УДК 510.635:004.891(045)

А.І. Вавіленкова, к.т.н., доц.
Національний авіаційний університет

СИСТЕМА ПОРІВНЯЛЬНОГО АНАЛІЗУ ТЕКСТОВИХ ДОКУМЕНТІВ

У матеріалах статті проаналізовано роботу існуючих інтелектуальних систем порівняльного аналізу електронних текстових документів, наведено порівняльну характеристику їх основних функціональних можливостей. Запропоновано інтелектуальну систему порівняльного аналізу текстових документів, що використовує нову методику пошуку текстових збігів, в основі якої лежить побудова логіко-лінгвістичних моделей текстової інформації. Розроблена система складається з п'яти блоків функціонування. Перший блок відповідає за подачу текстів на вхід системи. Другий – за ідентифікацію складових логіко-лінгвістичних моделей текстів, що передбачає створення множини логіко-лінгвістичних моделей усіх речень природної мови, що утворюють кожен текст. У третьому блоці для кожного з двох текстів будується його логіко-лінгвістична модель, що складається з лінгвістичної та семантико-синтаксичної складових. Далі відбувається об'єднання та заміна структурних компонентів логіко-лінгвістичних моделей, отриманих на попередньому блоці функціонування системи. П'ятий блок присвячено порівнянню логіко-лінгвістичних моделей текстових документів, внаслідок якого знаходиться відсоток збігу за стилістикою, за структурою складних синтаксичних частин, схожістю ключових слів та словосполучень, тематикою та унікальністю.

Ключові слова: *природна мова; логіко-лінгвістична модель; текстова інформація; обробка тексту; система порівняльного аналізу; інтелектуальна система.*

Постановка проблеми. Існуючі відкриті системи порівняльного аналізу текстової інформації, такі як «*Advego Plagiatus*», «*Shingles Expert*», «*Compare It!*», «*IsEqual*», «*Cognitive Dwarf*», а також системи, що здійснюють повнотекстовий пошук та аналітичну обробку текстів, містять в своїй основі спільні механізми вилучення знань з текстової інформації та базуються на статистичних методах [1–3].

Порівняння – це зіставлення об'єктів з метою виявлення спільних рис або різниці між ними. Прийом порівняння використовується в процесі узагальнення, коли необхідно виявити тотожності, збіги та протиріччя в об'єктах дослідження. Тут тотожність – це повноцінний збіг усіх ознак; збіг – узгодження ознак, починаючи з однієї; протиріччя – коли ознаки одних об'єктів відсутні в інших. Для здійснення порівняння необхідні ознаки, що визначають можливі відношення між об'єктами.

Одним із методів, що застосовується для виявлення кластерів документів, які мають схожі властивості лише за деякими ознаками, наприклад, словами чи зображеннями, є *бікластеризація*. Метод застосовується для здійснення запитів та індексації повнотекстових систем. Початкові дані являють собою матрицю, в якій рядки відповідають за слова, а стовпчики – за документи. Для кластеризації документів враховується кількість входжень слова до документа, загальна кількість документів та кількість документів, що містить певне слово [5, с. 149]. Тобто, слова можуть бути кластеризовані на основі документів, в яких вони зустрічаються. Кластери зручні для автоматичної побудови статистичних тезаурусів, уточнення запитів та автоматичної класифікації документів, проте здійснити змістовний аналіз тексту з використанням кластерів неможливо. Дослідження показали, що ані коди бібліотечних класифікаторів, ані назва текстового документа, ані множина слів, що найчастіше зустрічаються у тексті, у більшості випадків недостатньо адекватні або зовсім неадекватні його змісту [9, с. 15]. Тому при їх використанні як критерію добору текстів стандартний пошуковий сервер видає величезний обсяг інформації, більша частина якої немає ніякого відношення до тематики тексту, що підлягає аналізу.

Аналіз останніх джерел і публікацій. В основі проблеми порівняльного аналізу електронних текстових документів за змістом лежить протиріччя між існуючими моделями формального опису текстової інформації та алгоритмами доведення їх адекватності. Питанням екстракції знань із текстових документів присвячено багато праць вітчизняних та зарубіжних вчених у сфері комп'ютерної лінгвістики. Так, Апресян Ю.Д. у своїй праці «Лексична

семантика» намагається [4, с. 256] пояснити сутність процесу утворення тотожних семантичних конструкцій за рахунок вживання у текстах конверсивів, синонімів, антонімів тощо Нікітін М.В. пропонує моделі базової конверсії для представлення знань речень природної мови за допомогою різних синтаксичних конструкцій [10, с. 455]. Величезний внесок у автоматизацію вилучення знань із текстових документів зробив Широков В.А., створюючи лінгвістичні корпуси, що є незамінним інструментом у змістовному аналізі електронних текстів [11, с. 143].

Проте задача створення уніфікованої моделі представлення знань як основного способу аналізу контенту електронних текстових документів досі залишається невирішеною. Аналіз роботи існуючих відкритих систем порівняльного аналізу електронних документів дозволив здійснити їх порівняльну характеристику (табл. 1).

Таблиця 1

Характеристика систем порівняльного аналізу текстових документів

Параметр / Назва програми	Антиплагіат	StrikePlagiarism	Advego Plagiatus	Cognitive Dwarf
Концепція отримання результатів порівняльного аналізу	Показує ступінь унікальності тексту, джерела дублювання тексту, відсоток збігу, розширена оцінка відсотку збігу	Показує ступінь унікальності тексту, джерела дублювання тексту, відсоток збігу	Показує ступінь унікальності тексту, джерела дублювання тексту, відсоток збігу	Показує ступінь унікальності тексту, джерела дублювання тексту, відсоток збігу
Функціональні можливості	Пошук серед існуючих документів у базі даних	Порівняння текстових документів з існуючими у доступних пошукових системах	Порівняння текстових документів з існуючими у доступних пошукових системах	Порівняння текстових документів з існуючими у доступних пошукових системах
Мова	укр., рос., англ., франц., нім., чеська, польська	укр., рос., англ., франц., нім., чеська, польська	рос., англ.	рос., англ.
Використані методи	Метод "шинглів"	Метод "шинглів"	Метод "шинглів"	Метод "шинглів"
Операційна система	Windows	Windows	Windows	Windows
Формат файлів	txt, doc, rtf	txt, doc, rtf		
Швидкодія	15–20 хв.	4 хв. – декілька год.	15–20 хв.	15–20 хв.
% збігу	82,3 %	76,1 %	65 %	61,5 %

Із таблиці видно, що діапазон зміни відсотку збігу для різних систем і для однакових текстів великий і складає приблизно 20 %. Це говорить про неточність роботи існуючих систем порівняльного аналізу та необхідність розробки якісно нових алгоритмів екстракції знань із текстових документів.

Постановка завдання. Метою статті є опис та обґрунтування функціонала основних блоків системи порівняльного аналізу текстових документів (СПАТЛЛМ), розкриття основних принципів функціонування системи на основі нової методики побудови логіко-лінгвістичних моделей.

Викладення основного матеріалу. Інформаційна система є середовищем організації та зберігання інформації, а реалізація функцій неможлива без знання орієнтованої на неї інформаційної технології. Система СПАТЛЛМ створена на основі розробленої автором інформаційної технології порівняльного аналізу текстових документів [6, с. 98].

Блок № 1. Подача текстів на вхід системи. Середовищем розробки інтелектуальної системи порівняльного аналізу текстових документів СПАТЛЛМ є NetBeans IDE 8.0.2, що надає всі засоби, необхідні для створення професійних додатків робочого середовища, корпоративних, мобільних та *web*-додатків на різних мовах програмування. На цьому етапі функціонування системи до інтерфейсу користувача вводяться два тексти (рис. 1).

Блок № 2. Ідентифікація складових логіко-лінгвістичних моделей текстів. Кожен із двох текстів проходить етап членування (технічний етап), результатом виконання якого є множина речень природної мови.

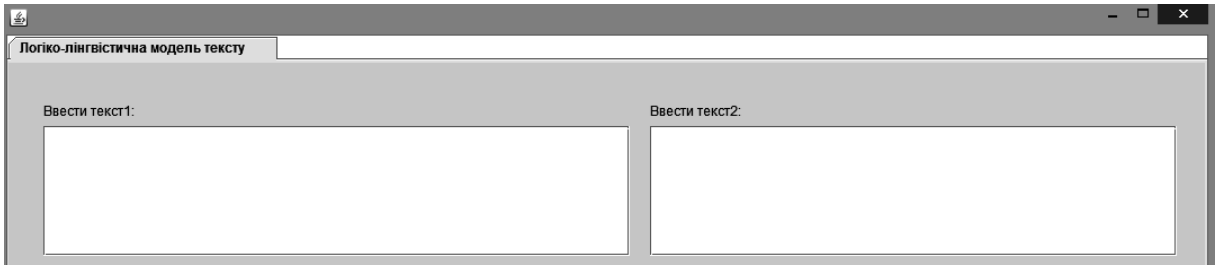

Рис. 1. Подача текстів на вхід системи СПАТЛЛМ

Просте речення природної мови представляється у вигляді атомарного предикату виду:

$$L(S) = p(x, g, y, q, z, r, h), \qquad (1)$$

де x – суб'єкт інваріантної до $L(S)$ логіко-лінгвістичної моделі; g – характеристика суб'єкта x; y – об'єкт логіко-лінгвістичної моделі; q – характеристика об'єкта y; p – відношення, що пов'язує суб'єкт x з об'єктом y логіко-лінгвістичної моделі; z – предмет p-го відношення між суб'єктом x та об'єктом y логіко-лінгвістичної моделі; r – характеристика предмету відношення; h – характеристик відношення p.

Геометрично логіко-лінгвістичну модель простого речення природної мови можна представити у вигляді паралелепіпеда, вершинами якого є предикат, суб'єкт, об'єкт, предмет відношень та їх характеристики відповідно (рис. 2).

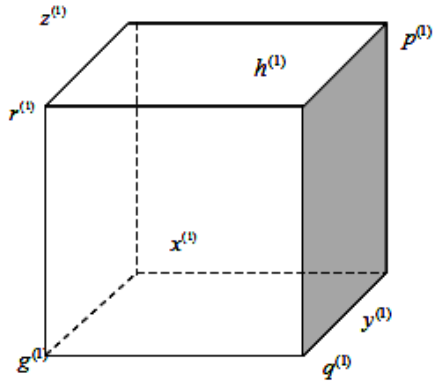

Рис. 2. Геометрична інтерпретація логіко-лінгвістичної моделі речення природної мови

Наприклад, для речення природної мови «*Українські вчені давно зробили значний внесок у світову лінгвістику*» логіко-лінгвістична модель має вигляд:

$L(S) = p(x, g, y, q, z, r, h)$,

$L(S) =$ зробили (вчені, українські, внесок, значний, лінгвістику, світову, давно).

Логіко-лінгвістична модель (1) будується на основі граматичних характеристик кожного зі слів, а також на основі продукційної моделі знань, що містить систему правил утворення словосполучень [6, с. 99]. Таким чином, етап ідентифікації передбачає побудову логіко-лінгвістичних моделей для кожного речення природної мови та надання інформації щодо логічних зв'язків в середині цих речень (рис. 3).

Блок № 3. Побудова логіко-лінгвістичних моделей текстів. *Логіко-лінгвістична модель текстового документа* – це абстрактна модель, яка об'єднує в собі основні властивості тексту та його складових частин, відображає основні взаємозв'язки між структурними компонентами, являє собою впорядковану четвірку та масив логіко-лінгвістичних моделей речень природної мови, що входять до тексту [7, с. 176].

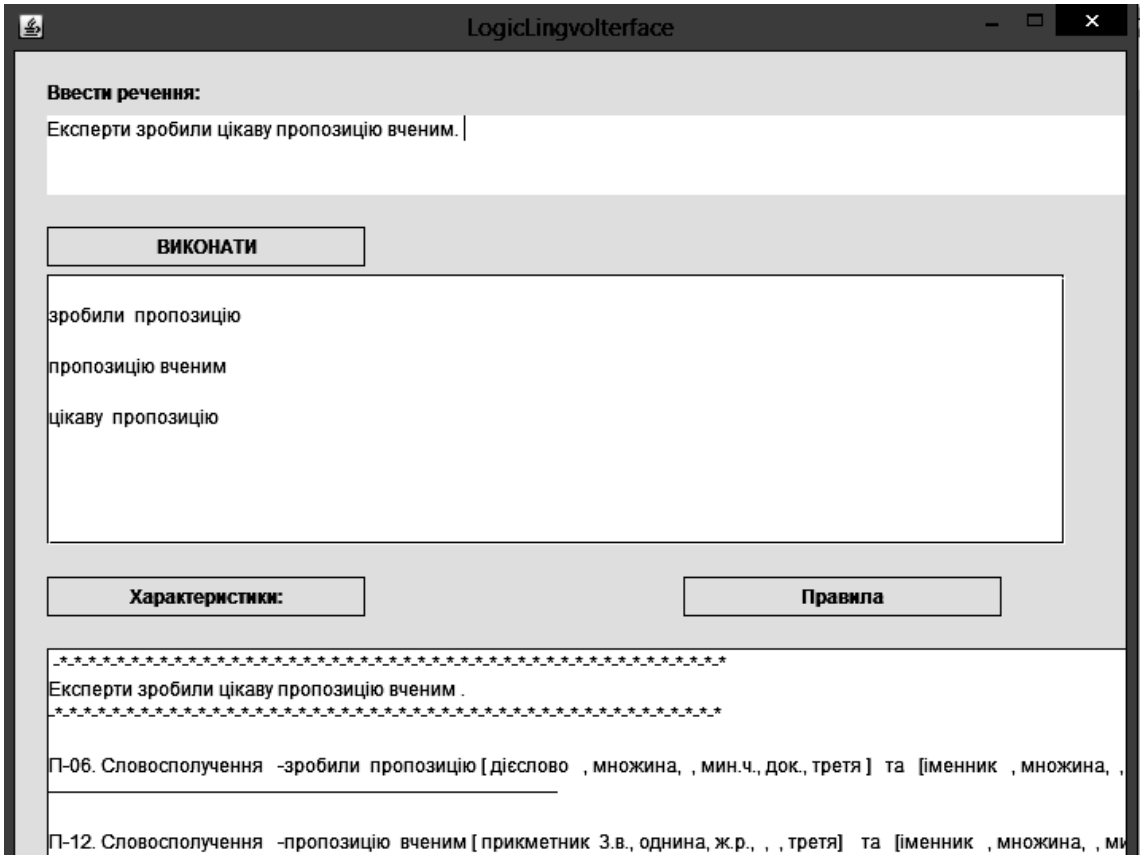

Рис. 3. Приклад встановлення зв'язків у реченні природної мови системою СПАТЛЛМ

Лінгвістична складова формального опису тексту:

$$t =< CQ, F, B, A >, \qquad (2)$$

де T – множина текстів; $t \in T$ – конкретний електронний текст із всієї множини текстів; $CQ = \{cq_1, ..., cq_i, ..., cq_n\}$ – множина існуючих типів текстів, $i = \overline{1, n}$, n – кількість типів; $F = \{f_1, ..., f_j, ..., f_m\}$ – множина складних синтаксичних частин тексту, $j = \overline{1, m}$, m – кількість складних синтаксичних частин; B – текстова база, що складається з набору ключових слів тексту та взаємопов'язаних пропозицій, і яку можна представити у вигляді трійки: $B =< K, SJ, D >$, K – множина ключових слів тексту; SJ – множина ключових словосполучень тексту; D – множина пропозицій [8, с.]; $A = \{a_1, ..., a_k, ..., a_q\}$ – множина абзаців тексту, $k = \overline{1, q}$, q – кількість абзаців. Кожен абзац, у свою чергу, описується трійкою: $a =< H, Y, R >$, $H = \{1, 2\}$ – множина типів зв'язків між реченнями (ланцюговий чи паралельний); $Y = \{1, 2, 3, 4, 5\}$ – множина

типів тематичних прогресій, що вжиті у абзаці $a_k \in A$; $R = \{1,2,3,4,5,6,7\}$ – множина рематичних домінант у абзаці $a_k \in A$.

Семантико-синтаксична складова формального опису тексту:

$$t' = \bigwedge_{\delta=1}^{N(t)} L^{S_\delta}, \tag{3}$$

де L^{S_δ} – логіко-лінгвістична модель речення S_δ, $\delta = \overline{1, N(t)}$, сформована за формулою (1); $N(t)$ – кількість речень у тексті t.

Після введення двох текстів система СПАТЛЛМ автоматично будує логіко-лінгвістичні моделі заданих текстів у формульному та природно мовному вигляді (рис. 4).

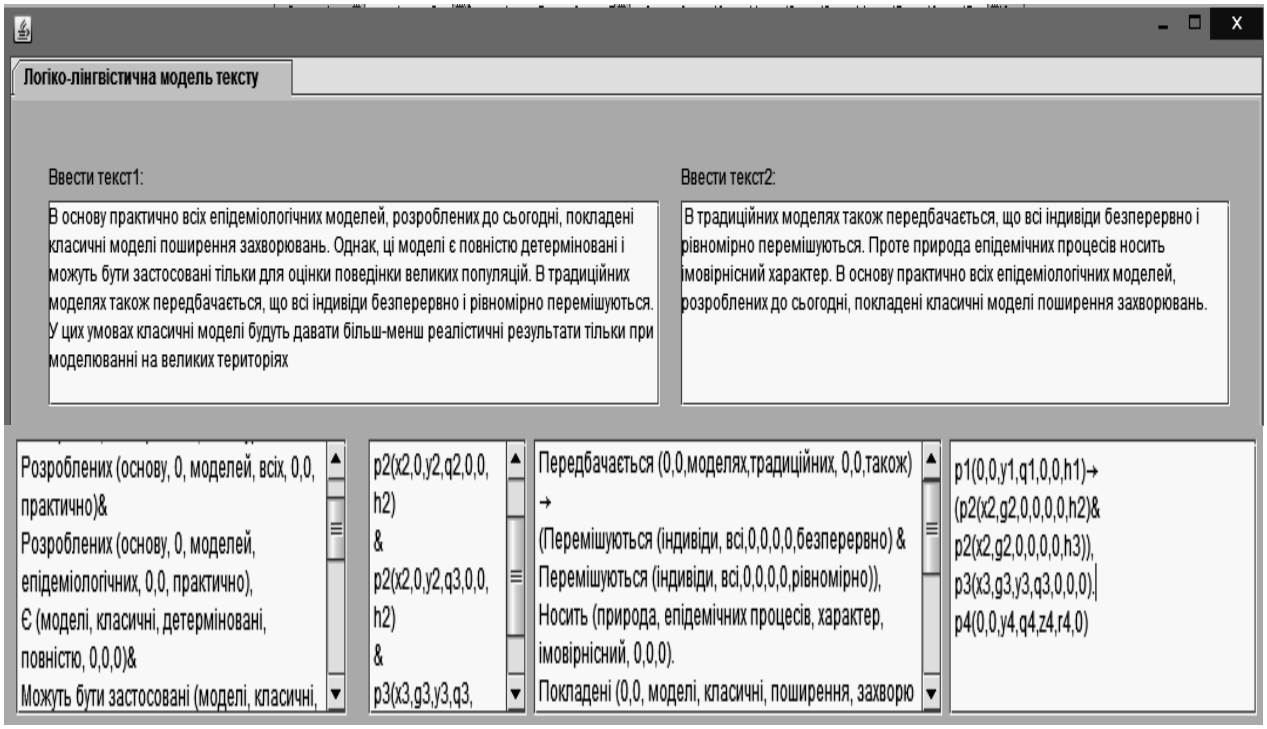

Рис. 4. Формування логіко-лінгвістичних моделей заданих електронних текстових документів системою СПАТЛЛМ

Блок № 4. Синтез логіко-лінгвістичних моделей речень природної мови. Виконання даного блока системою передбачає об'єднання та заміну структурних компонентів логіко-лінгвістичних моделей (3), отриманих на попередньому блоці функціонування системи. Це відбувається на основі виявлення способів логічного зв'язку між реченнями природної мови. Отже, масив логіко-лінгвістичних моделей речень текстових документів (3) набуває вигляду (4):

$$t'^{(\gamma)} = \begin{cases} L^{S_1(\gamma)} = \bigwedge\limits_{p \in P^{S_1}} \bigwedge\limits_{h \in H_p^{S_1}} L_p^{S_1(\gamma)}(h), \\ L^{S_2(\gamma)} = \bigwedge\limits_{p \in P^{S_2}} \bigwedge\limits_{h \in H_p^{S_2}} L_p^{S_2(\gamma)}(h), \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ L^{S_\delta(\gamma)} = \bigwedge\limits_{p \in P^{S_\delta}} \bigwedge\limits_{h \in H_p^{S_\delta}} L_p^{S_\delta(\gamma)}(h), \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ L^{S_{N(t)}(\gamma)} = \bigwedge\limits_{p \in P^{S_{N(t)}}} \bigwedge\limits_{h \in H_p^{S_{N(t)}}} L_p^{S_{N(t)}(\gamma)}(h), \end{cases} \tag{4}$$

Блок № 5. Порівняння логіко-лінгвістичних моделей текстів. Даний блок аналізує лінгвістичні складові логіко-лінгвістичних моделей введених текстів (2). На основі порівняння перших параметрів лінгвістичних складових *CQ* та *CQ* визначається відсоток збігу за стилістикою $\eta(CQ)$, почергове порівняння наступних параметрів дає відсоток збігу $\eta(F)$ – за структурою складних синтаксичних частин, $\eta(B)$ – за схожістю ключових слів та словосполучень, – за тематикою відповідно.

Перевірка низки умов, зокрема умов протиріччя та синонімії двох речень природної мови, можливість побудови імплікативних та пресуппозиційних зв'язків між атомарними предикатами, а також умови організації граматичного зв'язку між предикативними частинами речень дає змогу визначити загальний відсоток унікальності текстів $\eta(S)$ (рис. 5). Для цього система СПАТЛЛМ використовує метод порівняння логіко-лінгвістичних моделей простих речень природної мови та метод порівняння речень природної мови довільної складності.

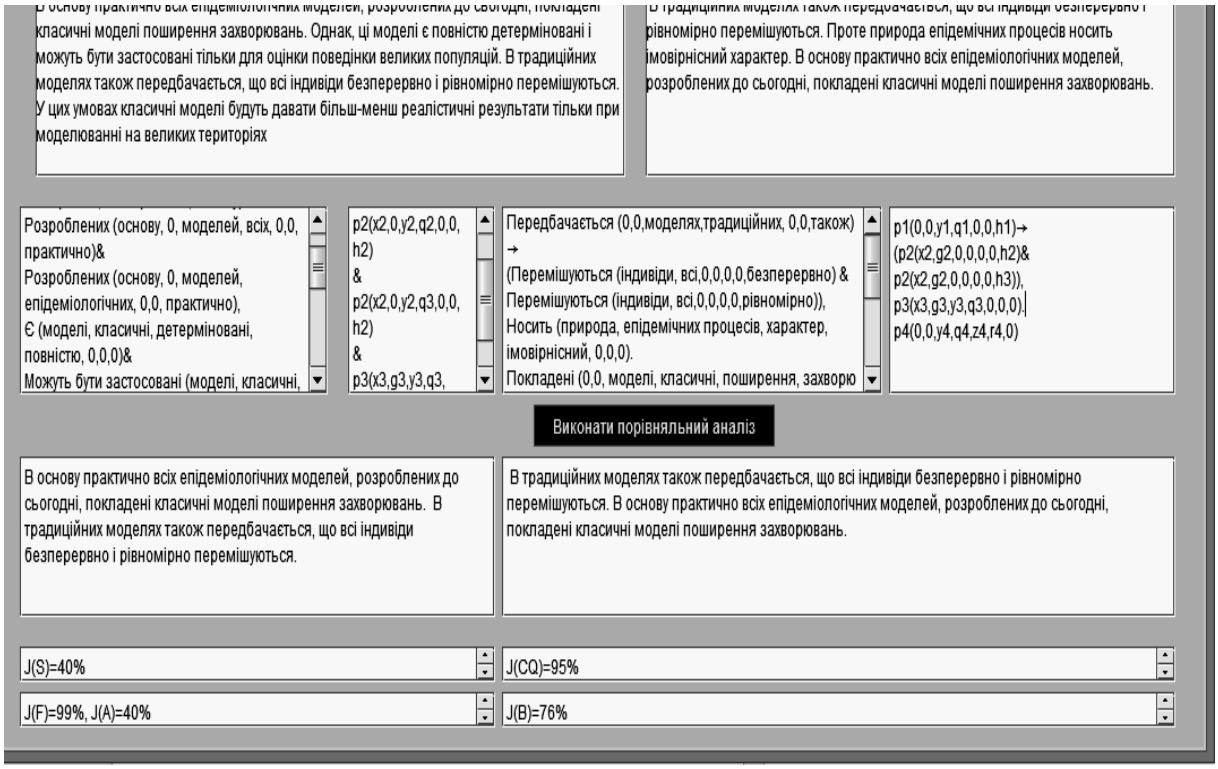

Рис. 5. Результат порівняння логіко-лінгвістичних моделей двох текстів

Висновки та перспективи подальших досліджень. Функціональні можливості інтелектуальної системи порівняльного аналізу текстових документів СПАТЛЛМ полягають у наданні вичерпної інформації щодо граматичних характеристик слів та логічних зв'язків між ними як у межах речення природної мови, так і у всьому електронному документі. СПАТЛЛМ дає можливість перетворити текстову інформацію, подану на вхід у вигляді файла з розширенням *.txt, у математичну модель для можливості застосування до неї методів комп'ютерної оброки. Інтелектуальна система порівняльного аналізу текстових документів СПАТЛЛМ дозволяє:

- отримувати з речення природної мови довільної складності набір словосполучень;
- отримувати граматичні характеристики кожного слова заданого речення;
- надавати вичерпну інформацію про граматичні параметри слів, навіть якщо для формування словосполучення підходить лише один набір таких параметрів;
- автоматично обирати коректні граматичні характеристики слова для формування словосполучення за наявності омонімії;
- надавати інформацію, за якими лінгвістичними правилами сформовано словосполучення;

- автоматично формувати логічні зв'язки між концептами речення української мови з наданням інформації про їх характеристики;
- автоматично порівнювати речення природної мови;
- обчислювати відсоток збігу з виявленням частин речення, що дублюються за змістом;
- надавати інформацію щодо умов тотожності речень природної мови;
- автоматично порівнювати електронні природномовні тексти;
- обчислювати відсоток збігу з виявленням частин текстів, що дублюються за змістом;
- надавати інформацію щодо умов тотожності природномовних текстів;
- надавати розширену оцінку тотожності текстів.

Система створена на основі розробленої інформаційної технології порівняльного аналізу текстових документів, дозволяє обчислити відсоток збігу між двома електронними документами, що порівнюються, з наданням детальної інформації про правила та умови тотожності, які спрацювали при обчисленні відсотку збігу, правила утворення синтаксичних конструкцій у введених текстах та надає розширену оцінку тотожності текстів, що порівнюються.

Список використаної літератури:

1. Advego Plagiatus – проверка уникальности текста, 2008–2016 [Електронний ресурс]. – Режим доступу : http://advego.ru/plagiatus/top/.
2. Он-лайн сервис проверки текста на уникальность, 2016 [Електронний ресурс]. – Режим доступу : www.text.ru.
3. Антиплагіатна Інтернет-система, 2015 [Електронний ресурс]. – Режим доступу : http://strikeplagiarism.com/ua/antiplagiarism-system/
4. *Апресян Ю.Д.* Лексическая семантика : в 2-х т. Т. 1 / *Ю.Д. Апресян*. – М. : Восточная литература, 1995. – 422 с.
5. Методы и модели анализа данных: OLAP и Data Mining / *А.А. Барсегян, М.С. Куприянов, В.В. Степаненко, И.И. Холод*. – СПб. : БХВ-Петербрг, 2007. – 384 с.
6. *Вавіленкова А.І.* Порівняльний аналіз речень природної мови за змістом / *А.І. Вавіленкова* // Математичні машини та системи. – 2015. – № 2. – С. 97–103.
7. *Вавиленкова А.И.* Основные принципы синтеза логико-лингвистических моделей / *А.И. Вавиленкова* // Кибернетика и системный анализ. – 2015. – Т. 51, № 5. – С. 176–185.
8. *Лайонз Дж.* Лингвистическая семантика : монография / *Дж.Лайонз.* – М. : Языки славянской культуры, 2003. – 400 с.
9. *Ландэ Д.В.* Поиск знаний в Internet. Профессиональная работа / *Д.В. Ландэ.* – М. : Диалектика, 2005. – 272 с.
10. *Никитин М.В.* Курс лингвистической семантики : учеб. пособие / *М.В. Никитин*. – 2-е изд. – СПб. : Изд-во РГПУ им. А.И. Герцена, 2007. – 819 с.
11. Лінгвістичні та технологічні основи тлумачної лексикографії / *В.А. Широков, В.М. Білоноженко, О.В. Бугаков та ін.* – К. : Довіра, 2010. – 295 с.

ВАВІЛЕНКОВА Анастасія Ігорівна – кандидат технічних наук, доцент кафедри комп'ютеризованих систем управління інституту Комп'ютерних інформаційних технологій Національного авіаційного університету.

Наукові інтереси:
- комп'ютерна лінгвістика;
- автоматизована обробка природної мови;
- експертні системи

Тел.: 066–751–65–01.

E-mail: a_vavilenkova@mail.ru.

Стаття надійшла до редакції 25.11.2015