,72).

MTLD вважається менш чутливим до довжини тексту, що робить його придатним для аналізу коротких есе, написаних студентами. Однак слід зазначити, що різниця в довжині порівнюваних текстів не має перевищувати 50 токенів, тобто цей показник не підходить для порівняння питань і відповідей до них у формі есе [28]. Однак, цей метод дає змогу порівнювати згруповані в ході тематичного аналізу та кластеризації тексти між собою в рамках одного питання та робити висновки щодо міри схожості між ними.

3. Експериментальні дослідження

На першому етапі дослідження ми провели тематичний аналіз корпусу EFCAMDAT, результати якого описано в нашій роботі [29]. Далі було проведено кластеризацію, результати якої ми проілюстровали в даній роботі кількома яскравими прикладами, але аналіз був проведений для всіх текстів, що мали позитивну оцінку, і порівняний із 128 питаннями/завданнями, представленими одним реченням. Нарешті, отримані результати аналізу були порівняні за тематичним аналізом та кластеризацією.

На другому етапі було проаналізовано лексичний склад відповідей на кожне зі 128 питань за допомогою розрахунків відсотка перетину кластерів і шуму та коефіцієнта Жаккара в межах одного питання та між питаннями, в межах кожного з 6 рівнів і між рівнями, а також зроблено лексичний аналіз за допомогою MDLT [30].

4. Результати кластеризації

4.1. Низький відсоток «шумності». Результати кластеризації у порівнянні з тематичним аналізом показують низький відсоток «шумів». Так лише відповіді на 16 із 128 питань мають рівень шуму вище 10%.

Крім того, шум має тенденцію до зростання на вищих рівнях. Наприклад, два питання рівня C1: «Interpreting a prophecy» та «Writing about future lifestyles, а також одне питання рівня C2 «Writing a visualization script» мають шум 100%, тобто вони не були розділені на кластери. Це може свідчити про те, що лексичні та семантичні поля концептів відповідей практично не перетинаються, тобто кожне питання по-особливому сприймається студентом, а текст відповіді має свій сенс, змістовні акценти та тему. Однак усі вони отримали позитивну оцінку від викладача і мають семантично відповідати поставленому питанню. Тим більше, що за результатами лексичного аналізу їх лексична складова не дуже відрізняється за кластерами. Тож, в таких випадках ми маємо визнати, що обрані моделі не справляються з розрізненням семантики.

Також високий рівень шуму (понад 50%) спостерігається в деяких питаннях нижчих рівнів, наприклад, у питаннях рівня A2 «Describing people in photos» та рівня B2 «Giving feedback about a colleague», що вказує на те, що ймовірна причина певної кількості «шуму» залежить не від рівня мови, а від формулювання питань з використанням певних лексичних елементів: загальних або специфічних, що ще раз підкреслює актуальність лексичного аналізу.

4.2. Кількість кластерів для одного питання. Метою кластеризації було також показати відсоток перетину кластерів за ключовими словами в рамках одного питання для кожного з 128 випадків. Ми порівняли 50 ключових слів кожного кластера з 100 ключовими

словами питання в цілому, а також співвідношення між кластерами та шумом на основі 50 ключових слів для всіх відповідей, що містять шум (32 питання зі 128).

Наприклад, вже згаданий випадок з високим рівнем шуму, такий як питання рівня A2 «Describing people in photos», яке містить 60% текстів із шумом у відповідях і розподіл інших текстів на 5 кластерів, є особливо цікавим, коли порівнюються кластери в межах одного питання.

На графіку (рис. 1) відносини між кластерами позначені синім кольором, відносини між кластерами та шумом – рожевим, а відносини між кластерами та 100 спільними ключовими словами у відповідях на питання рівня A2 «Describing people in photos» – зеленим.

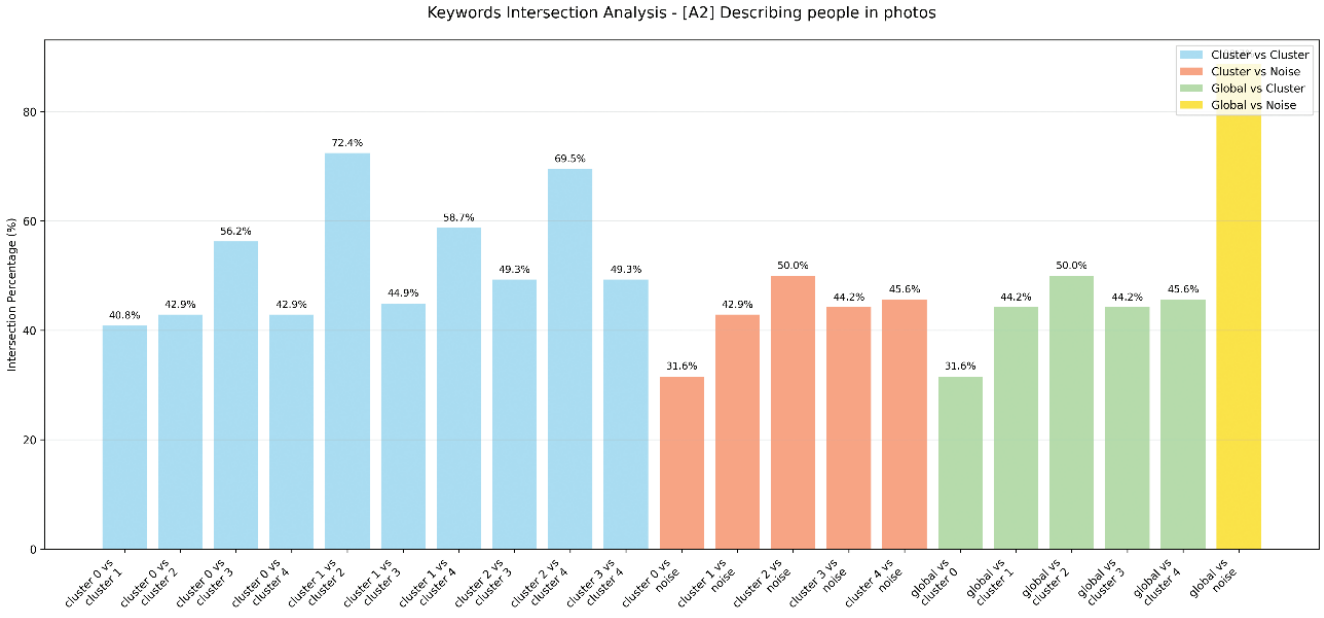

Рис. 1. Відношення кластерів та шуму між собою та до 100 загальних ключових слів питання рівня A2 «Describing people in photos»

Оскільки питання, що містять шуми, представляють найбільшу цікавість з точки зору лексичного складу, ми провели додатковий аналіз шумів, які поводяться як звичайні кластери. Було взято по 20 ключових слів з кожного з 16 питань, шум у яких перевищував 10%, класифіковано їх за TF-IDF-балом та значенням у тексті та порівняно з лексичним складом інших кластерів. Таким чином, ми отримали підтвердження того, що шуми поводяться як окремі кластери і що їх лексика дійсно відповідає поняттям питання.

Для наведеного питання рівня A2 «Describing people in photos», де шум становить 60%, можна бачити, що першим ключовим словом кластерів, і шуму в тому числі, може бути не лише лексема «photo», яка повторює значиме поняття питання, див. рис. 2. Це доводить, що лексичний склад як звичайних, так і шумних кластерів може варіюватися, але він має залишатися в межах одного лексико-семантичного поля, обмеженого основними поняттями питання, оскільки ми взяли всі роботи студентів, які отримали хорошу оцінку.

Однак, як і в природній мові, межі семантичної подібності не зводяться до синонімії, а також включають антонімію, гіперонімію, гіпонімію та інші явища семантичної спорідненості, включаючи асоціації, зміни в порядку слів за рангом у кожному кластері відповідатимуть індивідуальному представленню тексту автором.

cluster_id	texts size	top_keywords
1	759	wearing, photo, look, hair, black
0	93	photo, wearing, picture, friend, brazil
2	1000	wearing, photo, hair, look, friend
3	915	photo, picture, wearing, friend, look
4	155	photo, wearing, look, picture, shirt
Total	2922	Clusters in total: 5

Рис. 2. Top-5 ключових слів до кластерів питання рівня A2 «Describing people in photos»

Результати співвіднесення кластерів та шумів для кожного зі 128 питань корпусу EFCAMDAT було порівняно також на основі відсотка перетину та коефіцієнта Жаккара, а потім представлені у вигляді матриць перетину для кожного питання, див. рис. 3.

Аналізуючи матриці перетину кластерів і шумів у відсотках та матрицю перетину кластерів і шумів за коефіцієнтом Жаккара, складені для 128 питань і продемонстровані на прикладі відповідей на питання рівня A2 «Describing people in photos», ми дійшли висновку, що:

1) в цілому шуми в питаннях, де вони виділені, поводяться як окремі кластери;

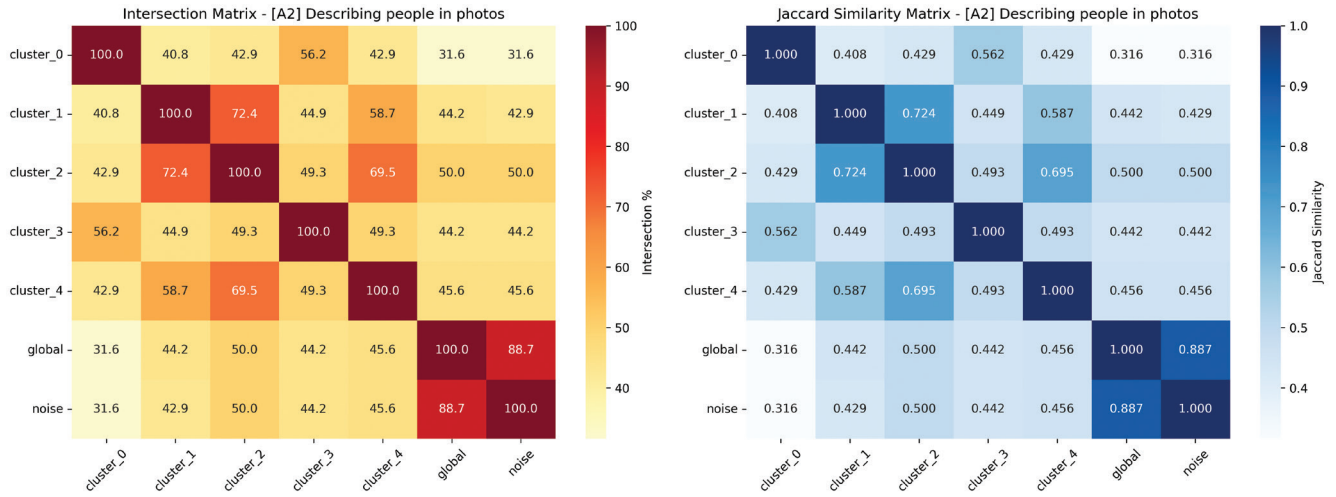

Рис. 3. Матриці перетину текстів 5 кластерів та шумів у порівнянні до 50 загальних ключових слів на основі відсотка перетину та коефіцієнта Джакарда для питання *A2 Describing people in photos*

2) лексичний склад кластерів в цілому дуже схожий і може варіюватися залежно від ваги окремих одиниць, що відображається в рейтингу слів серед 100 спільних ключових слів відповідей на питання;

3) лише відповіді на 10 питань зі 128 містять більше 10 кластерів, а в 30% випадків – більше 5 кластерів на питання.

Таким чином, розподіл кластерів можна вважати досить рівномірним. Він не залежить від рівня мови або кількості шумів, тому, на нашу думку, випадки, що відхиляються від загального розподілу, заслуговують на особливу увагу, зокрема на рівні лексико-семантичного аналізу.

4.3. Перетин кластерів за лексичним складом та рівнями. Хоча питання не повторюються на кожному з рівнів мови, нами було проведено порівняльний аналіз лексики за рівнями та групами рівнів, щоб побачити, наскільки лексичний склад повторюється між ними та за питаннями.

Результати показали, що, хоча найвищі показники лексичного перетину, а саме 85% для питань рівня C1 «Making a movie» і 65% для «Applying for sponsorship», стосуються питань високого рівня, лексичний склад відповідей все ж більше залежить від формулювання питання: більш конкретне чи абстрактне – ніж від самого рівня.

Згідно з рейтингом схожості кластерів за рівнем мови, що складає для рівня C2– 0,0484, C1– 0,0420, B1 – 0,0378, B2– 0,0369, A1 – 0,0764, A2 – 0,0151 і порівнюючи лексичний склад питань відповідних рівнів, треба відзначити, що:

– найнижчий рівень, A1, за класифікацією рівнів за схожістю, є вищим за A2, а також B1 порівняно з B2, за винятком останнього випадку групи B, де схожість є більшою;

– за кількістю тем рівні A1 і B2, а також B1 і C1 наближаються, а не ті, що є сусідніми;

– але за середньою схожістю кластерів рівнів найвищі, C1 і C2, вирізняються.

Також були проведені розрахунки 1) кількості нових слів, що додаються на кожному рівні мови; 2) кількості слів, що залишаються на попередньому рівні; 3) спільної кількості слів між рівнями та групами мови. Найбільш чітко представлена інформація для кожного рівня наведена на діаграмах Венна, див. рисунок 4, який ви можете переглянути на наступній сторінці. Діаграми показують певну нерівномірність у додаванні нового словника, особливо між проміжними рівнями, такими як A2 і B1, де до рівня B1 додається на 1500 нових слів менше, ніж до рівнів A2 і B2. Проте приблизно половина лексичних одиниць залишається спільною для рівнів A1-A2, A2-B1, B1-B2, B2-C1 і має тенденцію до збільшення до рівня C1. Значне зменшення кількості нових слів, унікальних для рівня C2, пояснюється тим, що цей рівень містить лише 8 питань, що значно звужує спектр тем, які розглядаються, і, як наслідок, кількість використовуваного словника.

Це пояснюється тим, що, незважаючи на відсутність повторення питань за рівнем вивчення мови, лексика рівномірно переходить з одного рівня на інший, групуючись за лексико-семантичними полями основних понять запропонованих завдань. Але в групі C вона асимілюється з вільним володінням мовою, і лексика цих рівнів є найбагатшою. Однак через обмежену кількість питань і, як наслідок, недостатню кількість текстів на рівні C2, ми не можемо повною мірою проаналізувати все розмаїття тем на цьому рівні.

Таким чином, ми доходимо висновку, що: 1) незалежно від того, чи йдеться про кластери, чи про шумні кластери, вони дуже схожі за лексичним складом; 2) порівняння лексичного складу різних кластерів не дозволяє відповісти на питання, чому в деяких питаннях (завданнях) розподіл відбувається за великою кількістю кластерів, в інших – за малою, в деяких – шум виділяється, в інших – ні.

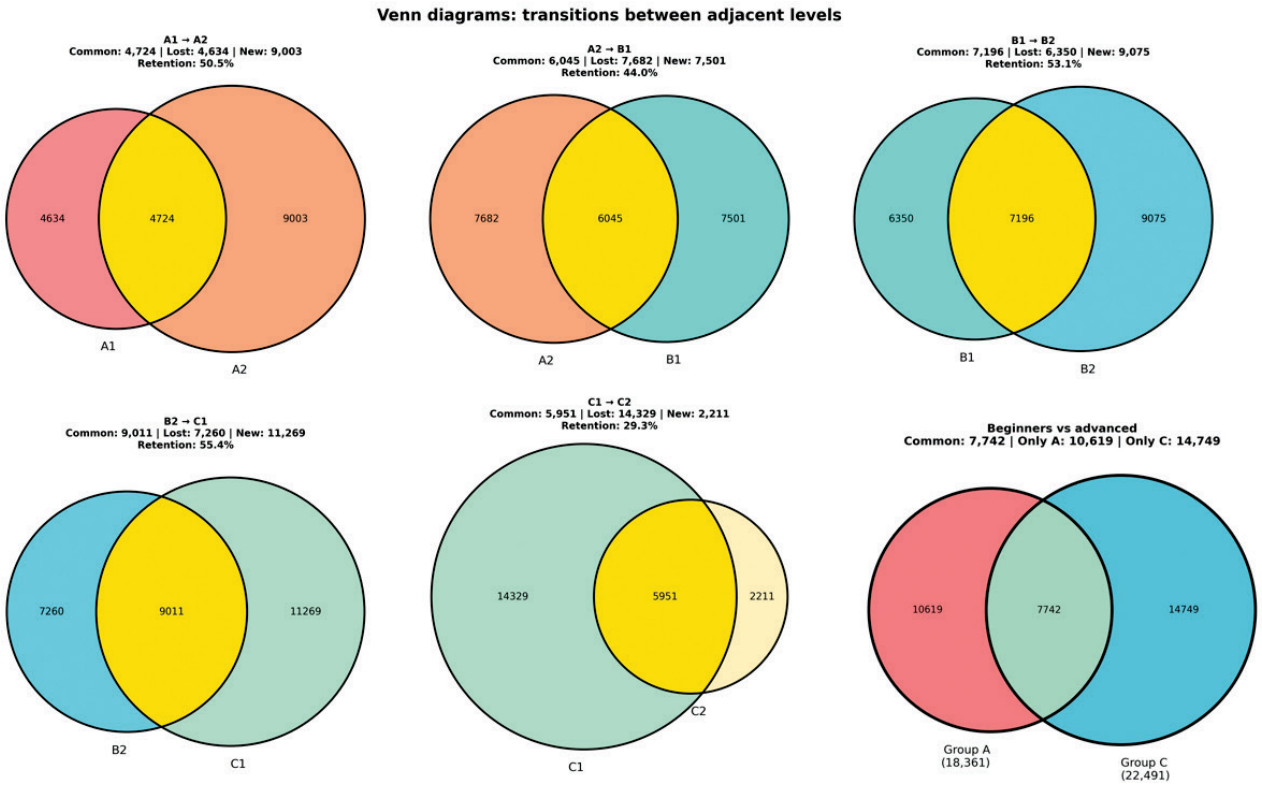

Рис. 4. Діаграми Венна та кількість лексики, що переходить до наступного рівня мови за рівнями та між групами рівнів A та C

5. Вимірювання лексичної різноманітності тексту

5.1. MTLD за рівнями мови. Для оцінки достовірності отриманих значень було проведено порівняння з типовими значеннями MDLT для різних рівнів мови, як зазначено в роботі Treffers-Daller et al. (2018). Наші значення MDLT для корпусу EFCAMDAT за рівнями A1 - 44,6, A2 – 49,4, B1 – 63,2, B1 – 65,8, C1 – 67,9, C2 – 69,5 відповідають діапазону типових значень, отриманих після лематизації, і показують збільшення лексичної різноманітності до рівнів C, навіть якщо на рівні C2 немає великої кількості текстів (див. рис. 5):

– порівняння загального MTLD рівня та середнього значення за документом показує зростання лексичного запасу до високих рівнів;

– порівняння MTLD за документами всередині рівня показує зменшення лексичного запасу до високих рівнів через зниження кількості документів і питань, особливо на рівні C2.

Крім того, було встановлено, що при застосуванні лематизації кількість неправильно написаних слів, що вважаються унікальними, значно зменшувалася із підвищенням рівня мови. Таким чином, ми змогли мінімізувати їх вплив на розрахунки.

5.2. MTLD за лексико-семантичними полями. Як зазначалося вище, питання корпусу EFCAMDAT не повторюються, тому було запропоновано проаналізувати відповіді на питання за лексемами, які можна віднести до одних і тих самих лексичних семантичних полів (ЛСП). Вручну було відібрано питання, що можна віднести до одних і тих самих ЛСП на основі значимих концептів питання, які ми умовно назвали «звички» і які присутні на всіх рівнях: A1 «Describing Your Familys Eating Habits» (концепт «habit»); A2 «Describing Routines» (концепт «routine»); B1 «Writing A Job Advertisement» (концепт «advertisement»); B2 «Setting Rules For Social Networking (концепт «rule»); C1 «Writing About Future Lifestyles» (концепт «lifestyle»); C2 2Following A Code Of Ethics» (концепт «code»). Лексика цих питань, як і всіх інших, була проаналізована за допомогою кластеризації по 20, 50 і 100 ключових слів на кластер, а також «мішків слів» у 4, 10 і 100 найчастіших словах в темі. Лексеми питань були порівняні з лексемами відповідей за рівнями. Аналіз 4 ключових слів BERTopic показує, що концепти ЛСП «звички» були присутні лише в декількох темах, що відповідають цим питанням, наприклад: тема 5 (working style, working style, habits, problems brought, ask improve work), тема 6 (working style, working style, habits, ask improve work, problems brought).

Цей результат також можна пояснити тим, що наявність певних концептів у запитаннях лише дає орієнтир для вибору у текстах відповідей певних лексем із відповідних лексико-семантичних полів. Самі концепти, взяті із запитань у формі речень, також варіюються залежно від того, як сформульовано запитання.

Таким чином, ми робимо висновок, що збільшення кількості проаналізованих слів з 4–10 до 20–50 підвищує ймовірність знайти поняття у відповідях. Але додаткове збільшення з 50 до 100 ключових слів кластера порівняно зі 100 ключовими словами, спільними для кожного питання, показує, що близькість до лексико-семантичних понять питань зменшується.

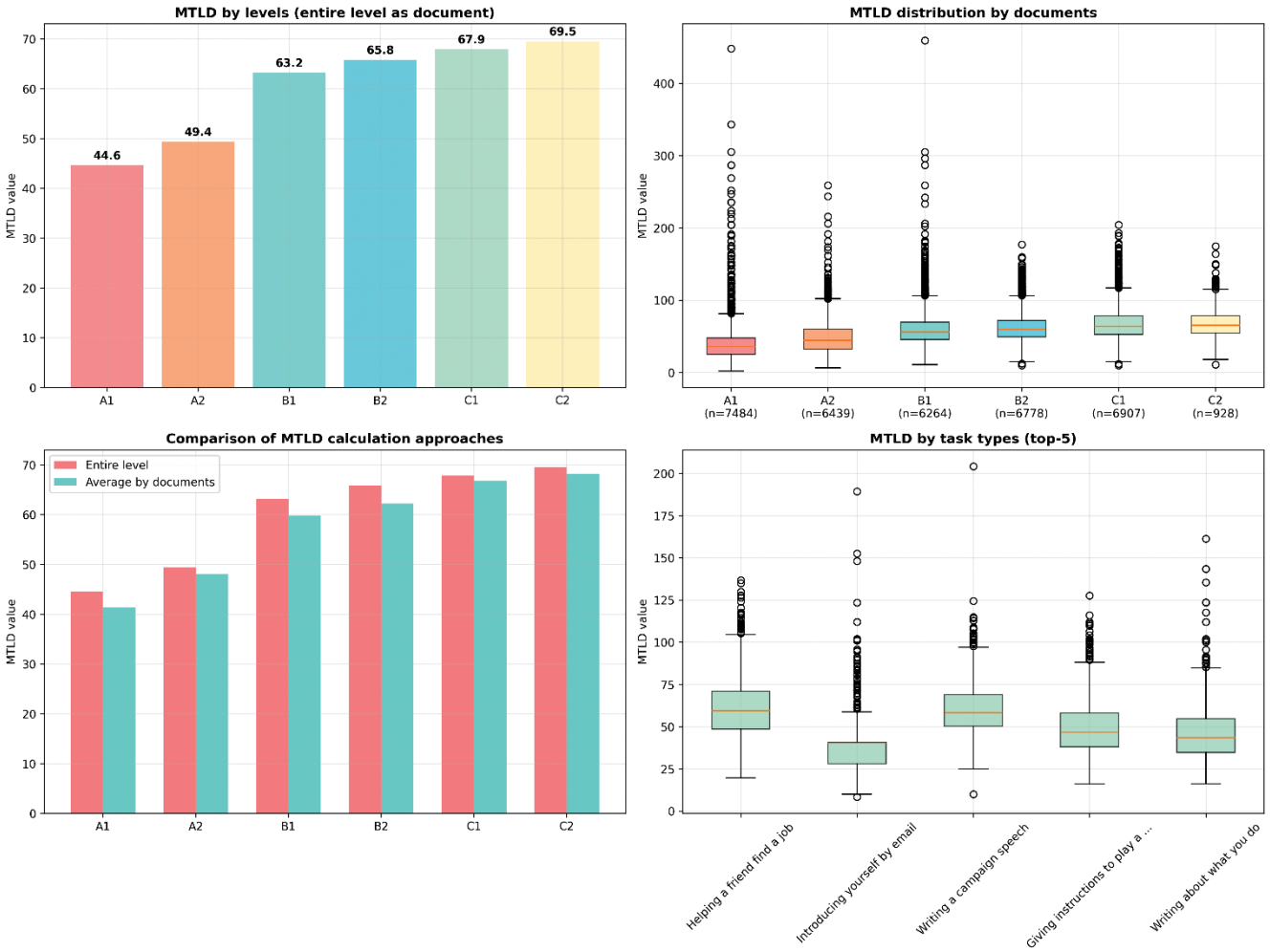

Рис. 5. Результати Measure of Textual Lexical Diversity (MTLD) за рівнями мови

5.3. MTLD за лексичним складом кластерів. Лексичний аналіз за допомогою методу MTLD [30] (Measure of Textual Lexical Diversity) порівняння ключових слів, виконаний додатково на прикладі завдання рівня C1 «Writing a campaign speech» (укр. Написання передвиборчої промови), показує, що алгоритм кластеризації, хоча і розділяє тексти відповідей на завдання на 3 окремі кластери, показані на рис. 6.

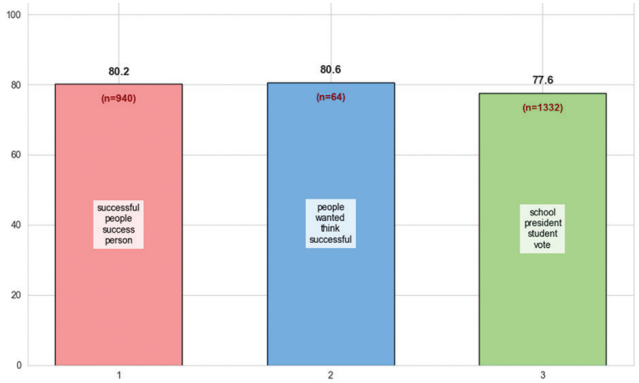

Рис. 6. Порівняння MTLD за кластерами для завдання рівня C1 «Writing a campaign speech»

Але семантично їх можна поділити на 2 теми, які умовно ми назвали:

– «Шкільні вибори»: (лексеми «school», «president», «student», «vote»). Причому цей кластер показав нижчий MTLD (~77,6), оскільки у відповідях використовувався досить обмежений, повторюваний словник.

– «Філософія успіху»: (лексеми «people», «success», «wanted», «successful»). Кластер з текстами, що використовували вказану лексику, показав значно вищий MTLD (80,2–80,6), оскільки абстрактні теми вимагають більш багатого словника.

Таким чином, ми дійшли висновку, що хоча лексичний аналіз складу кластерів допомагає знайти семантичні відмінності між текстами, представлення результатів лексико-семантичного аналізу у вигляді «мішків слів» залишає це на рівні припущень.

Висновки

На основі результатів, отриманих за допомогою моделей кластеризації, тематичного аналізу та індексу лексичної різноманітності текстів MDLT, а також з урахуванням обмежень корпусу EFCAMDAT, зокрема відсутності повторюваних завдань на різних мовних рівнях, ми дійшли таких висновків:

1) Аналіз MDLT показав, що між рівнями немає значних лексичних відмінностей, за винятком

орфографічних помилок на нижчих рівнях. Таким чином, лексичний склад відповідей на питання залежить головним чином від формулювання питання.

2) Щодо існування спільного лексичного та семантичного поля між реченням і текстом, слово саме по собі не є комунікативною синтаксичною одиницею, оскільки його сенс походить від речення, що є мінімальної одиниці мовлення. Отже, для семантичного аналізу краще розглядати контексти на прагматичному та дискурсивному рівнях, а не лише на лексичному чи синтаксичному.

3) У процесі семантичного розгортання речення в тексті аналіз лексичного складу кластерів моделлю all-MiniLM-L6-v2 виявився ефективнішим для групування текстів, ніж метод тематичного аналізу, проведений моделлю BERTopic.

4) Незалежно від того, чи йдеться про кластери, чи про кластери з шумом, їх лексичний склад дуже схожий.

5) Порівняння лексичного складу різних кластерів не дозволяє нам відповісти на питання:, чому в деяких завданнях розподіл відбувається на велику кількість кластерів, а в інших – на маленьку, і чому в деяких випадках шум виділяється, а в інших – ні. Ми припускаємо, що це залежить від формулювання питання.

6) Хоча лексичний аналіз складу кластерів допомагає виявити семантичні відмінності між текстами, представлення результатів у вигляді мішків слів залишає їх у стані гіпотези.

Тому для вирішення проблеми семантичного аналізу текстів в подальших дослідженнях планується використовувати контекстуалізовану векторну репрезентацію значення тексту у вигляді CLS-токену, отриманого з передостаннього шару нейронної мережі [31, 32, 33], та метрики тематичного аналізу текстів [34].

Подяки

Дослідження, описане в цій статті, провадилося в рамках написання дисертації, що виконується під керівництвом д.т.н. Чередніченко О.Ю. на кафедрі програмної інженерії та інтелектуальних технологій управління Національного технічного університету «Харківський політехнічний інститут» (Україна), та у межах проекту ANR-17-CE19-0016 CLEAR під керівництвом Наталії Грабар в лабораторії STL – Savoirs, Textes, Langage (UMR 8163 – CNRS) університету Ліля. Ми дякуємо за фінансову підтримку програмі PAUSE ANR (Франція), а також студенту магістратури Національного технічного університету «Харківський політехнічний інститут» Арсенію Лукашевському, за допомогу в розрахунках.

Список літератури:

[1] Wu, X., Nguyen, T., Zhang, D., Wang, W. Y., & Luu, A. T. (2025). FASTopic: Pretrained Transformer is a Fast, Adaptive, Stable, and Transferable Topic Model. Advances in Neural Information Processing Systems, 37, 84447-84481.

[2] Eertzen G. J., A. T. Lexopoulou A. T., Korhonen A. (2013). Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge open language database (EFCAMDAT), in: 31st Second Language Research Forum (SLRF).

[3] Jarvis, S. (2013). Capturing the Diversity in Lexical Diversity. Language Learning, 63(s1), 87-106.

[4] Nguyen, M. H., Tran, D. Q. (2021). Estimation in semantic similarity of texts, Journal of Information Science and Engineering 37 (2021) 617–633.

[5] Harispe, S., Ranwez, S., Montmain, J. (2022). Semantic similarity from natural language and ontology analysis, Springer Nature.

[6] Feng, Y., Bagheri, E., Ensan, F., Jovanovic, J. (2017). The state of the art in semantic relatedness: a framework for comparison, The Knowledge Engineering Review 32 (2017) 1–30.

[7] Andersen, P. B. (1990). A theory of computer semiotics: semiotic approaches to construction and assessment of computer systems, Vol. 3, Cambridge University Press.

[8] Jackson, H., Zé, E. (2000). Amvela, Words, Meaning, and Vocabulary, Continuum.

[9] Vakulenko, M. (2022). Semantic comparison of texts by the metric approach, Digital Scholarship in the Humanities 38 (2) (2022) 766–771.

[10] Lin, Y.-S., Jiang, Y., Lee, S.-J. (2014). A similarity measure for text classification and clustering, IEEE Trans. on Knowledge and Data Engineering 26 (2014) 1575–1590. doi:10.1109/TKDE.2013.19.

[11] Tytgat, J., Wisniewski, G., Betrancourt, A. (2024). Evaluation de la similarite textuelle : Entre semantique et surface dans les representations neuronales. In JEPTALNRECITAL.

[12] Warner, B., Chaffin, A., Clavié, B., Weller, O., Hallström, O., Taghadouini, S., ... Poli, I. (2024). Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, arXiv preprint arXiv:2412.13663.

[13] Ruder, S., Vulic, I., & Søgaard, A. (2019). A survey of cross-lingual word embedding models. Journal of Artificial Intelligence Research, 65(1):569–630.

[14] Grootendorst, M. R. BERTopic: Neural topic modeling with a class-based TF-IDF procedure, arXiv preprint arXiv:2203.05794 (2022).

[15] Blei, D. M., McAuliffe, J. D. (2010). Supervised Topic Models, in: Advances in Neural Information Processing Systems.

[16] Blei, D. M., Ng, A. Y., Jordan, M. I. (2003). Latent Dirichlet Allocation, Journal of Machine Learning Research 3.993–1022.

[17] McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: a validation study of sophisticated approaches to lexical diversity assessment. Behavior Research Methods, 42(2), 381–392.

[18] Bonvin, A., & Lambelet, A. (2017). Algorithmic and subjective measures of lexical diversity in bilingual written corpora: a discussion. Corela.

[19] Treffers-Daller, J., Parslow, P. & Williams, S. (2018). Back to basics: how measures of lexical diversity can help discriminate between CEFR levels. Applied Linguistics, 39 (3). pp. 302-327. ISSN 1477-450X.

[20] Laufer, B. et Nation, P. (1995). Vocabulary size and use : Lexical richness in L2 written production. Applied Linguistics, 16, 307-322.

[21] Duran, P., Malvern, D., Richards, B. et Chipere, N. (2004). Developmental trends in lexical diversity. Applied Linguistics, 25, 220-242.

[22] Huang, Y., J. Geertzen, R. Baker, A. Korhonen, and T. Alexopoulou (2017). The EF Cambridge Open Language Database (EFCAMDAT) : Information for Users, University of Cambridge and EF Education First.

[23] Issa, B., Jasser, M.B., Chua, H.N., & Hamzah, M. (2023). A Comparative Study on Embedding Models for Keyword Extraction Using KeyBERT Method. 2023 IEEE 13th International Conference on System Engineering and Technology (ICSET), 40-45.

[24] Egger, R., Yu, J. (2022). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts, Frontiers in Sociology 7 (2022).

[25] Babalola, O., Ojokoh, B., Boyinbode, O. (2024). Comprehensive Evaluation of LDA, NMF, and BERTopic's Performance on News Headline Topic Modeling, Journal of Computing Theories and Applications 2 (2024) 268–289.

[26] Malzer, C., & Baum, M. (2019). HDBSCAN(ε): An Alternative Cluster Extraction Method for HDBSCAN. ArXiv, abs/1911.02282.

[27] Travieso, G., Benatti, A., & Costa, L.D. (2024). An Analytical Approach to the Jaccard Similarity Index.

[28] Koizumi, R. (2012). Relationships Between Text Length and Lexical Diversity Measures : Can We Use Short Texts of Less than 100 Tokens ? Vocabulary Learning and Instruction, 1(1), 60-69.

[29] Grabar, N., Yurchenko, O., Cherednichenko, O., and Lukashevskyi, A. (2025). Exploring Semantic Similarity in English Learners' Texts through Topic Modelling. In CLW-2025: Computational Linguistics Workshop at 9th International Conference on Computational Linguistics and Intelligent Systems (CoLInS-2025), May 15–16, 2025, Kharkiv, Ukraine Vol-3976, p. 92-106.

[30] Yurchenko, O., Grabar, N., Cherednichenko, O. Analyse du vocabulaire appliquée au transfert sémantique phrase→texte. In: Actes de la conférence Extraction et Gestion des Connaissances (EGC 2026). Vol. RNTI-E-42. Toulouse: Cépaduès-Éditions, 2026. P. 373–374.

[31] Mollas, I., Bassiliades, N. et Tsoumakas, G. (2019). LioNets : Local Interpretation of Neural Networks through Penultimate Layer Decoding. arXiv : Learning.

[32] Bianchi, F., Terragni, S., et Hovy, D. (2020). Pre-training is a Hot Topic : Contextualized Document Embeddings Improve Topic Coherence. Annual Meeting of the Association for Computational Linguistics.

[33] Bouzina, S., De Rossi, D., Pavlov, V. G. et Moretti, S. (2024). Semantic Latency Mapping of Contextual Vector Embeddings in Transformer-Based Models.

[34] Terragni, S., Fersini, E., & Messina, E. (2021). Word Embedding-Based Topic Similarity Measures. International Conference on Applications of Natural Language to Data Bases.

Received (Надійшла) 22.01.2026

Accepted for publication (Прийнята до друку) 15.02.2026

Publication date (Дата публікації) 27.03.2026