Formální jazyk

Abeceda - slovo - jazyk

Abeceda je libovolná konečná množina

Prvky abecedy nazýváme **znaky / písmena / symboly**

Slovo (**řetězec**) nad abecedou Σ je libovolná konečná posloupnost znaků této abecedy.

Prázdné posloupnosti znaků odpovídá **prázdné slovo**, označované ε.

Počet členů posloupnosti v značíme $|v|$ a nazýváme **délkou slova**.

Počet výskytů znaku a ve slově v značíme $\#_a(v)$.

Jazyk nad abecedou Σ je libovolná množina slov nad Σ.

Množinu všech slov nad abecedou Σ značíme Σ^*,
množinu všech neprázdných slov Σ^+.

Jazyky nad Σ jsou tedy právě podmnožiny Σ^*.

Operace a relace nad slovy

Binární operace **zřetězení**, označována $\cdot$, která je definována předpisem $u.v = uv$

Operace zřetězení je asociativní, tj. $u.(v.w) = (u.v).w$ pro libovolná slova u, v, w.

ε se chová jako jednotkový prvek, tj. $u.\varepsilon = \varepsilon.u = u$ pro libovolné slovo u.

Slovo u je **podslovem** slova v, jestliže existují slova x, y taková, že $v = x.u.y$.

Pokud navíc $x = \varepsilon$, říkáme že slovo u je **předponou** (**prefixem**) slova v, což značíme $u \preceq v$. Je-li $y = \varepsilon$, nazveme u **příponou** (**sufixem**) slova v.

Unární operace i-**té mocniny** slova, která je definovaná induktivně pro každé $i \in \mathbb{N}_0$ takto: nechť Σ je libovolná abeceda, u libobolné slovo nad abecedou Σ. Pak

- $u^0 = \varepsilon$
- $u^{i+1} = u.u^i$

Operace nad jazyky

L je jazyk nad abecedou Σ, K je jazyk nad abecedou Δ
Výsledkem je vždy jazyk nad abecedou $\Sigma \cup \Delta$.

- Standardní množinové operace **sjednocení** ($\cup$), **průnik** ($\cap$) a **rozdíl** ($\smallsetminus$).
- **Zřetězením** jazyků K a L je jazyk $K.L = \{u.v \mid u \in K, v \in L\}$.

Platí $\quad \emptyset.L = L.\emptyset = \emptyset \quad$ a $\quad \{\varepsilon\}.L = L.\{\varepsilon\} = L$.

- i-**tá mocnina** jazyka L definována induktivně pro $i \in \mathbb{N}_0$:
  1. $L^0 = \{\varepsilon\}$
  2. $L^{i+1} = L.L^i$

$\emptyset^0 = \{\varepsilon\}$
$\emptyset^i = \emptyset$ pro libovolné $i \in \mathbb{N}$
$\{\varepsilon\}^j = \{\varepsilon\}$ pro libovolné $j \in \mathbb{N}_0$

- **Iterace** jazyka L je jazyk $L^* = \bigcup_{i=0}^{\infty} L^i$.

  $\emptyset^* = \{\varepsilon\}$

- **Pozitivní iterace** jazyka L je jazyk $L^+ = \bigcup_{i=1}^{\infty} L^i$.

  $\emptyset^+ = \emptyset$.

- **Doplněk** jazyka L je jazyk co–$L = \Sigma^* \smallsetminus L$.

- **Zrcadlovým obrazem** slova $w = a_1 \dots a_n$ nazýváme slovo $w^R = a_n \dots a_1$ ($\varepsilon^R = \varepsilon$).

  Zrcadlový obraz jazyka L definujeme $L^R = \{w^R \mid w \in L\}$.

Nechť $\mathcal{L}$ je třída jazyků a o je n-ární operace na jazycích. Řekneme, že $\mathcal{L}$ je **uzavřená** na o, pokud pro libovolné jazyky $L_1, \ldots, L_n$ patřící do $\mathcal{L}$ platí, že také jazyk $o(L_1, \ldots, L_n)$ patří do $\mathcal{L}$.

Aplikace

Konečná reprezentace jazyka

- potřeba konečné reprezentace
- co je konečná reprezentace
- automaty a gramatiky
- ??? existuje konečná reprezentace pro každý jazyk ???
- ??? jaké vlastnosti mají jazyky, které jsou konečně reprezentovatelné ???

Pojem gramatiky

Popis jazyka pomocí pravidel, podle kterých se vytvářejí všechny slova daného jazyka.

<věta> → <podmětná část><přísudková část>
<podmětná část> → <podstatné jméno>
<podstatné jméno> → JANA
<přísudková část> → <sloveso><předmětová část>
<sloveso> → ČTE
<předmětová část> → <podstatné jméno>
<podstatné jméno> → KNIHU

Zadání syntaxe vyšších programovacích jazyků — Backus-Naurova normální forma (BNF)

Definice 1. **Gramatika** $\mathcal{G}$ je čtveřice (N, Σ, P, S), kde

- N je neprázdná konečná množina **neterminálních symbolů** (stručněji: **neterminálů**).
- Σ je konečná množina **terminálních symbolů (terminálů)** taková, že $N \cap \Sigma = \emptyset$. Sjednocením N a Σ obdržíme množinu **všech symbolů** gramatiky, kterou obvykle označujeme symbolem V.
- $P \subseteq V^*NV^* \times V^*$ je konečná množina **pravidel**. Pravidlo (α, β) obvykle zapisujeme ve tvaru $\alpha \rightarrow \beta$ (a čteme jako "α přepiš na β").
- $S \in N$ je speciální **počáteční** neterminál (nazývaný také **kořen gramatiky**).

Příklad gramatiky

Gramatika $\mathcal{G} = (N, \Sigma, P, S)$ určuje

- relaci $\Rightarrow_{\mathcal{G}}$ **přímého odvození** na množině V^*

  $\gamma \Rightarrow_{\mathcal{G}} \delta$ právě když existuje pravidlo $\alpha \to \beta \in P$ a slova $\eta, \varrho \in V^*$ taková, že $\gamma = \eta\alpha\varrho$ a $\delta = \eta\beta\varrho$.

  Používá se i označení **krok odvození**.

- relaci $\overset{k}{\Rightarrow}_{\mathcal{G}}$ **odvození v** k **krocích** pro $k \in \mathbb{N}_0$
  - $\overset{0}{\Rightarrow}_{\mathcal{G}}$ je identická relace
  - $\overset{k+1}{\Rightarrow}_{\mathcal{G}} \;=\; \overset{k}{\Rightarrow}_{\mathcal{G}} \circ \Rightarrow_{\mathcal{G}}$

- relaci $\overset{\leq k}{\Rightarrow}_{\mathcal{G}}$ **odvození v nejvýše k krocích** pro $k \in \mathbb{N}_0$

$$\overset{\leq k}{\Rightarrow}_{\mathcal{G}} \;=\; \bigcup_{i=0}^{k} \overset{i}{\Rightarrow}_{\mathcal{G}}$$

- relaci $\Rightarrow_{\mathcal{G}}^{*}$ **odvození**

$$\Rightarrow_{\mathcal{G}}^{*} \;=\; \bigcup_{i=0}^{\infty} \overset{i}{\Rightarrow}_{\mathcal{G}}$$

  Relace $\Rightarrow_{\mathcal{G}}^{*}$ je reflexivní a tranzitivní uzávěr $\Rightarrow_{\mathcal{G}}$.

- relaci $\Rightarrow_{\mathcal{G}}^{+}$ **netriviálního odvození**

$$\Rightarrow_{\mathcal{G}}^{+} \;=\; \bigcup_{i=1}^{\infty} \overset{i}{\Rightarrow}_{\mathcal{G}}$$

  Relace $\Rightarrow_{\mathcal{G}}^{+}$ je tedy tranzitivní uzávěr relace $\Rightarrow_{\mathcal{G}}$.

Větná forma gramatiky $\mathcal{G}$ je každý řetěz z množiny V^*, který lze odvodit z počátečního neterminálu gramatiky.

Věta gramatiky $\mathcal{G}$ je každá větná forma, která obsahuje pouze terminály.

Jazyk generovaný gramatikou $\mathcal{G}$, $L(\mathcal{G})$ je množina všech vět gramatiky

$$L(\mathcal{G}) = \{w \in \Sigma^* \mid S \Rightarrow^* w\}.$$

Gramatiky $\mathcal{G}_1$ a $\mathcal{G}_2$ nazveme **jazykově ekvivalentní**, právě když generují tentýž jazyk, tj. $L(\mathcal{G}_1) = L(\mathcal{G}_2)$.

Konvence

Příklad

$$\mathcal{G} = (\{S, X\}, \{a, b\}, P, S)$$

$$P = \{ \begin{array}{lll} S & \to & abS \\ S & \to & bX \\ bbX & \to & babS \\ bbX & \to & \varepsilon \end{array} \}$$

Příklad

$\mathcal{G} = (\{S, X\}, \{a, b\}, P, S)$

$$P = \{ \begin{array}{lcl} S & \to & abS \\ S & \to & bX \\ bbX & \to & babS \\ bbX & \to & \varepsilon \end{array} \}$$

DÚ: Zjistěte, jaký jazyk generuje gramatika
$\mathcal{G} = (\{S, A\}, \{0, 1\}, P, S)$

$$
\begin{array}{llllllll}
P = \{ & S & \rightarrow & 1S & \mid & 0S & \mid & 101A \\
 & A & \rightarrow & 1A & \mid & 0A & \mid & \varepsilon \quad \}
\end{array}
$$

Chomského hierarchie gramatik

Klasifikace gramatik podle tvaru přepisovacích pravidel

typ 0 pravidla v obecném tvaru (**frázové gramatiky**)

typ 1 pro každé její pravidlo $\alpha \rightarrow \beta$ platí $|\alpha| \leq |\beta|$ s eventuelní výjimkou pravidla $S \rightarrow \varepsilon$, pokud se S nevyskytuje na pravé straně žádného pravidla (**kontextové gramatiky**)

typ 2 každé její pravidlo je tvaru $A \rightarrow \alpha$, kde $|\alpha| \geq 1$ s eventuelní výjimkou pravidla $S \rightarrow \varepsilon$, pokud se S nevyskytuje na pravé straně žádného pravidla (**bezkontextové gramatiky**)

typ 3 každé její pravidlo je tvaru $A \rightarrow aB$ nebo $A \rightarrow a$ s eventuelní výjimkou pravidla $S \rightarrow \varepsilon$, pokud se S nevyskytuje na pravé straně žádného pravidla (**regulární gramatiky**)

Chomského hierarchie jazyků

Hierarchie gramatik určuje hierarchii jazyků.

Jazyk L je typu 0 (rekursivně spočetný) pokud existuje gramatika $\mathcal{G}$ typu 0 taková, že $L(\mathcal{G}) = L$.

Analogicky: **kontextový, bezkontextový, regulární**

$\mathcal{L}_0$ třída všech rekursivně spočetných jazyků

$\mathcal{L}_1$ třída všech kontextových jazyků

$\mathcal{L}_2$ třída všech bezkontextových jazyků

$\mathcal{L}_3$ třída všech regulárních jazyků

$$\mathcal{L}_0 \supset \mathcal{L}_1 \supset \mathcal{L}_2 \supset \mathcal{L}_3$$

(Důkaz později)

Věta 1. Nad abecedou $\{a\}$ existuje jazyk, který není typu 0.

Důkaz.

- množina všech slov nad abecedou $\{a\}$ je spočetně nekonečná
- množina všech jazyků nad touto abecedou má proto mohutnost $2^{\aleph_0}$ (je tedy nespočetná)
- gramatik typu 0 nad abecedou $\{a\}$ je pouze spočetně mnoho:
  - buď M libovolná, ale pevně zvolená spočetná množina
  - b.ú.n.o. každá gramatika má neterminály z M
  - každá gramatika je slovo nad abecedou

$$M \cup \{a, \rightarrow, \varepsilon, \underline{(}, \underline{)}, \underline{\{}, \underline{\}}, \underline{,}\}$$

  - všech slov délky i nad touto abecedou je $\aleph_0^i = \aleph_0$ pro lib. $i \in \mathbb{N}$
  - **všech** slov nad touto abecedou je tedy spočetně mnoho *(sjednocení spočetně mnoha spočetných mn. je spočetné)* $\square$