**Дарчук Наталія Петрівна. Лінгвістичне забезпечення автоматичних систем аналізу українськомовного тексту (на прикладі системи автоматичного граматичного аналізу тексту АГАТ).- Дисертація д-ра філол. наук: 10.02.01, 10.02.21, Київ. нац. ун-т ім. Тараса Шевченка, Ін-т філології. - Київ, 2015.- 449 с.**

МІНІСТЕРСТВО ОСВІТИ ТА НАУКИ УКРАЇНИ

КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ТАРАСА ШЕВЧЕНКА

Інститут філології

УДК 811.161.2’42:004

**ДАРЧУК Наталія Петрівна**

**«ЛІНГВІСТИЧНЕ ЗАБЕЗПЕЧЕННЯ АВТОМАТИЧНИХ СИСТЕМ АНАЛІЗУ УКРАЇНСЬКОМОВНОГО ТЕКСТУ**

**(на прикладі системи автоматичного граматичного аналізу**

**тексту АГАТ)»**

Спеціальності 10.02.01 – українська мова

10.02.21 – структурна, прикладна та математична лінгвістика

Дисертація на здобуття наукового ступеня

доктора філологічних наук

Київ – 2015

**ЗМІСТ**

ВСТУП……………………………………………………………………………..6

РОЗДІЛ 1. ТРАДИЦІЙНА ЛІНГВІСТИКА – СТРУКТУРНО-МАТЕМАТИЧНА ЛІНГВІСТИКА – КОМП’ЮТЕРНА ЛІНГВІСТИКА – ТРИЄДИНА СУТНІСТЬ…………..…….………………………………………26

РОЗДІЛ 2. МОРФОЛОГІЧНЕ АНОТУВАННЯ УКРАЇНСЬКОГО ТЕКСТУ.49

2.1. Лінгвістичні засади створення автоматичного морфологічного

аналізу (АМА)……………………………………………………………………49

2.2. З історії створення АМА російської та української мов…..…………… 53

2.3. АМА з використанням словника основ …………………………………..65

2.4. Автоматичне зведення словоформ у парадигму слова (лематизація)…..69

2.5. Принципи і засади створення морфологічного аналізатора в автоматичній системі АГАТ..…………………………………………………..74

2.6. Обчислення омонімів у процесі конструювання лінгвістичних

процесорів ……. . . ……………………………………………………………...82

2.7. Контекстний аналіз ………………………………………………………...88

2.8.Конкорданс як джерело для контекстного аналізу ……………………….91

2.9. Висновки ……………………………………………………………………98

РОЗДІЛ 3. СИНТАКСИЧНЕ АНОТУВАННЯ УКРАЇНСЬКОГО ТЕКСТУ.108

3.1. Теоретичний синтаксис й автоматичний синтаксичний аналіз

(АСА)……………………………………………………………………… 110

3.2. Роль дистрибутивного методу і методу безпосередніх складників

у створенні АСА….…………………………………………………..........116

3.3. Принципи виокремлення словосполучень із тексту в системі АСА…...130

3.4. Граматика залежностей як спосіб представлення синтаксичної

структури речення........................................................................................137

3.4.1. Зображення присудка…………………………………………………... 151

3.4.2. Зображення підмета……………………………………………………...154

3.4.3. Зображення структур з однорідністю…………………………………..155

3.5. Дерева залежностей у системах АСА……………………………….........158

3.6. Автоматичне конструювання частотного словника сполучуваності української мови …………………………………………………………….....179

3.6.1. Досліджуваний матеріал та його систематизація……………………...181

3.6.2. Статистичне і лінгвістичне опрацювання матеріалу ………………...182

3.7. Висновки…………………………………………………………………...186

РОЗДІЛ 4. МОРФНЕ АНОТУВАННЯ ТЕКСТУ..……………………… …..193

4.1. Лінгвістичні засади автоматичного морфного сегментування ..…........193

4.2. Морфемно-словотвірна база як джерело лінгвістичних досліджень…..200

4.3. Висновки…………………………………………………………………...209

РОЗДІЛ 5. СЕМАНТИЧНЕ АНОТУВАННЯ УКРАЇНСЬКОГО ТЕКСТУ...211

5.1. Напрямки формалізації семантики…………………….............................211

5.2. Синтактико-семантичний аналіз. Машинна реалізація засобів синтаксичного і семантичного аналізу………………………………………..221

5.3. Структура семантичного опису в зарубіжних АОТ…………………….234

5.4. Види «розуміння» тексту комп’ютером………………………………….239

5.5. Семантичний компонент у системі розуміння тексту…………………..244

5.6. Словники як компонент опрацювання змісту тексту при АОТ………...252

5.7. Тезаурус як концептуальний компонент опису семантики………..……259

5.8. Принципи побудови бази даних та інформаційно-пошукова модель

(тезаурус) з юридичної термінології…… ……………………………………271

5.9. Формалізована методика конструювання комп’ютерного тезауруса

лінгвістичних термінів…………………………………………………………280

5.10. Теоретичні питання моделювання ідеографічного тезаурусу

української мови..................................................................................................297

5.11. Ідеографічний словник дієслів української мови………………………310

5.12. Ідеографічний словник української мови на синсетах………………...316

5.13. Висновки………………………………………………………………….327

РОЗДІЛ 6. УПРОВАДЖЕННЯ КОМП’ЮТЕРНИХ РЕСУРСІВ У СФЕРУ ЛІНГВІСТИЧНОЇ ПРОБЛЕМАТИКИ УКРАЇНСЬКОЇ МОВИ…………….335

6.1. Корпус української мови як джерело для мовознавчих досліджень….. 335

6.2.Поетичний словник в аспекті мовної картини світу……………………344

6.3. Лінгвістичні інформаційні ресурси та газетний політичний дискурс…367

6.3.1. Образ влади в сучасному газетному дискурсі………………………...374

6.3.2. Тропи в політичному дискурсі………………………………………...380

6.4. Висновки…………………………………………………………………...390

ЗАГАЛЬНІ ВИСНОВКИ……………………………………………………….394

СПИСОК ЛІТЕРАТУРИ……………………………………………………….406

ДОДАТКИ

**ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ**

АМА – автоматичний морфологічний аналіз

АОТ – автоматичне опрацювання текстів

АРТ – автоматичне розуміння тексту

АСА – автоматичний синтаксичний аналіз

АСЕА – автоматичний семантичний аналіз

АСМСА – автоматизована система морфемно-словотвірного аналізу

БД – база даних

БЗ – база знань

БС – безпосередній складник

ГЗ – граматика залежностей

ДС – дерево залежностей

ЗЗС – зона зв’язку слова

ІПС – інформаційно-пошукова система

КА – контекстний аналіз

КФЛ – квазіфлексія

ЛБД – лінгвістична база даних

ЛСВ – лексико-семантичний варіант

МБД – морфемна база даних

МКС – мовна картина світу

МП – машинний переклад

ПОД – пошуковий образ документа

ПЧ – предикативна частина

СА – синтаксичний аналіз

СД – смислова домінанта

СМ – семантичний множник

ТБД – термінологічний банк даних

ТЗ – тезаурус

ФА – флективний аналіз

ЧС – частотний словник

**ВСТУП**

Універсальним носієм знань є природна мова. Мова – не тільки форма вираження думок, а єдиний й найважливіший засіб змістової організації і представлення знань.Функція мовної системи – бути засобом породження, збереження і передавання інформації. А це означає, що на перший план комп’ютеризації людської діяльності виходять лінгвістичні аспекти, тому будь-які проблеми інформатики, штучного інтелекту, пов’язані з цими аспектами, сприяють усвідомленню як загальнонаукової, так і загальнодержавної важливості мовознавчої науки. Бодуен де Куртене ще у ХІХ ст., говорячи про завдання мовознавства, які необхідно розв’язувати у ХХ-му сторіччі, зауважував, що треба частіше застосовувати у мовознавстві кількісне, математичне мислення і таким чином наблизити його до наук більш точних [42].

У зв’язку з активними процесами інформатизації сучасного суспільства [217] зростає значення комп’ютерної лінгвістики, яка поєднує мовознавство – глибинні знання про мову – і кібернетику – комп’ютерні технології. Взаємодія людини і комп’ютера сприяє посиленню прикладного аспекту мовознавства – комп’ютерної лінгвістики [237]. Прикладні лінгвістичні завдання у переважній більшості є соціальним замовленням: передавання звукового мовлення каналами зв’язку, автоматичне розпізнавання мовлення, машинний переклад, автоматичне реферування, мережеве представлення даних, уніфікація і стандартизація термінології (у тому числі створення термінологічних баз даних і знань), науково-технічна лексикографія, навчальна лексикографія, проектування корпусів текстів і баз даних, розроблення автоматизованого робочого місця лінгвіста тощо [60; 61; 209]. Водночас комп’ютеризація поглиблює теоретичні проблеми мовознавства. Від цієї взаємодії виграють як теоретична лінгвістика, так і комп’ютерна. Як відзначав А.Єршов, один із піонерів теоретичного і системного програмування, людство накопичило необхідні прикладні знання, виробило способи їх одержання, але з’явилося нове завдання – безпосередньо перенести ці знання в машину. Вчений пропонує єдиний рецепт: експлікувати лінгвістичні знання, шукаючи об’єктивні методи опису мовних закономірностей, – експліцитний опис може бути програмованим й адекватно відтворюваним у комп’ютері. А. Єршов цитує видатного вченого-лінгвіста Ю. Апресяна, всі роботи якого стали класикою комп’ютерної лінгвістики: «Оскільки логічний автомат здатний діяти лише на основі абсолютно точних і вичерпних вказівок, він буде правильно встановлювати відповідність між смислами і текстами у тому випадку, якщо в нього буде закладено повний, достатній й експліцитний опис всіх мовних об’єктів і правил». А оскільки будь-яка теорія, в тому числі й лінгвістична, має відносний характер, то перед комп’ютерною лінгвістикою постає завдання виокремити частину теоретичного лінгвістичного знання і надати йому аксіоматичного характеру, властивого природничим наукам [101, с. 6].

Як зазначає О. Герд у своїй монографії, в якій узагальнено тридцятирічний досвід пошуків і розробок у галузі прикладної (комп’ютерної) лінгвістики, прикладна (комп’ютерна) лінгвістика – це єдність мовознавства, філософії та математики, однак, послуговуючись методами математики, вона все-таки залишається лінгвістикою з превалюванням у ній знань лінгвістичних. За своєю суттю прикладна лінгвістика – це наука математична: за строгістю поставлених завдань, структуризацією фактів, пошуком формальних моделей опису й адекватних комп’ютерних форм представлення даних [61, с. 8].

Великого значення для власне лінгвістичних досліджень набували системи штучного інтелекту [174], адже знання про мову, якими оперують лінгвісти, також задані у формі текстової інформації, опрацювання якої не менш трудомісткий і складний процес, аніж у будь-якій іншій галузі інтелектуальної діяльності людини (70 % свого робочого часу науковці-лінгвісти витрачають на підготовку лінгвістичних даних, пошук й опрацювання інформації). Перехід до автоматизованого (із залученням людини) способу збирання, опрацювання і збереження інформації сприятиме скороченню термінів виконання роботи, а база даних, на якій проводяться лінгвістичні експерименти, розшириться у рази. **Актуальність** дисертаційного дослідження полягає у створенні потужних лінгвістичних баз даних і програмних ресурсів для автоматичного опрацювання знакових одиниць у процесі їх функціонування у текстах на всіх рівнях мовної системи. Це забезпечить можливість та ефективність запровадження широкого спектра лінгвістичних досліджень, зокрема проектування сучасних комп’ютерних систем для оптимізації трудомісткої роботи не лише мовознавців, а й фахівців будь-яких предметних галузей, в яких інформація існує у вигляді текстів. Для мовознавців застосування комп’ютерних технологій є додатковою можливістю перевірки й уточнення наукових теорій і гіпотез.

**Зв’язок праці з науковими програмами, планами, темами**. Дисертаційне дослідження виконане в межах наукових тем "Актуальні проблеми філології" (номер державної реєстрації 02БФ044–01) та "Розвиток і взаємодія мов та літератур в умовах глобалізації" (номер державної реєстрації 06БФ044–01). Тема дисертації затверджена вченою радою Інституту філології Київського національного університету імені Тараса Шевченка (протокол № 4 від 6 грудня 2004 р.).

Інформаційні технології розвиваються досить швидко, з’являються нові можливості дослідження природної мови. Однак деякі принципи як теоретичного, так і практичного характеру вже усталилися, отже, можна вважати, що комп’ютерне мовознавство цілком забезпечує потреби порівняно нової предметної галузі – комп’ютерної лінгвістики, яка в Україні на сьогодні перебуває на стадії розвитку.

Оскільки основною формою представлення і збереження інформації є природна мова в усній і писемній формі, то ефективність використання комп’ютера залежить значною мірою від розв’язання проблеми автоматичного опрацювання текстів (АОТ), кінцева мета якого – розпізнавання їх змісту. В системах АОТ розглядаються два рівні залежно від глибини і складності процесу опрацювання: 1) формальне опрацювання, коли здійснюється перетворення фрагментів тексту без аналізу його змісту; 2) змістове (смислове) опрацювання, коли розпізнається зміст окремих елементів і логіко-семантичних відношень між ними для того, щоб побудувати семантичне представлення повідомлення. На цьому рівні виникає необхідність у використанні додаткової, семантичної інформації, яка експліцитно не виражена у тексті. Якщо перший рівень – формальне опрацювання, є основою всіх наявних інформаційних технологій у діючих системах АОТ, то другий рівень – поле для теоретичних та експериментальних досліджень.

Як у будь-якій теоретичній дисципліні, у комп’ютерному мовознавстві вирішується завдання побудови моделі об’єкта предметної галузі. Модель – це аналог об’єкта, штучно створений дослідником для пізнання об’єкта-оригінала. І тільки за наявності ізоморфності – структурної подібності – можна одержати адекватні висновки про будову і функціонування об’єкта-оригінала.

**Мета дисертації** – теоретичне й експериментальне обґрунтування лінгвістичних і процедурних засад інтегральної моделі семантико-граматичної взаємодії знакових одиниць у тексті та створення на цій основі автоматичної граматики для аналізу тексту української мови (АГАТ).

**Об’єктом** нашого дослідження є текст української мови, **предметом** –комп’ютерне моделювання граматичної та семантичної взаємодії знакових одиниць на всіх рівнях мовної системи. Мова – відносно відкрита система відносно відкритих підсистем, кожну з яких можна моделювати і встановлювати певні відношення між підсистемами. Як свідчить практика, більшість сучасних лінгвістичних процесорів модульного типу, адекватних поетапному розбиттю процесу аналізу, відповідає рівням мови: фонемному (одиниця фонема/алофон), морфемному (морфема/морф), лексичному (лексема/словоформа), синтаксичному (моделі словосполучень, моделі речень). Семантичний модуль не має рівневого відповідника в мовній системі, але він на семному рівні конструктивно завершує (на даному етапі розроблення) смисловий аналіз українського тексту. Як буде показано, всі одиниці модулів взаємодіють (АГАТ відображає їх послідовну роботу), але морфологічні, синтаксичні, семантичні характеристики обробляються різними алгоритмами і програмами з різним словниковим забезпеченням (див. рисунок 1).































Рисунок 1. Лінгвістичне словникове забезпечення системи АОТ АГАТ

Системно адекватним цьому є таке лінгвістичне забезпечення, яке має відкритий модульний характер, й у деяких випадках передбачається їхня сумісність для поповнення даних і коригування результатів (наприклад, у результаті автоматичного морфологічного аналізу обох етапів залишається кілька відсотків словоформ із незнятою омонімією, яка остаточно знімається на синтаксичному рівні). Система може виконувати окремі конкретні завдання аналізу (наприклад, тільки морфологічного). При цьому відкритість є головним фактором ефективного функціонування АОТ, гарантією того, що введення нової інформації не потребуватиме перебудови всієї системи.

Словникове забезпечення є інформаційною основою моделі і реалізується у вигляді ієрархічної, модульно сумісної і відкритої системи. Компоненти, що представляють лінгвістичну модель АОТ, це лінгвістичні процесори, які послідовно, один за одним обробляють вхідний текст. Вхід одного процесора – це вихід із другого.

У створеній системі виокремлюються такі модулі:

* Морфологічний аналіз. Побудова морфологічної анотації слів вхідного тексту.
* Синтаксичний аналіз. Виокремлення словосполучень. Побудова дерева залежностей усього речення.
* Морфемний аналіз. Членування вхідного тексту на морфи.
* Семантичний аналіз. Побудова тезауруса тексту.

Для кожного рівня розроблено метамову його представлення – константи і правила їх комбінування. На морфологічному рівні – константами є грамеми (рід, число, відмінок, час, особа, спосіб), на морфемному – тип морфа, на синтаксичному – тип словосполучення, тип зв’язку, на семантичному – семантичні категорії. Важливим фактором є спадковість у роботі модулів: наступний аналізатор покращує результати попереднього рівня. Наприклад, на синтаксичному рівні «дознімається» граматична і лексико-граматична омонімія, а семантичний аналізатор допомагає добудовувати синтаксичний граф.

У дисертації наведено приклади впровадження результатів дослідження.

**Джерельною базою дослідження** є Корпус текстів сучасної української мови обсягом у 24 млн. слововживань (Мовно-інформаційний портал. — Режим доступу: http://www.mova.info.); «Інверсійний словник української мови», укладений за матеріалами тлумачного «Словника української мови в 11-ти тт. (К., 1985), 2-томний словник-довідник І.Т. Яценка «Морфемний аналіз» (К., 1980-1981), «Кореневий гніздовий словник української мови» Є.А.Карпіловської (К., 2002), «Морфемний словник» Л.М.Полюги (К., 1983), Словник-довідник з правопису та слововживання» С.І.Головащука (К., 1989), Граматичний словник українських дієслів Л.А.Алексієнко (К.,1998), 2-томний «Англо-русский словарь по лингвистике и семиотике» А.М.Баранова (М., 1996-2001), «Словарь лингвистических терминов» О.С.Ахманової (М., 1966), «Великий тлумачний словник сучасної української мови» В.Т.Бусела (К., 2007), «Англо-русские термины по прикладной лингвистике и автоматической обработке текста» В.З.Дем’янкова (М., 1979), «Словник лінгвістичних термінів» Є.В.Кротевич (К., 1957), «Словарь лингвистических терминов» Ж.Марузо (М., 2004), «Французско-русский учебный словарь лингвистической терминологии» А.Г.Назаряна (М., 1989), словник-довідник «Основные понятия морфемики в терминах» В.Н.Немченка (Красноярск, 1985), 2-томний «Словарь поэтических образов» Н.В.Павлович (М., 1999), «Словник антонімів» Л.М.Полюги (К., 1987), «Словарь сочетаемости слов русского языка», П.Н.Денисової (М., 1978), 2-томний «Словник синонімів української мови» В.В.Бурячка (1999-2000), «Сучасний орфографічний словник української мови» В.В.Дубічинського (Х., 2009), «Толковый словарь русских глаголов» Л.Г.Бабенко (М., 1999), «Українська мова : енциклопедія» (К., 2007), «Український орфографічний словник» В.В.Чумака (К., 2009).

**Методи дослідження.**

Були застосовані такі методи**:** *описовий метод***,** якийполягає в інвентаризації та систематизації мовних одиниць: морфів – у словниках коренів та афіксів, лексем – у парадигматичних класах, словосполучень; *індуктивне* та *дедуктивне лінгвістичне**моделювання*: *індуктивне* – при ідеографічному описі термінологічної лексики для побудови інформаційно-пошукової системи термінів; *дедуктивне* – для побудови ідеографічного словника української публицистики і словника на синсетах; *метод компонентного аналізу* для аналізу значень лексем – від значень слів до системи категорій пізнання; *структурні методи* як основні методи створення комп’ютерної граматики, представлені *дистрибутивним аналізом*для зняття граматичної та лексико-граматичної омонімії, *аналізом за безпосередніми складниками* та *методом лінгвістичного моделювання*(побудова дерев залежностей); *алгоритмічний метод* для укладання алгоритмів і програм автоматичного опрацювання українськомовних текстів.

**Перед нами стояли такі завдання:**

1) Обґрунтування ролі структурно-математичної лінгвістики як теоретичної основи комп’ютерної граматики.

2) Розроблення загальної стратегії й тактики для інтегральної моделі АОТ української мови на рівні сучасних лінгвістичних та інформаційних технологій.

3) Укладання частиномовних лексиконів з необхідною граматичною та лексичною інформацією для автоматичного аналізу морфології, синтаксису і семантики української мови.

4) Розроблення лінгвістичної стратегії та правил для автоматичного морфологічного аналізу українських словоформ і зняття їх граматичної та лексико-граматичної омонімії.

5) Створення лінгвістичної стратегії і правил автоматичного морфного сегментування словоформ українського тексту.

6) Розроблення лінгвістичної стратегії і правил автоматичного синтаксичного аналізу (АСА); створення пакету алгоритмів і програм, реалізація і тестування синтаксичного процесора на текстах Корпусу української мови.

7) Автоматичне укладання словників сполучуваності дієслів, іменників, прикметників, прислівників, займенників, числівників та конкордансів до них із контекстами на базі текстів Корпусу української мови.

8) Розроблення лінгвістичної стратегії автоматичного семантичного аналізу (АСЕА). Автоматизоване укладання термінологічного тезауруса інформаційно-пошукового типу та ідеографічних словників різних частин мови (іменника, дієслова) як ресурсу семантичного аналізу українських текстів.

9) Демонстрація роботи комп’ютерного анотування українського тексту, результати якої викладено на мовно-інформаційному порталі [www.mova.info](http://www.mova.info).

**Наукову новизну роботи** вбачаємо в тому, що вперше в українській прикладній лінгвістиці запропоновано алгоритми і програми повного циклу автоматичного аналізу українського тексту. Експериментально доведено і статистично обґрунтовано ефективність граматичних стратегій морфологічного і синтаксичного аналізу та ідеографічного моделювання лексики української мови, які застосовуються в різних системах автоматичного опрацювання інформації.

**Теоретичне значення** дослідження полягає у розбудові нового вектора у сучасному мовознавстві – комп’ютерної лінгвістики, яка зорієнтована на новий адресат – штучний інтелект, на відміну від традиційної і структурно-математичної лінгвістики, адресатом яких є людина. Комп’ютерна лінгвістика виникла на методологічних засадах сучасного мовознавства і є його складовою. Цей мовознавчий напрямок розбудовується відповідно до свого предметного вектора у семіотичному аспекті з урахуванням адресата і підтримки інформаційними технологіями. Реалізація такого підходу потребувала адаптування теоретичних положень і відповідної стратегії й тактики лінгво-інформаційного моделювання. Створена нами система АОТ української мови з її лінгвістичним і комп’ютерним забезпеченням уможливила введення нового концепта «комп’ютерна граматика». Комп’ютерна граматика АГАТ (автоматичного граматичного аналізу текстів) є системним, строго впорядкованим, формалізованим, лінгво-статистичним інтегральним описом знакових одиниць української мови у вигляді алгоритмічних моделей штучного інтелекту.

**Практичне значення** дисертаційного дослідження пов’язане зі створенням унікального для української мови лінгвістичного і програмного забезпечення, яке дозволяє практично в автоматичному режимі на матеріалі текстів української мови: 1) будувати алфавітно-частотні словники лексем і словоформ із морфологічними характеристиками; 2) укладати алфавітно-частотні словники коренів та афіксів; 3) вилучати з кожного речення Корпусу словосполучення, 4) створювати електронний алфавітно-частотний словник сполучуваності української мови, якого немає в українській лексикології і який по завершенні дослідницької роботи буде викладений в Інтернеті на мовному порталі [www.mova.info](http://www.mova.info) для загального користування, 5) здійснювати повний синтаксичний аналіз речення у вигляді дерева залежностей; 6) будувати тезаурус іменників та дієслів конкретного тексту.

Існує велика кількість різних систем автоматичного опрацювання тексту. Вони відрізняються не тільки завданнями, а й принципами конструювання – інформаційними базами, словниковим забезпеченням. Це пояснюється, перш за все, розмежованістю колективів, відсутністю професійних зв’язків, а також відсутністю загальної теорії лінгвістичного забезпечення АОТ. О.Герд у цитованій монографії [61] ставить питання: що є критерієм істини у прикладній лінгвістиці (див. також [242])? Як відомо, не тільки окремі вчені, а й колективи, школи, групи роками сперечаються, доводячи істинність саме своєї концепції, своїх принципів аналізу, але таке сперечання не має перспектив, тому що кожний напрямок вносить щось нове, актуалізує якусь проблему. Для прикладної лінгвістики критерієм істини є практика, експеримент. Якщо поставлене завдання розв’язане, система працює, значить, вихідні постулати обрано правильно – у прикладній науці допускаються будь-які підходи. Пропонована дисертація не претендує на створення теорії глобального лінгвістичного процесора на українській мовній основі, який би забезпечив спілкування в системі «людина – машина – людина». Ми свідомі того, що для успішного вирішення такого завдання необхідно створити теоретичні передумови, які б дозволили сподіватися на побудову такої лінгвістичної моделі, яка була б адекватна мовному об’єкту і не суперечила теоретичним надбанням лінгвістики та емпіричним спостереженням над мовою і мовленням. Останні 10-15 років у комп’ютерній лінгвістиці спостерігається вплив таких факторів, як еволюція програмних засобів; прогрес теоретичної лінгвістики; залучення формального аналізу на всіх рівнях мовної системи, семантики, прагматики, що стимулює появу великих проектів міжнародного рівня – WORDNET, EURONET, UKRNET тощо, в яких акумульовано досягнення лінгвістичного комп’ютерного моделювання.

Не претендуючи на повний теоретичний опис, викладемо основи створення лінгвістичного забезпечення і визначимо ті вихідні теоретичні положення, за якими можна здійснити вибір оптимального варіанта для конкретних завдань АОТ українського тексту. Лінгвістичне забезпечення, необхідне для розв’язання різних завдань АОТ, у межах певної моделі повинно мати єдину основу і характеризуватися повнотою опису базових одиниць та їх особливостей у структурній та програмній реалізації. Звичайно, важливо при цьому тримати в полі зору систему універсалій для розроблення АОТ.

**Структура і обсяг роботи.**

Дисертація складається із вступу, шести розділів і висновків, списку літератури з 462 найменування і додатків. Загальний обсяг роботи – 707 стор., основний текст – 397 стор.

У **першому розділі** наводяться теоретичні міркування щодо взаємозв’язку традиційної лінгвістики, структурно-математичної лінгвістики і комп’ютерної лінгвістики як триєдиної сутності.

У **другому розділі** розглядаються лінгвістичні засади створення системи морфологічного аналізу АГАТ, різні підходи до його проектування. З урахуванням модуля зняття омонімії подається загальний алгоритм морфологічного аналізу українського тексту.

У **третьому розділі** подано опис синтаксичного процесора АГАТ: синтаксичні правила членування речення на словосполучення, принципи граматичної сегментації речення, а також загальний алгоритм синтаксичного аналізу.

У **четвертому розділі** подаються правила й алгоритм автоматичного сегментування словоформ тексту на морфи, описується сфера використання морфного сегментатора.

**П’ятий розділ** присвячено проблемам семантичного аналізу тексту: розглядаються напрямки формалізації семантики, різні принципи проектування семантичних процесорів в експериментальних і промислових системах, а також тезаурусотворення і тезаурус як концептуальний компонент опису семантики.

У **шостому розділі** розглядаються приклади використання і впровадження системи автоматичного опрацювання українських текстів АГАТ у різних аплікаціях на матеріалі Корпусу української мови. Корпус української мови – це електронне зібрання текстів, впорядковане, організоване й оформлене певним чином, призначене для наукового та практичного вивчення мови. За активної участі автора робота над ним ведеться з 2009 року, а обсяг станом на 1 листопада 2013 р. досяг 24 млн. словоформ. Лінгвістичний апарат корпусу, за яким здійснюється автоматичне опрацювання текстів, звичайно називається розміткою, або анотацією, тому у назвах розділів дисертації наявний цей термін (*морфологічна анотація, морфемна анотація* тощо). Корпус тим кращий, чим повніша і досконаліша його анотація. Корпус є полігоном, на якому ми відпрацьовували технологію, і водночас джерелом будь-якої інформації, яка стосується граматичних – морфологічних, синтаксичних, а також лексичних фактів. Для аналізу моделей залучалися тексти різних функціональних стилів – від художнього (поетичного), в якому реалізується вся палітра функцій мови, до газетно-публіцистичного, в якому поєднується інформативна функція з функцією впливу, і до наукового, для якого характерним є домінування інформативної функції.

У дисертацію включено дослідження і результати експериментальних робіт, більшість з яких авторські (другий розділ – співавторський: частка автора: АМА дієслова, прикметника, числівника, займенника. Третій розділ – авторський: створення бази даних і алгоритму виявлення словосполучень і встановлення типу синтаксичного зв’язку; Четвертий розділ – співавторський: частка автора: участь у створенні морфемної бази даних і авторська – база даних морфної сегментації словоформ; п’ятий і шостий розділи авторські. Програмне забезпечення системи АОТ АГАТ здійснив програміст В.М.Сорокін).

**Апробація результатів дисертації**. Основні висновки і наукові результати дисертації доповідалися на міжнародних конференціях :

1. Міжнародна конференція Ucrainica III: Současná ukrajinistika. Problémy jazyka, literatury a kultury. – Olomouc: Univerzita Palackého v Olomouci, 2008. – Чехія, м.Оломоуц, університет, міжнародна конференція «ІУ Оломоуцький симпозіум україністів. Сучасна україністика: Проблеми мови, літератури та культури» (28-30 серпня 2008 р.). – Доповідь на тему: «Образ влади в сучасному газетному дискурсі».

2. Міжнародна конференція «Пріоритети германського та романського мовознавства» 5 – 7 червня 2009 р. – Луцьк. – Доповідь на тему: Українсько-англійський тезаурус лінгвістичних термінів з інформаційно-пошуковою системою.

3. Міжнародні наукові читання, присвячені 70-річчю від дня народження чл.-кор. НАНУ Н.Ф.Клименко, на тему: «Мови та культури у новій Європі: контакти і самобутність».- Київ, 5–7 жовтня 2009 р. – Доповідь на тему: «Інформаційно-пошукова система з лінгвістичної термінології».

4. Міжнародна конференція з корпусної лінгвістики «Slovko - 2009», м.Братислава (Словаччина), 25 – 27 листопада. – Доповідь на тему: «Текстозорієнтована тезаурусна інформаційно-пошукова система з лінгвістичної термінології».

5. Міжнародна конференція, присвячена 80-річчю від дня народження проф. Ф.О.Нікітіної. Київ, 5 – 6 листопада 2009 р. – Доповідь на тему: «Електронний тезаурус лінгвістичних термінів».

6. Міжнародна конференція «Ucrainica IY: Soucasna ukrajinistica. Problemy jazyka, literatury a kultury». – Olomouc: Univerzita Palackeho v Olomouci, 25 -27 серпня 2010. – Доповідь на тему: «Електронний словник української лінгвістичної термінології тезаурусного типу».

7. Міжнародна конференція з корпусної лінгвістики «Slavicorp». Інститут польської мови, Варшава, 22 - 24. 11. 2010 р. – Доповідь на тему: «Корпус украинского языка».

8. Міжнародна конференція «Комп’ютерна лінгвістика: сучасне і майбутнє», 23 – 24 лютого 2012 р., Національний лігнгвістичний університет, Київ. – Доповідь на тему: «Морфологічне анотування Корпусу української мови».

9. Міжнародна конференція «Українська і слов’янська тлумачна та перекладна лексикографія: теорія, практика, перспективи». Інститут української мови НАНУ,18 – 19 жовтня 2012 р. – Доповідь на тему: «Корпусукраїнської мови як джерело для мовознавчих досліджень».

10. Міжнародна конференція «Мова як світ світів». Інститут філології, Національний університет імені Тараса Шевченка, 8 – 9 листопада 2012 р. – Доповідь на тему: «Автоматичний синтаксичний аналіз текстів Корпусу української мови».

11. 1-а Міжнародна конференція «Корпусні технології. Digital Humanities і сучасне знання» 17-20 жовтня 20113 р. Нижній Новгород. – Доповідь на тему: «Синтаксическое аннотирование украинского текста (результаты и перспективы)».

**ВИСНОВКИ**

1. Оскільки інформація організована засобами природної мови, її реальне засвоєння можливе лише за умови автоматичного смислового опрацювання текстів. Потреба в лінгвістичному забезпеченні обумовлена необхідністю створення систем «людина *–* машина *–* людина»: оперативна, зручна кооперація людини і машини повинна спиратися на природну мову. В соціальному плані значущість лінгвістичних проблем комп’ютеризації пов’язана з такими основними напрямками індустрії опрацювання знань, як збирання, зберігання, систематизація, поширення, інтерпретація інформації, для чого створюється спеціальне лінгвістичне забезпечення.
2. **Лінгвістичне забезпечення автоматизованих систем** – **с**укупність засобів для здійснення комп’ютеризації мовної діяльності – необхідне практично для будь-якої інтелектуальної діяльності людини. З технологічної точки зору йдеться про створення того чи іншого типу автоматичної системи опрацювання інформації, на вході і виході якої наявна текстова інформація природною мовою. Типи систем різноманітні й можуть бути спрямовані на моделювання різних мовних завдань, зокрема таких як діалогова взаємодія, стиснення інформації, реферування тексту, логічне опрацювання змісту, переклад іншою мовою тощо. Прикладні системи, які створює лінгвіст у цій галузі, – це **лінгвістично осмислені метамови** – моделі представлення знань, кожна з яких базується на **теоретичних положеннях мовознавства** і реалізується за допомогою методів **структурно-математичної лінгвістики**. Втілювана у прикладні завдання діалектична взаємодія «традиційна лінгвістика – структурно-математична лінгвістика – комп’ютерна лінгвістика» сприяє високому рівню лінгвістичного забезпечення автоматичних систем.
3. **Комп’ютерна граматика** – це системний, строго впорядкований, формалізований, лінгвостатистичний, інтегральний опис знакових одиниць певної мови у вигляді структурних моделей із необхідною і достатньою аналітикою для виконання завдань штучного інтелекту, які відтворюють та імітують дослідницьку діяльність лінгвіста. Комп’ютерна граматика АГАТ має такі особливості. Перш за все, у ній дотримано **рівневий** підхід – рівні взаємодіють між собою від нижнього до верхнього, кожний наступний рівень використовує результати аналізу попереднього. Друга особливість – **відкритість** стратифікаційної структури граматики, що є принциповим моментом, оскільки дозволяє досить вільно розширювати обсяг лінгвістичного забезпечення, ускладнювати словникове та модифікувати програмне забезпечення без перебудови всієї системи.
   * 1. Обов’язковим складником комп’ютерної граматики є **автоматичний морфологічний аналіз** (АМА) словоформ, тому що ані морфемний, ані синтаксичний, ані семантичний аналізи не можуть обійтися без визначення для словоформи її частини мови та словозмінних форм. До **завдань АГАТ-морфології** входять: автоматичне визначення для одиниць тексту граматичної інформації про місце їх у морфологічній системі мови; автоматична ідентифікація словоформ однієї лексеми.

3.1.2. Для створення АГАТ-морфології української мови в **теоретичному плані** виконувалися дослідження, пов’язані: 1) з принципами частиномовної класифікації в українській мові; 2) з формальним обґрунтуванням морфологічних граматичних значень; 3) з принципами опису формальних засобів, характерних для відповідних частин мови з їх морфологічними значеннями. Морфологічна аксіоматика була налаштована на можливість алгоритмічного оперування граматичними даними.

У **прикладному аспекті** створено словник квазіоснов, який налічує 210 тис. одиниць, і, відповідно, словник словоформ, які породжуються поєднанням інформації, взятої з таблиці основ і допоміжної таблиці, *–* близько 3,2 млн. слововживань, що забезпечує приписування морфологічної інформації словоформам практично на 97%. Методологічно АГАТ-морфологія української мови є автоматичним формально-морфологічним процесором з елементами морфолого-синтаксичного аналізу.

3.1.3. Особливу увагу при створенні АГАТ-морфології приділено визначенню мовленнєвих умов, у яких реалізуються актуалізовані граматичні значення одиниці-омоніма. В **теоретичному плані** здійснено дослідження словниковозорієнтованих умов виникнення граматичних і лексико-граматичних омонімів в українській мові, що дало можливість укласти Граматичний словник омоформ. У **прикладному аспекті** визначено мовленнєві умови для реалізації значень досліджуваної словоформи, сформульовані за допомогою лінгвістичного методу – контекстного аналізу. В основу КА покладено твердження про те, що багатозначні елементи мови функціонують у своїх конкретних значеннях у певному лексико-граматичному контексті. Реалізація цієї ідеї знайшла відображення у створенні автоматичного конкордансу, теоретичною основою якого є: 1) наявність таких визначників, за якими кожне значення словоформи (граматичне, лексичне) детермінується в контексті іншими словоформами, їх сполученнями або іншими текстовими ознаками; 2) текстоцентричний підхід до його створення: він укладається на певному масиві текстів для певної словоформи або лексеми. Такий словник-конкорданс вичерпно ілюструє використання певної лексеми і всіх її ЛСВ з лексико-граматичними значеннями.

3.2. **АГАТ-синтаксис комп’ютерної граматики української мови**  створювався як лінгвістичний процесор, налаштований на моделювання синтаксичної структури вхідного речення на рівні словосполучень (1-й етап) і дерева залежностей (2-й етап). Результат аналізу – синтаксична структура речення, яка є сукупністю даних про синтаксичні зв’язки слів / словоформ у словосполученні – мінімальній одиниці речення.

3.2.1. АГАТ-синтаксис базується на **формально-синтаксичній теорії** представлення речення. Це комплекс алгоритмічних операцій, що виконуються над ланцюжками інформації морфологічного характеру, представленими у вихідному тексті, для встановлення синтаксичних зв’язків між текстовими одиницями. Практична реалізація теоретичних положень здійснюється шляхом взаємодії двох структурних методів: для представлення синтаксичної структури речення в термінах словосполучень застосовано **метод безпосередніх складників**, а структури цілого речення – **дерево залежностей**. Алгоритмічно і програмно в синтаксичному модулі можна здійснювати **перехід від безпосередніх складників до дерева залежностей:** коренем дерева є дієслово-присудок, у вузлах речення розташовані словоформи, кожна дуга дерева, яка зв’язує пару вузлів, інтерпретується як підрядний зв’язок.

3.2.2. Теоретичне і прикладне значення АГАТ-синтаксису полягає також у тому, що, спираючись на теоретичний синтаксис, як і в поділі словосполучень на **іменникові, прикметникові, займенникові, числівникові, дієслівні та прислівников**і, можна автоматично виявляти тип сполучуваності – **предикативний, підрядний та сурядний**  – кожного повнозначного слова в тексті. За концепцією АСА виокремлення словосполучень передбачало попередній етап створення граматики валентностей з підграматиками для дієслова, іменника, ад’єктива, а також словника фразеологізмів. Створенням такої інформаційної бази у вигляді підграматик валентностей **були розширені можливості українського теоретичного синтаксису в одержанні** з текстів різного стильового і жанрового спрямування **інформації** про конструктивні можливості сполучуваності кожної частини мови та типові моделі для певної частини мови.

3.2.3. Встановлення за правилами АГАТ-синтаксису для кожного слова **підрядних, предикативних і сурядних** типів зв’язків є відтворенням загальної системи відношень між компонентами описуваної ситуації у реченні. Перспективу АГАТ-синтаксису вбачаємо в поєднанні його з семантикою, точніше – в поєднанні лексики і граматики, оскільки вплив семантики на сполучуваність загальновизнаний і види синтаксичного зв’язку між сполучуваними словами є похідними від їхньої семантики.

3.2.4. Розроблення АГАТ-синтаксису в межах АОТ пов’язане із **загальнотеоретичною** необхідністю вивчення сполучуваності лексичних одиниць, що відкриває можливість у сучасній україністиці досліджувати граматичну і лексичну валентність слів, моделювати типову сполучуваність класів слів, синонімію словосполучень різних структурних типів, спиратися на лексичну і граматичну валентність як критерій синонімічності, вивчати закони комбінаторики словосполучень різних типів і розрядів або розмежування вільних і фразеологічних словосполучень, взаємодію стійкості й ідіоматичності тощо. Незважаючи на те, що перераховані проблеми так чи інакше розглядалися в теоретичній граматиці, комп’ютерна лінгвістика відкриває нові перспективи в **дослідженні живої лінгвістичної реальності, якою є тексти**. Застосування АСА до Корпусу української мови дає можливість дослідникам української мови зрештою встановити синтаксичну і семантичну ємність такої одиниці, як словосполучення, а у прикладному плані – розроблюваний автоматичний синтаксичний модуль аналізу українського тексту – це механізм, за допомогою якого стає реальністю укладання принаймні двох словників: частотного словника словосполучень і частотного словника сполучень простих речень у складному.

3.2.5. За допомогою АГАТ-синтаксису комп’ютер «піднімається» ще на одну сходинку у процесі «розуміння» змісту тексту, наближаючись до вирішення кінцевого завдання АОТ – побудови його семантичного представлення. Якщо АГАТ-морфологія у термінах лексико-граматичних класів слів забезпечує «розуміння» комп’ютером денотативної інформації, що міститься у тексті, то АГАТ-синтаксис відкриває шлях до релятивної інформації, тобто до розуміння семантико-синтаксичної структури речення. Синтаксичні зв’язки не існують без семантичних. І якщо не зрозуміла синтаксична структура речення, не зрозумілий і його смисл.

3.3. **Теоретико-прикладне значення** **АГАТ-морфеміки** полягає в можливості здійснювати лінгвістичні дослідження морфемної та словотвірної структури, а саме: 1) укладати алфавітно-частотні словники всіх типів морфів на базі текстів різних стилів і жанрів; 2) встановлювати й об’єднувати аломорфи в морфему; 3) встановлювати системні і функціональні характеристики морфем; 4) автоматично конструювати морфемно-словотвірні гнізда.

3.3.1. Інформаційною основою морфного сегментатора АГАТ-морфеміки є дві бази даних – 170 тис. слів і 3,5 млн. словоформ, в яких кожне слово (словоформа) представлені у вигляді морфної моделі з інформацією про типи морфів, їх структурні відношення з іншими морфами. У базовому словнику омонімічним кореням (близько 3100 одиниць) та кореневим аломорфамприписуються індекси зі списку омонімічних коренів, а кореневому аломорфу (близько 2900) – інваріантна форма. Формалізація морфних структур слів через опис їх у термінах програмних процедур морфної бази даних дозволяє створити на основі цієї бази даних автоматизовану систему аналізу, здатну виконувати низку таких **прикладних завдань:** 1)групувати лексику за спільноафіксальними класами; 2) класифікувати лексику за морфними моделями; 3) створювати кореневі та афіксальні словники з урахуванням омонімії та аломорфії.

У **теоретичному плані** такий формалізований опис морфної структури передбачає моделювання структурних відношень морфів у двох площинах організації слова як мовного знака: у плані вираження та плані змісту.

3.3.2. Робота морфемного модуля АГАТ-морфеміки здійснюється самостійно й у зв’язку зі словотвірним модулем з метою автоматичної побудови словника морфемно-словотвірних гнізд. Для цього: 1) групується лексика у спільнокореневі вибірки за процедурою ідентифікації кореневих морфів, визначених у словах морфемної бази; 2) на базі вибірки всіх спільнокореневих слів будуються словотвірні гнізда як статті електронного словотвірного словника на основі теоретичних принципів словотвірної похідності.

3.3.3. Формалізований опис морфних структур, запропонований при укладанні електронного морфемного словника, дозволяє використати морф як інструмент у проведенні автоматичного морфемного аналізу інших лексикографічних систем.

3.3.4. Розроблені АГАТ-морфеміка разом з морфемно-словотвірною базою даних мають важливе **теоретичне значення** для вивчення правил взаємодії морфеміки і фонології, інвентарю словотвірних засобів української мови, семантики афіксів, афіксоїдів у широкій взаємодії з Корпусом української мови, який надає великий ілюстративний матеріал для вивчення і розв’язання цих питань.

4. У розробленні **АГАТ-семантики** української мови основна увага приділялася створенню ідеографічних словників тезаурусного типу – вони є інструментом прикладної лінгвістики, за допомогою якого моделюється семантична структура мови в парадигматичному аспекті.

4.1. До ідеографічного опису лексики та класифікації мовних даних застосовувалися два основні підходи: **індуктивний** і **дедуктивний**. Індуктивний метод побудови ТЗ полягав у моделюванні семантичних відношень у лексиці у вигляді ієрархії (від загального до часткового) і у вигляді семантичної мережі, у якій відсутнє мотивоване розташування лексем. Індуктивний метод застосовувався для побудови ІПС лінгвістичних термінів, а дедуктивний – для побудови ідеографічного словника української публіцистики і словника на синсетах.

4.2. Запропоновано концептуальну модель, на основі якої створено автоматизовану інформаційно-пошукову систему тезаурусного типу юридичних термінів з проблем усиновлення (2140 одиниць), що методологічно реалізована на тісному зв’язку **текст – тезаурус**.

4.3. На основі формалізованої **методики конструювання тезауруса (3400** термінів**)** укладено електронний словник лінгвістичних термінів із системою логічних відношень між ними у вигляді семантичної мережі **(9265** семантичних відношень**).** Мережеве представлення даних має **теоретичне значення**, а саме:дозволяє глибше проникнути в логічні відношення лінгвістичної метамови, точніше змоделювати аналізовану терміносистему. Розроблений текстозорієнтований інформаційно-пошуковий тезаурус із лінгвістичним апаратом і програмним забезпеченням має **прикладне значення**: він може бути використаний при створенні термінологічних банків даних і термінологічних банків знань, оскільки має значний обсяг термінологічної інформації. Засоби автоматизації відбору, збереження і пошуку інформації дають можливість збагачувати й розвивати словник лінгвістичних термінів.

4.4. Розроблено й описано принципи та методику укладання ідеографічного словника. На матеріалі Частотного словника української публіцистики створено ідеографічний словник іменників (13 тис. лексем) і дієслів (близько 6 тис. лексем).

Методика укладання списку онтологій полягає у створенні логічної моделі організації знань у вигляді фреймів, оскільки через терми і слоти краще структурується представлення знань і переконливішою стає позиція дослідника у членуванні позамовної дійсності. У **прикладному** аспекті ці словники в подальшому моніторингу газетних текстів виконуватимуть роль матриці, яка накладатиметься на ЧС публіцистичного стилю Корпусу текстів української мови, що створить, наприклад, базу даних для аналізу «мовної картини» суспільно-політичного життя України.

4.5. Укладено тезаурус синонімічних рядів слів української мови (14 тис. синонімічних рядів, які охоплюють понад 87 тис. слів), у якому кожне слово (разом із синонімічним рядом) описується як системно зумовлена одиниця і характеризується синонімічними, асоціативними та родо-видовими відношеннями. Базовою одиницею є не конкретне значення слова, а значення синонімічного ряду. **Теоретичне значення ТЗ** **на синсетах** полягає в розробленні принципів систематизації лексики української мови з урахуванням семантичних кореляцій (семантичних парадигматичних відношень). **Прикладне значення:** електронний словник синонімів української мови вбудовано в програмний лінгвістичний продукт Майкрософт-офіс.

4.6. З **теоретичного погляду** розроблені методики АГАТ-семантики є містком між семантичною структурою тексту і семантичною структурою словникового складу, оскільки слово з його лексичними значеннями є найпростішим мовним засобом номінації фрагментів дійсності (предметів, якостей, явищ, подій).

У **прикладному аспекті** запропоновані моделі спрямовані на вдосконалення методики побудови ТЗ, які дозволяють в автоматизованому режимі відносно легко будувати словники різного призначення, що значною мірою полегшує створення різних словникових ресурсів і внутрішніх словників автоматичного опрацювання інформації, а також відкриває можливості узагальнення, кластеризації різних значень, зокрема встановлювати семантичні відношення між частинами мови. В системі АОТ АГАТ розроблені принципи створення лінгвістичного ресурсу української мови у вигляді тезаурусів. Описуючи методики побудови ТЗ української мови, ми приділили увагу експериментам із застосування цих тезаурусів у різних завданнях опрацювання текстів для того, щоб показати, наскільки якісно можна вирішувати конкретні завдання на базі тезаурусних знань.

5. **Комп’ютерна граматика** взагалі і АГАТ-граматика зокрема, є динамічним механізмом, який є системою правил оперування граматичними значеннями з акцентом на форми їх вираження, за допомогою яких комп’ютеру відкривається доступ до денотативної і релятивної інформації, до елементарних смислових блоків у тезаурусах. **Комп’ютерні словник і граматика** – два тісно пов’язані й узгоджені компоненти структури мови. Їх узгодженість визначається спільністю основних функцій і збереженням у комп’ютерній пам’яті як мовних одиниць, готових до вживання, так і граматичних правил, за якими відповідно до завдання автоматично здійснюється аналіз тексту.

6. При побудові АГАТ-граматики перевага надавалася індуктивним методам дослідження з огляду на те, що взагалі процес пізнавальної діяльності лінгвіста по суті повинен бути індуктивним – від форми до значення, а потім може сполучатися з дедуктивним, який забезпечує процес переходу від значення до форми.

7. Досвід створення АГАТ-граматики і АГАТ-семантики показав, що розроблення архітектури повного циклу автоматичного аналізу тексту варто починати з верхніх мовних рівнів до нижніх, а вибудову лінгвістичного забезпечення – з нижніх до верхніх рівнів знакових одиниць. Послідовність рівнів аналізу спирається на нижній морфологічний рівень. При цьому кожній одиниці слід приписати інформацію, достатню для вирішення завдань свого рівня з перспективою використання її на вищих рівнях аналізу. Кожному рівневі відповідає алгоритм аналізу та комплекс даних, які будуть передбачені у словниковому забезпеченні. Модульне лінгвістичне забезпечення з відкритим характером передбачає сумісність окремих модулів, їх поповнення і коригування за умови взаємоузгодженості словникового і граматичного описів мовної інформації.

8. Запропоновані методи і принципи аналізу не є умоглядними – вони апробовані при створенні конкретних систем АОТ. Комп’ютерна граматика української мови, словникове та програмне забезпечення можуть бути ядром будь-якого комплексу, налаштованого на опрацювання текстової інформації. Результати дослідження, представлені в дисертації, можна розглядати, по-перше, з точки зору їхньої практичної цінності для систем АОТ української мови; по-друге, з точки зору оцінки лінгвістичних аналізів, принципи і методи яких базуються на традиційній і структурно-математичній лінгвістиці з прицілом на комп’ютерне моделювання і, ширше, на штучний інтелект. Ці завдання визначили і загальну спрямованість усього дослідницького апарату на пошук формальних ознак, за допомогою яких розроблені алгоритмічні граматики і семантика текстових одиниць.

9. При оцінці результатів досліджень складних систем, таких, як природна мова, штучний інтелект і под., особливий інтерес становлять дані про ступінь досягнутої на різних рівнях аналізу формалізації і дані про ступінь складності застосовуваних для цього процедур. У **теоретичному плані** цей аспект пов’язаний із методологічним питанням про межі формалізації явищ мови і мовлення та з питанням необхідності поєднання формальних і змістових методів дослідження.

10. При створенні комп’ютерної граматики довелося проаналізувати величезну кількість конкретних лінгвістичних фактів, і розгляд кожного з них вимагав заглиблення у багатовимірний простір лінгвістичної теорії. Якщо основою методології є єдність теоретичного і прикладного мовознавства і, відповідно, лінгвістичних процедур аналізу, можна сподіватися на хороші результати.

11. Комп’ютерна АГАТ-граматика і комп’ютерні словники були застосовані до Корпусу української мови, який є полігоном удосконалення і верифікації лінгвістичних моделей. Лінгвісти в усьому світі починають усвідомлювати важливість розробки **граматик** національної мови **для лінгвіста** – не для пересічного користувача шкільних або академічних граматик, а таких, які були б зорієнтовані і базувалися на матеріалі «активного» мовлення, представленого Корпусом української мови. Тому створення граматичних описів, які базуються на корпусних даних, набуває особливого значення. Проанотовані АГАТ-граматикою корпусні тексти призначені, в першу чергу, для виконання подібних завдань. Другою проблемою є створення **тлумачного словника для лінгвіста**, для чого треба розробити метамову опису лексичного значення слів, яка б передбачала розроблення 1) інвентарю семантичних дескрипторів (предметних, предикатних); 2) правил поєднання дескрипторів. І граматика, і тлумачний словник для лінгвіста неможливо створити без здобутків традиційної лінгвістики, у чому ми вбачаємо перспективу спільного розвитку і зближення традиційної, структурно-математичної і комп’ютерної лінгвістики.

12. У **методичному плані** всі результати дисертаційного дослідження використовуються при читанні студентам й аспірантам спецкурсів «Комп’ютерний морфологічний аналіз», «Комп’ютерний синтаксичний аналіз», «Комп’ютерний семантичний аналіз», «Семантичні мережі», «Бази даних і бази знань», «Термін у лінгвістичній інформатиці», «Політична лінгвістика», «Корпусна лінгвістика». Результати досліджень викладено для загального користування на мовно-інформаційному порталі [www.mova.info](http://www.mova.info).