1.7. ФОРМАЛНИ ЕЗИЦИ. ФОРМАЛНИ ГРАМАТИКИ

1. Формални езици - определение.

Формален език над азбуката Σ е всяко подмножество L на Σ* ($L \subseteq \Sigma^*$). Множеството Σ* е език над Σ. ∅ е език над азбуката Σ, който се нарича празен език. Празният език се различава от езика $L = \{\Lambda\}$, който се състои от празната дума Λ.

Сечение (обединение) на езиците L_1 и L_2 над азбуката Σ е език $L = L_1 \cap L_2$ ($L = L_1 \cup L_2$), където L_1 и L_2 са множество от думи над азбуката Σ.

Произведение на езиците L_1 и L_2 е език

$$L = L_1 L_2 = \{u \in \Sigma^* \mid u = vw, v \in L_1, w \in L_2\},$$

т.е. множество от всички думи, което се представя като конкатенация между думите на L_1 и L_2.

Свойства: $L_1 L_2 \neq L_2 L_1$; $L_1(L_2 L_3) = (L_1 L_2) L_3$.

n-та степен на език: $L^n = \underbrace{LL...L}_{n-\text{пъти}}$, като $L^0 = \{\varepsilon\}$.

Итерация на L: $L^* = \bigcup_{n \geq 0} L^n$;

Положителна итерация на L: $L^+ = \bigcup_{n > 0} L^n$.

Пример:

Над азбука Σ = {0, 1, 2, *a*, *b*} се дефинират езиците $L_1 = \{\varepsilon, 01, a1\}$ и $L_1 = \{\varepsilon, 02, 1b\}$, тогава

$L_1 \cap L_2 = \{\varepsilon\}$;

$L_1 \cup L_2 = \{\varepsilon, 01, a1, 02, 1b\}$;

$L_1 L_2 = \{\varepsilon, 01, a1, 02, 1b, 0102, 011b, a102, a11b\}$.

Пример:

Азбука: $\Sigma = \{a, b, c\}$:

$L_1 = \left\{ \underbrace{\Lambda, a, ab, aab, bc, abc}_{6\,\text{думи}} \right\}$ - краен език (броят на елементите е краен);

$L_2 = \{a^n | n \geq 0\}$ - безкраен език (броят на елементите е безкраен).

Ако $L_1 \subseteq L_2$ то

- L_1 е подезик на L_2 (езикова съвместимост нагоре);

- L_2 е подезик на L_1 (езикова съвместимост надолу);
- ако думите в двата езика съвпадат, т.е. множествата от низове съвпадат: ($L_1 = L_2$), то съществува езикова съвместимост.

Описание на езиците:

- чрез изреждане (при малка мощност на множеството от думи (езика) $|L|$);
- чрез правила на образуване (при голяма мощност на множеството от думи (езика) $|L|$ или при безкрайни езици).

За всички езици броят на правилата, по които се извършва описанието на елементите на езика е краен.

Формалните езиците се описват с генерираща формална граматика.

2. Формална пораждаща граматика - определение

Формалната пораждаща граматика е наредена четворка

$$G = \{\Sigma, N, S, P\},$$

където Σ е крайното множество (азбука) от *терминални* символи ($\Sigma \in L$ - запазени думи или символи над езика L); N – крайното (непразно) множество (азбука) от *нетерминалните символи* ($N \notin L$ - символи, подлежащи на дефиниране чрез синтаксиса на езика), наречени *променливи*; S – началният символ (променлива) на граматиката ($S \in N$); P – крайно множество от *правила* от вида $\alpha \rightarrow \beta$ (продукции) за пораждане на елементи (думи от езика) - наредени двойки ($\alpha\beta$), за които $\alpha, \beta \in (\Sigma \cup N)^*$.

Свойства:

-$\Sigma \cap N = \varnothing$, т.е. Σ и N нямат общи символи;

-$\Sigma \cup N = D$ - обединението от множествата от терминални и нетерминални символи - речникът на езика;

-$\alpha \in (D)^+ = (\Sigma \cup N)^+ = (\Sigma \cup N)^* \backslash \{\Lambda\}$ - множеството от всички думи над D без празния низ (α има поне един нетерминален символ);

-$\beta \in (\Sigma \cup N)^*$ - множеството от всички думи (низове), образувани над азбуката. Следователно липсват продукции от вида $\{\Lambda\} \rightarrow \beta$ и $\alpha \rightarrow \{\Lambda\}$.

За група от правила $\alpha \rightarrow \beta$, $\alpha \rightarrow \gamma$, $\alpha \rightarrow \delta$,...$\alpha \rightarrow k$ се използва означението $\alpha \rightarrow \beta | \gamma | \delta | ... | k$.

Формалната граматика G генерира дума $w \in \Sigma^*$ над езика L, ако съществува крайна редица от думи:

$$w_0 \rightarrow w_1 \rightarrow w_2 \rightarrow ... \rightarrow w_n | n \geq 1.$$

Този израз дефинира *изводът* на думата $w_n = w$ от началния символ $w_0 = S$. Ако в P има правило $\alpha \to \beta$, то думата $w_{i+1} = \alpha_1 \beta \alpha_2$ се получава от всяка поддума от вида $w_i = \alpha_1 \alpha \alpha_2$, като α се замени с β. Броят на непосредствени изводи $w_i \to w_{i+1}$ определя *дължината на извода* на $w_1 \to w_n$.

Множеството от всички думи w над Σ, за които $S \to w$, се нарича формален език $L(G)$ над Σ, породен от граматиката G. Граматиките G_1 и G_2 са еквивалентни, ако $L(G_1) = L(G_2)$.

Пример:

1. Нека $G_1 = \{\Sigma, N, S, P\}$ е пораждаща граматика, където $\Sigma = \{a, b,\}$, $N = \{S, A, B\}$, S – началният символ; P се състои от правилата:
$S \to A \mid B$;
$A \to aA \mid a$;
$B \to bB \mid b$.

G_1 поражда всички думи от вида $\alpha = a^n, n \geq 1$ или $\beta = b^n, n \geq 1$ от формалния език $L(G_1)$ след прилагане един път на правило $S \to A \mid B$ и $n-1$ пъти правило $A \to aA \mid a$ или $B \to bB \mid b$, т.е.

$$L(G_1) = \{a^n \mid n \geq 1\} \cup L(G_1) = \{b^n \mid n \geq 1\}.$$

2. Нека $G_2 = \{\Sigma, N, S, P\}$ е пораждаща граматика, където $\Sigma = \{a, b,\}$, $N = \{S, A, B\}$, S – началният символ; P се състои от правилото:
$S \to aA \mid bB \mid a \mid b$, $A \to aA \mid a$, $B \to bB \mid bI$.

G_2 поражда всички думи от вида $\alpha = a^n, n \geq 1$ или $\beta = b^n, n \geq 1$, т.е.

$$L(G_2) = \{a^n \mid n \geq 1\} \cup L(G_2) = \{b^n \mid n \geq 1\}.$$

Следователно G_1 и G_2 са еквивалентни, т.е $L(G_1) = L(G_2)$.

3. Нека $G_3 = \{\Sigma, N, S, P\}$ е пораждаща граматика, където $\Sigma = \{a, b,\}$, $N = \{S\}$, S – началният символ; P се състои от правилата:
$S \to aSb \mid ab$.
G_3 генерира думата $a^3 b^3$ в съответствие с извода:

$$S \to aSb \to aaSbb \to aaabbb,$$

където два пъти се прилага правилото $S \to aSb$ и един път правилото $S \to ab$.

В общия случай граматиката G_3 генерира езика

$$L(G_3) = \{a^n b^n | n \geq 1\}.$$

3. Класификация на пораждащите граматики по Чомски

В съответствие със структурата на правилата граматиките се подразделят на:

- Граматика – **общ вид – тип 0**:

Правило: $\alpha A\beta \to w$, $\alpha, \beta, w \in (\Sigma \cup N)^*$, $A \in N$.

- Граматика – к**онтекстна – тип 1:**

Правило: $\alpha A\beta \to \alpha w\beta$, $w \neq \varepsilon$, $\alpha, \beta, w \in (\Sigma \cup N)^*$, $A \in N$: A може да се замести с w само ако съществува контекст α - β, където α е ляв контекст, а β десен контекст на A.

Тъй като $|\alpha w\beta| \geq 1$, то празната дума не се поражда в тази граматика. При необходимост се добавя правилото $S \to \varepsilon$, като получената граматика отново е контекстна. Правилото $S \to \varepsilon$ се прилага еднократно и не влияе върху останалите думи, различни от Λ и породени от граматиката.

Пример:

Граматиката $G_4 = \langle \{a, b, c\}, \{S, A, C\}, S, P \rangle$ с множество от правила от вида $P = \{S \to aSAC | abc, S \to abC, CA \to BA, BA \to BC, BC \to AC, bA \to bb, C \to c\}$ е контекстна и поражда език

$$L(G_4) = \{a^n b^n c^n | n \geq 1\}$$

след прилагане $n-1$ пъти първото правило, един път - второто правило, третото, четвъртото и петото правило в посочената последователност, n-1 пъти правила – шестото правило, n пъти седмото правило.

Същият език се генерира от G_4 и с множеството от правила от вида
$P = \{S \to aSAC | abc, cA \to Ac, bA \to bb, C \to c\}$

- Граматика – **безконтекстна – тип 2:**

Правило: $A \to w$, $w \in (\Sigma \cup N)^+$, $A \in N$. Добавя се правилото $S \to \varepsilon$ за пораждане на празната дума Λ.

Пример:

Граматиката $G_5 = \langle \{a, b\}, \{S, A, B\}, S, P \rangle$ с множество от правила от вида

$P = \{S \to AB,\ A \to aA|a,\ B \to bB|b\}$ е безконтекстна и поражда език $L(G_5) = \{a^n b^m | n, m \geq 1\}$.

- Граматика – **автоматна – тип 3:**

Правило: $A \to aB|a$, $a \in \Sigma$, $A, B \in N$. Добавя се правилото $S \to \varepsilon$ за пораждане на празната дума Λ.

Пример:

Граматиката $G_6 = \langle \{a,b\}, \{S, A, B\}, S, P \rangle$ с множество от правила от вида $P = \{S \to aA|bB|a|b,\ A \to aA|a,\ B \to bB|b,\ S \to \varepsilon\}$ е автоматна и поражда език $L(G_6) = \{a^n | n \geq 0\} \cup \{b^n | n \geq 0\}$.

Всяка автоматна граматика е безконтекстна, всяка безконтекстна граматика е контекстна, всяка контекстна граматика е граматика от общ вид, т.е. {G – тип 3}⊂{G – тип 2}⊂{G – тип 1}⊂{G – тип 0}. Тези включвания се наричат йерархия на Чомски за пораждащите граматики. Езикът, породен от граматика от тип *i* (*i* = 0, 1, 2, 3) се нарича език от тип *i*, съответно за *i* = 0, 1, 2, 3 - език от общ вид, контекстен, безконтекстен, автоматен. Породените езици образуват йерархия на Чомски.

4. Видове продукции

В съответствие със структурата на рекурсивната продукция се дефинират следните видове продукции:

- Двойно рекурсивна продукция:

$x \to \alpha x \beta$ (*x* се получава от себе си, като αβ са низове);

- Ляво рекурсивна продукция:

$x \to x\beta$ (*x* се получава чрез добавяне на низ отдясно).

- Дясно рекурсивна продукция:

$x \to \alpha x$ (*x* се получава чрез добавяне на низ отляво).

В зависимост от предназначението си продукциите се подразделят на следните видове:

Лексика: правила на множеството P за пораждане на думи;

Синтаксис: правила на множеството P за пораждане на изрази;

Семантика: допълнителни правила, свързани със смисъла на породените думи или изрази.

ФОРМАЛНИ ГРАМАТИКИ

Пример:

Дадена е азбуката $\Sigma = \{a, b, c\}$. Да се съставят граматики, които пораждат езиците

1. $L = \{a^i b^i c^i \mid i \geq 1\}$;
2. $L = \{a^i b^k c^k \mid i, k \geq 1\}$;
3. $L = \{a^i b^k c^l \mid i, k, l \geq 1\}$.

1. Контекстна граматика

За езика $L = \{a^i b^i c^i \mid i \geq 1\}$ се дефинира граматиката *G*. Въвежда се азбука на терминалните символи Σ={*a*, *b*, *c*} и азбука на нетерминални символи *N* с нетерминалните символи S, B, C, където S е началният символ. Множество от правила *P*:

$$P_1 : S \to aSBC \underbrace{\to}_{i-1} a^{i-1} S (BC)^{i-1}$$

Например при $i = 4$: $S \to aSBC \to a^3 S (BC)^3$

$$P_2 : S \to abC \underbrace{\to}_{1} a^i bC (BC)^{i-1} = a^i bCBC (BC)^{i-2}$$

$$P_3 : CB \to BC \underbrace{\to}_{(i-1)\,\text{пъти}} a^i bBCC (BC)^{i-2} = a^i bBCCBC (BC)^{i-3} = a^i bB^{i-1} C^i$$

$$P_4 : bB \to bb \underbrace{\to}_{i-1} a^i b^i C^i = a^i b^i C^i$$

$$P_5 : C \to c \underbrace{\to}_{i-\text{пъти}} a^i b^i c^i$$

Следователно контекстната граматика *G* поражда езика $L = \{a^i b^i c^i \mid i \geq 1\}$.

2. Безконтекстна граматика

За езика $L = \{a^i b^k c^k \mid i, k \geq 1\}$ се дефинира граматиката

$$G = \langle \{S, A\}, \{a, b, c,\}, S = \{S\}, P \rangle$$

с множество от правила *P*:

$$P_1 : S \rightarrow aS \underbrace{\rightarrow}_{i-1} a^{i-1}S\,;$$

$$P_2 : S \rightarrow aA \underbrace{\rightarrow}_{1} a^{i}A\,;$$

$$P_3 : A \rightarrow bAc \underbrace{\rightarrow}_{k-1} a^{i}b^{k-1}Ac^{k-1}\,;$$

$$P_4 : A \rightarrow bc \underbrace{\rightarrow}_{1} a^{i}b^{k}c^{k}\,.$$

Следователно граматиката *G* поражда езика $L = \{a^i b^k c^k \,|\, i,k \geq 1\}$.

3. Регулярна автоматна граматика

За езика $L = \{a^i b^k c^l \,|\, i,k,l \geq 1\}$ се дефинира граматиката

$$G = \langle \{S, A, B\}, \{a, b, c,\}, S = \{S\}, P \rangle$$

с множество от правила *P*:

$$P_1 : S \rightarrow aS \underbrace{\rightarrow}_{i-1} a^{i-1}S\,;$$

$$P_2 : S \rightarrow aA \underbrace{\rightarrow}_{1} a^{i}A\,;$$

$$P_3 : A \rightarrow bA \underbrace{\rightarrow}_{k-1} a^{i-1}b^{k-1}A\,;$$

$$P_4 : A \rightarrow bB \underbrace{\rightarrow}_{1} a^{i}b^{k}B\,;$$

$$P_5 : B \rightarrow cB \underbrace{\rightarrow}_{l-1} a^{i}b^{k}c^{l-1}B\,;$$

$$P_6 : B \rightarrow c \underbrace{\rightarrow}_{1} a^{i}b^{k}c^{l}\,.$$

Следователно автоматната граматика поражда езика $L = \{a^i b^k c^l \,|\, i,k,l \geq 1\}$.