

УДК 681.3

Ю.М.Щербина

Національний університет „Львівська політехніка”,
кафедра “Інформаційні системи та мережі”

ПРЕДМЕТ МАТЕМАТИЧНОЇ ЛІНГВІСТИКИ

© Щербина Ю. М., 2002

Deals with the preconditions of the appearance and main principles of the mathematical linguistics and different points of view on its subject were analyzed. The program of corresponding educational course was proposed. The context has been revealed.

Розглянуто передумови виникнення та основні положення математичної лінгвістики, проаналізовано різні погляди на її предмет. Запропоновано програму відповідного навчального курсу та розкрито її зміст.

Вступ. У зв'язку з поширенням інтелектуальних інформаційних систем обробки інформації та систем перекладу з однієї природної мови на іншу зростає необхідність підготовки інженерних кадрів, які володіють точними методами у мовознавстві та можуть кваліфіковано і свідомо використовувати весь спектр сучасного програмного забезпечення у цій галузі. Застосування математичних методів у лінгвістиці зумовлено двома причинами.

По-перше, розвиток теорії і практики мовознавства вимагає введення все більш точних і об'єктивних методів для аналізу мови і тексту. Одночасно використання математичних прийомів при систематизації, вимірюванні та інтерпретації лінгвістичного матеріалу у поєднанні з якісною інтерпретацією результатів дозволяє мовознавцям глибше проникнути у таємниці побудови мови і утворення тексту.

По-друге, контакти мовознавства, які постійно розширюються, з іншими науками (акустикою, фізіологією вищої нервової діяльності, кібернетикою та обчислювальною технікою) можуть здійснюватись тільки при використанні математичної мови, яка має високий ступінь загальності та універсальності для різних гілок знань. Особливо наполегливо математизується мовознавство у зв'язку з використанням природної мови в інформаційних і управлінських системах людина – машина – людина. В існуючих системах машинного перекладу, автоматичного анотування, людино-машинного діалогу будь-яке повідомлення на природній мові перекодовується в математичну мову комп'ютера.

Застосування математичних методів у мовознавстві має на меті замінити звичайно дифузну, інтуїтивно сформульовану і таку, що не має повного розв'язання, лінгвістичну задачу однією або декількома простішими, логічно сформульованими математичними задачами, які мають алгоритмічний розв'язок. Таке розділення складної лінгвістичної проблеми на простіші задачі, які алгоритмізуються, ми будемо називати *математичною експлікацією* лінгвістичного об'єкта або явища.

Математична експлікація цікава не тільки з суто пізнавального і теоретичного погляду. Вона абсолютно необхідна при вирішенні прикладних питань, пов'язаних з

аналізом і синтезом усної мови або інформаційною переробкою текстів на комп'ютері. Математична експлікація лінгвістичних об'єктів застосовується не тільки при розв'язуванні на комп'ютері нескладних, хоча і важливих задач такого типу, як упорядкування частотних і алфавітних словників або послівного і пооборотного машинного перекладу, але також при складанні і реалізації таких евристичних алгоритмів штучного інтелекту як семантичний переклад або тезаурусне реферування тексту.

Відзначимо, що абстрактні моделі використовуються при вивченні мови дуже давно – з тих часів, як існує граматики. Такі елементарні поняття граматики, як підмет, присудок, відмінок, рід тощо, є достатньо абстрактними конструкціями, які належать до фактів мови саме як моделі. Усвідомити їх абстрактний характер важко лише через їх звичність. За своїм характером ці поняття досить близькі до математичних. Тут доречно зауважити, що на базі понять підмету і присудку було вироблено центральне поняття сучасної математичної логіки – *поняття предиката*.

По суті, створення абстрактних моделей завжди було головним засобом *теоретичного* вивчення мови. Але значення цього факту було усвідомлено лише на початку ХХ століття, коли у працях основоположників так званої структурної лінгвістики виникла концепція мови як абстрактної знакової системи, у якій визначальна роль належить не матеріальній природі знаків, а співвідношенням між ними. Тому засобом пізнання законів мови є побудова абстрактних моделей її структури і вивчення самих цих моделей.

З вказаної концепції природно було зробити висновок, що мову потрібно вивчати засобами, що є близькими до тих, якими користується математика під час аналізу своїх формальних систем; адже ці системи, як і мова, характеризуються тим, що в них важливими є тільки відношення між об'єктами, а не матеріальна природа останніх. По суті, математичні системи – це спеціальні різновиди мови, які вирізняються особливо чіткою структурою. Іншими словами, математика може виявитись придатною *метамовою* для вивчення природних мов і тим самим стати універсальною мовою лінгвістики. Саме це і відбулось з виникненням математичної лінгвістики.

У середині 50-х років ХХ століття визначились основні принципи трактування лінгвістичних понять і математична лінгвістика почала бурхливо розвиватись. Прискоренню її розвитку істотно сприяла та обставина, що як раз тоді у зв'язку з появою обчислювальних машин і швидким зростанням потоку наукової інформації була поставлена задача автоматизованого перекладу та інформаційного пошуку. Ця задача привернула до лінгвістики увагу спеціалістів у галузі точних наук і поставила до неї вимоги, які неможливо було задовольнити без різкого підвищення рівня строгості лінгвістичних понять.

Комбінаторна і квантитативна лінгвістика. Вибір математичного апарату у лінгвістичних дослідженнях – питання не просте і дискусійне [1, 3, 7, 8]. Його розв'язання першочергово залежить від того, як визначається предмет і основні поняття мовознавства.

Один погляд полягає в тому, що в математичну лінгвістику входить тільки такий математичний апарат, який є *специфічним* для лінгвістики в тому смислі, що він належить до найбільш важливих сторін функціонування мови і розробляється *спеціально* у зв'язку із застосуванням до мови. З такого погляду математична лінгвістика входить у той комплекс математичних дисциплін, який сьогодні прийнято називати „дискретною математикою”. На основі застосування апарату дискретної математики у теоретичному мовознавстві сформувалась галузь, яка умовно називається *комбінаторною* лінгвістикою [1, 8]. Головним

напрямок тут є вивчення граматик, які породжують текст. При цьому граMATика розуміється як скінченна множина детермінованих правил, а мова розглядається як нескінченна кількість регулярних ланцюжків слів, які породжуються цією граMATикою. Формальний апарат теорії породжувальних граматик має математико-логічний характер; по суті ця теорія є складовою частиною так званої теорії числень, яка, своєю чергою, є однією із відгалужень математичної логіки. Проблематика цієї теорії має багато спільного з проблематикою теорії алгоритмів і теорії автоматів. Початок теорії породжувальних граматик був закладений у роботах американського лінгвіста Н. Хомського. Інший напрям тут полягає у побудові таких описів мов, у яких вихідним матеріалом слугує деяка сукупність мовних одиниць і за допомогою формальних операцій над цією сукупністю одержуються певні „граматичні характеристики мови” (Такий підхід у певному сенсі протилежний підходу теорії породжувальних граматик.) Методи цього напрямку – алгебраїчні і комбінаторні, але і тут є точки перетину з теорією алгоритмів і теорією автоматів [1]. Серед інших напрямків можна відзначити теорію категоріальних граматик [1]; ці граMATики, подібно до породжувальних, належать до кожної мовної одиниці певного вигляду опису її структури, але при цьому вони є не породжувальними, а *розпізнавальними* системами [1]. Важливим розділом математичної лінгвістики є теорія підпорядкування між частинами речення.

Інший погляд полягає у тому що математичне мовознавство не може обмежитись детерміністською, „некількісною” експлікацією лінгвістичних об'єктів.

По-перше, це обмеження утруднює перетворення нечітких лінгвістичних множин, елементи яких мають імовірнісні ваги приналежності, у чіткі множини природних мов. У той же час вказане перетворення становить основу машинної обробки тексту і автоматичного розпізнавання смислу [7, 8].

По-друге, за такого обмеження поза сферою застосування математичних методів залишається акустико-фізіологічна та психолінгвістична проблематика мовотворення, а також стилістика та історія мови, під час вивчення яких широко застосовуються не стільки комбінаторні, скільки кількісні виміри.

У зв'язку з цим у теоретичному мовознавстві виділяється галузь, яка умовно називається *квантитативною* лінгвістикою [8]. Її основу становлять такі „кількісні” розділи математики, як математична статистика, теорія ймовірностей, теорія інформації, математичний аналіз.

Для того, щоб правильно оцінити співвідношення комбінаторних і кількісних математичних методів під час описання мови і тексту, пропонується розглянути таку загальну схему мовної діяльності і текстотворення [8].

Породження тексту визначається, з одного боку, системою мови і нормою, що обмежує її дію, а, з іншого – зовсім незалежною від мови зовнішньою ситуацією.

Якщо погодитись з тим, що система мови є механізмом, який породжує тексти без жодних імовірнісних обмежень [1, 2, 8], то стане зрозумілим, що експлікація цієї системи повинна здійснюватись за допомогою тих „некількісних” розділів математики, про які ми говорили вище.

Розглядаючи мову як некількісну систему, комбінаторна лінгвістика намагається описати механізм текстотворення за допомогою тих же засобів некількісної „математики”. Таке описання здійснюють контекстно-вільні граMATики, тобто граMATики, які не враховують контекстних обмежень на вживання окремих лінгвістичних одиниць та їхніх

сполук. У зв'язку з цим контекстно-вільні граматики породжують багато ланцюжків, які не є реальними реченнями даної мови. Щоб добитись породження реальних текстів, необхідно перейти до більш сильних контекстно-залежних граматик. Такі граматики можна побудувати за умови, що до елементів системи мови застосовуються імовірнісні оцінки, а сама мова розглядається як некількісна твірна система, функціонування якої регулюється імовірнісними обмеженнями, які закладені в нормі [8].

Щодо тексту, то він є лінійним ланцюжком відокремлених один від одного (дискретних) символів (букв, складів, слів). Кожний із символів зустрічається у тексті з певною частотою і має особливі валентності, тобто лінгвістичні властивості сполучатись з іншими символами. Ці властивості лінгвістичних одиниць у тексті експлікуються в термінах теорії ймовірностей і математичної статистики. До результатів імовірнісно-статистичного описання, які розглядаються разом з даними лінгвістично-психологічного експерименту, може бути застосований апарат теорії інформації, за допомогою якого вдається кількісно оцінити як структурну організацію тексту, так і смислову інформацію, яка в ньому міститься.

Отже, прихильники другого погляду на предмет математичної лінгвістики [8] роблять висновок, що математична експлікація центральної проблеми сучасного мовознавства „система мови – норма – текст” може бути здійснена на основі застосування методів як комбінаторної, так і квантитативної лінгвістики.

Особливостями багатьох конструкцій математичної лінгвістики є те, що вони можуть застосовуватись до мовних явищ різної конкретної природи, або, як часто говорять, до різних „рівнів” мови. Так, за допомогою породжувальних граматик можна однаково успішно описувати як синтаксичний рівень мови (правила утворення речень із слів), так і морфологічний (правила утворення слів із морфем).

У зв'язку з розробкою лінгвістичних аспектів штучного інтелекту виникає необхідність формального опису зовнішніх ситуацій, які стимулюють породження тексту. Для описання таких ситуацій використовується як кількісна, так і комбінаторна методика.

Щодо моделювання неперервних змін мови у часі (діахронічна лінгвістика), географічному просторі (діалектологія), а також у соціально-професійному і художньому континуумі (соціолінгвістика та стилістика), то доцільно використовувати поняття нескінченної множини, граничного переходу, неперервності, тобто поняття, що становлять основу математичного аналізу.

Завдання вивчення дисципліни „Математична лінгвістика” та її зміст. Викладені вище міркування було покладено в основу програми курсу „Математична лінгвістика” для студентів четвертого курсу, які навчаються за базовим напрямком „Філологія” за фахом „Прикладна лінгвістика”. За навчальним планом на вивчення курсу виділено 15 годин лекцій і 45 годин лабораторних занять. При відборі матеріалу ми вважали необхідним ознайомити студентів з основними напрямками як квантитативної, так і комбінаторної лінгвістики, а також показати деякі практичні застосування. З квантитативної лінгвістики вибрано теми, пов'язані з використанням теорії ймовірностей, математичної статистики та теорії інформації: розглядаються виміри смислової інформації слів та надлишковості тексту, описи функцій розподілу в тексті складів, слів, словосполучень і граматичних класів, методи побудови статистичних моделей тексту та визначення імовірнісних характеристик норм мови. З тематики комбінаторної лінгвістики розглядається теорія формальних породжувальних граматик Н.Хомського, класифікація мов,

представлення регулярних мов детермінованими та недетермінованими скінченними автоматами, а також регулярними виразами. Вивчаються властивості контекстно-вільних мов.

Ми виходили з того, що в результаті вивчення дисципліни студенти повинні знати:

- означення та методи визначення смислової інформації слів та надлишковості тексту;
- означення та методику побудови функцій розподілу в тексті складів, слів, словосполучень і граматичних класів;
- методи побудови статистичних моделей тексту та імовірнісних характеристик норм мови;
- означення породжувальних граматики та їх класифікацію за Н. Хомським;
- означення поняття скінченного автомата та його використання для адекватного представлення регулярних мов;
- властивості контекстно-вільних мов, представлення виведення у вигляді дерева граматичного розбору;
- можливості застосування контекстно-вільних граматики;

повинні вміти

- визначати смислову інформацію слів та надлишковість тексту;
- будувати функції розподілу в тексті складів, слів, словосполучень і граматичних класів;
- визначати імовірнісні характеристики норм мови;
- визначати тип породжувальних граматики за Н. Хомським;
- здійснювати виведення;
- на простих прикладах визначати мову, яка задається граматиною, і за мовою будувати граматику;
- будувати скінченний автомат, що сприймає задану регулярну мову;
- використовувати апарат скінченних автоматів для пошуку у тексті;
- давати лінгвістичну інтерпретацію основних теоретичних понять курсу.

Нижче наведено робочу програму курсу з коротким переліком тем і основних підтем.

Лекції (15 годин).

N п/п	Назва розділів, тем	Години
1	Квантитативна лінгвістика. Ймовірність та інформація лінгвістичних подій. Множина лінгвістичних об'єктів. Лінгвістична подія. Ймовірність елементарної лінгвістичної події. Ймовірності складних лінгвістичних подій. Інформаційні виміри в тексті.	2
2	Ймовірнісне моделювання породження тексту та його складових одиниць. Повторення незалежних випробувань у тексті. Випадкова лінгвістична величина, її характеристики і функції розподілу. Закони розподілу, які моделюють утворення мовних одиниць тексту. Поняття про застосування закону великих чисел у лінгвістиці.	2

N п/п	Назва розділів, тем	Години
3	Первинна статистична обробка тексту. Статистична сукупність лінгвістичних об'єктів та її організація. Варіаційні ряди лінгвістичних ознак. Дослідження лінгвістичних варіаційних рядів за допомогою емпіричних моментів.	2
4	Статистичні моделі тексту та ймовірнісні характеристики норми мови. Точечна оцінка параметрів генеральної лінгвістичної сукупності. Оцінка математичного сподівання за допомогою довірчого інтервалу та статистична параметризація стилів. Довірчі інтервали для ймовірності якісної лінгвістичної ознаки.	2
5	Комбінаторна лінгвістика. Породжувальні граматики. Основні означення, класифікація за Н. Хомським. Деревя виведення. Граматичний аналіз.	2
6	Скінченні автомати. Означення скінченного автомата (детермінованого і не детермінованого). Способи задання скінченних автоматів. Застосування скінченних автоматів для пошуку у тексті.	2
7	Розпізнавання мов. Регулярні множини. Теорема Кліні. Регулярні множини і регулярні граматики. Лема про накачування для регулярних мов. Лема про накачування для контекстно-вільних мов. Поняття про застосування контекстно-вільних мов.	3

Лабораторні роботи (45 годин).

N п/п	Назва розділів, тем	Години
1	Ймовірність та інформація лінгвістичних подій. Інформаційні виміри у тексті.	4
2	Ймовірнісне моделювання породження тексту та його складових одиниць. Закони розподілу, які моделюють утворення мовних одиниць тексту.	6
3	Первинна статистична обробка тексту. Статистична сукупність лінгвістичних об'єктів та її організація. Варіаційні ряди лінгвістичних ознак. Дослідження лінгвістичних варіаційних рядів за допомогою емпіричних моментів.	7
4	Статистичні моделі тексту та ймовірнісні характеристики норми мови.. Оцінка математичного сподівання за допомогою довірчого інтервалу та статистична параметризація стилів. Довірчий інтервал для дисперсії та середнього квадратичного відхилення. Довірчі інтервали для ймовірності якісної лінгвістичної ознаки.	8
5	Породжувальні граматики. Основні означення, класифікація за Н. Хомським. Деревя виведення. Граматичний аналіз.	4
6	Скінченні автомати (детерміновані і недетерміновані). Означення, способи завдання. Застосування скінченних автоматів для пошуку в тексті.	8
7	Розпізнавання мов. Регулярні множини. Теорема Кліні. Регулярні множини і регулярні граматики. Лема про накачування для регулярних мов. Лема про накачування для контекстно-вільних мов. Поняття про застосування контекстно вільних-мов.	8

Характеристика змісту курсу за темами. Коротко зупинимось на аналізі пропонованої тематики. Теми 1 – 4 репрезентують матеріал, що стосується квантитативної лінгвістики, а теми 5 – 7 відображають головні аспекти комбінаторної лінгвістики.

1. Основою всіх індуктивних досліджень у мовознавстві є спостереження за поведінкою і ознаками лінгвістичних об'єктів, що вивчаються. Це спостереження може здійснюватись також через експеримент або кількісне вимірювання. Здійснення кожного такого спостереження (досліді або вимірювання) називається *випробуванням*. Сукупність умов, за яких здійснюється дане випробування, називають *комплексом умов*. Результатом лінгвістичного випробування є *лінгвістична подія*.

Мірою можливості появи лінгвістичної події A при виконанні комплексу умов є ймовірність $P(A)$ цієї події. Для мовознавства важливими є три означення ймовірності: а) означення, яке ґрунтується на суб'єктивній кількісній оцінці можливості події; б) класичне означення ймовірності; в) статистичне означення ймовірності.

Розглядаються правила обчислення ймовірностей складних лінгвістичних подій, які широко використовуються для прогнозування подій у різного роду ймовірнісно-лінгвістичних, інженерно-лінгвістичних та інформаційних задачах.

2. При дослідженні механізмів породження тексту результати окремого лінгвістичного випробування не мають великого значення. Вивчення взаємодії системи, норми та ситуації експлікується за допомогою моделей теорії ймовірностей, які передбачають здійснення масового експерименту, при якому одна і та сама лінгвістична подія повторюється багато разів. Ці випробування, які повторюються, утворюють серії, в кожній з яких подія з'являється або не з'являється певну кількість разів.

Вибір тої чи іншої моделі опису тексту залежить від способу побудови імовірнісно-лінгвістичного випробування і, зокрема, від того, як організовано вибір з тексту окремих його одиниць. Квантитативне мовознавство широко використовує метод серійного спостереження. Суть його полягає в тому, що лінгвістичні одиниці вибираються з тексту групами фіксованої довжини: наприклад, по десять фонем, по сто речень або словоформ тощо. Лінгвістичні одиниці, які утворюють серію, обов'язково повинні міститись у тексті одна за одною, вони можуть вибиратись і через певний інтервал.

При розв'язуванні багатьох теоретичних та інженерно-лінгвістичних задач часто потрібно знати ймовірність появи тої чи іншої кількості певних лінгвістичних одиниць у серії. Якщо лінгвістичні випробування, які утворюють серію, розглядаються як незалежні, то ми можемо здійснювати необхідне прогнозування за допомогою розроблених у теорії ймовірностей трьох систем незалежних випробувань: простої, поліноміальної та пуасонівської.

При проведенні лінгвістичного досліді ми постійно зустрічаємось з випадковими величинами. У теорії ймовірностей відомі десятки законів розподілу випадкової величини. Розглядаються закони, які можуть вибиратись у ролі найбільш адекватних математичних моделей породження тексту і мовних одиниць, що його утворюють.

3. Дослідження тексту за допомогою імовірнісних моделей може бути здійснено за умови, що зроблено первинну статистичну обробку тексту і до її результатів застосовані спеціальні критерії переходу до імовірнісної моделі. Розглядаються методи організації статистичного досліді тексту, варіаційні ряди лінгвістичних ознак, їх графічне зображення та статистичні характеристики.

Кожний варіаційний ряд узагальнює результати статистичного експерименту над текстом обмеженої довжини, котрий виступає в ролі часткової вибірки, яку взято з деякої генеральної сукупності (мова в цілому, стиль, підмова, мова письменника). Звичайно

вибіркова текстова сукупність є цікавою тільки в тому, що вона може розглядатись як *модель*, яка відображає імовірнісні властивості досліджуваної генеральної сукупності текстів та норми відповідної мови.

4. Перехід від статистичної моделі вибірки тексту до імовірнісних характеристик норм мови пов'язаний з трьома задачами.

По-перше, за характеристиками θ^* варіаційного ряду необхідно числово оцінити приховані від прямого спостереження параметри θ відповідного розподілу генеральної сукупності, тобто параметри, які виступають у ролі імовірнісних характеристик норми мови.

По-друге, за даними варіаційного ряду потрібно оцінити характер генерального розподілу.

По-третє, за наявності числових оцінок параметрів генерального розподілу, визначити, який об'єм досліджуваного тексту дасть достатньо надійні результати

5. Слова у природній мові (українській, англійській тощо) можуть об'єднуватись у реченнях різними способами. Граматика відповідної мови вказує, як скомбінувати слова, щоб одержати правильне речення. У теорії формальних граматики концентрують увагу лише на *синтаксисі*, або формі речення і не беруть до уваги *семантику* або значення (смісл).

Синтаксис у природних мовах дуже складний. Не можливо точно формально визначити всі правила синтаксису природної мови. Дослідження в галузі перекладу з однієї мови на іншу привели до концепції *формальної мови*, яка, на відміну від природної, описується точно визначеною множиною правил синтаксису. Правила синтаксису є важливими не тільки в лінгвістиці, що вивчає природні мови, але також і при вивченні мов програмування. Ми взяли за основу таке означення (породжувальної) граматики.

Граматика фразових структур $G = (V, T, S, P)$ – це четвірка, де V – словник, $T \subset V$ – множина термінальних елементів, S – початковий (стартовий) символ та P – множина продукцій (правил виведення). Множина $V \setminus T$ позначається через N . Елементи множини N називаються нетермінальними символами, $S \in N$. Кожна продукція з P повинна містити принаймні один нетермінальний символ у своїй лівій частині.

Продукція (правило виведення) записується у формі $z_1 \rightarrow z_2$; це означає, що у довільному ланцюжку елементів з V ланцюжок z_1 може бути замінений на ланцюжок z_2 . Тут доречно відзначити, що цю заміну здійснювати не обов'язково.

Нехай $w_0 = lz_0r$ та $w_1 = lz_1r$ – ланцюжки з елементів V . Якщо $z_0 \rightarrow z_1$ є продукцією граматики, то кажуть, що w_1 *прямо виводиться* з w_0 і пишуть $w_0 \Rightarrow w_1$. Якщо $w_0, w_1, \dots, w_n, n \geq 1$ – ланцюжки з елементів V , такі, що $w_0 \Rightarrow w_1, w_1 \Rightarrow w_2, \dots, w_{n-1} \Rightarrow w_n$, то кажуть, що w_n виводиться з w_0 і пишуть $w_0 \Rightarrow w_n$. Послідовність кроків, які потрібні для одержання w_n з w_0 , називається *виведенням*.

Мовою $L(G)$, що породжується граматикою G , називається множина таких ланцюжків термінальних символів, які виводяться зі стартового символу S :

$$L(G) = \left\{ w \in T^* : S \Rightarrow w \right\},$$

де через T^* позначено множину всіх ланцюжків символів з T .

Граматика фразової структури класифікуються за типами продукцій, які дозволені. Ми розглядаємо класифікацію, запропоновану Н. Хомським. У таблиці через $|z|$ позначено довжину ланцюжка z , а через λ – порожній ланцюжок.

Типи граматик

Тип	Обмеження на продукції $z_1 \rightarrow z_2$
0	немає обмежень
1	$ z_1 \leq z_2 $, або $z_2 = \lambda$
2	$z_1 = A$, де A – нетермінальний символ
3	$z_1 = A$ та $z_2 = aB$ або $z_2 = a$, де $a \in T, A \in N, B \in N$, або продукція $S \rightarrow \lambda$

З наведеної таблиці випливає, що будь-яка граматика типу 3 є граматикою типу 2, будь-яка граматика типу 2 є граматикою типу 1 і будь-яка граматика типу 1 є граматикою типу 0 (але не навпаки).

Граматика типу 2 називається *контекстно-вільною* граматикою, оскільки нетермінальний символ A у лівій частині її продукції може бути замінений щоразу як він зустрічається у ланцюжку, незалежно від оточення. Мова, яка породжується граматикою типу 2, називається *контекстно-вільною* мовою.

Іншим (альтернативним по відношенню до даного в таблиці) означенням граматик типу 1 є те, що кожна продукція повинна мати вигляд $xAy \rightarrow xwy$, (тобто тут $z_1 = xAy$, $z_2 = xwy$) де $A \in N$, $x, y, w \in V^*$, $w \neq \lambda$, або $z_1 \rightarrow \lambda$. Еквівалентність цих двох означень неочевидна і доводиться окремо [6]. У цьому означенні ланцюжки x та y можуть розглядатись як контекст, у якому A може замінятись на w , тому граматика типу 1 називається *контекстно-залежною* граматикою. Мова, яка породжується граматикою типу 1, називається *контекстно-залежною* мовою.

Граматика типу 3 називаються *регулярними*. Мова, яка породжується регулярною граматикою, називається *регулярною* мовою.

Виведення в мові, яка породжена контекстно-вільною граматикою, може бути зображено графічно з використанням упорядкованого кореневого дерева, яке тут називається *деревом граматичного розбору* (*derivation*, або *parse*, *tree*). Корінь цього дерева відповідає стартовому символу. Внутрішні вершини репрезентують нетермінальні символи, які зустрічаються у виведенні. Якщо зустрічається продукція $A \rightarrow w$, де w – ланцюжок, вершина, яка репрезентує A , має синами вершини, які представляють кожний символ з w у порядку зліва направо.

Ми тут докладно навели потрібні означення, оскільки в літературі наявні певні відмінності у трактуванні основних понять [1-6, 9, 10].

6. Скінченні автомати у нашому курсі розглядаються як засіб представлення регулярних мов. Скінчений автомат має множину станів і „керування”, яке переводить його з одного стану в інший у залежності від „вхідних даних”. Класи автоматів істотно розрізняються за типом цього керування. Керування може бути *детермінованим* у тому

сміслі, що автомат може бути в кожний момент часу в одному стані, і *недетермінованим*, тобто автомат може бути одночасно у декількох станах або перехід не визначений. Виявляється, що додавання недетермінізму не дозволяє визначати мови, які б не можна було визначити за допомогою детермінованих скінченних автоматів. Разом з цим, недетерміновані автомати виявляються дуже ефективними у застосуваннях. Одне із практичних застосувань – пошук у тексті – розглядається як приклад. Розглядається також спосіб переходу від недетермінованого скінченного автомата до детермінованого, котрий після цього може бути виконаний на звичайному комп'ютері.

7. У темі „Розпізнавання мов”, зокрема, продовжується вивчення регулярних мов. Тут представлений ще один важливий спосіб їх опису за допомогою такої алгебраїчної нотації, як *регулярний вираз*. Регулярні вирази задають *регулярні множини*. Наступна теорема показує еквівалентність регулярних виразів і скінченних автоматів.

Теорема Кліні. Множина є регулярною тоді і тільки тоді, коли вона розпізнається скінченим автоматом.

Скінченні автомати і регулярні вирази можуть бути використані як засоби для доведення деяких важливих властивостей регулярних мов [9, 10].

У цій же темі розглядаються дві важливі теореми, які дають необхідні умови регулярності і контекстно вільності мови – це, відповідно, *лема про накачування для регулярних мов* і *лема про накачування для контекстно-вільних мов*. У вітчизняній літературі замість терміна „накачування” часто використовується термін „розростання”, проте перший, на нашу думку, краще відображає суть справи.

У заключній частині курсу коротко розглядаються деякі практичні застосування контекстно-вільних мов. Контекстно-вільні граматики були введені Н.Хомським як спосіб опису природних мов, але їх виявилось недостатньо. Проте зі збільшенням кількості прикладів використання понять, які визначаються рекурсивно, зростала і потреба у контекстно-вільних граматиках як способі опису прикладів таких понять.

1. Гладкий А.В. *Лекции по математической лингвистике*. Новосибирск. 1966.
2. Гладкий А.В. *Формальные грамматики и языки*. М.: Наука. 1973.
3. Гладкий А.В., Мельчук И.А. *Элементы математической лингвистики*. М.: Наука, 1969.
4. Гросс М., Липтен А. *Теория формальных грамматик*. М.: Мир, 1971.
5. Кузнецов О.П., Адельсон-Вельский Г.М. *Дискретная математика для инженера*. Гл. 7, 8. М.: Энергоатомиздат, 1988.
6. Кук Д., Бейз Г. *Компьютерная математика*. М.: Наука, 1990. Гл. 8, 9.
7. Пиотровский Р.Г. *Информационные измерения языка*. М.: Наука, 1968.
8. Пиотровский Р.Г., Бектаев К.Б., Пиотровская А.А. *Математическая лингвистика*. К.: Высшая школа, 1977.
9. Хопкрофт Д., Мотвани Р., Ульман Д. *Введение в теорию автоматов, языков и вычислений*. Вильямс, 2002.
10. K.Rosen. *Discrete Mathematics and its Applications*// McGraw-Hill. – 1988. – Р. 10.