

ПЕРЕДМОВА

Повний та всебічний опис мови та мовлення потребує врахування і якісних, і кількісних характеристик лінгвістичних об'єктів. Використання кількісних методів у мовознавстві має об'єктивні підстави: очевидним є вплив кількісних характеристик на якісні (кількість голосних морфем, наприклад, впливає на звукову якість слів); можна обчислити кількість абзаців, речень, слів, букв, звуків та інших кількісних характеристик тексту, частоти різних елементів мови у мовленні підпорядковані ймовірно-статистичним законам. Прикладна лінгвістика виникла наприкінці 20-х років ХХ ст. З появою комп'ютерів у цій галузі активізували дослідження щодо практичного застосування накопичених лінгвістичних знань, завдяки чому виникають і з середини ХХ ст. стрімко розвиваються математична лінгвістика, комп'ютерна лінгвістика, машинний переклад тощо. Однією з перших в Україні фахівців з прикладної лінгвістики почала готувати кафедра прикладної лінгвістики Національного університету "Львівська політехніка".

Математична лінгвістика як складова частина прикладної лінгвістики. Прикладна лінгвістика – галузь мовознавства, об'єктом якої є розроблення методів і засобів застосування у соціальній практиці людини результатів фундаментальних лінгвістичних досліджень, спрямованих на пізнання онтологічних властивостей мови. Термін *прикладна лінгвістика* виник наприкінці 20-х років ХХ ст., її складовими є комп'ютерна (машинна, обчислювальна, інженерна) лінгвістика, математична лінгвістика, комунікативна лінгвістика, лінгводидактика (викладання іноземної мови), термінознавство, перекладознавство, контрастивна лінгвістика, статистична лінгвістика, структурна лінгвістика тощо.

Комп'ютерна лінгвістика (машинна, обчислювальна, інженерна лінгвістика) займається застосуванням комп'ютерів з метою моделювати

функціонування мови в певних умовах, використанням комп'ютерів (технологій і програм опрацювання даних) для вирішення лінгвістичних завдань, а також розробляє лінгвістичні аспекти комп'ютеризації. У широкому розумінні до комп'ютерної лінгвістики зараховують усе, що пов'язано з використанням комп'ютерів у мовознавстві. *Структурна лінгвістика* – сукупність поглядів на мову та методи її дослідження, яка вважає мову знаковою системою з чітко виділеними структурними елементами і прагненням до строгого, формального опису мови (її вважають попередницею комп'ютерної та математичної лінгвістики). *Контрастивна лінгвістика* (зіставна лінгвістика) – порівняння двох, рідше кількох мов для виявлення їх подібності або відмінностей на всіх рівнях мовної структури з метою типологічної класифікації мов. Вона виникла і активно розвивалася у 50-ті роки ХХ ст.

Математична лінгвістика виникла в 50-ті роки ХХ ст., підґрунтям для її появи стала необхідність в уточненні основних лінгвістичних понять; потреба в уведенні точніших та об'єктивніших методів для аналізу та синтезу мови і тексту; поява міжпредметних зв'язків з іншими галузями – акустикою, фізіологією вищої нервової діяльності, кібернетикою та обчислювальною технікою, які потребують спілкування мовою математики; з розвитком можливостей комп'ютерних технологій також виникла потреба, зокрема, у машинному перекладі та автоматизованому інформаційному пошуку. Власне *математичну лінгвістику* – галузь науки на межі *мовознавства* та *математики*, яка вивчає можливості застосування математичних методів для опису та дослідження природних і штучних мов, для пояснення лінгвістичних подій, – вважають теоретичним підґрунтям прикладної лінгвістики. Підкреслюючи спільність поняттєвого апарату, математичну лінгвістику іноді розділяють на галузі мовознавства та математики, а також зазначають, що в частині *використання* розроблених математичних моделей для опису будови природних мов математична лінгвістика належить до такої галузі досліджень, як *штучний інтелект*. Виділимо специфічні завдання мовознавства й прикладної лінгвістики зокрема, які вирішує математична лінгвістика за допомогою взаємодії відповідних математичних методів.

• Побудовані за допомогою обернених тригонометричних функцій моделі можна використовувати для опису структурних зсувів (наприклад, появи або зникнення лінгвістичних одиниць) в області лексикології, морфології, фонології, синтаксису та стилістики.

- У роботах з діахронічної лінгвістики часто наводяться таблиці, що містять кількісні дані про виявлення досліджуваного явища на різних етапах його історії. Такі таблиці відображають динаміку лінгвістичного процесу, який можна поглибленіше аналізувати за допомогою елементарних функцій, графічно, а також із застосуванням аналітичних моделей досліджуваного процесу.

- Класична глотохронологія має на меті наближено датувати процес розходження діалектів та споріднених мов, а також кількісно оцінити ступінь їхньої спорідненості, й для цього використовує статистичні методи.

- Методи математичної статистики також допоможуть у створенні словників (зокрема частотних і статистичних), автоматичних словників, тезаурусів (лексикографія), у вирішенні завдань стилеметрії та інформаційного пошуку.

- Постійне розширення сфери діяльності сприяє безперервному збільшенню термінологічного словника людини, і для лексикографічної практики, а особливо під час побудови систем машинного перекладу та реферування, важливо мати прогноз кількісного зростання термінології в різних галузях знань. Таке прогнозування дало б можливість свідомо планувати випуск та ефективне поповнення електронних політехнічних і галузевих словників та довідників.

- Для визначення швидкості змін у мові та мовленні, а також їхніх кількісних оцінок використовують поняття похідної.

- Тригонометричні функції та поняття границі дають змогу моделювати циклічний і стрибкоподібний характер лінгвістичних процесів.

- Можна також описати процеси приросту та накопичення нових мовних елементів, використовуючи теорію рядів та інтегрування.

- Приблизні оцінювання вимірів обмежень, що накладають на використання лінгвістичних одиниць система та норми мови, можна виконувати за допомогою комбінаторики.

- Лінгвістичні дослідження щодо підготовки систем машинного перекладу та інформаційного пошуку вимагають, зокрема, знання обсягу вибірки, необхідного, щоб забезпечити із заданою ймовірністю появу певної лінгвістичної події хоча б раз. Визначити

достатній обсяг вибірки необхідно для встановлення достовірності зроблених висновків.

- Створення машинних фондів мов.

Математична лінгвістика також бере участь у вирішенні таких універсальних проблем прикладної лінгвістики:

- розроблення формальних моделей природних і штучних мов;
- вирішення питань практичної транскрипції та транслітерації;
- дешифрування невідомих писемностей;
- усний/письмовий переклад, розроблення систем автоматичного перекладу;

- авторська та часова атрибуція твору;
- вивчення процесів і формулювання правил утворення нових назв товарів, виробів тощо;

• створення систем стенографії, систем письма для сліпих;

• судова і кримінальна лінгвістика;

• лінгводидактика: вивчення рідної та іноземної мов, розроблення відповідних лінгвістичних методик (навчання дітей та дорослих, емігрантів тощо);

- розроблення раціональної та стабільної орфографії;
- автоматичне розпізнавання і синтез мови;

• розроблення автоматизованих систем опрацювання текстової інформації: автоматичне зняття полісемії, автоматичне анотування, індексування і реферування, синтаксичний аналіз, компресія текстів, інтелектуальний аналіз тексту з метою виділення неявного змісту (“читання між рядками”);

- лінгвістичне забезпечення автоматизованих систем керування;
- комп’ютеризація навчання та автоматизація викладання за допомогою комп’ютера: автоматизовані системи навчання, системи дистанційного навчання, автоматизоване робоче місце перекладача і викладача;

• розв’язання задач прикладного термінознавства: систематизація та автоматизація терміносистем певних галузей, урахування динамічності їхнього розвитку, укладання термінологічних словників, стандартизація й уніфікація науково-технічної термінології тощо;

- створення автоматизованих редакційно-видавничих систем.

Предмет математичної лінгвістики. Структура курсу. Цей навчальний посібник містить опис математичних методів, необхідних для ґрунтовного висвітлення фундаментальних засад математичної лінгвістики, що стосуються комбінаторної лінгвістики, а також можливостей застосування описаних методів для розв'язування прикладних задач мовознавства. Для закріплення знань запропоновано запитання й тестові завдання із ключем, а також тридцять варіантів завдань трьох рівнів складності для самостійної роботи. Матеріал сформовано так, щоб розкрити й описати методологію вирішення завдань мовознавства за допомогою відповідних математичних методів.

Комбінаторна лінгвістика вивчає та описує лінгвістичні явища за допомогою методів “некількісної” математики – теорії формальних грамастик, теорії автоматів, мереж Петрі.

Комбінаторна лінгвістика – напрям у мовознавстві, який вивчає синтагматичні зв'язки мовних одиниць та їхні комбінаторні властивості. Комбінаторна лінгвістика – синтез двох областей. По-перше, це синтагматика, що являє собою аспект мови, який визначає мовні правила сполучуваності однорівневих одиниць. По-друге, комбінаторика, змістом якої є складання і вивчення комбінацій слів, підпорядкованих певним комунікативним завданням за конкретних умов їх реалізації, які можна утворити із заданої кількості слів.

У межах комбінаторної лінгвістики, залежно від вибраної одиниці аналізу, виділяють такі розділи, як комбінаторна фонологія, комбінаторна морфологія, комбінаторна морфеміка і комбінаторна лексикологія.

Актуальність цього напряму зумовлена такими причинами: 1) потребою класифікаційно-систематизованого опису сполучувальних властивостей мовних одиниць; 2) інтересом до дослідження функціонально-мовного аспекту мови; 3) необхідністю вивчення когнітивного аспекту сполучуваності; 4) розширенням сфери викладання іноземних мов, що спричиняє потребу в зіставно-типологічних дослідженнях сполучуваності в різних мовах; 5) необхідністю створення словників комбінаторного типу в різних сферах вживання мови тощо.

Зародження комбінаторної лінгвістики зумовлено появою синтагматики – сполучуваності, основаної на лінійних відносинах одиниць мови.

Незважаючи на великий інтерес до вивчення комбінаторних властивостей слова, єдиної загальноновизнаної теорії сполучуваності поки що не створено. Сьогодні комбінаторна лінгвістика впритул взаємодіє із синтагматикою різних видів. По-перше, із синтаксичною синтагматикою, суть якої полягає у вивченні синтаксичної сполучуваності одиниць мови або закономірностей з'єднання граматичних класів слів у мові. Синтаксична синтагматика – це сукупність правил сполучуваності одиниць у синтаксичні конструкції різних складності та протяжності. Закони і правила синтагматики діють у всіх сферах синтаксису. На синтаксичному рівні одиниці мови вивчають тільки в плані вираження.

Метою комбінаторної лінгвістики є теоретичний опис і вирішення проблем, пов'язаних із синтагматичними відносинами і комбінаторними властивостями одиниць мови, природою сполучуваності, синтезом семантики і сполучуваності знаків мови, функціями сполучуваності й, особливо, її прикладним аспектом.

З появою комп'ютерів виникла потреба у створенні засобів спілкування між людиною та машиною. Алгоритмічні мови, які організують таке спілкування, повинні задовольняти певні вимоги: бути зручними та зрозумілими для людини, а також сприйматися машинами.

У посібнику розглянуто процес моделювання за допомогою формальних граматик і автоматів лексичного етапу роботи компілятора, коли під час аналізу вхідного тексту в ньому виділяють найпростіші конструкції мови – ідентифікатори, символи, числа, символи операцій, службові слова тощо. Описано і другий етап роботи компілятора, коли тіло синтаксичного аналізатора – це діаграма переходів відповідного детермінованого скінченного автомата. Для задання синтаксису алгоритмічної мови використовують формальні грамматики, які Н. Хомські ввів як формалізацію одного із розділів лінгвістики – синтаксису. Засоби, які уможливають перетворення тексту з мови людини на мову машини, називають трансляторами.

У роботі компілятора (транслятора компілювального типу) виділяють три етапи: лексичний аналіз, синтаксичний аналіз, генерування вихідного тексту.

Комплексне завдання для закріплення знань, отриманих у цьому розділі, полягає у реалізації лексичного аналізатора мови програмування.

Теоретичну частину організовано у формі скінченного автомата, який розрізняє такі класи лексем: числа, рядкові та символні константи, коментарі, зарезервовані слова, оператори, розділові знаки, ідентифікатори.

Основою для методів формалізації мови є теорія формальних граматики. Скінченні автомати розглядатимемо як абстрактні моделі найпростіших пристроїв оброблення даних.

Простежити зв'язок між мовами, граматиками й автоматами допомагають певні задачі. Задають тип ідентифікатора – це ланцюжок довжини 5, який може складатися із цифр, великих або малих літер українського алфавіту. Треба побудувати скінченний автомат з виходом, що розпізнає ланцюжок заданого типу, і лише його, з певними обмеженнями (наприклад, символи з парними номерами – цифри). Необхідно побудувати граматику, що породжує такий і лише такий ланцюжок, знайти можливу кількість комбінацій елементів ідентифікатора такого типу. Комплексне (групове) завдання для закріплення знань, отриманих у цьому розділі, полягає у реалізації лексичного аналізатора мови програмування. Теоретичну частину організовано у формі скінченного автомата, який розрізняє такі класи лексем: числа, рядкові та символні константи, коментарі, зарезервовані слова, оператори, розділові знаки, ідентифікатори.

У розділі, що стосується лінгвістичного аналізу текстової інформації, текст розглядають як автономну структуру, вивчають його структуру, категорії та складові, а також правила побудови зв'язного тексту. Реалізація таких завдань передбачає застосування методів лінгвістичного аналізу для попереднього розбору багатьох текстів та порівняння результатів. Об'єктом лінгвістичного аналізу тексту може бути, наприклад, науковий, офіційно-діловий, публіцистичний чи художній тексти.

Визначені під час лінгвістичного аналізу логічні правила побудови зв'язного тексту можна використати для розв'язування задач опрацювання та синтезу природної мови, що є важливою проблемою створення систем штучного інтелекту. Для моделювання асинхронних інформаційних потоків у системах опрацювання даних використовують модифікацію мережевої моделі – розфарбовану мережу Петрі. Введення поняття кольору дає змогу запровадити нове правило взаємодії між складовими

мережі (дуга певного кольору, може пропустити тільки фішки відповідного кольору), коли із кожним кольором пов'язана конкретна семантична інтерпретація.

Застосування лінгвістичного аналізу дає змогу виділити правила побудови зв'язного тексту на основі аналізу текстів різного стилю. У машинній побудові зв'язних текстів, а також у реалізації машинного синтезу мови знайшли застосування мережі Петрі, пов'язані із моделюванням причинно-наслідкових зв'язків.