PROGRAMSKI PREVODIOCI

- Formalni jezici i formalne gramatike -

Azbuka

- Simbol (znak ili slovo) je osnovni (nedeljivi) element jezika.
- Apstraktna (formalna) azbuka, ili samo azbuka, je svaki konačan neprazan skup elemenata V.
- Npr. Azbuku V čine sledeći simboli:
  - Mala i velika slova abecede A,a,B,b,C,c ...
  - Specijalni znaci +, -,*, :=, ...
  - Reči kao što su **if**, **while, class**, ...

Reči

- Niz (niska ili reč) – konačan broj redom napisanih simbola azbuke V.
- Niz koji ne sadrži nijedan simbol naziva se prazna reč i označava sa ε

- Primer: V= {a,b,c}
- Reči: ε, a, b, c, aa, bb, cc, ab, ac, abc, aabc,

- Reči su uređeni nizovi tako da je ab $\neq$ ba

Formalna definicija reči

1. ε reč nad azbukom V
2. Ako je x reč azbuke V i ako je a element azbuke V tada je i xa reč azbuke V.
3. y je reč nad azbukom V ako i samo ako je dobijen pomoću pravila 1. i 2.

Za označavanje reči koriste se obično završna mala slova abecede: *u, v, w, x, y, z*

Dužina reči

- Dužina reči – broj simbola u nizu
- Oznaka: $|x|$ je dužina reči x.
- Za $x = abc$ $|x| = 3$.
- $|\varepsilon| = 0$

Operacije nad rečima: Spajanje (konkatenacija) proizvod reči

- Ako su x i y dve reči azbuke V, proizvod ili spajanje reči je operacija kojom se stvara nova reč tako što se na jednu reč nadovezuje druga reč.

  $$x = aA\ ,\quad y = ab$$

  $$z = xy = aAab$$

- ε je neutralni element za operaciju množenja (nadovezivanja) reči.

  $$\varepsilon x = x\varepsilon = x$$

Operacije nad rečima: Eksponent

$$\underbrace{xxx\ldots x}_{\text{n puta}} = x^n$$

$$x^i = x^{i-1}x$$

$$x^0 = \varepsilon$$

$$x^1 = x$$

$$x^2 = xx$$

$$x^3 = xxx$$

Delovi reči

prefiks reči x	Niz koji se dobija kada se izbaci nula ili više simbola na kraju reči *x*, pr. ban je prefiks reči banana.
sufiks reči x	Niz koji se dobija izbacivanjem nula ili više početnih simbola reči *x*. nana je sufiks reči banana
podniz reči x	reč koja se dobija kada se izbaci neki prefiks i/ili neki sufiks reči *x*. ana je podniz reči banana Svaki prefiks i svaki sufiks reči *x* su podnizovi reči *x*, dok svaki podniz reči *x* ne mora da bude ni sufiks ni prefiks reči *x*. Za svaku reč *x* i *x* i ε su prefiksi, sufiksi i podnizovi reči *x*

Delovi reči

Pravi prefiks, pravi sufiks i pravi podniz reči *x*	Svaki neprazan niz *y* koji je prefiks, sufiks ili podniz reči *x*, takav da je *x* različito od *y*.
Podsekvenca reči *x*	Svaki niz koji se dobija izbacivanjem nula ili više sukcesivnih simbola iz reči *x*. baa je podsekvenca niza banana

Formalni jezik

- Formalnim jezikom L nad azbukom V naziva se bilo koji skup reči nad tom azbukom.

- Prema ovoj definiciji formalni jezik je i prazan skup reči kao i skup { ε } koji sadrži samo reč ε.

Primeri jezika

1. Neka azbuku V čine sva slova naše azbuke:

   V={а,б,в,г,д,ђ,ж,з,и,...,ш}

   Sve reči srpskog jezika predstavljaju jedan formalni jezik definisan nad ovom azbukom.

Operacije nad jezicima

OPERACIJA	DEFINICIJA OPERACIJE
Unija jezika L i M L U M	$L \bigcup M = \{x \mid x \in L \vee x \in M\}$
Nadovezivanje konkatenacija L i M LM	$LM = \{xy \mid x \in L \wedge y \in M\}$
Potpuno zatvaranje L*	$L^* = \bigcup_{i=0}^{\infty} L^i$
Pozitivno zatvaranje L^+	$L^+ = \bigcup_{i=1}^{\infty} L^i$

V*

V = {a,b,c}

V* = {ε,a,b,c,aa,bb,cc,ab,ac,bc,ca,cb,abc,…}

$V^+ = V^* \setminus \varepsilon$

Formalni jezik nad azbukom V je bilo koji podskup skupa V*.

$$L \subseteq V^*$$

L^T

- Transponovana reč reči x u oznaci x^T definiše se na sledeći način:
  1. $\varepsilon^T = \varepsilon$
  2. $(xa)^T = ax^T$

Pr. Ako je $x = abc$ tada je $x^T = cba$

Transponovani jezik jezika L je skup svih transponovanih reči jezika L.

Primer

Neka je $L=\{A, B, C,\ldots,Z, a, b, c, \ldots, z\}$ i
$D= \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$.
Kako se slova azbuke mogu posmatrati kao reči dužine 1, onda je svaki od skupova L i D i formalni jezik.

$L \cup D$	skup slova i cifara
LD	skup svih reči koje se sastoje od slova iza kog stoji cifra
L^4	skup svih četvoroslovnih reči
L^*	skup svih nizova slova uključujući i ε
$L(L \cup D)^*$	skup nizova slova i cifara koji počinju slovom.
D^+	skup svih nizova od jedne ili više cifara

Opis i prepoznavanje jezika

- Formalna gramatika je sredstvo za opis jezika na konačan način.
- Gramatika jezika opisuje kako se generišu reči koje pripadaju određenom jeziku.

- Prepoznavanje jezika je problem utvrđivanja da li određena reč pripada jeziku opisanom zadatom gramatikom.
- Ovaj problem se rešava pomoću uređaja za prepoznavanje jezika ili automata

Opis jezika (primer sa proslog casa)

1. Svaki identifikator je izraz
2. Svaka konstanta je izraz
3. Ako su *izraz1* i *izraz2* izrazi tada je su izrazi i:

   izraz1+ izraz2

   izraz1* izraz2

   (izraz1)

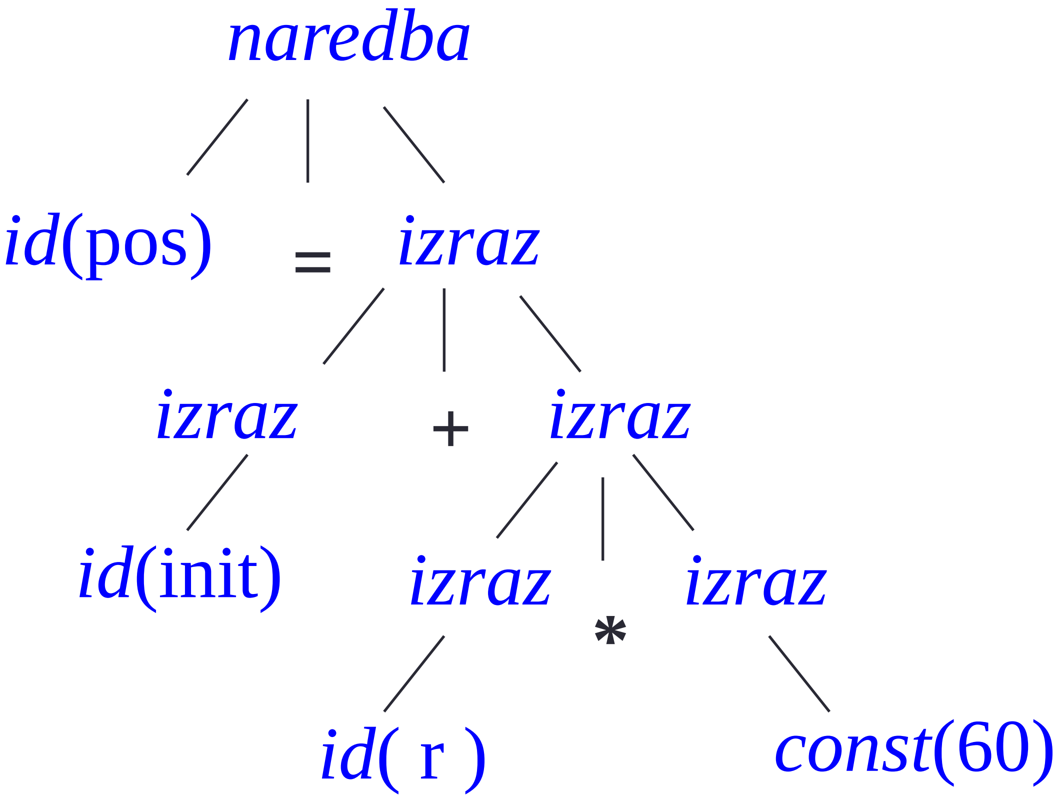
Opis jezika (primer sa proslog casa)

1. Ako je *id1* identifikator i ako je *izraz1* izraz tada je *id1 := izraz1* *naredba.*
2. Ako je *izraz1* izraz i *naredba1* naredba tada su naredbe i:

   while (*izraz1*) do *naredba1*

   if (*izraz1*) then *naredba1*

Sintaksno stablo



Elementi gramatike

- Terminalni simboli – Osnovni simboli od kojih se satoje reči jezika. Npr. ključne reči nekog programskog jezika **if**, **then**, **else** su terminalni simboli.
- Neterminalni simbol – Pomoćni (sintaksni) simboli kojima se označavaju skupovi reči. Neterminali se uvode da bi se lakše definisao jezik koji se generiše gramatikom, kao i da bi se lakše definisala hierahijska struktura jezika.
- Startni simbol – Neterminalni simbol iz kojeg se izvode sve reči jezika koji se definiše.
- Produkciono pravilo (smena) – definiše način na koji se stvaraju nizovi koji mogu da se sastoje od neterminala i terminala. U opštem slučaju pravila su oblika:

  x ::= y ili x -> y

Notacija

- Terminalni simboli:
  - Slova abecede a,b,c,...
  - Simboli operatora *, +, -, ..
  - Specijalni znaci: (,), <, >, …
  - Cifre 0,1,2, ...9.
  - reči napisane boldiranim fontom kao što su **id** ili **if.**
- Neterminalni simboli:
  - Velika slova A, B, C, ...
  - Reči napisane malim slovima italik: *expr*, *stmt* ili između zagrada: <expr>, <stmt>
- Produkciona pravila
  - x ::= y ili x -> y

Direktno izvođenje i direktna redukcija

$$z = z_1 x z_2, \quad x \to y, \quad z' = z_1 y z_2$$

Niz z' je direktno izveden iz niza z. Ovaj postupak se naziva direktno izvođenje i označava se sa:

$$z_1 x z_2 \Rightarrow z_1 y z_2$$

Niz z' se redukuje na niz z.

Važi i: $x \Rightarrow y, \qquad x \Rightarrow x$

Izvođenje

$$x \Rightarrow u_1, \quad u_1 \Rightarrow u_2, \ldots, u_n \Rightarrow y$$

Kažemo da se reč y izvodi iz reči x, i da se y redukuje na x. Izvođenje je višestruko primenjeno direktno izvođenje i označava se sa: $x \xrightarrow{*} y$

Formalne gramatike
Noam Chomsky

G = (V_n, V_t, S, P)

Važi: $V = V_n \cup V_t \quad i \quad V_n \cap V_t = \phi$

P je skup smena oblika:

$$x \to y, \quad gde\ je \quad x \in V^* V_n V^* \wedge y \in V^*$$

Jezik: $L(G) = \{w \mid S \xrightarrow{*} w, w \in V_t^*\}$

Gramatika – primer 1

$G = (\{A, B, C, D\}, \{a, b\}, A, P)$

$P:$

1. $A \to CD$
2. $C \to aCa$
3. $C \to bCB$
4. $BD \to bD$
5. $Ba \to aB$
6. $Bb \to bB$
7. $C \to \varepsilon$
8. $D \to \varepsilon$

$$L(G) = \{ww \mid w \in \{a, b\}^*\}$$

Jezik L(G) sadrži samo reči parne dužine, pri čemu je prva polovina reči jednaka drugoj. Primer izvođenja:

$$A \xrightarrow{1} CD \xrightarrow{2} aCaD \xrightarrow{3} abCBaD \xrightarrow{7}$$
$$abBaD \xrightarrow{5} abaBD \xrightarrow{4} ababD \xrightarrow{8} abab$$

Gramatika – primer 2

1. Svaki identifikator je izraz
2. Svaka konstanta je izraz
3. Ako su *izraz1* i *izraz2* izrazi tada je su izrazi i:

    izraz1+ izraz2

    izraz1* izraz2

    (izraz1)

Gramatika – primer 2

1. Svaki identifikator je izraz
2. Svaka konstanta je izraz
3. Ako su *izraz1* i *izraz2* izrazi tada je su izrazi i:

   izraz1+ izraz2

   izraz1* izraz2

   (izraz1)

$$G = (V_n, V_t, S, \mathrm{P}), V_t = \{\mathbf{id}, \mathbf{const}, +, *, (,)\}, V_n = \{izraz\}, S = izraz$$

P:

$$\begin{aligned}
izraz &\rightarrow \mathbf{id} \\
izraz &\rightarrow \mathbf{const} \\
izraz &\rightarrow izraz + izraz \\
izraz &\rightarrow izraz * izraz \\
izraz &\rightarrow (izraz)
\end{aligned}$$

Tipovi gramatika
Gramatike tipa 0

G = (V_n, V_t, S, P) u kojoj su sve smene iz skupa P oblika:

$$x \to y, \quad gde\ je \quad x \in V^* V_n V^* \wedge y \in V^*$$

Primer:

$$V_n = \{S\} \quad i \quad V_t = \{0,1\}$$

$$P = \{S \to 0S1, S \to 01\}$$

$$L(G) = \{0^n 1^n \mid n \geq 1\}.$$

Primer izvođenja:

$$S \to 0S1 \to 00S11 \to 000S111 \to \ldots \to 0^n 1^n$$

Gramatike tipa 1. Konteksna gramatika

$$Za \quad x \rightarrow y \quad vazi \quad |y| \geq |x|$$

Kako je |x| >= 1 sledi da je i |y| >= 1, što znači da na desnoj strani pravila ne može da bude prazan niz ε.

Gramatike tipa 2. Beskonteksne gramatike

Ako u gramatici G svaka smena ima oblik:

$$A \rightarrow y, \quad A \in V_n, \quad y \in V^*$$

Za ove gramatike se koristi i naziv *Bekusova normalna forma* BNF i najčešće se koriste za opis sintakse programskih jezika.

Primer: Sledeća gramatika definiše proste aritmetičke izraze:

$$G = (\{E, A\}, \{(,), +, -, *, /, \mathbf{id}\}, E, P)$$

$$P:$$

$$E \rightarrow EAE \mid (E) \mid -E \mid \mathbf{id}$$

$$A \rightarrow + \mid - \mid * \mid /$$

Gramatike tipa 3. Regularne gramatike

Gramatika G je gramatika tipa 3 ako je svaka njena smena oblika:

$$A \to aB \lor A \to a, \quad A, B \in V_n \land a \in V_t \bigcup \{\varepsilon\}$$

Za ove gramatike se koriste još i nazivi *Regularne gramatike*, *Gramatike sa konačnim brojem stanja* i *Automatne gramatike*.

Služe za opis leksičkih elemenata jezika.

Relacije između jezika

$$L(3) \subseteq L(2) \subseteq L(1) \subseteq L(0)$$

Gramatike tipa 0 su najopštije i praktično su sinonim za algoritam. Sva algoritamska preslikavanja se mogu opisati gramatikama tipa 0.

Gramatike tipa 3 pokrivaju najuži skup jezika ali je za ove jezike najjednostavnije rešiti problem prepoznavanja. Prepoznaju se pomoću konačnih automata.

Rečenične forme

- Svi nizovi koji nastaju u postupku generisanja jezika su rečenične forme tog jezika.
- Sve rečenične forme jezika se redukuju na startni simbol.

$$G = (\{E, A\}, \{(,),+,-,*,/,\mathbf{id}\}, E, P)$$

$$P: \quad E \to EAE \mid (E) \mid -E \mid \mathbf{id}; \quad A \to + \mid - \mid * \mid /$$

$$E \to \boxed{EAE} \to \boxed{E+E} \to \boxed{id+E}$$

Normalne forme gramatika

- Dve gramatike su ekvivalentne ako generišu isti jezik.
- Pod normalnim formama podrazumevamo standardni način zadavanja skupa ekvivalentnih gramatika.

NF za konteksne gramatike

Ako je G konteksna gramatika onda postoji njoj ekvivalentna konteksna gramatika G1 u kojoj svaka smena ima oblik:

$$xAy \rightarrow xry \quad A \in V_n, \quad x, y \in V^*, \quad r \in V^+$$

Na osnovu ovog svojstva izveden je i naziv konteksne gramatike. Vidi se da se neterminal *A* zamenjuje nizom *r* samo ako se nađe u kontekstu reči *x* i *y*.

NF za beskonteksne gramatike (1)

Svaka beskonteksna gramatika G ima ekvivalentnu gramatiku G1 u kojoj svaka smena ima jedan od sledećih oblika:

$$S \to \varepsilon, \quad A \to BC, \quad A \to a$$

$$A, B, C \in V_n, \quad a \in V_t$$

Ova normalna forma se često naziva Normalna forma Čomskog.

NF za beskonteksne gramatike (2)

Svaka beskonteksna gramatika G ima ekvivalentnu gramatiku G1 u kojoj svaka smena ima jedan od sledećih oblika:

$$S \to \varepsilon, \quad A \to a, \quad A \to aB, \quad A \to aBC$$

$$A, B, C \in V_n, \quad a \in V_t$$