# Kontekstno-neodvisne gramatike za kodiranje in stiskanje podatkov

Janez Podlogar

Univerza v Ljubljani, Fakulteta za matematiko in fiziko

21. 11. 2022

# Kodiranje in kod

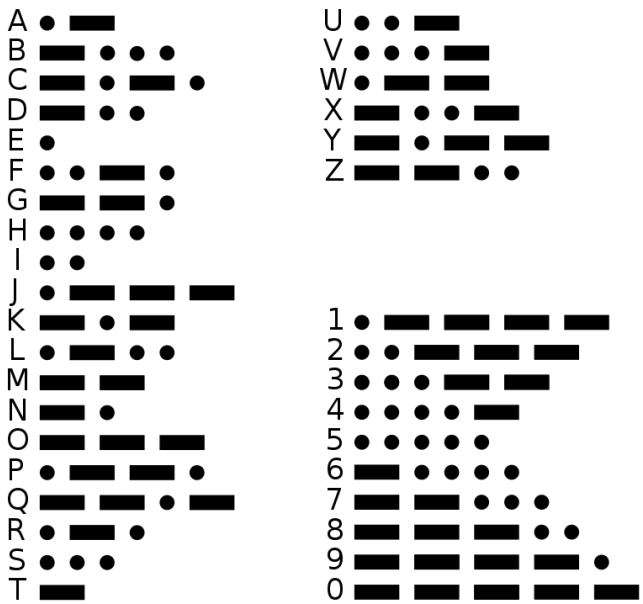
Zapis informacije v neki obliki ni primeren za vsakršno rabo. Spreminjanje zapisa sporočila imenujemo *kodiranje*, sistemu pravil, po katerem se kodiranje opravi, pa *kod*.

# Primer kodiranja

*Morsejeva abeceda* je kodiranje črk, števil in ločil s pomočjo zaporedja kratkih in dolgih signalov:

- Dolžina kratkega signala je ena enota.
- Dolgi signal je trikrat daljši od kratkega signala.
- Razmik med signali znotraj črke je tišina dolžine kratkega signala.
- Razmik med črkami je tišina dolga tri kratke signale oz. en dolgi signal.
- Presledek med besedami je tišina dolga sedmih kratkih signalov.

# Morsejeva abeceda




Slika: Mednarodna Morsejeva abeceda

# Abeceda in nizi na abecedi

## Definicija

*Abeceda* je končna neprazna množica $\Sigma$. Elementom abecede pravimo *črke*. *Množica vseh končnih nizov abecede* $\Sigma$ označimo z $\Sigma^*$ in vključuje tudi prazen niz, ki ga označimo z $\varepsilon$. *Dolžino niza* $w$ označimo z $|w|$ in je enaka številu črk v nizu $w \in \Sigma^*$. *Jezik na abecedi* $\Sigma$ je poljubna podmnožica množice $\Sigma^*$.

# Abeceda in nizi na abecedi

## Definicija

*Abeceda* je končna neprazna množica $\Sigma$. Elementom abecede pravimo *črke*. *Množica vseh končnih nizov abecede* $\Sigma$ označimo z $\Sigma^*$ in vključuje tudi prazen niz, ki ga označimo z $\varepsilon$. *Dolžino niza* $w$ označimo z $|w|$ in je enaka številu črk v nizu $w \in \Sigma^*$. *Jezik na abecedi* $\Sigma$ je poljubna podmnožica množice $\Sigma^*$.

## Primer

Naj bo $\Sigma = \{a, b, c\}$ abeceda, potem je

$$ab \in \Sigma^*, \quad cababcccababcccab \in \Sigma^*.$$

# Kodiranje in dekodiranje

## Definicija

*Kodiranje nizov abecede* $\Sigma$ je injektivna funkcija $\kappa\colon \Sigma^* \to \Sigma_c^*$, kjer je $\Sigma_c$ *kodirna abeceda* in $\kappa(w)$ imenujemo *koda niza* $w$. *Dokodiranje kodiranja* $\kappa$ je funkcija $\kappa^{-1}\colon C \subseteq \Sigma_c^* \to \Sigma^*$, da velja

$$\forall w \in \Sigma^*\colon \kappa^{-1}(\kappa(w)) = w$$

# Formalizirzacija Morsejeve abecede

Abecedi sta

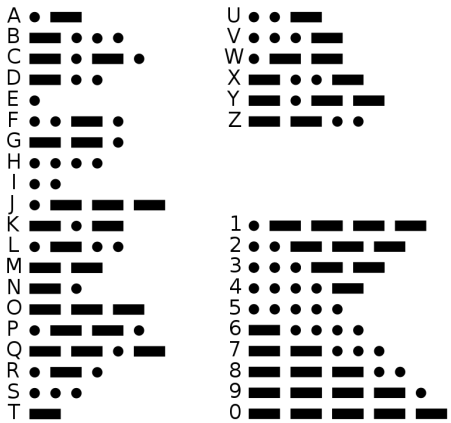
$$\Sigma = \{\mathrm{A}, \mathrm{B}, \ldots, \mathrm{Z}\} \cup \{0, 1, \ldots, 9\} \cup \{\_\},$$
$$\Sigma_c = \{\cdot, -, \_\}.$$

Definirajmo kodno funkcijo črk abecede $\kappa\colon \Sigma \to \Sigma_c^*$, ki vsakei črki iz abecede $\Sigma$ priredi niz črk kodirne abecede $\Sigma_c$. Za niz $w = a_1 a_2 \ldots a_n \in \Sigma^*$ definiramo kodirno funkcijo $K$ po črkah

$$K(w) = \kappa(a_1)\_\kappa(a_2)\_ \cdots \kappa(a_n).$$

# Formalizirzacija Morsejeve abecede

Vrednosti funkcije $\kappa$ so določene s tabelo




Dodatno presledek med besedami _ kodiramo v šest kratkih enot tišine

$$\kappa(\_) = \_\_\_\_$$

# Stiskanje podatkov

## Definicija

*Stiskanje* je kodiranje $K$ za katerega velja

$$\exists n \in \mathbb{N}\, \forall w \in \Sigma^* : |w| \geq n \implies |K(w)| \ll |w| .$$

# Zgled stiskanja niza $w$

Za abecedo vzemimo $\Sigma = \{a, b, c\}$ in poglejmo niz

$$w = cababcccababcccab.$$

# Zgled stiskanja niza $w$

Za abecedo vzemimo $\Sigma = \{a, b, c\}$ in poglejmo niz

$$w = cababcccababcccab.$$

Uvedemo novi spremenljivki $A = ab$ in $B = ccc$.

# Zgled stiskanja niza $w$

Za abecedo vzemimo $\Sigma = \{a, b, c\}$ in poglejmo niz

$$w = cababcccababcccab.$$

Uvedemo novi spremenljivki $A = ab$ in $B = ccc$. Potem je

$$w = cAABAABA.$$

# Zgled stiskanja niza $w$

Za abecedo vzemimo $\Sigma = \{a, b, c\}$ in poglejmo niz

$$w = cababcccababcccab.$$

Uvedemo novi spremenljivki $A = ab$ in $B = ccc$. Potem je

$$w = cAABAABA.$$

Uvedemo novo spremeljivko $C = AAB$.

# Zgled stiskanja niza $w$

Za abecedo vzemimo $\Sigma = \{a, b, c\}$ in poglejmo niz

$$w = cababcccababcccab.$$

Uvedemo novi spremenljivki $A = ab$ in $B = ccc$. Potem je

$$w = cAABAABA.$$

Uvedemo novo spremeljivko $C = AAB$. Potem je

$$w = cCCA.$$

# Zgled stiskanja niza $w$

Prešnji postopek napišemo na sledeč način s pomočjo produkcijskih pravil

$$S \to cCCA,$$
$$A \to ab,$$
$$B \to ccc,$$
$$C \to AAB.$$

# Kontekstno-neodvisne gramatike

## Definicija

*Formalna gramatika* $G$ so pravila, ki nam iz abecede $\Sigma$ tvorijo jezik, označimo ga z $L(G)$.

# Kontekstno-neodvisne gramatike

## Definicija

*Formalna gramatika* $G$ so pravila, ki nam iz abecede $\Sigma$ tvorijo jezik, označimo ga z $L(G)$.

## Definicija

*Kontektsno-neodvisna gramatika* je četverica $G = (V, \Sigma, P, S)$, kjer je $V$ končna množica *spremenljivk*, abeceda $\Sigma$ množica *končnih simbolov* tako, da $\Sigma \cap V = \emptyset$, $P \subseteq V \times (V \cup \Sigma)^*$ relacija, ki ji pravimo *produkcijsko pravilo* in $S \in V$ *začetna spremenljivka*.

# Kontekstno-neodvisne gramatike

## Definicija

Naj bo $G = (V, \Sigma, P, S)$ kontekstno-neodvisna gramatika. Naj bodo $\alpha$, $\beta, \gamma \in (V \cup \Sigma)^*$ nizi spremenljivk in končnih simbolov, $A \in V$ spremenljivka 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 $\alpha$ *porodi* $\beta$, če je $\alpha = \beta$ ali če za $k \geq 0$ obstaja zaporedje $\alpha_1, \alpha_2, \ldots \alpha_n \in (V \cup \Sigma)^*$ tako, da

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

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

# Kontekstno-neodvisne gramatike

## Posledica

*Jezik kontekstno neodvisne gramatike $G$ je*

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

# Zgled stiskanja niza $w$ z zapisom gramatike

Formalizirajmo gramatiko iz prejšnjega primera. Gramatiko smo generirali z nizom $w = cababcccababcccab$.

# Zgled stiskanja niza $w$ z zapisom gramatike

Formalizirajmo gramatiko iz prejšnjega primera. Gramatiko smo generirali z nizom $w = cababcccababcccab$. Dobimo $G_w = (V, \Sigma, P, S)$, kjer je

$$V = \{S, A, B, C\},$$
$$\Sigma = \{a, b, c\},$$
$$P = \{S \to cCCA, A \to ab, B \to ccc, C \to AAB\},$$
$$S = S.$$

# Zgled stiskanja niza $w$ z zapisom gramatike

Formalizirajmo gramatiko iz prejšnjega primera. Gramatiko smo generirali z nizom $w = cababcccababcccab$. Dobimo $G_w = (V, \Sigma, P, S)$, kjer je

$$V = \{S, A, B, C\},$$
$$\Sigma = \{a, b, c\},$$
$$P = \{S \to cCCA, A \to ab, B \to ccc, C \to AAB\},$$
$$S = S.$$

Vidimo, da $G_w$ ustreza naši definiciji kontekstno-neodvisne gramatike in res kodira $w$, saj je

$$L(G_w) = \{w\}$$