Фізико-технічний інститут *кафедра ММЗІ*

Теорія складності **Осінь 2021**

Матеріал 1: Теорія формальних мов

1.2 Теорія формальних мов

Означення 1.1. *Множиною натуральних чисел* $\mathbb{N}$ називають множину невід'ємних цілих чисел $\{0, 1, 2, \ldots\}$.

Зауваження. Іноді множина натуральних чисел визначається без значення 0, але сучасний підхід використовує саме означення 1.1. Наприклад, як в офіційному міжнародному стандарті ISO 80000-2:2019, який визначає математичні позначення та символи. З метою уникнення непорозумінь використовують також додаткові позначення: $\mathbb{N}^0 = \mathbb{N}_0 = \mathbb{N} = \{0, 1, \ldots\}$ і $\mathbb{N}^+ = \mathbb{N}_1 = \mathbb{N}_{>0} = \mathbb{Z}^+ = \{1, 2, \ldots\}$.

Теорія формальних мов є розділом математики, який вивчає формальні мови, методи їх визначення, операції з ними, їх класифікацію та аналіз. Теорія формальних мов дуже часто є складовою частиною теорії автоматів, оскільки автомати є одним з основних інструментів роботи з формальними мовами, але ця теорія є однією з базових, які визначають фундамент математичних дисциплін, і також часто використовується іншими теоріями — наприклад, теорією обчислюваності, теорією алгоритмів та математичною логікою.

Означення 1.2. *Алфавітом* формальної мови називають довільну скінченну або нескінченну непорожню множину елементів, а елементи цієї множини називають *символами*.

Зауваження. Зауважимо, що означення алфавіту формальної мови наводиться дещо по різному в різних джерелах. Сучасний підхід вимагає, щоб алфавіт формальної мови був непорожньою та скінченною множиною, навіть якщо про це не було явно сказано, оскільки саме такі алфавіти використовуються в абсолютній більшості випадків. Але все ж таки в деяких випадках розглядаються нескінченні алфавіти, які можуть бути і незліченими множинами. Тому будемо вважати означенням алфавіту формальної мови означення 1.2, як найбільш загальне, але далі в межах цього курсу, якщо явно не сказано іншого, будемо розглядати виключно непорожні та скінченні алфавіти формальних мов, щоб уникнути в формулюваннях постійної вимоги скінченності алфавіту.

Зазвичай, алфавіт позначають літерою Σ, але можуть використовуватися й інші позначення.

Означення 1.3. *Словом* чи *скінченним словом* (*над алфавітом* Σ) називають довільну скінченну послідовність (рядок) символів (алфавіту Σ). Кількість символів алфавіту, які утворюють слово x, називають *довжиною* слова x і позначають як $|x|$. Кількість входжень довільного символу алфавіту $a \in \Sigma$ в слово x позначають як $|x|_a$ (іноді для цього використовують позначення $\#_a(x)$). Рядок, який не містить жодного символу алфавіту Σ, також є словом над алфавітом Σ, яке називають *порожнім словом*, позначають літерою ε (за умови, що символ ε не належить алфавіту Σ) та яке має довжину 0. За допомогою Σ^n, $n \in \mathbb{N}$, позначають множину всіх слів над алфавітом Σ довжини n. За допомогою Σ^* позначають множину всіх слів над алфавітом Σ, $\Sigma^* = \bigcup\limits_{n \in \mathbb{N}} \Sigma^n$. Через Σ^+ позначають множину всіх непорожніх слів над алфавітом Σ, $\Sigma^+ = \Sigma^* \setminus \{\varepsilon\}$.

***Зауваження*.** Фактично слово над алфавітом є впорядкованим набором символів алфавіту. Порожнє слово ε є словом над довільним алфавітом та є єдиним словом, довжина якого дорівнює 0. Довжина слова визначається через кількість символів алфавіту, але мається на увазі кількість всіх символів, що утворюють слово. Наприклад, слово 1011 над алфавітом $\{0, 1\}$ містить лише два різних символи алфавіту: 0 та 1, але загальна кількість символів алфавіту, які утворюють слово 1011 дорівнює 4. Отже, довжина слова 1011 над алфавітом $\{0, 1\}$ дорівнює 4. Більш того, довжина слова залежить від алфавіту, над яким визначено слово. Так, довжина слова 1011 над алфавітом $\{00,\ 01,\ 10,\ 11\}$ дорівнює 2.

Останній приклад також показує, що символи алфавіту в загальному випадку можуть бути достатньо різними елементами. Наприклад, розглянемо алфавіт $\{0, 1, 101, 010, 1010101\}$. Слово 1010101 над цим алфавітом представляється за допомогою різних впорядкованих наборів символів — наприклад, за допомогою одного символу 1010101; трьох символів 0 та чотирьох символів 1; два представлення за допомогою символів 1, 101 і 010 та інші. В такому випадку навіть довжина слова 1010101 є невизначеною. Подібні випадки розглядаються в спеціальному розділі абстрактної алгебри, а далі, якщо явно не сказано іншого, будемо розглядати виключно алфавіти, над якими кожне слово має унікальне представлення символами алфавіту (тобто, для алфавіту виконується умова унікальності представлення кожного слова).

Приклад 1.1. Розглянемо декілька прикладів алфавітів.

а) За допомогою алфавіту $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ можна побудувати будь-яке натуральне число.

б) Одним з найбільш вживаних алфавітів є алфавіт $\{0, 1\}$, який називають *двійковим алфавітом* (англ. binary alphabet), а слова над двійковим алфавітом називають *двійковими рядками* (англ. binary string). За допомогою слова над двійковим алфавітом можна представити довільний математичний об'єкт, використовуючи певну схему кодування.

в) Прикладом нескінченного алфавіту є множина всіх натуральних чисел, а словами над цим алфавітом також є натуральні числа, можливо, з провідними нулями в записі. Для цього алфавіту умова унікальності представлення кожного слова не виконується.

Вправа 1.1. Наведіть приклад нескінченного алфавіту, для якого виконується умова унікальності представлення кожного слова.

Окрім скінченних послідовностей символів алфавіту іноді розглядають також нескінченні послідовностей символів.

Означення 1.4. Довільну нескінченну послідовність (рядок) символів (алфавіту Σ) називають *ω-словом* чи *нескінченним словом* (*над алфавітом* Σ). За допомогою Σ^ω позначають множину всіх нескінченних слів над алфавітом Σ. За допомогою Σ^∞ позначають множину всіх скінченних та нескінченних слів над алфавітом Σ, $\Sigma^\infty = \Sigma^* \cup \Sigma^\omega$.

Далі в абсолютній більшості випадків будемо розглядати тільки слова, які є скінченними послідовностями символів, якщо явно не буде сказано про використання ω-слів чи нескінченних слів.

Над словами визначають такі операції.

- *Обернення* слова — це унарна операція, результатом застосування якої до слова $x = \alpha_1\alpha_2\ldots\alpha_n$ над алфавітом Σ, де $n \in \mathbb{N}$ і $\alpha_1, \alpha_2, \ldots, \alpha_n \in \Sigma$, є слово над тим же алфавітом Σ, яке позначають як x^R та визначають рівністю $x^R = \alpha_n\ldots\alpha_2\alpha_1$. Якщо результатом обернення слова є те саме слово, то таке слово називають *паліндромом*.
- *Конкатенація* слів — це бінарна операція, визначена для слів над одним алфавітом. Результатом *конкатенації* слів $x = \alpha_1\alpha_2\ldots\alpha_n$ та $y = \beta_1\beta_2\ldots\beta_m$ над алфавітом Σ, де $n, m \in \mathbb{N}$ та $\alpha_1, \alpha_2, \ldots, \alpha_n$, $\beta_1, \beta_2, \ldots, \beta_m \in \Sigma$, називають слово над тим же алфавітом Σ, яке визначається рівністю $\alpha_1\alpha_2\ldots\alpha_n\beta_1\beta_2\ldots\beta_m$, і яке позначають як $x||y$, а іноді як $x \cdot y$ або xy, якщо це не суперечить іншим позначенням.
- *Кратна конкатенація* слова або *піднесення до степеня* n, де $n \in \mathbb{N}, n > 1$, — це унарна операція, результатом застосування якої до слова x є слово $x^n = \underbrace{x\ldots x}_{n \text{ разів}}$, при чому $x^1 = x$, $x^0 = \varepsilon$ для довільного слова x і $\varepsilon^k = \varepsilon$ для довільного натурального числа $k \in \mathbb{N}$.
- *Ліве ділення* слів — це часткова бінарна операція, визначена для певних слів над одним алфавітом. Операція лівого ділення слів над алфавітом Σ визначена тільки для тих слів x та y над алфавітом Σ, для яких існує таке слово z над алфавітом Σ, що $x = yz$. Тоді результатом лівого ділення слова x на слово y є слово z над алфавітом Σ, яке називають *лівою часткою від ділення* слова x на слово y, і яке позначають як $y\backslash x$.

- *Праве ділення* слів — це часткова бінарна операція, визначена для певних слів над одним алфавітом. Операція правого ділення слів над алфавітом Σ визначена тільки для тих слів x та y над алфавітом Σ, для яких існує таке слово z над алфавітом Σ, що $x = zy$. Тоді результатом правого ділення слова x на слово y є слово z над алфавітом Σ, яке називають *правою часткою від ділення* слова x на слово y, і яке позначають як y/x.

Наслідок. *1*) порожнє слово ε є нейтральним елементом відносно операції конкатенації слів, тобто $\varepsilon x = x\varepsilon = x$ для довільного слова x, зокрема, $\varepsilon\varepsilon = \varepsilon$;

2) операція конкатенації слів не є комутативною в загальному випадку;

3) конкатенація слів є асоціативною, тому дужки при багаторазовому використанні не потрібні;

4) оскільки слово може складатися з одного символу алфавіту, то операції обернення, конкатенації та кратної конкатенації є застосовними до символів алфавіту, зокрема будь-яке непорожнє слово є символом алфавіту або результатом конкатенації символів алфавіту;

5) множина всіх непорожніх слів утворює напівгрупу відносно операції конкатенації;

6) множина всіх слів утворює моноїд відносно операції конкатенації (з нейтральним елементом ε);

7) довжина результату конкатенації слів дорівнює сумі довжин операндів;

8) операції лівого та правого ділення слів є в певному розумінні оберненими операціями до операції конкатенації слів, але результат застосування цих операцій не збігається з конкатенацією з результатом обернення слова.

Приклад 1.2. Слово 101010 над двійковим алфавітом є результатом застосування операції конкатенації до слів 101 та 010. Результатом обернення слова 101010 є слово 010101, а $101\backslash 101010$ =010 і $010/101010$ =101. Операція лівого ділення слів є незастосовною до слів 101010 та 010, а операція правого ділення слів є незастосовною до слів 101010 та 101.

Означення 1.5.

- Слово $x \in \Sigma^*$ називають *префіксом* (англ. prefix) слова $y \in \Sigma^*$, якщо існує слово z над алфавітом Σ таке, що $y = xz$. Якщо при цьому $x \neq y$, то такий префікс x називають *власним префіксом* (англ. proper prefix) слова y (в деяких джерелах додатково вимагається, щоб власний префікс також був непорожнім словом).
- Слово $x \in \Sigma^*$ називають *суфіксом* (англ. suffix) слова $y \in \Sigma^*$, якщо існує слово z над алфавітом Σ таке, що $y = zx$. Якщо при цьому $x \neq y$, то такий суфікс x називають *власним суфіксом* (англ. proper suffix) слова y (в деяких джерелах додатково вимагається, щоб власний суфікс також був непорожнім словом).
- Слово $x \in \Sigma^*$ називають *підсловом* (англ. substring) слова $y \in \Sigma^*$, якщо існують слова z та w над алфавітом Σ такі, що $y = zxw$. Якщо при

цьому $x \neq y$, то таке підслово x називають *власним підсловом* (англ. proper substring) слова y. Для довільних слів $x, y \in \Sigma^*$ кількість різних входжень, можливо з перекриттям, слова x у слово y позначають як $|y|_x$ або $\#_x(y)$. Для довільного слова $x = \alpha_1\alpha_2 \ldots \alpha_n$ над алфавітом Σ, де $n \in \mathbb{N}$ і $\alpha_1, \alpha_2, \ldots, \alpha_n \in \Sigma$, через $x[i]$, $i \in \mathbb{N}$, $1 \leq i \leq n$, позначають символ α_i, а через $x[i..j]$, $i, j \in \mathbb{N}$, $1 \leq i < j \leq n$, позначають підслово $\alpha_i\alpha_{i+1} \ldots \alpha_j$.

- Слово $x \in \Sigma^*$ називають *підпослідовністю* (англ. subsequence) слова $y \in \Sigma^*$, якщо $y = \alpha_1\alpha_2 \ldots \alpha_n$, де $n \in \mathbb{N}$ і $\alpha_1, \alpha_2, \ldots, \alpha_n \in \Sigma$, а $x = \alpha_{k_1}\alpha_{k_2} \ldots \alpha_{k_m}$, де $m \in \mathbb{N}$, $m \leq n$, $k_1, k_2, \ldots, k_m \in \mathbb{N}$, $1 \leq k_1 < k_2 < \cdots < k_m \leq n$, позначають це як $x \preccurlyeq y$.
- Слово $x \in \Sigma^*$ називають *межею* (англ. border) слова $y \in \Sigma^*$, якщо $x \neq y$ і слово x є одночасно і префіксом, і суфіксом слова y, при цьому слово y називають *x-обмеженим*. Якщо додатково виконується умова, що $|x| \leq \frac{|y|}{2}$, то слово y називають *неперетинно x-обмеженим*, а, якщо $|x| > \frac{|y|}{2}$, то слово y називають *перетинно x-обмеженим*.

Наслідок. *1*) Якщо слово $x \in \Sigma^*$ є підсловом слова $y \in \Sigma^*$, то воно також є підпослідовністю слова y.

2) Бінарні відношення на множині слів, які визначають, що перше слово є підсловом другого слова і перше слово є підпослідовністю другого слова відповідно, є передпорядками.

Приклад 1.3. *а*) Слова ε, 1, 1011 є префіксами, а слова ε, 1, 01 є суфіксами слова 10111101 над двійковим алфавітом.

б) Всі різні підслова слова 0110 над двійковим алфавітом складають множину $\{\varepsilon, 0, 1, 01, 10, 11, 011, 110, 0110\}$.

в) Фактично утворення підпослідовності слова можна розглядати як викреслювання деякої кількості символів слова, можливо нульової, без змін порядку символів, які залишаються. Всі різні підпослідовності слова 0110 над двійковим алфавітом складають множину $\{\varepsilon, 0, 1, 00, 01, 10, 11, 010, 011, 110, 0110\}$.

г) Слово 101 є межею слів 10101 і 10111101 над двійковим алфавітом.

Означення 1.6. *Формальною мовою* (або *мовою*) над скінченним алфавітом називають довільну підмножину множини всіх слів над алфавітом. Для довільного скінченного алфавіту Σ множину, яка не містить жодного слова, називають *порожньою мовою* і позначають як $\varnothing$. Для довільного скінченного алфавіту Σ мови $\varnothing$ і Σ^* називають *тривіальними*, а всі інші мови називають *нетривіальними*.

Зазвичай мови задають

- за допомогою наведення всіх слів, які входять до цієї мови;
- використовуючи певні властивості, які мають всі слова з цієї мови, при цьому таких властивостей не має жодне слово, яке не входить до мови;

- за допомогою граматик.

Оскільки мови є множинами слів, то над мовами, як над множинами, визначені будь-які теоретико-множинні операції: об'єднання, перетину, доповнення, різниці та теоретико-множинні відношення рівності множин та включення.

Приклад 1.4. Для двійкового алфавіту мови $\{01, 10\}$ і $\{\varepsilon, 01, 10\}$ є різними, як і мови $\varnothing$ і $\{\varepsilon\}$.

Означення 1.7. Нехай L_1 та L_2 є мовами над довільним алфавітом Σ, тобто $L_1, L_2 \in \Sigma^*$.

- *Перетином* мов L_1 та L_2 над алфавітом Σ називають бінарну операцію, результатом якої є мова над алфавітом Σ, що складається з множини слів $\{x \in \Sigma^* \mid x \in L_1, x \in L_2\}$, і яку позначають як $L_1 \cap L_2$.

- *Об'єднанням* мов L_1 та L_2 над алфавітом Σ називають бінарну операцію, результатом якої є мова над алфавітом Σ, що складається з множини слів $\{x \in \Sigma^* \mid x \in L_1 \text{ або } x \in L_2\}$, і яку позначають як $L_1 \cup L_2$.

- *Різницею* мов L_1 та L_2 над алфавітом Σ називають бінарну операцію, результатом якої є мова над алфавітом Σ, що складається з множини слів $\{x \in \Sigma^* \mid x \in L_1 \text{ і } x \notin L_2\}$, і яку позначають як $L_1 \setminus L_2$.

- *Конкатенацією* мов L_1 та L_2 над алфавітом Σ називають бінарну операцію, результатом якої є мова над алфавітом Σ, що складається з множини слів $\{xy \in \Sigma^* \mid x \in L_1, y \in L_2\}$, і яку позначають як L_1L_2.

- *Кратною конкатенацією* мови L_1 над алфавітом Σ або *піднесенням до степеня* n, де $n \in \mathbb{N}$, $n > 1$, називають унарну операцію, результатом застосування якої є мова над алфавітом Σ, що складається з множини слів $\{x_1 \ldots x_n \in \Sigma^* \mid x_1, \ldots, x_n \in L_1\}$, при чому $L_1^1 = L_1$, $L_1^0 = \{\varepsilon\}$ для довільної мови L_1.

- *Доповненням* мови L_1 над алфавітом Σ називають унарну операцію, результатом якої є мова над алфавітом Σ, що складається з множини слів $\{x \in \Sigma^* \mid x \notin L_1\}$, і яку позначають як $\overline{L_1}$.

- *Оберненням* мови L_1 над алфавітом Σ називають унарну операцію, результатом якої є мова над алфавітом Σ, що складається з множини слів $\{x \in \Sigma^* \mid x^R \in L_1\}$, і яку позначають як L_1^R.

- *Замиканням Кліні (англ. Kleene)* (або *зіркою Кліні*) мови L_1 над алфавітом Σ називають унарну операцію, результатом якої є мова над алфавітом Σ, що складається з множини слів $\{x \in \Sigma^* \mid \text{існує таке натуральне число } n \in \mathbb{N}_0\text{, що } x \in L_1^n\}$, і яку позначають як L_1^*.

- *Додатним замиканням Кліні* (або *плюсом Кліні*) мови L_1 над алфавітом Σ називають унарну операцію, результатом якої є мова над алфавітом Σ, що складається з множини слів $\{x \in \Sigma^* \mid$ існує таке число $n \in \mathbb{N}_1$, що $x \in L_1^n\}$, і яку позначають як L_1^+.

- *Діленням* мов L_1 та L_2 над алфавітом Σ називають бінарну операцію, результатом якої є мова над алфавітом Σ, яку називають *часткою* і яка складається з множини слів $\{x \in \Sigma^* \mid$ існує таке слово $y \in L_2$, що $xy \in L_1\}$, і яку позначають як L_1/L_2.

- *Змішуванням* мов L_1 та L_2 над алфавітом Σ називають бінарну операцію, результатом якої є мова над алфавітом Σ, яка складається з множини слів $\{x \in \Sigma^* \mid$ існують $n \in \mathbb{N}_1$ та $y_1, \ldots, y_n \in L_1$, $z_1, \ldots, z_n \in L_2$ такі, що $x = y_1 z_1 y_2 z_2 \ldots y_n z_n\}$.

Операції замикання Кліні, об'єднання мов та конкатенації мов називають *регулярними операціями* над мовами.

Наслідок. Для довільної мови L_1 виконуються тотожності $L_1^* = \bigcup_{n \in \mathbb{N}_0} L_1^n$, $L_1^+ = \bigcup_{n \in \mathbb{N}_1} L_1^n$ і $L_1^* = L_1^+ \cup \{\varepsilon\}$.

Зауваження. З метою скорочення запису операцій над мовами дозволяється для множин, які містять лише одно слово, опускати дужки { і }. Таким чином через запис $\{0,1\}^* 0 \{0,1\}^*$ (який позначає результат застосування операції конкатенації до трьох мов) можна позначити мову або множину всіх слів над двійковим алфавітом, які містять хоча б один символ 0.

Означення 1.8. Впорядкований набір слів $(x_1, \ldots, x_n)$, $n \in \mathbb{N}_1$, називають *факторизацією* слова x, якщо $x = x_1 \ldots x_n$, і позначають це як $x = (x_1, \ldots, x_n)$. Для мови L_1 впорядкований набір слів $(x_1, \ldots, x_n)$, $n \in \mathbb{N}_1$, називають L_1-*факторизацією* слова x, якщо $x = x_1 \ldots x_n$ і всі слова $x_1, \ldots, x_n$ належать мові L_1. L_1-факторизацію множини слів називають *повною*, якщо кожне слово цієї множини має L_1-факторизацію. L_1-факторизацію множини слів називають *унікальною*, якщо кожне слово цієї множини має щонайбільше одну L_1-факторизацію.

Означення 1.9. *Гомоморфізмом* слів над алфавітом Σ_1 в слова над алфавітом Σ_2 називають таке відображення $h : \Sigma_1^* \to \Sigma_2^*$, що для довільних слів $x, y \in \Sigma_1^*$ має місце тотожність $h(xy) = h(x)h(y)$.

Наслідок. Для довільного гомоморфізму $h : \Sigma_1^* \to \Sigma_2^*$ виконується рівність $h(\varepsilon) = \varepsilon$ і будь-який результат відображення можна обчислити, знаючи значення відображення h на всіх символах алфавіту Σ_1.

Приклад 1.5. Гомоморфізмом є відображення $h : \{a, b\}^* \to \{0, 1\}^*$, визначене за допомогою значень $h(a) = 010$ і $h(b) = 11$. Так, наприклад, $h(aabba) = 0100101111010$.

Для довільного гомоморфізму $h : \Sigma_1^* \to \Sigma_2^*$ і мови $L_1 \subseteq \Sigma_1^*$ через $h(L_1)$ позначають мову $\{h(x) \mid x \in L_1\} \subseteq \Sigma_2^*$. Також довільний гомоморфізм $h : \Sigma^* \to \Sigma^*$ визначає унарну операцію над мовами над алфавітом Σ. Для довільного гомоморфізму $h : \Sigma_1^* \to \Sigma_2^*$ і мови $L_2 \subseteq \Sigma_2^*$ через $h^{-1}(L_2)$ позначають мову $\{x \mid h(x) \in L_2\} \subseteq \Sigma_1^*$.