

УДК 621.3: 811.161.2

Дудник В. Ю., аспірант

Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»

АВТОМАТИЗАЦІЯ РОЗВ'ЯЗКУ АНАФОРИ У СИСТЕМАХ АНАЛІЗУ ПРИРОДНОМОВНИХ ТЕКСТІВ ДЛЯ УКРАЇНСЬКОЇ МОВИ

Дудник В. Ю. Автоматизація розв'язку анафори у системах аналізу природномовних текстів для української мови. У статті автори роблять детальний аналіз сучасних методів дослідження питання розв'язку анафори у системах аналізу текстів з української мови. Автори зосереджують увагу на тому, що у зв'язку із збільшенням обсягів доступної для людей інформації, важливу роль відіграє інформаційний пошук, вилучення інформації, машинний пошук. Але, незважаючи на практичну значущість даної теми, необхідних для роботи науковців, автоматизованих систем аналізу текстів з української мови існує не так багато, і вони значно поступаються місцем у порівнянні з якістю ручної роботи лінгвістів. Таким чином, постає необхідність до побудови нових або вдосконалення існуючих систем аналізу з метою ефективного вирішення питань анафори у текстах з української мови. Анафору необхідно розуміти як певне відношення залежності значення одного мовного вираження від іншого, тобто розглядається не повторення висловлювання вцілому, а тільки про спрощену форму, яка зустрічається у тексті, який був наведений раніше у необхідному мовному вираженні і є цілком логічним та легким для сприймання читачем. Висвітлено питання побудови моделей та методів для інтелектуальної обробки природномовних текстів на основі використання автоматизованих систем. Пропонується загальне введення в лінгвістичну проблематику, опис суті проблеми і аналітичного апарату, що використовується в сучасних теоріях для її дослідження. Стаття ілюструє способи міждисциплінарної взаємодії між філософами, програмістами, логіками, лінгвістами у рамках сучасних тенденцій при дослідженні української мови. Проведений аналіз літературних джерел, які широко висвітлюють питання автоматизації лінгвістичних процесів та основних підходів при виборі методу дослідження питання анафори природномовних текстів. У процесі інтелектуальної діяльності науковці-лінгвісти удосконалюють існуючі методи, технічні засоби для проведення автоматизації та обробки текстових даних. З цієї причини виникає необхідність у дослідженні існуючих систем обробки текстових масивів та побудови конструктивних алгоритмів, розширень на основі комбінованих методів аналізу.

Ключові слова: українська мова, лінгвістика, анафора, семантичний аналіз, морфологічний аналіз, відношення, автоматизовані системи, алгоритм.

Дудник В. Ю. Автоматизация решения анафоры в системе анализа естественно-мовных текстов для украинского языка. В статье авторы делают детальный анализ современных методов исследования вопроса развязку анафоры в системах анализа текстов с украинского языка. Авторы сосредоточивают внимание на том, что в связи с увеличением объемов доступной для людей информации, важную роль играет информационный поиск, извлечение информации, машинный поиск. Но, несмотря на практическую значимость данной темы, необходимых для работы ученых, автоматизированных систем анализа текстов с украинского языка существует не так много, и они значительно уступают место по сравнению с качеством ручной работы лингвистов. Таким образом, возникает необходимость к построению новых или совершенствования существующих систем анализа с целью эффективного решения вопросов анафоры в текстах по украинскому языку. Анафору необходимо понимать как некое отношение зависимости значения одного языкового выражения от другого, то есть рассматривается не повторение высказывания в целом, а только об упрощенной форме, которая встречается в тексте, который был приведен ранее в необходимом языковом выражении и является вполне логичным и легким для восприятия читателем. Освещены вопросы построения моделей и методов для интеллектуальной обработки естественно-текстов на основе использования автоматизированных систем. Предлагается общее введение в лингвистическую проблематику, описание сути проблемы и аналитического аппарата, используемого в современных теориях для ее исследования. Статья иллюстрирует способы междисциплинарного взаимодействия между философами, программистами, логиками, лингвистами в рамках современных тенденций при исследовании украинского языка. Проведенный анализ литературных источников, которые широко освещают вопросы автоматизации лингвистических процессов и основных подходов при выборе метода исследования вопроса анафоры естественно-текстов. В процессе интеллектуальной деятельности ученые-лингвисты совершенствуют существующие методы, технические средства для проведения автоматизации и обработки текстовых данных. По этой причине возникает необходимость в исследовании существующих систем текстовых массивов и построения конструктивных алгоритмов, расширений на основе комбинированных методов анализа.

Ключевые слова: украинский язык, лингвистика, анафора, семантический анализ, морфологический анализ, отношение, автоматизированные системы, алгоритм.

Dudnik V. Y. Automation of anaphora solution in the system of analysis of natural text for the Ukrainian language. In the article, the authors make a detailed analysis of modern methods of studying the problem of the anaphora solution in systems of analysis of texts from the Ukrainian language. The authors focus on the fact that in connection with the increase in the volume of information accessible to people, information search, information retrieval, machine search plays an important role. But, despite the practical significance of this topic, there is not much that is needed for scientists, automated systems for analyzing texts from the Ukrainian language, and they are significantly inferior to the quality of manual work of linguists. Thus, there is a need to build new or improve existing analysis systems in order to effectively address the issues of anaphora in Ukrainian language texts. An anaphora must be understood as a certain relation of the dependence of the value of one language expression from another, that is, not a repetition of the expression in the whole, but only a simplified form that occurs in the text that was given earlier in the required language expression, is quite logical and easy to read by the reader. The

questions of the construction of models and methods for the intellectual processing of natural language texts based on the use of automated systems are discussed. The general introduction to the linguistic perspective, the description of the essence of the problem and the analytical apparatus used in modern theories for its study are offered. The article illustrates the ways of interdisciplinary interaction between philosophers, programmers, logicians, and linguists within the framework of modern tendencies in the study of the Ukrainian language. The analysis of literary sources, which widely covers the issues of automation of linguistic processes and basic approaches when choosing the method of research, is an anaphora of natural language texts. In the process of intellectual activity, linguists improve existing methods, technical means for automation and processing of text data. For this reason, there is a need for the study of existing systems for processing text arrays and the construction of constructive algorithms, extensions based on combined analysis methods.

Keywords: ukrainian language, linguistics, anaphora, semantic analysis, morphological analysis, relations, automated systems, algorithm.

Вступ. Активний розвиток всесвітньої інформаційно-комунікаційної системи зумовлює вдосконалення існуючих інтелектуальних лінгвістичних систем, зокрема найбільш актуальним є питання розробки засобів інтелектуального автоматизованого опрацювання природномовних текстів [1,4,5].

Дослідження питання автоматизації лінгвістичних процесів є на даний час передовим у більшості науковців, це можна пояснити тим, що у всесвітньо-глобальних мережах почав накопичуватися великий арсенал текстових масивів, у зв'язку з цим виникає необхідність до розбудови технологій обробки інформації та вдосконалення процесу автоматизації опрацювання текстових масивів, перш за все для того щоб описати алгоритм знаходження анафоричних зв'язків для систем аналізу природномовних текстів, з метою дослідження впливу семантичних ознак на якість вирішення питання анафори для українськомовних текстів [2].

Існуючі системи для опрацювання природномовних текстів потребують детального аналізу тексту де відтворення мовних одиниць відбувається точно та однозначно. Здійснюючи дослідження україномовних текстів виникає проблема при проведенні аналізу існуючих відношень між словами, оскільки виокремити окремі слова та дослідити їх морфологічні ознаки має змогу велика кількість систем аналізу і це не становить великих труднощів [6] (система Apache UIMA, WordNet, Text Mining, TQAS).

Анафора охоплює широке коло питань для проведення дослідження та практичної реалізації результатів, а саме провідні організації такі як Google, IBM, Microsoft займаються вивченням питання анафори та шляхів її розв'язку, але вони зосереджені на проведенні наукових досліджень в рамках англійської мови, розглядаючи вітчизняні наукові праці можна простежити повільну тенденцію розвитку систем аналізу для автоматичного впровадження розв'язання питань анафори у природномовних текстах для української мови. Виходячи з цього, окреслений напрям дослідження представляє інтерес для детального вивчення питання розв'язання анафори та шляхів реалізації у системах аналізу з метою здійснення автоматизації процесів [3,5].

Аналіз останніх досліджень і публікацій. Вивченням питання автоматичного семантичного аналізу тексту, а саме дослідженням питання анафори займалися такі науковці як : А. Анісімов, Н.Д. Арутюнова, О. Большакова, Є. Гінзбург, І. Замаруєва, Л. Компанцева, А.А. Кибрик, О. Кургаєв, Н. Леонтьєва, О. Палагін, Е.В. Падучева , Д. Ланде, О. Литвиненко, Ю. Марчук, О. Малащук, Р. Мітков, К. Jones та ін. [1,2]. Варто відмітити, що семантика тексту характеризується недостатньою формалізацією та одночасно дає поштовх до створення лінгвістичних методів, моделей, де завдяки певним алгоритмам буде можливість виділити з текстового масиву необхідні смислові елементи, наукові досягнення у даному напрямі можна знайти у роботах: Р. Даревича, А. Лучик, Т. Гранкіна, Ю. Рашкевича, О. Селіванова [5,6]. Описані науковцями-лінгвістами шляхи вирішення питання автоматизації текстового аналізу мають важливе практичне використання під час проведення інформатизації освітніх процесів, а саме – реалізації оцінювання якості навчання це можна пояснити тим що система оцінювання знань тісно пов'язана із проведенням семантичного аналізу та виявлення ступеня релевантності даних текстових відповідей вказаних студентами до необхідних зразкових текстів навчальних предметів.

Дослідження методів аналізу природномовних текстів є досить актуальним напрямом в колі сучасних науковців про що свідчить поява, останнім часом, великої кількості публікацій по даній тематиці [1,3, 7]. Основні підходи до розв'язання питання визначення анафори детально описані у роботах [2,4,5], де проводяться дослідницькі експерименти на основі використання методів машинного навчання, а саме методу опорних векторів, який полягає у тому, що для вказаного набору даних, кожен з яких відноситься до однієї з двох категорій, виконується

побудова моделі завдяки опорно-векторної машини, яка класифікує нові надходження даних до однієї чи іншої категорії використовуючи бінарні лінійні класифікатори.

Провівши детальний аналіз наукових праць, можна з впевненістю стверджувати, що вкрай непростою та практично значущою задачею сучасних інтелектуальних лінгвістичних систем є вдалий вибір методів для розв'язання питання анафори в українській мові, оскільки кожен з них має власні переваги та недоліки при практичному використанні [4].

Вагоме місце в наукових публікаціях сьогодення відводиться правильному виборі методів розв'язання питання анафори природномовних текстів та їх успішної реалізації в системах аналізу, оскільки інтенсивний розвиток технічного прогресу провокує появу та стрімкий розвиток нових засобів з метою розв'язання більшості лінгвістичних задач, вдосконалення програмних середовищ з надвисокою точністю обробки україномовних текстів [2].

Виходячи з цього виникає необхідність досліджувати вирішення питання анафори із використанням існуючих систем аналізу для потреб сучасної лінгвістики та окреслювати шляхи подальшого розвитку в даному напрямі.

Формулювання цілей статті (постановка завдання). Дослідити питання знаходження анафоричних зв'язків для систем аналізу природномовних текстів та побудови алгоритмів для їх автоматичної реалізації у сучасній інтелектуальній лінгвістиці.

Виклад основного матеріалу дослідження. Розглянемо шляхи реалізації семантичних процесів на основі використання систем аналізу. Науковці-лінгвісти показують практичну значущість необхідності проведення семантичного аналізу тексту, оскільки з усього змісту текстового матеріалу здійснюють вибірку та відтворюють її у деяку модель, яка дасть змогу проаналізувати змістовну близькість досліджуваного тексту, під час розв'язання питання задачі пошуку враховують близькість пошуку та документу [3].

Під час проведення семантичного аналізу масив синтаксем для кожного речення представляється у вигляді неоднорідної семантичної мережі, яку вперше запропонував Г.С. Осиповий, яка відображується синтаксемами при вершинах, семантичні зв'язки представляють собою ребра. У більшості випадків семантичний аналіз тексту зосереджений на роботу із іменними синтаксемами, які утворюються у результаті проведення морфологічного та синтаксичного аналізів. Іменна синтаксема подається у вигляді іменної або прийменникової групи де слово є керуючим об'єктом, вказану синтаксему можна охарактеризувати морфологічною формою – прийменником, врахування категоріально-семантичного класу іменника, відмінку досліджуваного слова від якого вона була похідна. Лінгвістичний аналізатор тексту дає змогу визначити морфологічну форму у текстовому масиві даних та встановити категоріально-семантичний клас.

Відомі випадки, коли синтаксему вдається детально розглянути завдяки аналізу синтаксичної функції, яку вона виконує у реченні та врахуванням синтаксичного значення. В процесі семантичного аналізу тексту потрібно виокремити значення іменних синтаксем, які відображають основний зміст тексту. При виникненні питання неоднозначності встановлення морфологічної форми та категоріально-семантичного класу іменної синтаксеми потрібно додатково використати контекст – віддієслівний іменник (дієслово), де іменна синтаксема входить до складу речення, зручним є використання спеціальних вбудованих словників, які вказують найбільш вживані сполучення дієслова із потрібними синтаксемами.

Використання комп'ютерно-лінгвістичних методів для розв'язання задач аналізу природномовних текстів, зокрема питання анафори є найбільш перспективним для проведення дослідницьких робіт у даному напрямі. Виникає потреба до створення єдиної класифікації існуючих лінгвістичних методів та виокремлення основних підходів для того щоб можна було ввести єдину термінологію, а це в свою чергу дасть змогу створити лінгвістичні шаблони та відповідно автоматизувати роботу завдяки використанню ЕОМ. Побудова шкали лінгвістичних методів сприятиме ефективнішому використанню систем аналізу, отримані результати характеризуватимуться чіткістю, точністю, однозначністю, крім того, можна буде сформулювати певні набори ознак того чи іншого класу методів [1,6].

Наведемо основні підходи, які використовуються при розв'язанні великої кількості лінгвістичних завдань, а також задач прикладного характеру. Відомо чотири основні підходи – морфологічний, статистичний, синтаксичний, семантичний. Наведена класифікація не є фіксованою і більшість науковців сьогодення у своїх наукових доробках вказують на те що можливо на перспективу побудова системного аналізу змішаних типів на основі комбінування

існуючих методів, оскільки більшість прикладних задач не можливо конкретно віднести до певного методу.

Морфологічний підхід являє собою велику кількість сучасних методів, які ґрунтуються на перетворенні тексту морфологічними способами: проведення дослідження граматичних класів слів, процес нормування вхідного текстового масиву, збільшення запиту завдяки словоформам та ін. При опрацюванні текстових баз даних виникає необхідність до відшукування певних документів за вказаним змістом. Традиційні підходи базуються на наявності входження слів у документ, досить часто не здатні забезпечити чіткий та достовірний відбір інформації, це пов'язано з тим, що формулювання потрібного запиту не є точним та зрозумілим за вказаними ключовими словами. До причин виникнення неоднозначного пошуку можна віднести: відсутність необхідної теоретичної бази знань користувача, присутність синонімічних слів, орфографічні помилки, які зустрічаються як у пошукових запитах так і самому текстовому масиві даних. Вирішити описану проблему запропонувала компанія RCO, яка дає змогу збільшувати запити близькими, подібними до суті словами, які містяться в тексті документів, за якими користувач здійснює пошук [5].

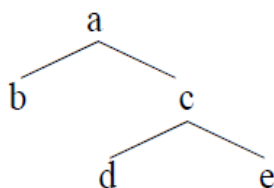
Асоціативно-статистичні методи тісно пов'язані із математично-статистичними методами, оскільки в їх основі лежить математичний апарат, який дає змогу використовувати лінгвістичні моделі та опрацьовувати результати статистичного аналізу, на відміну із теоретичною обробкою текстових документів. Побудова рефератів у своїй основі містить орієнтований граф, вершинами якого виступають потрібні теми, які користувач виокремив у проаналізованому тексті, а дуги представлені зв'язками між ними. Кожна вершина має значимість, тобто вагу та частоту повторення теми, а дуга має силу зв'язку та частоту використання зв'язку у тексті. Дослідник може вказувати вручну мінімально допустимі значення по вазі, нижче яких система аналізу інформації не включатиме знайдені теми до семантичної мережі.

Побудова асоціативних зв'язків між вказаними користувачем темами відбувається із врахуванням частоти їх одночасної появи в одному реченні. Відомою системою для побудови рефератів є розроблена у дослідницькому центрі Xerox, система Inxight Summarizer, яка у своїй роботі використовує алгоритм виділення тих речень із усього масиву тексту, які мають статистично вагоме значення, а також беруть до уваги слова-підказки. Розробники даної системи побудували один із найбільш точних, на даний час, для проведення оцінки якості реферату. Успіх використання вказаної системи полягає в тому, що відразу використовуються найдієвіші статистичні алгоритми, а на завершення роботи алгоритму користувач отримував тісний зв'язок між отриманими результатами роботи.

Застосування синтаксичного підходу для автоматизації аналізу природно мовних текстів використовується рідко, оскільки задача проведення детального синтаксичного аналізу для флективних мов (українська, російська, польська) до цих пір є не до кінця розв'язною, на відміну від аналітичних мов (англійська, французька, італійська), це можна пояснити тим, що флективні мови у порівнянні із аналітичними мають складнішу структуру. Прийнято проводити для флективних мов неповний аналіз синтаксичного розбору структури речень.

Використання користувачами семантичного підходу базується на перебудові структури початкового тексту у внутрішню модель для того щоб подати дані відповідної системи. Внутрішні системи прийнято називати «семантичне подання», «семантичний граф» і т.д. Робота семантичних алгоритмів побудована таким чином, щоб виконувати порівняння «семантичної структури» новостворених документів, зі структурою, яку надала відповідна система аналізу.

У лінгвістиці, де досліджується не так побудова логічної форми речення, скільки їх синтаксична структура, важливою умовою для анафори прийнято вважати с- командування між антецедентом і анафоричним виразом. Відношення с-командування необхідно розуміти як відношення яке відтворює структуру речення. Прийнято вважати α с-командує β тоді, і тільки тоді якщо β є сестрою α або сестра α домінує над β , описаний алгоритм можна зобразити у вигляді схеми [4]:



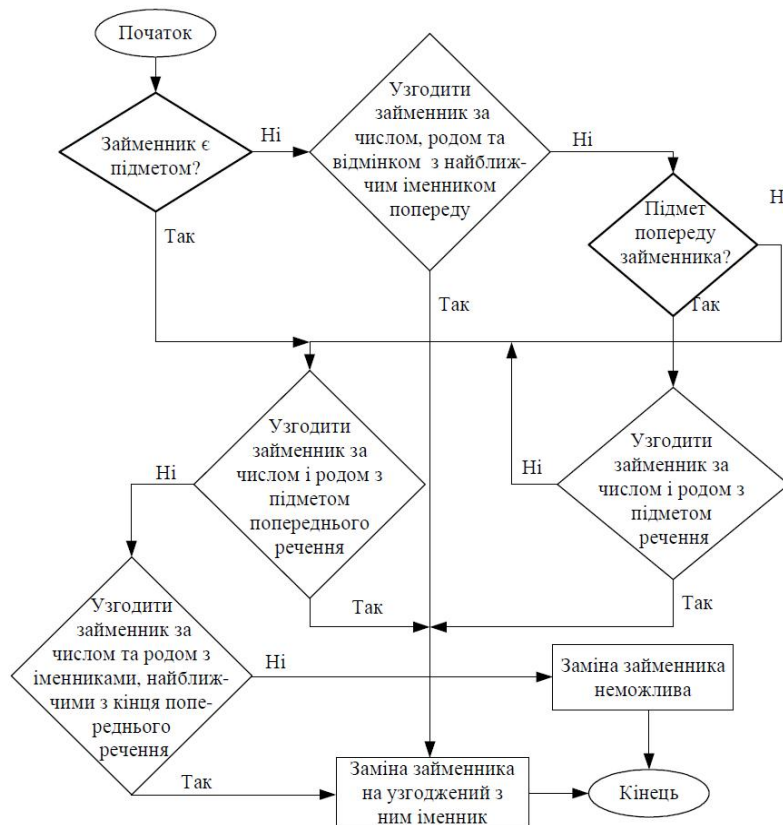
Наведена схема показує, що *b* *c*-командує *c*, *d*, *e*; *c* *c*-командує *b*; *d* і *e* *c*-командують один одним; *a* не *c*-командує нічим. Тобто, для того щоб вирази мали анафоричну залежність, прийнято вважати, що антецедент повинен *c*-командувати тим виразом, значення якого у тій чи іншій мірі від нього залежить. При розгляді двох слів науковці використовують термін референційно пов'язані, це означає, що вони ціле направлені на спільний предмет. Для кращого розуміння питання вирішення анафори розглянемо наступне речення:

Марійка прийшла до дитячого садка, і вона миттю побігла гратися до дітей

У даному випадку займенник «вона» вказує на конкретне ім'я «Марійка». Слово, на котре вказує займенник прийнято називати анцендентом, а утворена конструкція називається анафорою. Прийнято вважати, що система обробки текстової інформації вирішує питання анафори, якщо при проведенні аналізу система здатна розібратися, що «Вона» - це «Марійка».

При практичній реалізації процесу автоматизації розв'язку анафори науковці стикаються із труднощами, які полягають у тому що під час попередньої обробки тексту необхідно виконати певну послідовність дій: встановлення частин мови, визначення морфологічних характеристик слів, створення груп, проведення синтаксичного та поверхнево-семантичного дослідження. На кожному із описаних етапів під час проведення автоматизації процесів використовується певна точність у реалізації, відбувається накопичення похибок, які в свою чергу негативно впливають на кінцевий результат. У зв'язку з цим підвищення якості процесів автоматизованих систем аналізу є досить проблематичним [3].

Сучасні науковці активно займаються відшукуванням рішення для встановлення анафоричних зв'язків та побудови алгоритмів з метою реалізації в системному аналізі, що призведе до автоматизації ручної роботи лінгвістів. Ключовим питанням залишається необхідність виконати заміну займенників у текстовому масиві даних на іменники (відповідні поняття). Проведення заміни стосується до таких видів займенників як: особові, присвійні, зворотні, вказівні, відносні. Для того щоб побудувати алгоритм заміни необхідно використовувати дані які наведені у базі синтаксису природномовних об'єктів, завдяки яким буде можливість виконати аналіз закономірностей використання займенників у природній українській мові залежно від типу та відповідних характеристик займенника. Загальноприйнята блок-схема для реалізації описаного алгоритму виглядає наступним чином [2]:



Для того щоб провести заміну займенників необхідно мати вихідні дані, які будуть відповідати детальній граматичній характеристиці слів, їх дослідник отримує в результаті проведення морфологічного аналізу, встановлення синтаксично-семантичних зв'язків.

Практична реалізація даного алгоритму показує, що як правило, поняття, до якого співвідноситься займенник є співставлень найближче слово, яке в повній мірі відповідає граматичним характеристикам, воно може стояти попереду, або входити до ядра цього чи раніше наведених речень.

Правильно побудована, з технічної точки зору структура речення, дає змогу відшукати іменник у реченні, який можна буде спів ставити займеннику. Процедура заміни займенників дає змогу встановлювати нові зв'язки між наведеними поняттями у конкретному реченні та всім текстом вцілому. Після проведення заміни займенників видається за доцільне проведення повторного синтаксичного аналізу речень із займенниками.

Підсумовуючи вище сказане, відмітимо, що з метою оптимізації великого класу задач лінгвістики використовують комп'ютерну обробку природномовних текстів, комп'ютерна реалізація яких на практиці відкриває нові перспективи для вдосконалення традиційних методів аналізу україномовних текстів.

Висновки. Проведено детальний аналіз наукових результатів у галузі розробки автоматизованих систем аналізу для текстів на українській мові, а саме розв'язання питання анафори. Вказане спонукає до комплексного поєднання методів в системах аналізу інформації з різних аспектів та рівнів лінгвістичного аналізу.

Встановлено, що найдієвішими методами для проведення обробки природно мовних текстів є системи аналізу на основі використання морфологічного, статистичного, синтаксичного, семантичного підходів та їх комбінації, що дає змогу досліджувати приховані зв'язки всередині природно мовних текстів. Розглянуто основні підходи для аналізу текстових масивів. Вказана широка специфіка їх застосування та можливості подальшого розширення існуючих методів. Описана детальна реалізації процесу автоматизації розв'язку анафори, для наочного розуміння подана блок-схема.

Стає зрозумілим, що створення єдності використання лінгвістичних методів, а саме особливостей, способу реалізації, формування основних підходів, шляхів застосувань дасть змогу системам аналізу структурувати інформацію та перейти на автоматизацію подібних процесів.

Таким чином, виникає інтерес до подальшого вивчення, вдосконалення існуючих методів для здійснення автоматизації процесів розв'язання питань анафори природномовних українських текстів з метою розширення класу прикладних лінгвістичних задач.

Анісімов А.В. Марченко О.О. Нагорний В.А Створення керуючого простору синтаксичних структур природної мови // Вісник Київського університету. Сер. фіз.-мат. науки.- 2011.-Вип. 1- С. 159-169.

Вихованець І.Р. Теоретична морфологія української мови: Академічна грамматика української мови / І.Р. Вихованець, К.Г. Городенська. – К.: Унів. вид-во «Пульсари», 2012. – 400 с.

Комарницька О. І. Моделі штучного інтелекту для порівняння текстової інформації за змістом /О. І. Комарницька // Лінгвістичні студії : зб. наук. праць / Донецький нац. ун-т ; гол. ред. А. П. Загнітко. – К. ; Вінниця : ДонНУ, 2015. – Вип. 30. – С. 135–139.

Палагин, А. В. Знание-ориентированные информационные системы с обработкой естественно-языковых объектов: основы методологии и архитектурно-структурная организация [Текст] / А. В. Палагин, С. Л. Кривой, Н. Г. Петренко // Управляющие системы и машины. – 2016. – № 3. – С. 42-55.

Evans G. Pronouns, quantifiers, and relative clauses (I) // Canadian Journal of Philosophy 7: 467–536. References are to the 2003 reprint in Gareth Evans, Collected Papers. Oxford: Clarendon Press, 2003. P. 76–152.

Protopopova E.V., Bodrova A.A., Volskaya S.A., Krylova I.V., Chuchunkov A.S., Alexeeva S.V., Bocharov V.V., Granovsky D.V. Anaphoric annotation and corpus-based anaphora resolution: an experiment, computational linguistics and intellectual technologies. Dialog-2014: sb. tr. Mezhdunar. nauch. konf. po kompyuter. lingvistike [Proc. Annual Int. Science Conf. on Computer Linguistics "Dialogue-2014"]. 2014, iss. 13 (20), pp. 562–571

Mitkov R. Anaphora resolution / Ruslan Mitkov // The Oxford Handbook of Computation Linguistics. – Oxford University Press, 2003. – P. 267.