– *метод сценариев*. Идея метода состоит в том, чтобы представить диаграмму в виде конечного автомата и тестировать его согласно определению и его свойствам [4].

Модуль обработки данных

В данном модуле происходит непосредственная идентификация ошибок, а также получение результата о правильности диаграммы в целом.

Пользовательский интерфейс

Тут происходит оптимизация результатов и реализация удобного пользовательского интерфейса.

Выводы и предложения. Рассмотрев указанные продукты, позволяющие осуществлять верификацию диаграмм, можно утверждать, что TAUG2 от Telelogic, Enterprise Architect и Visual Paradigm более эффективные и продвинутые по сравнению с AgroUML,так как они полностью завершены и поддерживают стандарт UML 2.0. Однако AgroUML является бесплатным, кроссплатформенным и открытым, что существенно упрощает доступ к нему. Тем не менее его незавершенность, а также отсутствие оптимальных поддержек многих диаграмм не позволяет использовать его как основное средство для проектирования диаграмм.

На основании рассмотренных вышесуществующих средств верификации можно предположить, что они не являются полноценными и не предоставляют полной картины корректности диаграммы. В связи с этим предлагается новое средство верификации UMLTester, которое является бесплатным, обеспечивает верификацию несколькими методами и позволяет максимально точно оценить корректность диаграммы.

Список использованных источников

1. *Режим* доступа: http://argouml.tigris.org/.
2. *Режим* доступа: http://www.visual-paradigm.com/.
3. *Макгрегор Дж.* Тестирование объектно-ориентированного программного обеспечения : практическое пособие / Дж. Макгрегор, Д. Сайкс ; пер. с англ. – К. : ООО «ТИД ”ДС“», 2002. – 432 с.
4. *Хоар Ч.* Взаимодействующие последовательные процессы / Ч. Хоар ; пер. с англ. – М. : Мир, 1989. – 264 с.

УДК 004.82(045)

А.І. Вавіленкова, канд. техн. наук
Національний авіаційний університет, м. Київ, Україна

ПРОГРАМНЕ ЗАБЕЗПЕЧЕННЯ ДЛЯ ВИЯВЛЕННЯ ТЕКСТОВИХ ДОКУМЕНТІВ, ІДЕНТИЧНИХ ЗА ЗМІСТОМ

А.И. Вавиленкова, канд. техн. наук
Национальный авиационный университет, г. Киев, Украина

ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ДЛЯ ВЫЯВЛЕНИЯ ТЕКСТОВЫХ ДОКУМЕНТОВ, ИДЕНТИЧНЫХ ПО СМЫСЛУ

A.I. Vavilenkova, Candidate of Technical Sciences
National Aviation University, Kyiv, Ukraine

SOFTWARE FOR DETECTION OF TEXT DOCUMENTS IDENTICAL IN CONTENT

Проаналізовано основні методи, що лежать в основі відкритого програмного забезпечення з виявлення дублікатів електронних документів, зазначено їх недоліки: відсутність компоненти семантичного та змістовного аналізу текстів. Запропоновано систему автоматизованого формування логіко-лінгвістичних моделей як допоміжний механізм вилучення змісту з речень природної мови, на основі якого можна вирішити проблему екстракції знань з текстової інформації.

Ключові слова: *текстові документи, семантичний аналіз, логіко-лінгвістичні моделі, аналіз тексту, дублікати, природна мова.*

Осуществлен анализ основных методов, которые лежат в основе открытого программного обеспечения по выявлению дубликатов электронных документов, определены их недостатки: отсутствие компоненты семантиче-

ского и смыслового анализа текстов. Предложена система автоматизированного формирования логико-лингвистических моделей как вспомогательный механизм извлечения содержания из предложений естественного языка, на основе которого можно решить проблему экстракции знаний из текстовой информации.

Ключевые слова: *текстовые документы, семантический анализ, логико-лингвистические модели, анализ текста, дубликаты, естественный язык.*

The basic techniques underlying open source software to detect duplicates, butchers electronic documents indicated their shortcomings: lack of semantic component and component of content analysis of texts. The system of automated creation of logic-linguistic model as an auxiliary mechanism for extracting the contents of the sentences of natural language, from which you can solve the problem of extracting knowledge from text information.

Key words: *text documents, semantic analysis, logic-linguistic models, text analysis, duplicate, natural language.*

Постановка проблеми. Щоразу під час перевірки дипломних, курсових, дисертаційних робіт, звірки різноманітних нормативних документів виникає потреба у здійсненні їх порівняння для знаходження збігів та суперечностей з метою виявлення і запобігання використанню ідентичних текстів. З проблемою визначення збігів у текстових документах зіштовхуються у тих сферах діяльності, де кінцевим результатом є текстовий документ. Це, в першу чергу, освіта, наука, законотворчість, патентування, інноваційна та інша діяльність, пов'язана із захистом інтелектуальної власності.

З появою сьогодні величезної кількості веб-документів у мережі Інтернет виникають проблеми у функціонуванні пошукових машин. Релевантність інформації, отриманої на запит навіть найдосконаліших пошукових систем *Google* та *Yandex*, дуже низька. Отримані на запити відповіді містять велику кількість копій веб-документів.

Все це говорить про те, що існуючі алгоритми пошуку текстових збігів не здатні опрацьовувати електронні документи за змістом. Більшість методик, на основі яких працюють сучасні пошукові машини, базуються на таких статистичних методах, як метод «шинглів», алгоритм Карпа-Рабіна, алгоритм Кнута-Морріса-Пратта, метод лексичних сигнатур. Це дозволяє виявити в електронних документах ідентичні за написанням фрагменти, проте не дає можливості говорити про тотожність документів, якщо текстова інформація перефразована, у тексті вжиті синоніми, омоніми, інверсний порядок слів і т. д.

Таким чином, для виявлення текстових документів, ідентичних за змістом, необхідно розроблення якісно нових алгоритмів лінгвістичного та семантичного аналізу. Зокрема, існуюча у лінгвістів теоретична база аналізу семантико-синтаксичних структурних одиниць тексту слабо структурована. Тому розроблення програмного забезпечення для знаходження змістовних текстових збігів потребує, в першу чергу, формалізації інформації щодо побудови текстів, формування у них логічних зв'язків та створення методів автоматичного семантичного аналізу.

Аналіз останніх досліджень і публікацій. Дослідження показали, що ні коди бібліотечних класифікаторів, ні назва текстового документа, ані множина слів, які найчастіше трапляються у тексті, у більшості випадків недостатньо адекватні або зовсім неадекватні його змісту. Тому під час їх використання як критерій добору текстів стандартний пошуковий сервер видає величезний обсяг інформації, більша частина якої немає ніякого відношення до тематики тексту, що підлягає аналізу [1].

Перші теоретичні спроби формалізувати процес здійснення семантичного аналізу тексту робить В.А. Звегінцев [2], а згодом Е.В. Попов у своїй роботі [3], де формулює принципи побудови неструктурованого та структурованого семантичного графу речення. У роботі [4] В.А. Звегінцев намагається узагальнити всі відомі знання з когнітивних аспектів мови, які публікують такі іноземні вчені, як Дж. Лакофф, Ч. Філлмор, Д. Шпербер, Д. Уілсон, Т.А. ван Дейк.

У книзі І.А. Мельчука [5] проведено глибокий аналіз тексту, описано етапи побудови моделі «зміст-текст», обговорено семантичне представлення висловлювань та універсальні правила перефразування. Роботи провідних вітчизняних науковців [6; 7] у

сфері лексикографії та комп'ютерної лінгвістики описують теорію семантичних станів мовних одиниць.

Виділення не вирішених раніше частин загальної проблеми. Всі перераховані вище матеріали носять теоретичний характер. А в основі відкритого програмного забезпечення щодо пошуку дублікатів серед електронних документів відсутнє таке, в основі якого лежало б змістовне оброблення текстової інформації.

Мета статті. Метою роботи є аналіз методів, що лежать в основі відкритого програмного забезпечення з виявлення ідентичних текстових документів; розроблення принципів функціонування систем аналізу електронних текстів за змістом; демонстрація роботи системи автоматизованого формування логіко-лінгвістичних моделей текстової інформації.

Виклад основного матеріалу. Існуючі відкриті системи порівняльного аналізу текстової інформації, такі як *Advego Plagiatus*, *Shingles Expert*, *Compare It!*, *IsEqual*, а також системи, що здійснюють повнотекстовий пошук та аналітичне оброблення текстів, містять у своїй основі спільні механізми вилучення знань з текстової інформації та базуються на статистичних методах. Розглянемо основні принципи функціонування програмних систем пошуку дублікатів серед електронних документів.

Порівняння – це співставлення об'єктів з метою виявлення спільних рис або різниці між ними. Прийом порівняння використовується у процесі узагальнення, коли необхідно виявити тотожності, збіги та суперечності в об'єктах дослідження. Тут тотожність – це повноцінне співпадіння всіх ознак; збіг – співпадіння ознак, починаючи з однієї; суперечності – 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 володіють схожими властивостями 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 число входжень слова до документа, загальна кількість документів та кількість документів, що містить певне слово [1]. Таким чином слова можуть бути кластеризовані на основі документів, в яких вони трапляються. Кластери зручні для автоматичної побудови статистичних тезаурусів, уточнення запитів та автоматичної класифікації документів, проте здійснити змістовний аналіз тексту з використанням кластерів неможливо.

Advego Plagiatus – програмний продукт, що дозволяє здійснити семантичний аналіз тексту он-лайн, знаходячи при цьому семантичне ядро, а також дає можливість перевірити електронний документ на унікальність відносно веб-документів, які знаходяться у відкритому доступі в мережі Інтернет.

Нехай на вхід системи подано текст (рис. 1). Результат перевірки введеного тексту на унікальність (рис. 2) містить відсоток знайдених збігів, а також наведено посилання на сайти, де є схожий текст. Але, якщо проаналізувати введений текст детальніше, можна побачити, що збіги знайдено лише у назві літератури, на яку наведене посилання. Таким чином, змістовних збігів система не повинна була знайти. Це означає, що здійснено не семантичний розбір за змістом, а системою *Advego Plagiatus* використано статистичний метод пошуку тотожних виразів.

Пошук нечітких дублікатів передбачає кластеризацію документів за схожістю їх певних характеристик, а реалізація алгоритму складається з таких кроків:

– канонізація кроків – на цьому етапі текст очищується від непотрібних слів, що не несуть зміст під час порівняння, тобто текст приводиться до канонічної форми;

– розбиття тексту на шингли;

– знаходження контрольних сум – унікальних чисел, кожному з яких ставиться у відповідність деякий текст та функція його обчислення. Потім із всієї множини контрольних сум (їх кількість дорівнює кількості слів у документі мінус $(w-1)$, де w – число слів у шинглі) відбираються лише ті, які діляться на певне вибране завчасно число;

– пошук однакових послідовностей – один шингл, який співпав під час відбору, приблизно відповідає наперед заданому числу однакових частин у повному тексті.

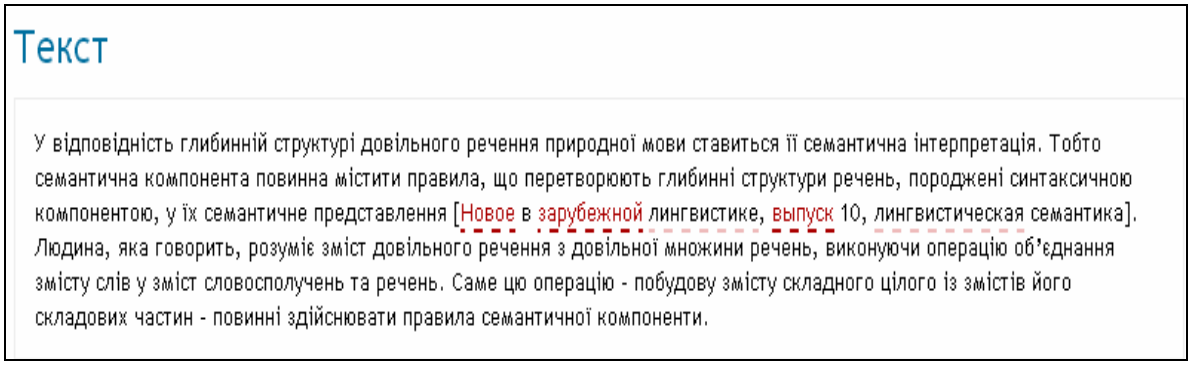

Рис. 1. Введення тексту для аналізу програмою Advego Plagiatus

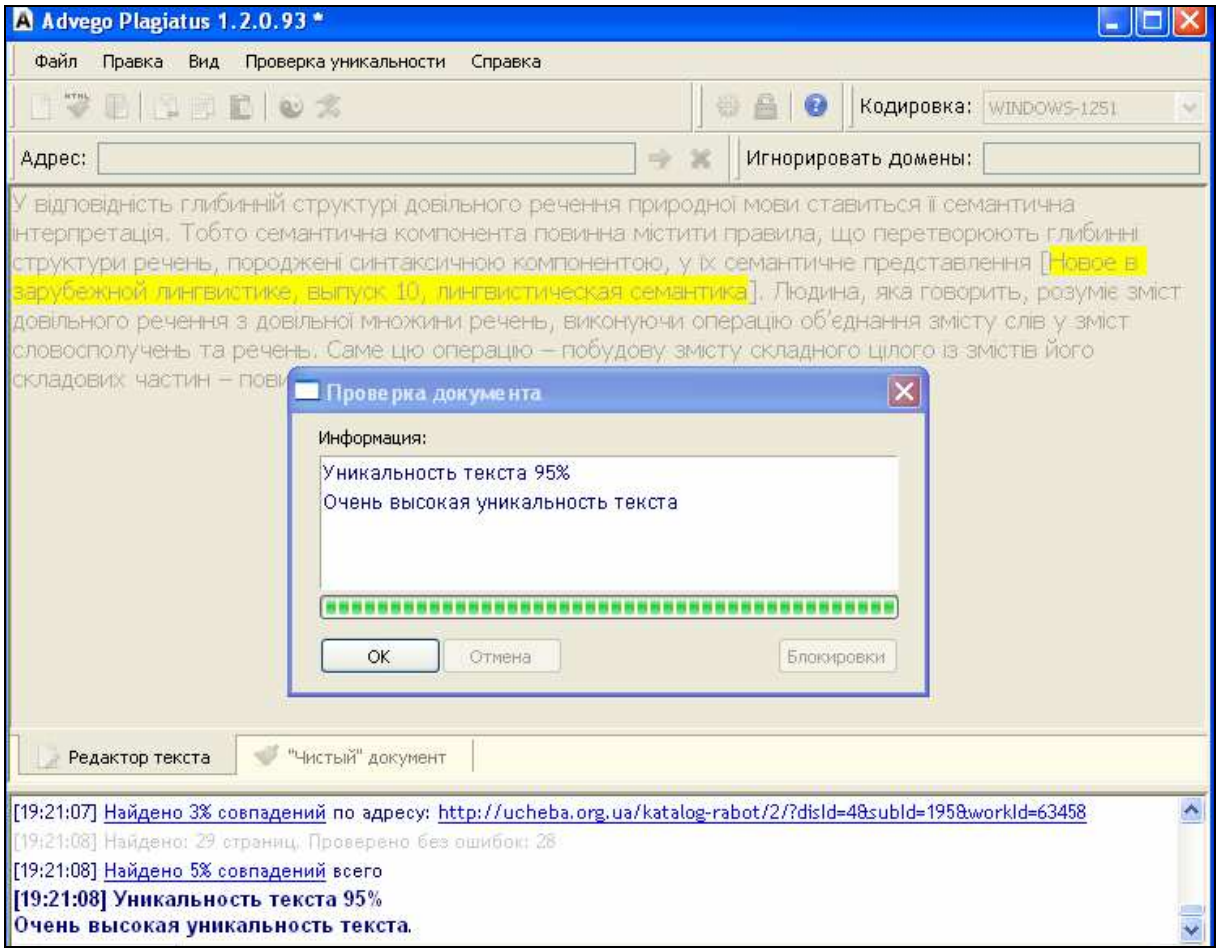

Рис. 2. Семантичне ядро введеного тексту в програму Advego Plagiatus

Пошук нечітких дублікатів лежить в основі функціонування системи порівняльного аналізу *Shingles Expert* (рис. 3).

На вхід системи подано два абсолютно різних за змістом тексти, в яких часто трапляються такі слова, як «контент» та «система управління контентом». Саме завдяки виявленню цих ключових слів в обох текстах система, яка працює на основі статистично-

го методу шинглів, видала відсоток збігів – 35 %. Експеримент показав, що тексти не аналізуються за змістом.

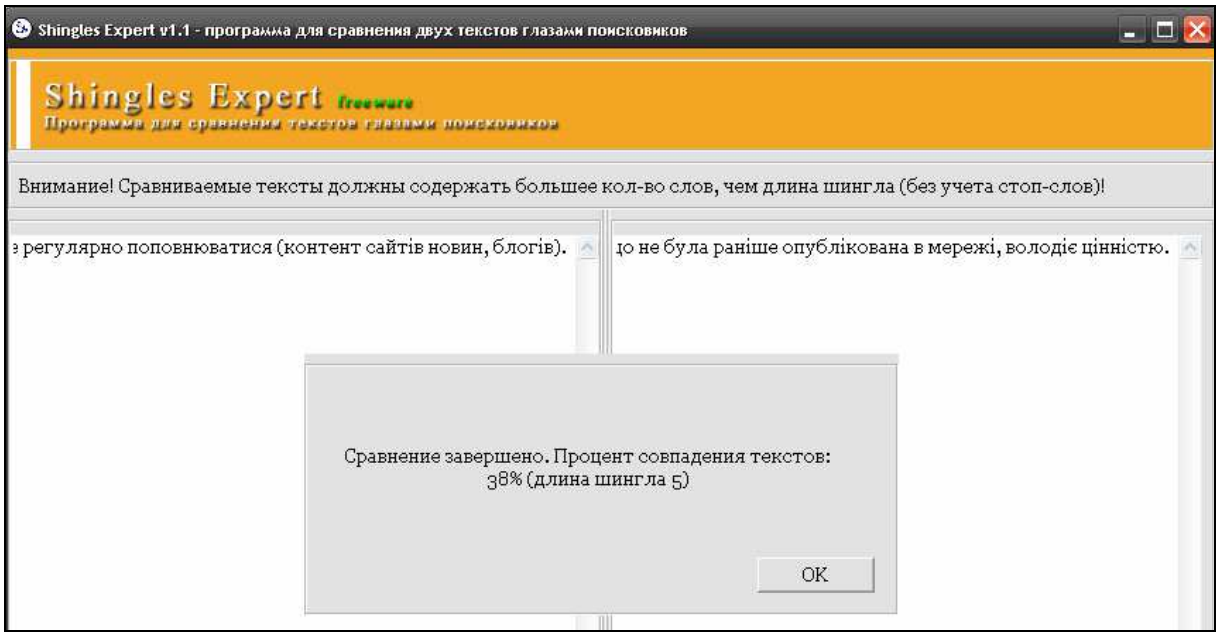

Рис. 3. Результат порівняння текстів системою Shingles Expert

Лексико-синтаксичні шаблони являють собою характерні вирази (словосполучення або звороти), конструкції із відповідних елементів природної мови. Такі шаблони дозволяють побудувати семантичну модель тексту. Передбачається, що саме за допомогою шаблонів (зразків) можна описати лексичні відношення в документі. Такий метод використовує ієрархію шаблонів, що складаються з індикаторів частин мови та групових символів.

Таким чином, аналіз програмного забезпечення, що використовується для виявлення текстових документів, показав недосконаліть методів, які лежать в основі сучасних систем автоматичного аналізу. Отже, існує потреба у створенні такого програмного забезпечення, яке б використовувало методи глибокого лінгвістичного оброблення. Це дасть можливість порівнювати текстові документи за змістом.

Програмним забезпеченням, яке дозволяє вилучати зміст з речень природної мови є система автоматизованого формування логіко-лінгвістичних моделей (САФЛЛМ). Вона являє собою прикладну програму, яка виступає засобом формування бази знань аналітичної системи порівняльного аналізу. Тобто система САФЛЛМ виконує функції допоміжного механізму, що використовується для визначення набору логіко-лінгвістичних моделей ключових речень, які максимально точно відображають зміст тексту, що підлягає подальшому аналізу, і вирішує проблему екстракції знань з текстової інформації. В основу роботи цієї системи покладено відповідність між формулами логіки предикатів та концептами, що належать реальному світу. Таким чином, САФЛЛМ є прикладним засобом, який використовується в системі порівняльного аналізу електронних текстів, утворюючи для неї базу знань у вигляді набору логіко-лінгвістичних моделей, що відображають семантико-синтаксичну структуру речень.

Система автоматизованого формування логіко-лінгвістичних моделей функціонує таким чином.

1. На вхід системи користувачем вводиться речення природної мови (рис. 4). Під час натискання кнопки «Модель» на екран буде виведено логіко-лінгвістичну модель – результат перетворення текстової інформації у формулу.

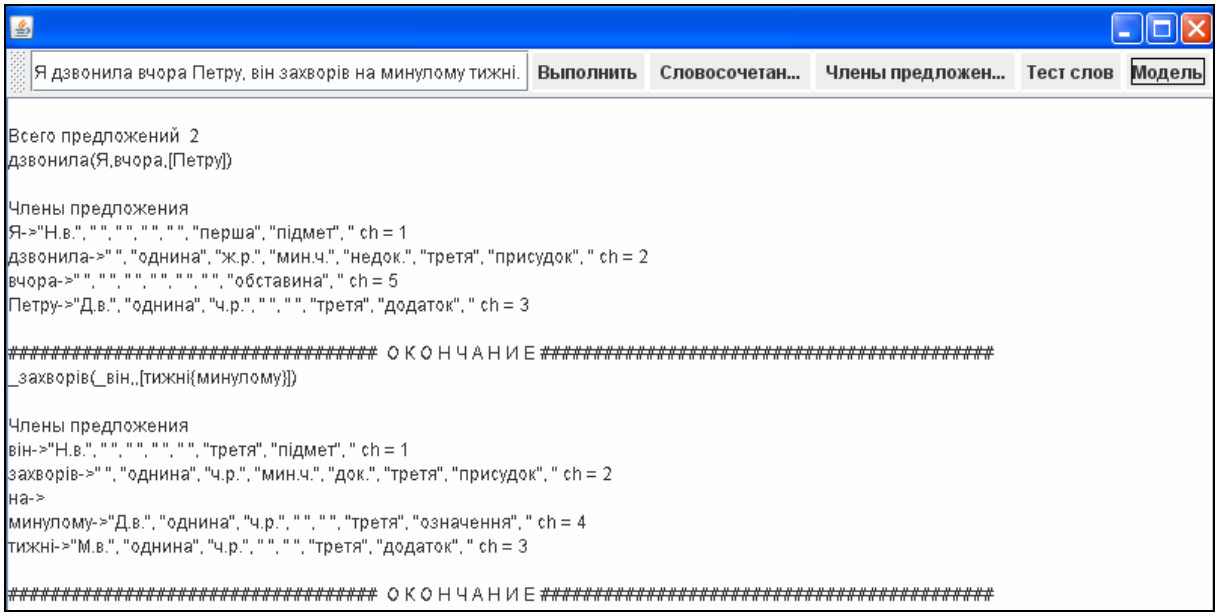

Рис. 4. Результат роботи САФЛЛМ

2. З рис. 4 видно, що система розпізнала введене речення як складне і розбила його на два, для кожного з яких створено формулу, де

– «дзвонила» – предикат, що відображає зміст першого речення;

– «я» – предикатна змінна (суб'єкт), що знаходиться у предикативному відношенні з предикатом «дзвонила»;

– «вчора» – предикатна константа, що вказує на ознаку дії;

– «Петру» – предикатна змінна (аргумент);

– «захворів» – предикат, що відображає зміст другого речення;

– «він» – предикатна змінна (суб'єкт), що знаходиться у предикативному відношенні з предикатом «захворів»;

– «тижні» – предикатна змінна (аргумент);

– «минулому» – предикатна константа, що вказує на ознаку аргументу «тижні».

3. Під час натискання кнопки «Виконати» система виводить на екран характеристики кожного слова (простого елемента формальної системи) (рис. 5).

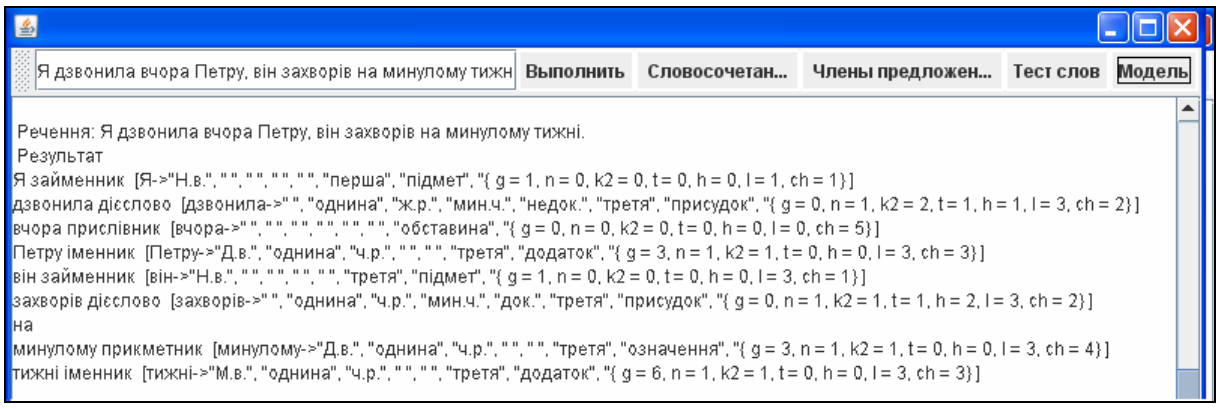

Рис. 5. Результат роботи САФЛЛМ після натискання кнопки «Виконати»

4. Кнопка «Словосполучення» дасть змогу отримати всі складні елементи формальної системи (рис. 6).

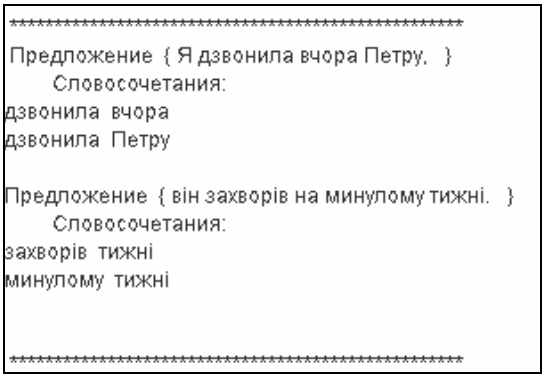

Рис. 6. Результат роботи САФЛЛМ після натискання кнопки «Словосполучення»

5. Кнопка «Члени речення» дасть можливість отримати синтаксичні ролі простих елементів формальної системи (рис. 7):

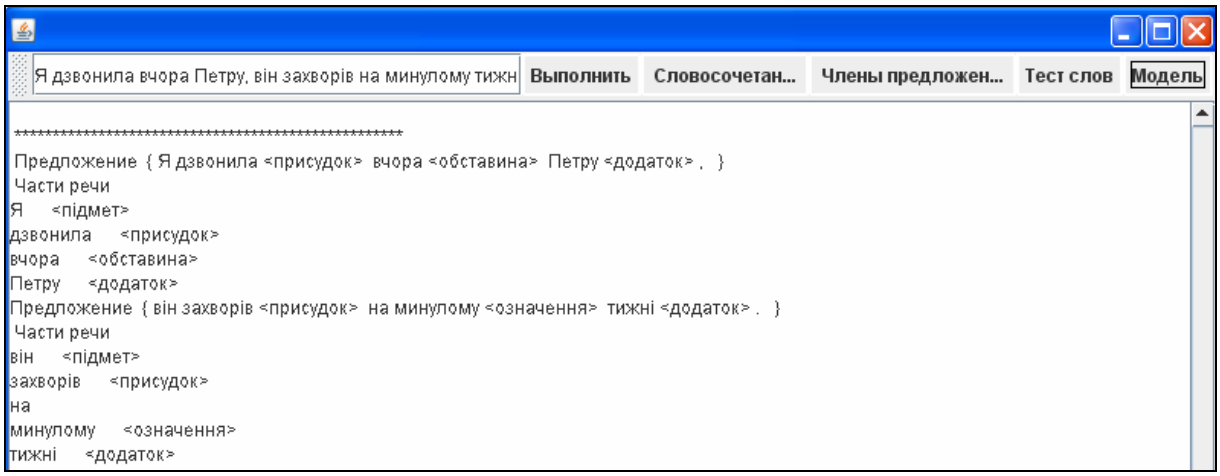

Рис. 7. Синтаксичні ролі слів у введеному реченні

Отримані внаслідок автоматизованої побудови логіко-лінгвістичні моделі близькі до природної мови, вирішують проблему структурної лінгвістики – завдання опису того, як будь-який текст, написаний конкретною мовою, може бути породжений з одиниць мови за допомогою кінцевого набору формальних правил роботи з цими одиницями (словами). Побудова таких моделей надалі може використовуватися для порівняння текстів, машинного перекладу, вилучення знань з текстової інформації та відшукання в ній суперечностей.

В основу роботи системи САФЛЛМ покладено метод автоматизованої побудови логіко-лінгвістичних моделей текстової інформації. На відміну від статистичних методів, він дозволяє виявити зміст речень і використовує при цьому набір правил побудови зв’язків у реченнях природної мови. Для флективних мов набір правил обмежений, тому цю інформацію можна формалізувати. База знань системи САФЛЛМ побудована на основі використання множини продукцій, кожна з яких відповідає одному з правил. Так, САФЛЛМ використовує тридцять одне правило формування словосполучень, тридцять два правила визначення граматичної основи речень та двадцять одне правило визначення типів складних речень для української мови.

Використання інформації щодо можливих типів зв’язків між реченнями, абзацами, частинами документа дозволить створити аналогічний набір правил побудови будь-якого текстового документа. А отримані практичні результати роботи системи САФЛЛМ дозволяють висувати гіпотези відносно того, як породжується мова людиною. Якщо певна модель достатньо простим і логічним способом породжує фрази природної мови, то можна припустити, що аналогічним чином працює і мозок людини.

Висновки і пропозиції. Використання технологій оброблення текстової інформації є перспективним напрямом для вирішення завдань оцінювання відповідності програмного забезпечення вимогам безпеки. Це дозволить підвищити ефективність проведення, скоротити час та вартість сертифікаційних досліджень, покращить процедуру сертифікації [8]. У сфері бізнесу необхідні системи прогнозування популярності характеристик тих чи інших продуктів (машин, літаків, засобів телекомунікації, побутової техніки та ін.), системи, що відслідковуватимуть політику гарантій та постійних клієнтів. Здійснення аналітичного оброблення текстової інформації необхідне в медицині для постановки діагнозів. Пошукові системи використовують алгоритми пошуку дублікатів текстових документів з метою індексування тільки унікальних веб-ресурсів. Тобто майже всі галузі знань, науки і техніки через складність своєї системної організації потребують інтелектуального оброблення даних. Автоматичне оброблення природної мови базується на послідовному аналізі мови як ієрархічної системи. Етапи аналізу природномовного тексту відповідають рівням структури мови: лексичний, морфологічний, синтаксичний, семантичний та прагматичний аналіз тексту. Тому лише методи оброблення текстової інформації, реалізація яких передбачає здійснення всіх етапів аналізу, можуть претендувати на проведення змістовного аналізу електронних документів.

Список використаних джерел

1. *Ландэ Д. В.* Поиск знаний в Internet. Профессиональная работа / Д. В. Ландэ. – М. : Вильямс, 2005. – 272 с.

2. *Звегинцев В. А.* Новое в зарубежной лингвистике. Вып. 10. Лингвистическая семантика / В. А. Звегинцев. – М. : Прогресс, 1981. – 568 с.

3. *Попов Э. В.* Общение с ЭВМ на естественном языке / Э. В. Попов. – М., 1982. – 360 с.

4. *Звегинцев В. А.* Новое в зарубежной лингвистике. Вып. 23. Когнитивные аспекты языка / В. А. Звегинцев. – М. : Прогресс, 1988. – 320 с.

5. *Мельчук И. А.* Опыт теории лингвистических моделей «СМЫСЛ-ТЕКСТ» / И. А. Мельчук. – М. : Школа «Языки русской культуры», 1999. – 346 с.

6. *Широков В. А.* Семантичні стани мовних одиниць та їх застосування в когнітивній лексикографії / В. А. Широков // Мовознавство. – 2005. – № 3-4.

7. *Широков В. А.* Інформаційна теорія лексикографічних систем / В. А. Широков. – К. : Довіра, 1998. – 331 с.

8. *Беляков И. А.* Применение интеллектуальных технологий в процессе сертификации программного обеспечения / И. А. Беляков, М. А. Еремеев // Молодой ученый. – 2011. – № 11, т. 1. – С. 23-31.