KONTEKSTNO-NEODVISNE GRAMATIKE ZA KODIRANJE IN STISKANJE PODATKOV

JANEZ PODLOGAR

1. Kontekstno-neodvisne gramatike

V jezikoslovju pravopis določa pravila o rabi črk in ločil. S slovnico poimenujemo sistem pravil za tvorjenje povedi in sestavljanje besedil. Slovenska slovnica, Slovenski pravopis in Slovar slovenskega knjižnega jezika natančno določajo Slovenski knjižni jezik, ki je poglavitno sredstvo javnega in uradnega sporazumevanja v Sloveniji.

Podobno je formalna gramatika sistem pravil, ki nam pove kako iz dane abecede tvorimo nize. Gramatika nam torej določa neko podmnožico nizev, ki jo imenujemo formalni jezik. Gramatike in formalni jeziki imajo široko teoretični in praktično uporabo. Uporabljajo se za modeliranje naravnih jezikov, so osnova programskih jezikov, formalizirajo matematično logiko in sisteme aksiomov ter se uporabljajo tudi za kompresijo podatkov.

Definicija 1.1. *Abeceda* je končna neprazna množica Σ. Elementom abecede pravimo *črke*. *Množica vseh končnih nizov abecede* Σ je

$$\Sigma^* = \{a_1a_2a_3\cdots a_n \mid n \in \mathbb{N}_0 \wedge \forall i : a_i \in \Sigma\},$$

kjer za $n = 0$ dobimo prazen niz, ki ga označimo z ε. *Množica vseh končnih nizov abecede brez praznega niza* označimo s Σ^+. *Dolžino niza* w označimo z $|w|$ in je enaka številu črk v nizu $w \in \Sigma^*$. *Množico vseh nizov dolžine* ℓ, kjer je ℓ pozitivno celo število, označimo s Σ^ℓ. *Jezik na abecedi* Σ je poljubna podmnožica množice Σ^*.

Definicija 1.2. Naj bo Σ abeceda. Naj bo $*$ asociatna binarna operacija ma množici vseh končnih nizov Σ^* tako, da je prazen niz ϵ nevtralen element in za niza $w, u \in \Sigma^*$ velja

$$w * u = w_1w_2\cdots w_nu_1u_2\cdots u_m,$$

kjer sta $w_1w_2\cdots w_n$ in $u_1u_2\cdots u_m$ predstavitvi nizov w in u z črkami abecede Σ. $(\Sigma^*, *)$ je prost monoid nad Σ.

Opomba 1.3. Dvočlena operacija $\circ$ na množici A je preslikava

$$A \times A \to A,$$
$$(x, y) \mapsto x \circ y.$$

Monoid $(A, \circ)$ je neprazna množica A z dvočleno asociativno operacijo $\circ$, ki ima nevtralni element. Ime prosti monoid izhaja iz Teorije kategorij.

Definicija 1.4. *Kontektsno-neodvisna gramatika* je četverica $G = (V, \Sigma, P, S)$, kjer je V končna množica *nekončnih simbolov*; abeceda Σ množica *končnih simbolov* tako, da $\Sigma \cap V = \emptyset$; $P \subseteq V \times (V \cup \Sigma)^*$ celovita relacija, elementom relacije pravimo *produkcijska pravila*; in $S \in V$ *začetni simbol*.

Date: 29. januar 2023.

Opomba 1.5. Relacija $P \subseteq A \times B$ je celovita, če velja

$$\forall x \in A \; \exists y \in B \colon (x, y) \in P.$$

Naj bo $G = (V, \Sigma, P, S)$ kontekstno-neodvisna gramatika. Ker je relacija P celovita, za vsak nekončni simbol $A \in V$ obstaja končen niz nekončnih in končnih simbolov $\alpha \in (V \cup \Sigma)^*$, da je (A, α) produkcijsko pravilo. Torej $(A, \alpha) \in P$, kar pišemo

$$A \to \alpha.$$

Definicija 1.6. Naj bo $G = (V, \Sigma, P, S)$ kontekstno-neodvisna gramatika. Naj bodo α, β, $\gamma \in (V \cup \Sigma)^*$ nizi nekončnih in končnih simbolov, $A \in V$ nekončni simbol ter naj bo $(A, \beta) \in P$ produkcijsko pravilo, označimo ga z $A \to \beta$. Pravimo, da se $\alpha A \gamma$ *prepiše s pravilom* A v $\alpha\beta\gamma$, pišemo $\alpha A \gamma \Rightarrow \alpha\beta\gamma$. Pravimo, da α *izpelje* β, če je $\alpha = \beta$ ali če za $n \geq 0$ obstaja zaporedje $\alpha_1, \alpha_2, \ldots \alpha_n \in (V \cup \Sigma)^*$ tako, da

$$\alpha \Rightarrow \alpha_1 \Rightarrow \alpha_2 \Rightarrow \ldots \Rightarrow \alpha_n \Rightarrow \beta,$$

pišemo $\alpha \overset{*}{\Rightarrow} \beta$.

Posledica 1.7. *Jezik kontekstno neodvisne gramatike G je*

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

Opomba 1.8. Ime kontekstno-neodvisna gramatika izvira iz oblike produkcijskih pravil. Na levi strani produkcijskega pravila mora vedno stati samo spremenljivka. Torej vsebuje samo pravila oblike

$$A \to \alpha,$$

kjer je $A \in V$ in $\alpha \in (V \cup \Sigma)^*$. Ne sme pa vsebovati pravila oblike

$$\alpha A \gamma \to \alpha\beta\gamma,$$

kjer je $A \in V$ in so $\alpha, \beta, \gamma \in (V \cup \Sigma)^*$, saj je možnost uporabe pravila odvisno od konteksta nekončnega simbola A. Kontekst določa niza α in β, ki se nahajata neposredno pred in po nekončnim simbolom A.