

## ПЕРЕДМОВА

Мета цього курсу – ознайомити читача з деякими основними моделями і результатами, використовуваними в теоретичній інформатиці. Не дивно, що вони належать до математики, а не до будь-якої іншої галузі знань – адже в науці про комп'ютери саме математичні абстракції є найбільш плідними.

Розглянуті тут ідеї і результати належать до теорії формальних мов, граматики і автоматів. По суті, ця теорія описує деякі обмежені абстрактні машини, здатні виконувати певні операції з рядками. Наприклад, скінченний автомат може з'ясувати, чи містить деякий файл певне слово, а автомат з магазинною пам'яттю здатний визначити, чи правильна система вкладених круглих, квадратних і фігурних дужок.

Як підказує сама назва курсу, основним об'єктом розгляду є формальна мова – довільна множина скінченних послідовностей символів, взятих з деякої скінченної множини (такі послідовності називаються словами).

У першому розділі подано класифікацію формальних мов відповідно до ієрархії Хомського: автоматні мови, контекстно-вільні (або безконтекстні) мови, контекстні (або контекстно-залежні) мови, мови типу 0. Якщо виключити мови, що містять порожнє слово, то названі класи вкладені один в одного (тут їх перераховано за зростанням).

У розділах 2–3 розглянуто найвужчий з цих класів – клас автоматних (або праволінійних, або регулярних) мов, що володіє багатьма властивостями, важливими для програмних додатків. Саме автоматні мови покладено в основу лексичних аналізаторів, що входять до структури компіляторів і інтерпретаторів мов програмування. Лексичному аналізу присвячено розділ 4.

У розділах 5–6 розглянуто контекстно-вільні мови, різні їх форми представлення та способи еквівалентного перетворення з однієї форми на іншу. Контекстно-вільні мови використовують для визначення сучасних мов програмування (а також мов пошукових запитів тощо). Основну частину синтаксису описано в формалізмі, відомому як форма Бекуса–Наура. По суті, це спосіб записувати контекстно-вільні граматики. Інша важлива область використання контекстно-вільних граматики – розширювана мова розмітки (XML) і мови опису структури XML-документа (наприклад, мова DTD).

Два класи з ієрархії Хомського, що залишилися – контекстні мови і мови типу 0 – у цьому курсі вивчаються менш докладно, оскільки вони стосуються

швидше теорії складності обчислень і теорії алгоритмів. Також тут наведено еквівалентні визначення цих класів у термінах автоматів.

Розділи 7 та 9 містять основні результати теорії детермінованих контекстно-вільних мов і формальні визначення, що демонструють зв'язок цієї теорії зі низхідними та висхідними синтаксичними аналізаторами.

Говорячи “формальна мова”, ми будемо говорити насамперед про штучні мови, створені людьми для певних задач, наприклад, мови програмування. Однак між штучними та природними мовами фундаментальних розбіжностей не існує: природні мови часто характеризуються доволі складними граматичними правилами, що жорстко формалізують їх, тоді як в найдосконаліших штучних мовах іноді виникають такі непередбачувані аспекти поведінки, неоднозначне розуміння яких стає проблемою.

Вивчаючи мови, ми повинні мати на увазі такі три аспекти.

Перший аспект – *синтаксис* мови. Мова – це деяка множина “слів”, де слово – це скінченна послідовність “букв” – символів деякого зафіксованого алфавіту. Терміни “слово”, “буква” можна визначати по-різному. Так, “буквами” можуть бути окремі символи української мови чи паскаль-програми, а “слова” – відповідно їхні скінченні послідовності, наприклад, “крокодил” – в українській мові, “integer” – в паскалі. Такі слова називають лексемами. Водночас “буквами” можна вважати окремі слова, тоді “словами» стають речення української мови чи програми мовою паскаль, наприклад. Якщо зафіксована послідовність “букв”, то не кожне таке слово буде лексемою мови, наприклад, “кракодил” не є лексемою української мови, а iff не є лексемою мови паскаль. Тому в мовах існує множина правил правильної побудови лексем мови, яка називається *синтаксисом* мови (слово синтаксис походить від давньогрецького *syn* – разом та *taxis* – порядок, тобто “порядок складання разом”). Оскільки кожне “слово” характеризується певною структурою, специфічною саме для цієї мови, то необхідно розробити механізми перерахування або породження “слів”, а з іншого боку, розробити механізми, які б допомагали визначити, чи є “слово” лексемою, тобто чи належить воно цій мові. Саме механізми породження та розпізнавання “слів” і вивчає класична теорія формальних мов.

Другий аспект – *семантика* мови. Семантика передбачає зіставлення із “словом” певного значення, змісту (від давньогрецького *sema* – знак, позначення та *semanticos* – означає). Наприклад, записуючи математичну

формулу, ми повинні дотримуватись певних синтаксичних правил (розташування дужок та знаків операцій тощо), але, крім цього, формула має визначений для нас зміст.

Мова – це засіб спілкування та передавання інформації. Для того, щоб нас розуміли, слова мови повинні бути правильно побудовані та нести потрібний зміст – мати правильний синтаксис та семантику. Додамо, що і для опису семантики мови розроблено формальні системи, проте семантику формалізувати значно важче, ніж синтаксис.

Третій аспект – *прагматика* мови. Прагматика (від давньогрецького *pragma* –діло, дія та *pragmateia* – діяльність) пов'язана з цілями носія мови, наприклад, отримання грошей за виступ. Прагматика пов'язана з соціально-психологічними аспектами мови та не входить до розгляду теорії, викладеної у цьому посібнику.

Основними поняттями, що розглядаються у посібнику, є:

- формальні мови та породжуючі граматики;
- скінченні автомати та автомати з магазинною пам'яттю, їх роль у синтезі та аналізі мови;
- класи граматик і мов, алгоритми синтезу та аналізу для різних класів мов;
- процес трансляції формальних мов.