УДК 004.83

**С.Д. ПОГОРІЛИЙ, А.А. КРАМОВ**

# ОЦІНЮВАННЯ КОГЕРЕНТНОСТІ ТЕКСТУ ЗА ДОПОМОГОЮ ПОБУДОВИ ГРАФУ СЕМАНТИЧНОЇ ТА ЛЕКСИКО-ГРАМАТИЧНОЇ УЗГОДЖЕНОСТІ СЛОВОСПОЛУЧЕНЬ РЕЧЕНЬ

**Анотація.** Запропоновано метод оцінювання когерентності текстів за допомогою побудови графу семантичної та лексико-граматичної узгодженості словосполучень речень. Виконано експериментальну перевірку ефективності методу на англомовному корпусі. Отримані результати розрахованих метрик пропонованого методу перевищують відповідні значення інших сучасних підходів. Метод може бути застосований до іншомовних текстів шляхом заміни лінгвістичних моделей відповідно до особливостей певної мови.

**Ключові слова:** обробка природної мови, оцінювання когерентності тексту, двочастковий граф словосполучень, метод розрахунку когерентності текстів на основі графу, лексико-граматична узгодженість речень.

## ВСТУП

Обробка природної мови (natural language processing, NLP) є одним із напрямків досліджень у галузі штучного інтелекту. Розв'язання більшості задач з обробки природної мови потребує використання людських ресурсів (експертних знань), тобто задачі цього типу не можна розв'язати за допомогою визначеного алгоритму. До класу таких задач варто віднести розпізнавання та синтез мовлення, синтаксичний аналіз тексту, виявлення плагіату, оцінювання тональності тексту тощо. При цьому постає задача формалізації текстів природної мови та виявлення закономірностей між їхніми компонентами відповідно до очікуваного вихідного результату. Зважаючи на постійний приріст потужності обчислювальних ресурсів, для розв'язання відповідних задач застосовують різноманітні комбіновані методи машинного навчання та комп'ютерної лінгвістики [1]. Отже, з'являється можливість виконати навчання моделі на попередньо сформованому корпусі (сукупності текстової інформації) для подальшого її використання на тестовій вибірці. Однак неоднорідність текстової інформації (різна структура, довжина речень, смислова залежність наступних речень від попередніх) та різноманітність її вмісту ускладнюють процес проєктування та розрахунок параметрів моделей машинного навчання. У зв'язку з цим розв'язання задач обробки природної мови, що здійснюють аналіз семантичних та граматичних властивостей тексту, як і раніше, є актуальним. До задач такого типу відносять оцінювання когерентності тексту.

Відповідно до означення [2] під когерентністю тексту розуміють взаємозв'язок його компонент у граматичний та лексичний способи. Когерентність тексту передбачає послідовну передачу основної ідеї читачу в межах цього тексту, що робить його зрозумілішим та простішим для сприйняття. Цей комунікативний зв'язок між автором та читачем досягається за допомогою семантичної цілісності тексту. Іншим критерієм когерентного тексту є наявність структурної узгодженості його складових (речень та словосполучень). Перевірку ступеня дотримання зазначених критеріїв можна здійснити шляхом оцінювання когерентності тексту. Методи оцінювання когерентності тексту застосовуються в різних сферах, пов'язаних з обробкою текстової інформації: генерація тексту [3], написання інструкцій, аналіз медичних записів, автоматизований пошук даних [4] тощо.

У статті запропоновано метод оцінювання когерентності текстів на основі графу (graph-based method) за допомогою аналізу семантичного та лексико-граматичного зв'язку словосполучень тексту; здійснено експериментальну перевірку ефективності пропонованого методу на корпусі англомовних текстів порівняно з іншими методами.

## СУЧАСНІ МЕТОДИ ОЦІНЮВАННЯ КОГЕРЕНТНОСТІ ТЕКСТІВ

Наявні методи оцінювання когерентності текстів ґрунтуються на методології машинного навчання. У 2008 році в роботі [5] був запропонований метод Entity Grid, за яким було здійснено аналіз зміни ролі сутності в реченні. Виявлення закономірностей виконувалося шляхом попереднього формування векторів ознак та застосування методу опорних векторів. Ідея відстеження ролі сутності в межах тексту була також використана в роботі [6] щодо методу Entity Graph. Суть цього методу полягає у формуванні двочасткового графу зв'язку ролей сутностей і речень та подальшій побудові проєкційного графу тексту. З появою моделей семантичного векторного представлення слів з'явилися методи на основі аналізу семантичної схожості елементів тексту. У роботах [7, 8] запропоновано підхід до проєктування нейронних мереж з рекурентними та згортковими шарами відповідно. При цьому вхідними даними мережі є векторне представлення речень тексту у семантичному просторі. Суть методу Semantic Similarity Graph [9] полягає у побудові графу тексту за допомогою розрахунку семантичної близькості речень. У роботі [10] здійснено побудову онтологічної системи та анотації тексту, а вихідним результатом є оцінка когерентності анотації, розрахована на основі результатів аналізу екстрагованих концептів. Метод, що ґрунтується на прогнозуванні позиції речення в тексті, був запропонований у роботі [11]. Прогнозування виконується за допомогою попереднього векторного представлення речень з використанням рекурентного шару нейронної мережі та з подальшою класифікацією їхньої позиції (частини тексту з місцезнаходженням речення).

Користувачу потрібно розуміти причину формування вихідного результату, щоб отримати можливість для аналізу та подальшого вдосконалення тексту. Відповідними методами є Entity Graph, Semantic Similarity Graph та метод, описаний у роботі [11], оскільки вони надають змогу виконати візуальне відображення зв'язку між компонентами тексту. З огляду на переваги та недоліки розглянутих методів, пропонований метод повинен відповідати таким критеріям:

- передбачати можливість візуалізації зв'язку між компонентами тексту;
- забезпечувати одночасний аналіз семантичних та лексичних властивостей тексту;
- бути адаптованим для застосування до іншомовних текстів.

## ПРОПОНОВАНИЙ МЕТОД

Для відстеження процесу формування вихідної оцінки потрібно визначити компоненти тексту, над якими власне виконуються операції для розрахунку міри когерентності. Відповідно до інших методів оцінювання когерентності тексту пропонується розглядати цілісність тексту на рівні речень. Розрахунок міри схожості речень здійснюється з використанням сучасних методів за допомогою аналізу семантичної узгодженості слів чи іменних фраз (сутностей). У цій роботі запропоновано розглядати взаємозв'язок речень на рівні словосполучень. Представлення речення за допомогою набору екстрагованих словосполучень має такі переваги:

• здійснюється автоматизована фільтрація стоп-слів та допоміжних елементів речення;

• виконується перевірка структурної (граматичної) узгодженості слів речення, адже у випадку некоректно сформованого речення зменшується ймовірність успішної екстракції словосполучень.

**ЕКСТРАКЦІЯ СЛОВОСПОЛУЧЕНЬ РЕЧЕННЯ**

Універсальний пошук словосполучень у тексті не можна виконати за допомогою заздалегідь визначеного алгоритму. Шаблонами, реалізованими у вигляді регулярних виразів, можна скористатися лише як допоміжним засобом через неоднорідність структури текстової інформації. Принцип побудови речення може бути різним залежно від особливостей мови чи стилістики тексту. Доцільним є використання засобів синтаксичного аналізу тексту для виявлення зв'язку між словами речення з подальшим аналізом отриманої залежності. У роботі [12] запропоновано метод екстракції інформації (open information extraction — open IE) з англомовних текстів за допомогою аналізу синтаксичного розбору речення. Перевагою цього методу є відсутність потреби у формуванні шаблонів. Принцип його роботи полягає у поділі речення на незалежні фрагменти (потенційні словосполучення) з подальшим стисканням для вилучення зайвих слів. Розглянемо процес формалізації отриманих словосполучень. Представимо речення $S$ як множину слів (токенів)

$$S = \{t_1, t_2, \ldots, t_N\}, \tag{1}$$

де $N$ — кількість слів речення. Результатом застосування методу open IE до речення $S$ є множина кортежів

$$C = \{C_1, C_2, \ldots, C_K\}, \tag{2}$$

де $K$ — кількість екстрагованих кортежів. Кожен кортеж містить такі елементи:

$$C_i = (Obj, Sub, Rel), \ \ i \in I = \{1, 2, \ldots, K\}, \tag{3}$$

де *Obj* — «об'єкт» (головний елемент), *Rel* — «відношення» (залежність між головним та підрядним елементами), *Sub* — «суб'єкт» (підрядний елемент). Кожен елемент кортежу представлений множиною слів речення $\{t \mid t \in S\}$. «Об'єкт» містить слова батьківської сутності, а «суб'єкт» — підрядної. Слова елемента «відношення» вказують на тип зв'язку між «об'єктом» та «суб'єктом». Об'єднавши елементи кортежу $C_i$, отримуємо словосполучення $X_i = Obj \cup Rel \cup Sub$. Отже, речення $S$ можна представити як множину словосполучень

$$X = \{X_1, X_2, \ldots, X_K\}. \tag{4}$$

На рис. 1 зображено приклад екстракції словосполучень з речення «Presley was born in Tupelo and relocated to Memphis».

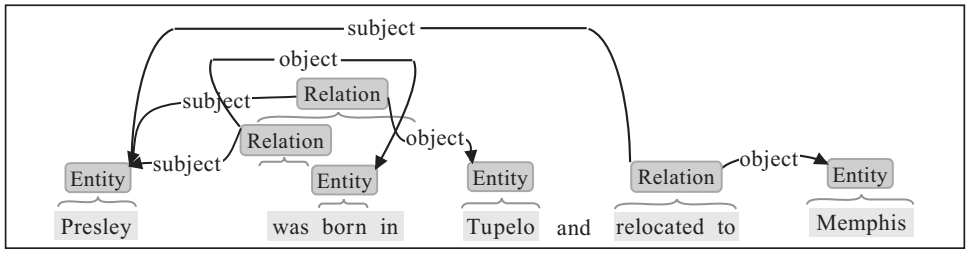


*Рис. 1.* Приклад екстракції словосполучення з речення за методом open IE

З наведеного вище прикладу можна екстрагувати таку множину словосполучень:

$$X = \{X_1, X_2\},$$
$$X_1 = \{\text{Presley, was, born, in, Tupelo}\}, \quad (5)$$
$$X_2 = \{\text{Presley, relocated, to, Memphis}\}.$$

**РОЗРАХУНОК МІРИ СХОЖОСТІ РЕЧЕНЬ**

Для оцінювання взаємозв'язку речень слід аналізувати не тільки їхню семантичну схожість, але й логічну послідовність розташування в тексті. Одним з індикаторів логічного зв'язку між реченнями є наявність спільних кореферентних об'єктів — сутностей, що посилаються на один елемент. Виявлення кореферентних об'єктів [13] надає змогу враховувати зв'язок між реченнями незалежно від їхнього розташування в тексті. Тому пропонується застосувати метод пошуку кореферентних пар до всього тексту. Отримані групи об'єктів записуються в тимчасову пам'ять для подальшого використання.

Розглянемо речення $S_i$ та $S_j$, представлені множинами словосполучень $X^i = \{X^i_1, X^i_2, \ldots, X^i_{|X^i|}\}$ і $X^j = \{X^j_1, X^j_2, \ldots, X^j_{|X^j|}\}$ відповідно. З множин словосполучень $X^i$ та $X^j$ побудуємо повнозв'язний орієнтований двочастковий граф $K_{ij}$ — граф семантичної та лексико-граматичної узгодженості словосполучень речень $S_i$ та $S_j$. Приклад графу $K_{ij}$ зображено на рис. 2.

Розглянемо детальніше процес формування ваг ребер графу $K_{ij}$. Вагу ребра розраховують як відношення кількості спільних елементів відповідних словосполучень до загальної кількості унікальних елементів

$$\text{lex}(X^i_l, X^j_m) = \frac{\text{common}(X^i_l, X^j_m)}{\text{unique}(X^i_l, X^j_m)}. \quad (6)$$

У разі виявлення кореферентних об'єктів (у прикладі, наведеному вище, слово «there» є взаємозамінним зі словом «Memphis») значення $\text{lex}(X^i_l, X^j_m)$ встановлюють рівним одиниці. У такий спосіб підвищують значимість кореферентного зв'язку порівняно з наявністю спільних термінів. Після встановлення значень ваг усіх ребер графу $K_{ij}$ лексико-граматичну узгодженість речень $S_i$ та $S_j$ обраховують у такий спосіб:

$$\text{lex}(S_i, S_j) = \frac{\sum\limits_{l \in \{1,2,\ldots,|X^i|\}} \sum\limits_{m \in \{1,2,\ldots,|X^j|\}} \text{lex}(X^i_l, X^j_m)}{|X^i|\,|X^j|}. \quad (7)$$

Розглянемо процес оцінювання семантичної схожості речень $S_i$ та $S_j$. Здійснимо векторне представлення речень в семантичному просторі та використаємо косинусну відстань між відповідними векторами $\mathbf{s}^i$ і $\mathbf{s}^j$ як міру схожості

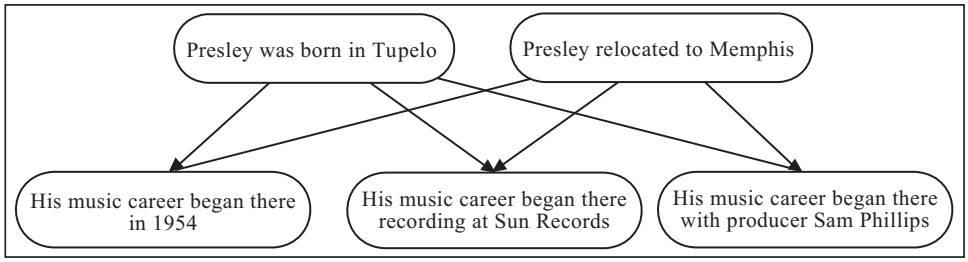


*Рис. 2.* Приклад двочасткового графу семантичної та лексико-граматичної узгодженості речень $S_i$ та $S_j$

речень. Перетворення вигляду речень на векторний можна виконати за допомогою попередньо навченої моделі семантичного представлення слів чи речень (Word2Vec [14], Doc2Vec [15], fastText [16] тощо). Отже, семантичну схожість речень $S_i$ та $S_j$ можна розрахувати у такий спосіб:

$$\text{sem}(S_i, S_j) = \frac{\mathbf{s}^i \cdot \mathbf{s}^j}{||\mathbf{s}^i|| \, ||\mathbf{s}^j||}. \tag{8}$$

Загальна міра схожості речень повинна одночасно враховувати семантичну та лексико-граматичну узгодженість речень. Уведемо регулятивний параметр $\alpha \in [0,1]$ для аналізу впливу цих компонент на вихідну оцінку когерентності тексту. До того ж, слід взяти до уваги відстань між реченнями. Загальну міру схожості речень $S_i$ та $S_j$ обраховують як

$$\text{sem}(S_i, S_j) = \frac{(1-\alpha)\text{sem}(S_i, S_j) + \alpha \, \text{lex}(S_i, S_j)}{|i-j|}, \tag{9}$$

де $|i-j|$ — фактор врахування відстані між реченнями в тексті.

### ПРЕДСТАВЛЕННЯ ТЕКСТУ ЗА ДОПОМОГОЮ ГРАФУ

Розглянемо текст $T$ як множину речень $T = \{S_1, S_2, \ldots, S_M\}$, де $M$ — кількість речень тексту. Побудуємо орієнтований граф $G = (V, E)$, де $V$ — множина вершин, що інтерпретують речення тексту $T$ (потужність множини $V$ дорівнює кількості речень тексту $M$); $E$ — множина ребер. Ребра встановлюють між усіма вершинам графу. Вага ребра $e_{ij} \in E, i \neq j$, дорівнює загальній мірі схожості відповідних речень $\text{sem}(S_i, S_j)$. Когерентність тексту $T$ розраховують як середнє арифметичне значення ваг усіх ребер графу $G$

$$\text{Coherence}(T) = \frac{\sum\limits_{i,j \in \{1,2,\ldots,M\}, i \neq j} \text{weight}(S_i, S_j)}{M}. \tag{10}$$

### ЕКСПЕРИМЕНТАЛЬНА ПЕРЕВІРКА МЕТОДУ

Перевірку ефективності пропонованого методу виконано за допомогою розрахунку точності розв'язання задач розрізнення документів (document discrimination task) та вставки (insertion task) [17]. Перевірочну вибірку англомовних текстів сформовано з корпусу OntoNotes Release 5.0 (LDC2013T19) [18]. Екстракцію словосполучень з речень тексту (метод open IE) та пошук кореферентних пар виконано за допомогою прикладного програмного інтерфейсу Stanford CoreNLP. Як семантичну модель представлення елементів тексту обрано модель Word2Vec, натреновану на множині текстів GoogleNews. Для порівняння метрик пропонованого методу з іншими методами вирішено виконувати аналіз текстів, що використовувалися в роботі [9]. Насамперед, це стосується задачі вставки, адже у випадку збільшення кількості речень зменшується ймовірність коректного розпізнавання тексту. Отже, для розв'язання задачі розрізнення документів використано всі тексти корпусу OntoNotes Release 5.0; для розв'язання задачі вставки відібрано тексти із середньою кількістю речень, що дорівнює семи.

### ПЕРЕВІРКА ЕФЕКТИВНОСТІ ЗАСТОСУВАННЯ ПРОПОНОВАНОГО МЕТОДУ ДО АНГЛОМОВНИХ ТЕКСТІВ

Для дослідження впливу семантичної та лексичної компонент на загальну оцінку когерентності тексту, точності розв'язання задач розрізнення документів та вставки розраховано для різних значень регулятивного параметра $\alpha$ з кроком 0.1 ($\alpha \in [0,1]$). Максимальне значення точності розв'язання задач розрізнення

**Таблиця 1.** Порівняння результатів точності розв'язання задач розрізнення документів та вставки для англомовних текстів

| Метод і значення регулятивних параметрів | Задача розрізнення документів | Задача вставки |
|---|---|---|
| PAV | 0.774 | 0.356 |
| SSV | 0.676 | 0.346 |
| MSV | 0.741 | 0.327 |
| Entity Grid | 0.845 | 0.346 |
| Entity Graph | 0.725 | 0.260 |
| Пропонований метод, $\alpha = 0.4$ | 0.900 | 0.370 |

документів (0.900) і задачі вставки (0.370) отримано для значення регулятивного параметра $\alpha = 0.4$. Отже, для оцінювання когерентності англомовних текстів доцільно одночасно враховувати семантичні та лексичні властивості тексту: семантична та лексична складові ваг ребер графу є рівнозначними компонентами для розрахунку міри схожості речень. Врахування лексичної складової надає змогу підвищити точність методу, що підтверджує необхідність використання пошуку кореферентних об'єктів для оцінювання когерентності англомовних текстів.

У табл. 1 наведено результати точності розв'язання задач розрізнення документів та вставки з використанням різних методів для англомовних текстів [9]. Виконано порівняння пропонованого методу з методами PAV, SSV, MSV, Entity Graph та Entity Grid. Максимальні значення точності розв'язання задач отримано для пропонованого методу зі значеннями регулятивного параметра $\alpha = 0.4$. Результати свідчать про доцільність застосування пропонованого методу на основі графу для оцінювання когерентності англомовних текстів.

## ВИСНОВКИ

У роботі запропоновано метод оцінювання когерентності текстів на основі графу за допомогою аналізу семантичних та лексичних властивостей тексту на рівні словосполучень. На основі аналізу отриманих результатів експериментальної перевірки ефективності методу можна зробити такі висновки:

- виявлення кореферентних об'єктів надає змогу підвищити точність методу за рахунок відстеження зв'язку між віддаленими компонентами тексту під час його послідовного оброблення (імітації процесу читання тексту);
- найвищу точність пропонованого методу отримано для значення регулятивного параметру $\alpha = 0.4$. Отже, одночасне врахування семантичних та лексичних властивостей тексту є доцільним. Підвищити точності аналізу семантичної складової можна за рахунок використання інших моделей семантичного представлення елементів тексту;
- аналіз речень за допомогою екстрагованих словосполучень дає можливість виконати перевірку структурної узгодженості речення та здійснити оцінювання семантичної схожості речень на рівні отриманих класів об'єктів і зв'язків між ними;
- отримані результати порівняльного аналізу точності пропонованого методу та інших методів свідчать про можливість його використання для оцінювання когерентності англомовних текстів. Пропонований метод можна застосовувати для іншомовного корпусу за умови попереднього навчання та заміни відповідних лінгвістичних моделей (семантичне представлення тексту, екстракція словосполучень, виявлення кореферентних об'єктів).

СПИСОК ЛІТЕРАТУРИ


1. Kurdi M. Natural language processing and computational linguistics 2: Semantics, discourse and applications. John Wiley & Sons, 2018. 316 p.
2. Poulimenou S., Stamou S., Papavlasopoulos S., Poulos M. Short text coherence hypothesis. *Journal of Quantitative Linguistics*. 2016. Vol. 23, Iss. 2. P. 191–210. https://doi.org/10.1080/09296174.2016.1142328.
3. Marchenko O., Radyvonenko O., Ignatova T., Titarchuk P., Zhelezniakov D. Improving text generation through introducing coherence metrics. *Cybernetics and Systems Analysis*. 2020.Vol. 56, N 1, P. 13–21. https://doi.org/10.1007/s10559-020-00216-x.
4. Pogorilyy S., Kramov A. Automated extraction of structured information from a variety of web pages. *Proc. 11th International Conference of Programming UkrPROG 2018* (22–24 May 2018, Kyiv, Ukraine). Kyiv, Ukraine, 2018. P. 149–158.
5. Barzilay R., Lapata M. Modeling local coherence: an entity-based approach. *Computational Linguistics.* 2008. Vol. 34, N 1, P. 1–34. https://doi.org/10.1162/coli.2008.34.1.1.
6. Mesgar M., Strube M. Normalized entity graph for computing local coherence. *Proc. TextGraphs-9: the workshop on Graph-based Methods for Natural Language Processing* (29 October 2014, Doha, Quatar). Doha, Quatar, 2014. P. 1–5. https://doi.org/10.3115/v1/w14-3701.
7. Li J., Hovy E. A model of coherence based on distributed sentence representation. *Proc. 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (25–29 October 2014, Doha, Quatar). Doha, Quatar, 2014. P. 2039–2048, 2014. https://doi.org/10.3115/v1/d14-1218.
8. Cui B., Li Y., Zhang Y., Zhang Z. Text coherence analysis based on deep neural network. *Proc. 2017 ACM on Conference on Information and Knowledge Management (CIKM'17)* (6–10 November 2017, Singapore, Singapore). Singapore, Singapore, 2017. P. 2027–2030. https://doi.org/10.1145/3132847.3133047.
9. Putra J., Tokunaga T. Evaluating text coherence based on semantic similarity graph. *Proc. TextGraphs-11: the Workshop on Graph-based Methods for Natural Language Processing* (3 November 2017, Vancouver, Canada). Vancouver, Canada, 2017. P. 76–85. 2017. https://doi.org/10.18653/v1/w17-2410.
10. Giray G., Ünalír M. Assessment of text coherence using an ontology-based relatedness measurement method. *Expert Systems.* 2019. Vol. 37, N. 3. P. 1–24. https://doi.org/10.1111/exsy.12505.
11. Bohn T., Hu Y., Zhang J., Ling C.X. Learning sentence embeddings for coherence modelling and beyond. *Proc. Recent Advances in Natural Language Processing* (2–4 September 2019, Varna, Bulgaria). Varna, Bulgaria, 2019. P. 151–160. https://doi.org/10.26615/978-954-452-056-4_018.
12. Angeli G., Premkumar M.J.J., Manning C. Leveraging linguistic structure for open domain information extraction. *Proc. 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (Vol. 1: Long Papers) (26–31 July 2015, Beijing, China). Beijing, China, 2015. P. 344–354. https://doi.org/10.3115/v1/p15-1034.
13. Pogorilyy S., Kramov A. Coreference resolution method using a convolutional neural network. *Proc. 2019 IEEE International Conference on Advanced Trends in Information Theory (ATIT)* (18–20 December 2019, Kyiv, Ukraine). Kyiv, Ukraine, 2019. P. 397–401. https://doi.org/10.1109/ATIT49449.2019.9030596.
14. Le Q., Mikolov T. Distributed representations of sentences and documents. *Proc. 31st International Conference on Machine Learning* (21–26 June 2014, Beijing, China). Beijing, China, 2014. P. 1188–1196.
15. Mikolov T., Sutskever I., Chen K., Corrado G., Dean J. Distributed representations of words and phrases and their compositionality. *Proc. 26th International Conference on Neural Information Processing Systems* (5–8 December 2013, Lake Tahoe, Nevada, USA). Lake Tahoe, Nevada, USA, 2013. P. 3111–3119.

16. Mikolov T., Grave E., Bojanowski P., Puhrsch C., Joulin A. Advances in pre-training distributed word representations. *Proc. Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (7–12 May 2018, Miyazaki, Japan). Miyazaki, Japan, 2018. P. 52–55.

17. Pogorilyy S., Kramov A. Method of the coherence evaluation of Ukrainian text. *Data Recording, Storage & Processing*. 2018. Vol. 20, N 4. P. 64–75. https://doi.org/10.35681/1560-9189.2018.20.4.178945.

18. OntoNotes Release 5.0. Linguistic Data Consortium, Catalog.ldc.upenn.edu, 2020. URL: https://catalog.ldc.upenn.edu/LDC2013T19.

*Надійшла до редакції 13.03.2020*

**С.Д. Погорелый, А.А. Крамов**

**ОЦЕНКА КОГЕРЕНТНОСТИ ТЕКСТА С ПОМОЩЬЮ ПОСТРОЕНИЯ ГРАФА СЕМАНТИЧЕСКОЙ И ЛЕКСИКО-ГРАММАТИЧЕСКОЙ СОГЛАСОВАННОСТИ СЛОВОСОЧЕТАНИЙ ПРЕДЛОЖЕНИЙ**

**Аннотация.** Предложен метод оценки когерентности текстов с помощью построения графа семантической и лексико-грамматической согласованности словосочетаний предложений. Осуществлена экспериментальная проверка эффективности метода на англоязычном корпусе. Полученные результаты рассчитанных метрик предложенного метода превышают соответствующие значения других современных подходов. Метод может быть применен к текстам других языков путем замены лингвистической модели в соответствии с особенностями конкретного языка.

**Ключевые слова:** обработка естественного языка, оценка когерентности текста, двудольный граф словосочетаний, метод расчета когерентности текстов на основе графа, лексико-грамматическая согласованность предложений.

**S.D. Pogorilyy, A.A. Kramov**

**ASSESSMENT OF TEXT COHERENCE BY CONSTRUCTING THE GRAPH OF SEMANTIC, LEXICAL AND GRAMMATICAL CONSISTENCY OF PHRASES OF SENTENCES**

**Abstract.** The graph-based method of coherence evaluation of texts based on the analysis of semantic, grammatical, and lexical consistency of sentence phrases has been suggested. The experimental verification of the efficiency of the method has been performed on the English-language corpus. The metrics obtained can indicate that the suggested method outperforms other state-of-the-art approaches. The method can be applied to other languages by replacing the linguistic models according to the features of a certain language.

**Keywords:** natural language processing, evaluation of text coherence, bipartite graph of phrases, graph-based method of coherence assessment of texts, lexical and grammatical consistency of sentences.

**Погорілий Сергій Дем'янович,**
доктор техн. наук, професор, завідувач кафедри Київського національного університету імені Тараса Шевченка, e-mail: sdp77@i.ua.

**Крамов Артем Андрійович,**
аспірант Київського національного університету імені Тараса Шевченка,
e-mail: artemkramovphd@knu.ua.