Strukturální rozpoznávání

Strukturální rozpoznávání

- obsah
  - hierarchický strukturální popis
  - systém strukturálního rozpoznávání
    - teorie gramatik
  - volba popisu
    - výběr primitiv
    - výběr gramatiky
  - syntaktická analýza
  - inference gramatik
    - inference kanonické regulární gramatiky
    - inference kanonické gramatiky formálních derivací
  - stochastické gramatiky
  - vliv syntaktických deformací
    - syntaktická analýza s opravou chyb

Metody rozpoznávání

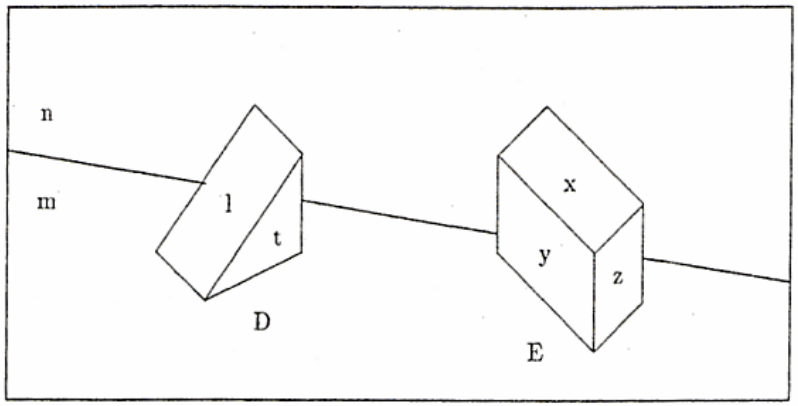
- příznakové rozpoznávání
  - vzor/obraz → bod $\boldsymbol{x} = (x_1, \ldots, x_n)$ v n-dimenzionálním příznakovém prostoru
  - klasifikace → rozdělení prostoru na části (1 část = 1 třída) při minimalizaci kriteriální funkce

- strukturální rozpoznávání
  - Pavlidis: „identifikace ideálního, podle kterého byl analyzovaný obraz stvořen“
  - obraz/vzor → popsán pomocí primitiv a vztahů mezi nimi
    - primitiva = základní popisné elementy
  - rozpoznávání → 1. klasifikace do tříd
    → 2. popis obrazů pomocí jeho částí a vztahů mezi nimi

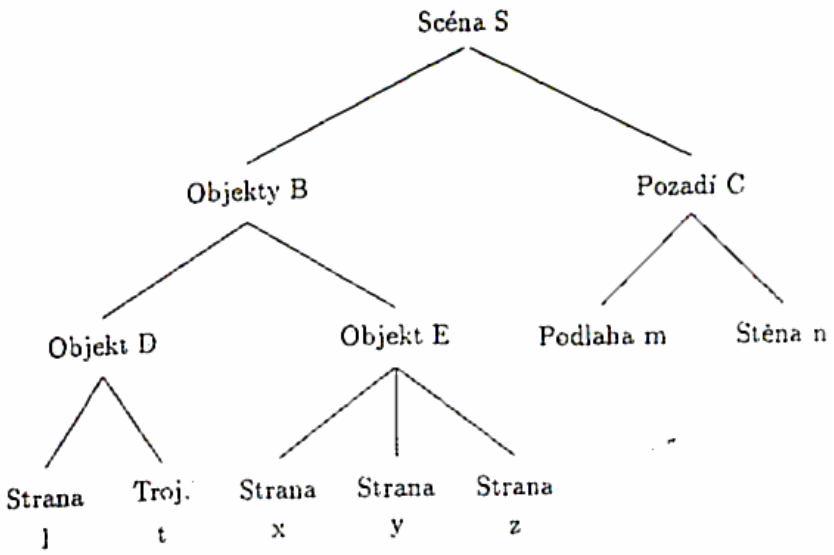
Hierarchický strukturální popis

Hierarchický strukturální popis

- obraz



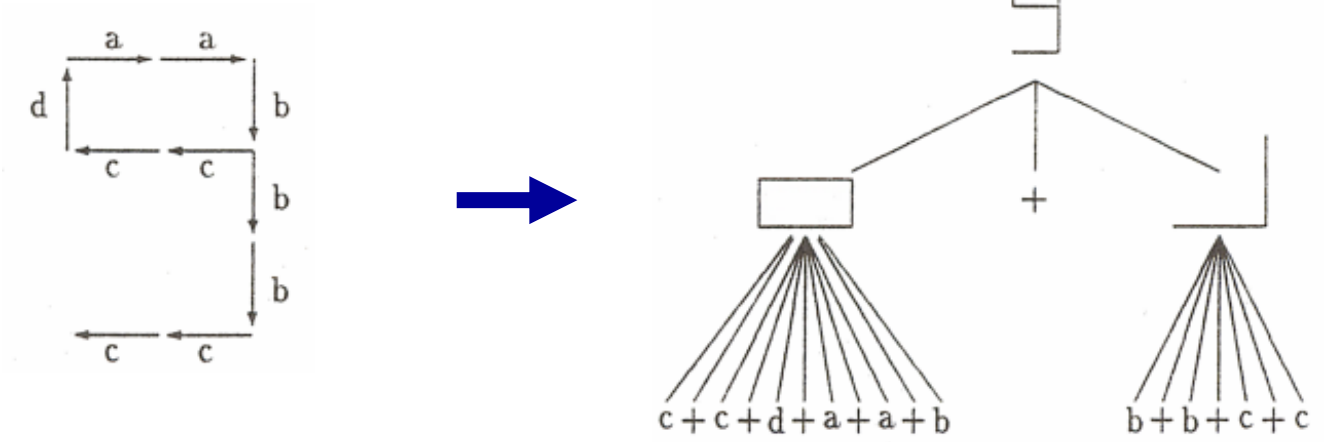
- hierarchická struktura



Hierarchický strukturální popis – příklad

- příklad
  - primitiva

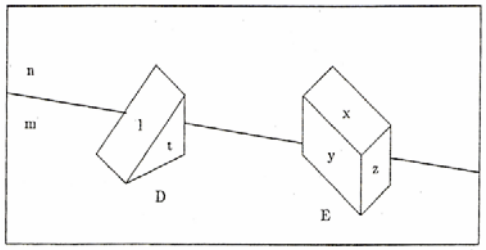
  - hierarchický strukturální popis číslice „devět“

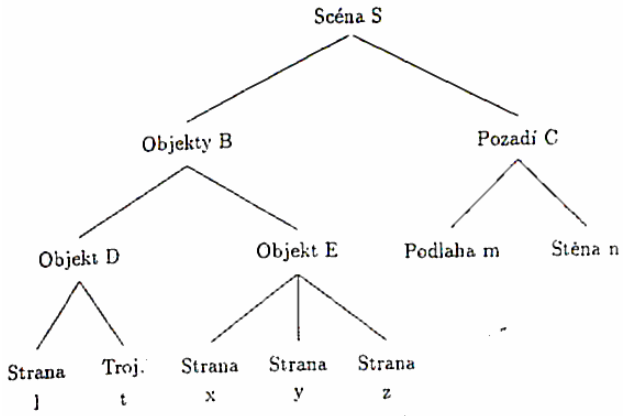


Hierarchický strukturální popis

- hierarchická strukturální informace = primitiva + syntaktická pravidla
  - primitiva = minimální kvalitivní charakteristiky
  - pravidla dekompozice = přípustné dekompozice složitých obrazů

→ teorie formálních jazyků





→ gramatika

S → BC
B → DE
D → lt
C → mn
E → EE
E → xyz

- co všechno generuje tato gramatika?

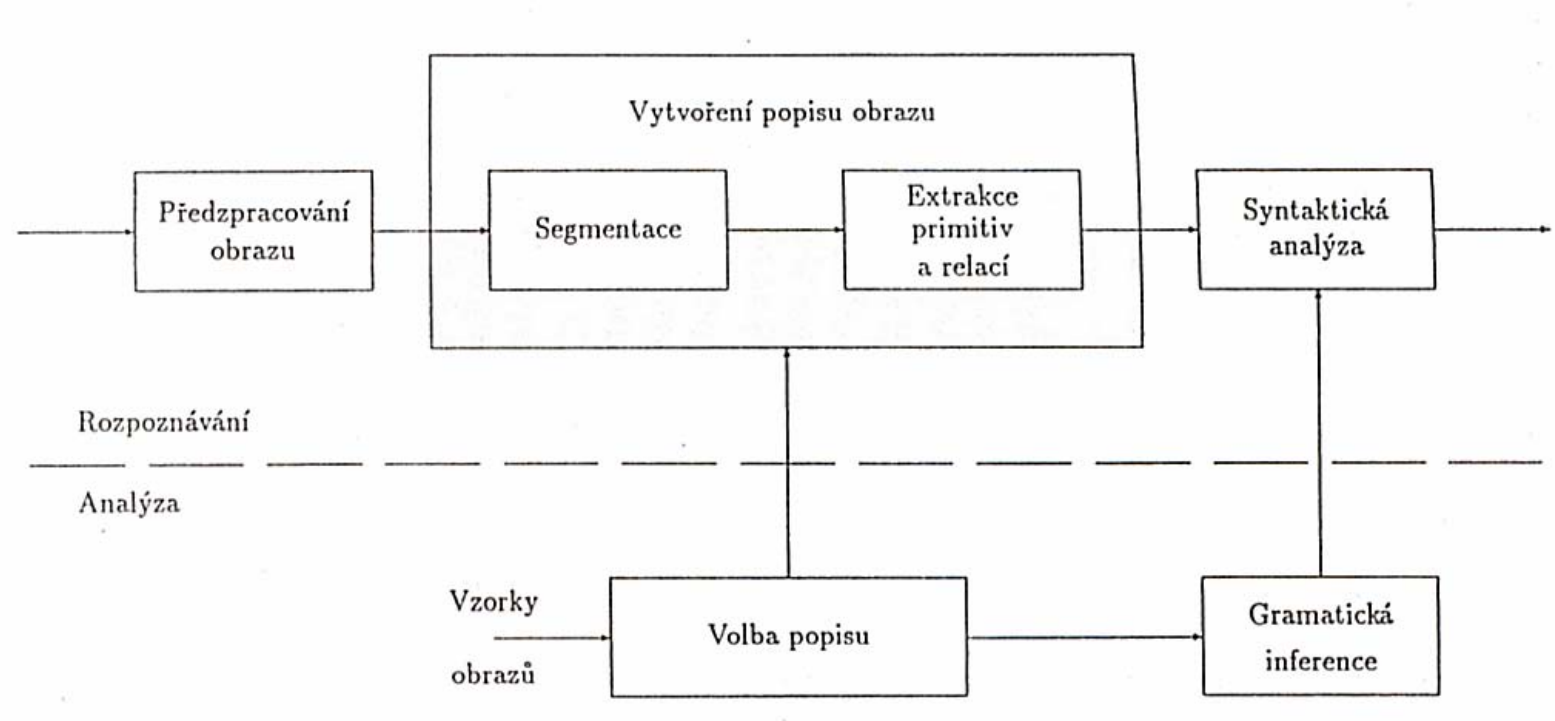
Systém strukturálního rozpoznávání

Strukturální rozpoznávání – idea

- idea strukturálního rozpoznávání
  - 1. detekce primitiv a vztahů mezi nimi
  - 2. vytvoření reprezentace vzoru → řetězec primitiv
  - 3. syntaktická analýza řetězce
    - vzor generován gramatikou? → ANO → strukturální popis vzoru

- klasifikační třídy
  - 1 třída = 1 gramatika

Systém strukturálního rozpoznávání



Teorie gramatik

Teorie gramatik – připomenutí (1)

- gramatika $G = (V_N, V_T, S, P)$
  - V_N ... množina neterminálů
  - V_T ... množina terminálů
  - S ... počáteční neterminál
  - P ... množina přepisovacích pravidel $\alpha \rightarrow \beta$
    - $\alpha, \beta \in (V_N \cup V_T)$
    - α obsahuje alespoň jeden neterminál

- gramatiky a strukturální popis
  - terminály → primitiva
  - neterminály → strukturálně jednodušší části vzoru/obrazu
  - počáteční neterminál → analyzovaný vzor

Teorie gramatik – připomenutí (2)

- Chomského hierarchie
  - bez omezení
  - kontextové gramatiky: $\alpha_1 A \alpha_2 \rightarrow \alpha_1 \beta \alpha_2$
    - $A \in V_N$
    - $\alpha_1, \alpha_2, \beta \in (V_N \cup V_T)^*$
    - $\beta \neq \lambda$
  - bezkontextové gramatiky: $A \rightarrow \beta$
    - $A \in V_N$
    - $\beta \in (V_N \cup V_T)^+$
  - regulární gramatiky: $A \rightarrow \alpha B$, $A \rightarrow \alpha$
    - $A, B \in V_N$
    - $\alpha \in V_T$
- jazyk generovaný gramatikou G
  - $L(G) = \{x \mid x \in V^*_T : S \Rightarrow^* x\}$

Volba primitiv pro popis vzorů

Výběr primitiv (1)

- požadavky
  - 1. požadovaný popis pomocí primitiv
  - 2. snadná detekovatelnost a rozpoznatelnost primitiv ve vzorech

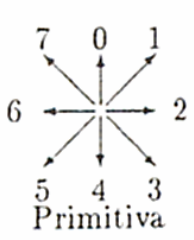
- příklad 1
  - odlišit rovnostranné trojúhelníky od ostatních obrazů

  - primitiva
    - *x* ... horizontální jednotkový segment
    - *y* ... jednotkový segment se sklonem 120°
    - *z* ... jednotkový segment se sklonem -120°

  → rovnostranné trojúhelníky
    - $L = \{x^n y^n z^n, n=1,2...\}$

Výběr primitiv (2)

- Freemanův řetězový kód
  - popis čárových obrazců, obrysů, ...
  - segment → oktalové číslo podle na sklonu segmentu

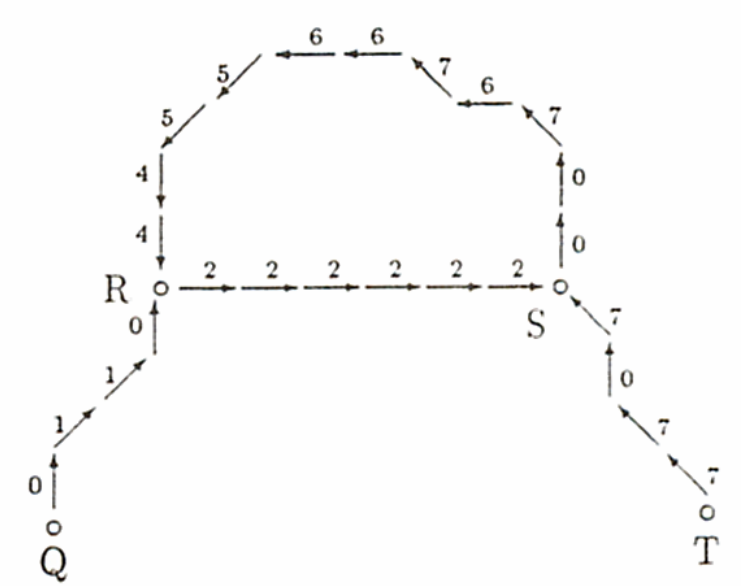


  - výhody
    - otáčení obrazu o 45° → přičtení násobku oktalových čísel ke každému číslu řetězce
    - umožňuje měření délek

- použití Freemanova kódu
  - popis strukturálních vztahů v ručně psaných písmenech

Výběr primitiv (2)

- použití Freemanova řetězového kódu
  - popis strukturálních vztahů v ručně psaných písmenech



řetězec	počáteční uzel	koncový uzel
0110	Q	R
222222	R	S
00767665544	S	R
7707	T	S

Detekce primitiv

- detekce primitiv
  - typicky se používají příznakové metody

- příklad: Giese – analýza elektroencefalogramů (EEG)
  - 100-sekundové záznamy aktivity mozku ve 4 kanálech
  - segmentace křivky po sekundě → úsek = primitivum
  - v úseku extrahováno 17 příznaků
    - dominantní frekvence, střední hodnota signálu v pásmu 0-4Hz, střední hodnota signálu v pásmu 5-7Hz, ...
  - shluková analýza
    - zjištěno 7 tříd primitiv

→ reprezentace EEG pomocí 4 řetězců po 100 primitivech

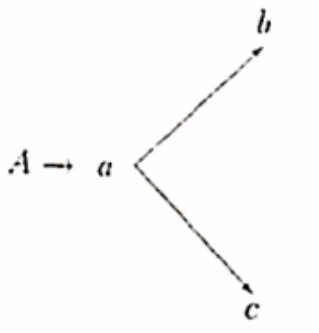
Výběr gramatiky

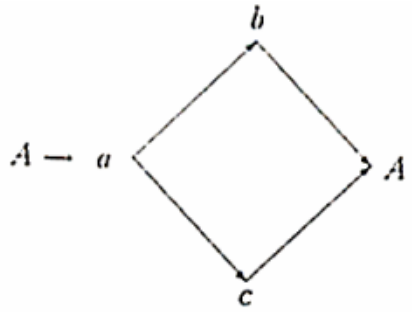
Výběr gramatiky (1)

- jednodimenzionální gramatiky
  - čím složitější jazyk
    → silnější vyjadřovací prostředek (+)
    → složitější automat → delší doba klasifikace (-)
      - gramatiky typu 0 → Turingův stroj → halting problem
  - kompromis: vyjadřovací síla jazyka × efektivnost analýzy jazyka
- vícedimenzionální gramatiky
  - pavučinové gramatiky
    - generují orientované grafy (uzly = terminály)
  - stromové gramatiky
    - generují stromy

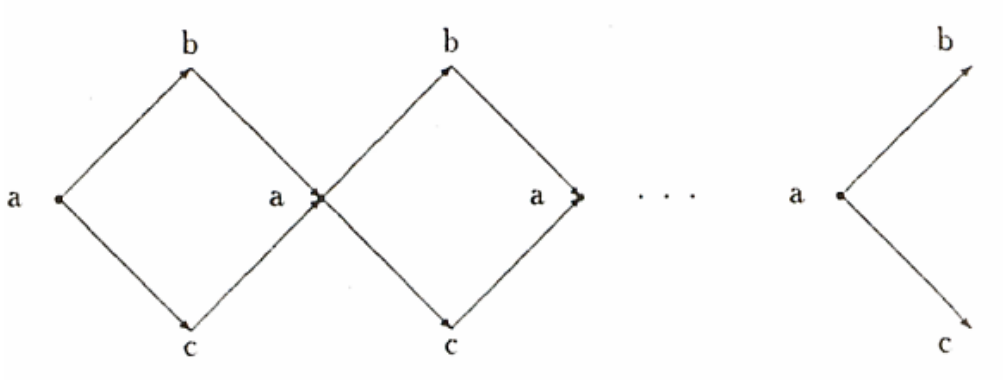
Výběr gramatiky (2)

- příklad pavučinové gramatiky
  - G = ({A}, {a,b,c}, A, P)



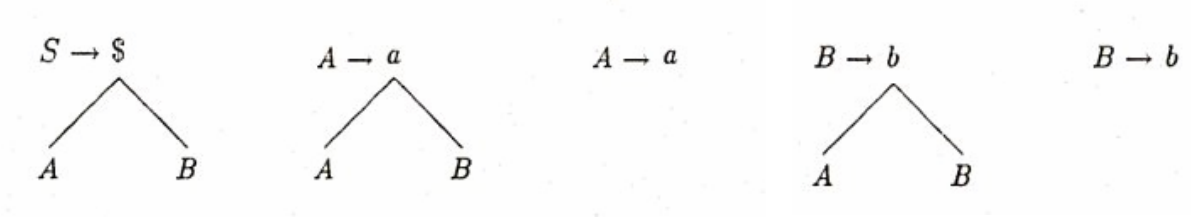


  - jazyk generovaný pavučinovou gramatikou ?

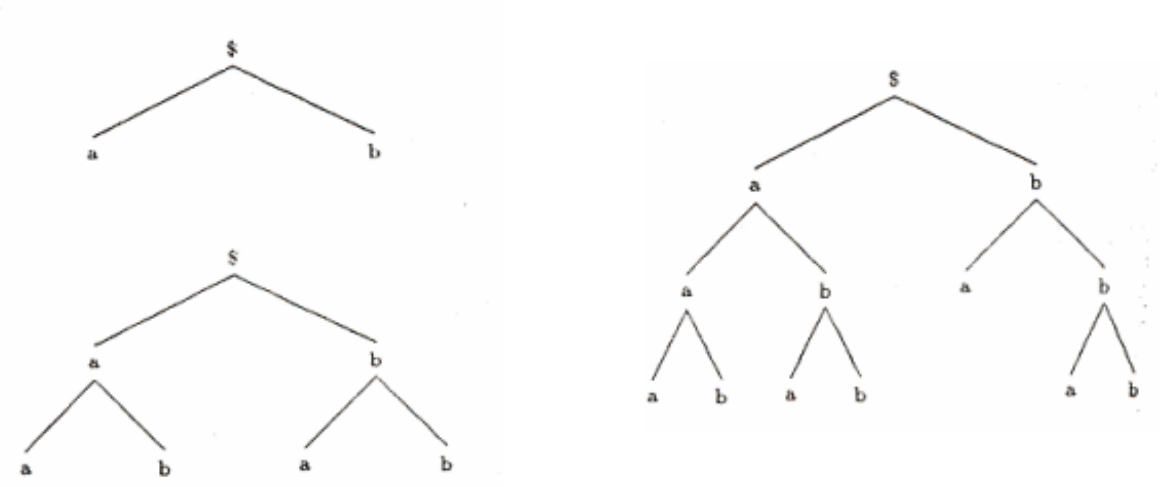


Výběr gramatiky (3)

- příklad stromové gramatiky
  - G = ({S, A, B}, {a,b,$}, S, P)



  - vygenerované stromy



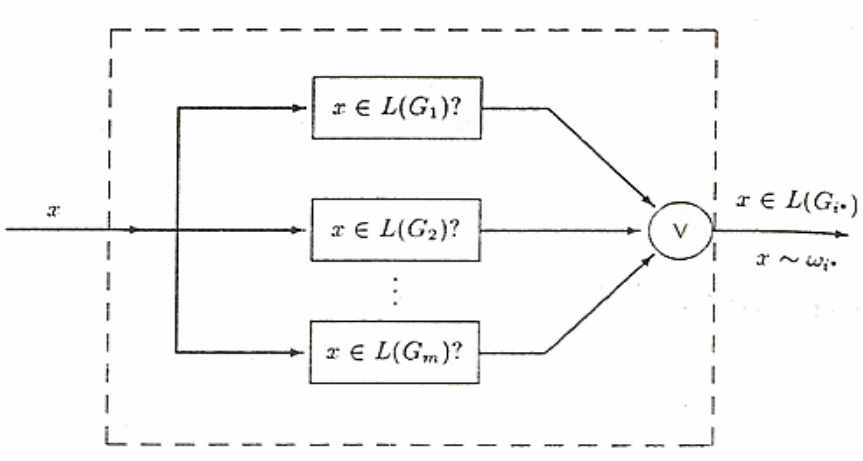
Syntaktická analýza

Syntaktická analýza (1)

- klasifikace do tříd $\omega_1, \ldots, \omega_c$
  - třída ω_i = gramatika G_i
  - řetězce generované G_i = vzory třídy ω_i

- předpoklad
  - $L(G_i) \cap L(G_k) = \emptyset$ pro $i \neq k$

- klasifikace neznámého vzoru x
  - hledání gramatiky G_{i*}: $x \in L(G_{i*})$

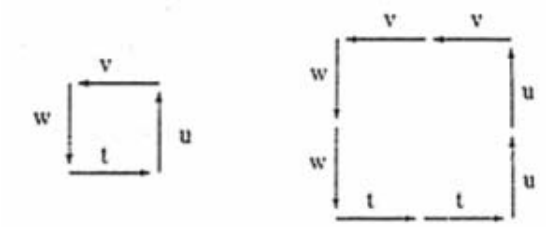


Syntaktická analýza (2)

- regulární gramatika
  - konstrukce konečného automatu
  - snadná syntaktická analýza
- bezkontextová gramatika
  - syntaktická analýza shora dolů nebo zdola nahoru
- kontextová gramatika
  - velmi složité
  - typicky náhrada kontextové gramatiky → bezkontextové gramatiky s řízeným přepisováním

Syntaktická analýza (3)

- příklad
  - $L = \{ t^n u^n v^n w^n, n \in N \}$ → kontextový jazyk



  - bezkontextová gramatika s řízeným přepisováním

č. prav.	pravidlo	úspěch	neúspěch
1	$S \rightarrow tAB$	$\{2,3\}$	$\emptyset$
2	$A \rightarrow tAC$	$\{2,3\}$	$\emptyset$
3	$A \rightarrow D$	$\{4\}$	$\emptyset$
4	$C \rightarrow w$	$\{5\}$	$\{6\}$
5	$D \rightarrow uDv$	$\{4\}$	$\emptyset$
6	$B \rightarrow w$	$\{7\}$	$\emptyset$
7	$D \rightarrow uv$	$\emptyset$	$\emptyset$

Inference gramatik

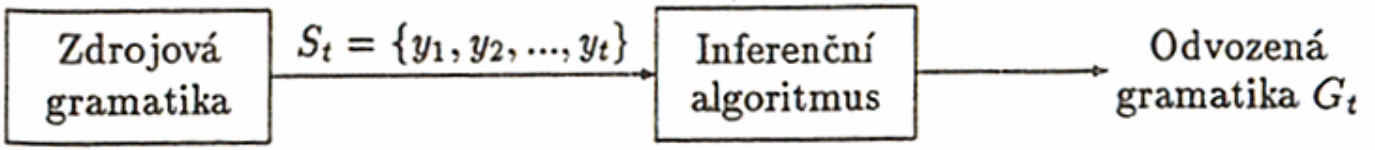
Inference gramatik

- inference gramatik
  - odvození gramatik na základě trénovacích dat

- vstup
  - trénovací množina S_t

    $S_t = S^+ \cup S^- = \{y_1^+, \ldots, y_{t+}^+\} \cup \{y_1^-, \ldots, y_{t-}^-\}$
  - S^+ … vzory patřící do L
  - S^- … vzory nepatřící do L

- gramatická inference



Inference gramatik

- odvozená gramatika G_t kompatibilní
  - pro každý $y^+ \in S^+ \rightarrow y^+ \in L(G_t)$
  - pro každý $y^- \in S^- \rightarrow y^- \notin L(G_t)$

- množina S^+ strukturálně úplná
  - každé pravidlo z neznámé zdrojové gramatiky → použito při generování alespoň jednoho vzoru z S^+

- automatická inference gramatik
  - otevřený problém
  - automatické odvození → existuje pro nejjednodušší regulární a bezkontextové gramatiky
  - obecná metoda není zatím známa

  - 1. inference kanonické regulární gramatiky
  - 2. inference kanonické gramatiky formálních derivací

Inference kanonické regulární gramatiky

Inference kanonické regulární gramatiky

- inference kanonické regulární gramatiky
  - vstup: $S^+ = \{x_1, \ldots, x_t\}$
  - výstup: regulární gramatika $G_C = (V_{NC}, V_{TC}, S, P_C)$

- Krok 1
  - najít všechny různé terminály v $S^+ \rightarrow$ vytvoří V_{TC}
- Krok 2
  - pro každý vzor $x_i = a_{i1} \ldots a_{in}$ $(x_i \in S^+)$ vytvořit pravidla

$$S \rightarrow a_{i1} Z_{i1}$$
$$Z_{i1} \rightarrow a_{i2} Z_{i2}$$
$$\ldots\ldots$$
$$Z_{i,n-2} \rightarrow a_{i,n-1} Z_{i,n-1}$$
$$Z_{i,n-1} \rightarrow a_{in}$$

každé Z_{ij} představuje nový neterminál

Inference kanonické regulární gramatiky – příklad (1)

- regulární gramatika *G* .... neznámá

  G = ({S,A,B,C}, {a,b}, S, P)

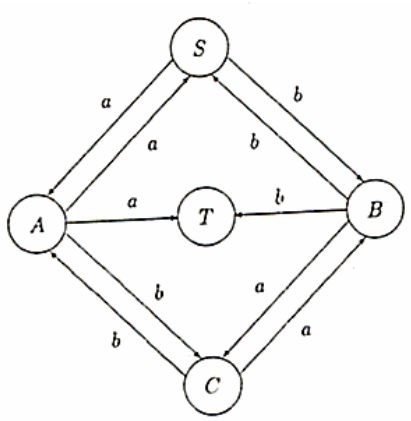
  – pravidla

  $S \rightarrow aA$ $A \rightarrow a$ $B \rightarrow b$ $C \rightarrow aB$

  $S \rightarrow bB$ $A \rightarrow aS$ $B \rightarrow bS$ $C \rightarrow bA$

  $A \rightarrow bC$ $B \rightarrow aC$

- konečný automat této gramatiky *G*



- trénovací množina

  $S^+ = \{abab, bbaa, baba, aabb\}$

  → S^+ strukturálně úplná

Inference kanonické regulární gramatiky – příklad (2)

- odvozená kanonická gramatika $G_C = (V_{NC}, V_{TC}, S, P_C)$
  - $V_{NC} = \{S, Z_1, Z_2, Z_3, Z_4, Z_5, Z_6, Z_7, Z_8, Z_9, Z_{10}, Z_{11}, Z_{12}\}$
  - $V_{NC} = \{a,b\}$
  - P_C

$S \rightarrow aZ_1$ $\quad$ $Z_1 \rightarrow bZ_2$

$S \rightarrow bZ_4$ $\quad$ $Z_4 \rightarrow bZ_5$

$S \rightarrow bZ_7$ $\quad$ $Z_7 \rightarrow aZ_8$

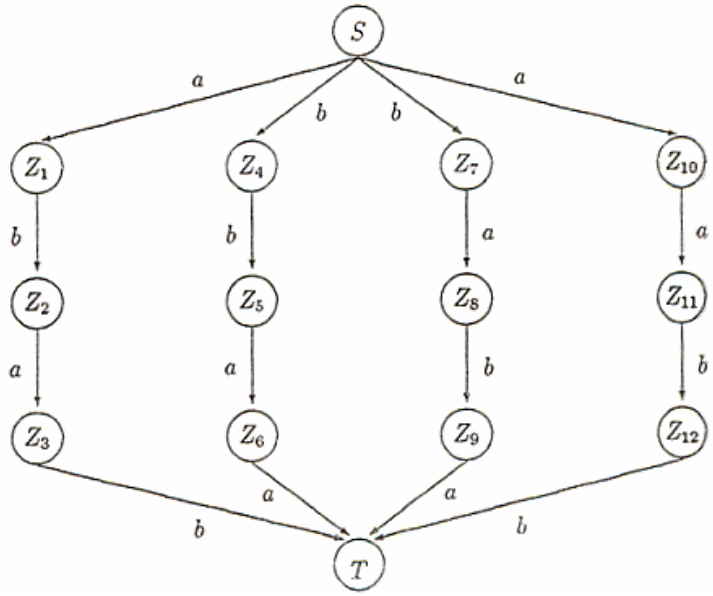
$S \rightarrow aZ_{10}$ $\quad$ $Z_{10} \rightarrow aZ_{11}$

$Z_2 \rightarrow aZ_3$ $\quad$ $Z_3 \rightarrow b$

$Z_5 \rightarrow aZ_6$ $\quad$ $Z_6 \rightarrow a$

$Z_8 \rightarrow bZ_9$ $\quad$ $Z_9 \rightarrow a$

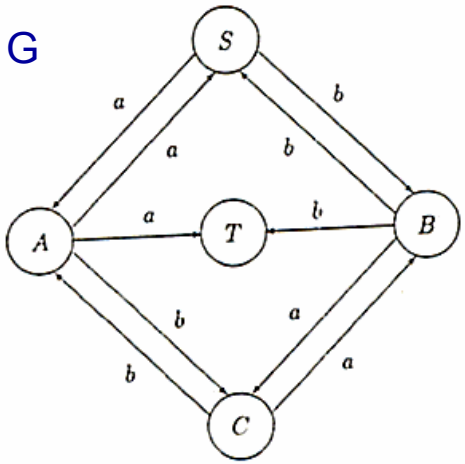
$Z_{11} \rightarrow bZ_{12}$ $\quad$ $Z_{12} \rightarrow b$



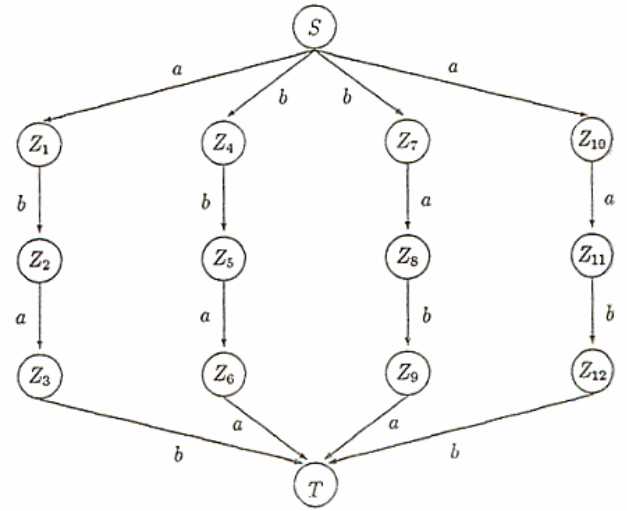
Inference kanonické regulární gramatiky – příklad (3)

- porovnání původní (neznámé) a odvozené gramatiky

G



G_C



→ obě popisují jiný jazyk

- *L(G)* nekonečný
- $L(G_C)$ konečný → generuje jen řetězce z trénovací množiny

- obecně platí
  - $L(G_C) = S^+$
  - $S^+ \subseteq L(G)$

Inference kanonické regulární gramatiky

- kanonická regulární gramatika
  - velké množství neterminálů (-)
  - některé neterminály ekvivalentní (-)
    - nevíme ale které to jsou
    - řešení
      - hledat skupiny vzájemně ekvivalentních neterminálů
      - skupinu nahradit jediným neterminálem

      → redukce neterminálů a počtu pravidel
  - generuje výhradně řetězce z S^+ a žádné jiné (-)
    - syntaktická analýze odmítne vzory strukturálně podobné trénovacím vzorům

      → gramatika neumí „zobecňovat“
    - řešení → syntaktická analýza s opravou chyb

Inference kanonické gramatiky formálních derivací

Inference kanonické gramatiky formálních derivací

- inference kanonické gramatiky formálních derivací
  - jiný algoritmus odvození gramatiky

- formální derivace množiny řetězců *A* pro symbol *a*

  $D_\lambda A = A$

  $D_a A = \{x \mid ax \in A, x \in V_T^*\}$

- vstup
  - množina řetězců S^+
- výstup
  - gramatika G_{CD}

Inference kanonické gramatiky formálních derivací – algoritmus

- Krok 1
  - vytvoření množiny $U = \{U_2, ..., U_i\}$ různých formálních derivací z S^+, které nejsou rovny λ nebo $\emptyset$
  - navíc do U přidána $U_1 = D_\lambda S^+ = S^+$

- Krok 2
  - vytvoření počátečního symbolu $S = U_1$
- Krok 3
  - vytvoření množiny terminálů V_T z trénovací množiny řetězců S^+
- Krok 4
  - vytvoření množiny neterminálů $V_N = U$
- Krok 5
  - vytvoření pravidel $\rightarrow$ pro všechna $a \in V_T$, $U_i, U_k \in V_N$

    $U_i \rightarrow aU_k$ pokud $D_a U_i = U_k$

    $U_i \rightarrow a$ pokud $D_a U_i = \{\lambda\}$

Inference kanonické gramatiky formálních derivací – příklad (1)

- vytvořit gramatiku podle trénovací množiny S^+

  S^+ = *{abab, baba, aaabba, bbabab, aaabab, bbbaba}*

- algoritmus inference kanonické regulární gramatiky

  → gramatika G_C 27 neterminálů !

- algoritmus kanonické gramatiky formálních derivací

  → konstrukce gramatiky G_{CD}

  → po konstrukci 13 neterminálů

Inference kanonické gramatiky formálních derivací – příklad (2)

- Krok 1 – vytvoření množin formálních derivací

$D_\lambda S^+ = S^+ = \{abab, baba, aaabba, bbabab, aaabab, bbbaba\} = U_1$
$D_a U_1 = \{bab, aabba, aabab\} = U_2$
$D_b U_1 = \{aba, babab, bbaba\} = U_3$
$D_a U_2 = \{abba, abab\} = U_4$
$D_a U_3 = \{ba\} = U_5$
$D_b U_2 = \{ab\} = U_6$
$D_b U_3 = \{abab, baba\} = U_7$
$D_a U_4 = \{bba, bab\} = U_8$
$D_b U_5 = \{a\} = U_9$
$D_a U_6 = \{b\} = U_{10}$
$D_a U_7 = \{bab\} = U_{11}$
$D_b U_8 = \{ba, ab\} = U_{12}$
$D_b U_7 = \{aba\} = U_{13}$
$D_b U_{10} = \{\lambda\}$
$D_a U_9 = \{\lambda\}$

$D_a U_{12} = \{b\} = U_{10}$
$D_a U_{13} = \{ba\} = U_5$
$D_b U_{12} = \{a\} = U_9$
$D_b U_{11} = \{ab\} = U_6$

Inference kanonické gramatiky formálních derivací – příklad (3)

- Krok 2 – počáteční symbol

  $S = U_1$

- Krok 3 – množina terminálů

  $V_T = \{a, b\}$

- Krok 4 – množina neterminálů

  $V_N = \{S=U_1, U_2, \ldots, U_{13}\}$

- Krok 5 – množina pravidel

pro všechna $a \in V_T$, $U_i, U_k \in V_N$

$U_i \rightarrow aU_k$ pokud $D_a U_i = U_k$

$U_i \rightarrow a$ pokud $D_a U_i = \{\lambda\}$

$S \rightarrow aU_2$	$S \rightarrow bU_3$
$U_2 \rightarrow aU_4$	$U_3 \rightarrow aU_5$
$U_2 \rightarrow aU_6$	$U_3 \rightarrow bU_7$
$U_4 \rightarrow aU_8$	$U_5 \rightarrow bU_9$
$U_6 \rightarrow aU_{10}$	$U_7 \rightarrow aU_{11}$
$U_8 \rightarrow bU_{12}$	$U_7 \rightarrow bU_{13}$
$U_{10} \rightarrow b$	$U_9 \rightarrow a$
$U_{12} \rightarrow aU_{10}$	$U_{13} \rightarrow aU_5$
$U_{12} \rightarrow bU_9$	$U_{11} \rightarrow bU_6$

Inference kanonické gramatiky formálních derivací

- kanonická gramatika G_{CD} formálních derivací
  - generuje právě množinu S^+
  - obsahuje méně neterminálů
    - než kanonická regulární gramatika
  - množina odvozených gramatik nemusí obsahovat zdrojovou (neznámou) gramatiku

Inference gramatik

- shrnutí metod inference gramatik
  - inference kanonické regulární gramatiky
  - inference kanonické gramatiky formálních derivací

  - gramatická inference založená na *k*-“koncích“ řetězců
    - využívá kanonické gramatiky formálních derivací
  - inference gramatik s informátorem
    - využívá dodatečných informací od člověka
  - heuristické metody
  - ....

Stochastické gramatiky

Stochastické gramatiky

- problém v praxi
  - vzor lze generovat gramatikami různých tříd → $L(G_i) \cap L(G_k) \neq \emptyset$
  - příčiny
    - strukturální podobnost vzorů z různých tříd
    - šum
    - nepřesnost ve předzpracování (detekce primitiv, ...)
    - nepřesnost při konstrukci gramatik

- stochastické gramatiky $G = (V_N, V_T, S, P)$
  - přepisovací pravidla $\alpha_i \xrightarrow{p_{ik}} \beta_{ik}$
    - $\alpha_i, \beta_{ik} \in (V_N \cup V_T)$
    - α_i obsahuje alespoň jeden neterminál
    - p_{ik} ... pravděpodobnost spojená s aplikací pravidla

$$0 \leq p_{ik} \leq 1 \quad \wedge \quad \sum_k p_{ik} = 1$$

  - pravď. odvození řetězce = součin pravď. použitých pravidel v odvození

Stochastické gramatiky – příklad

- příklad
  - stochastická gramatika $G = (\{S,A,B\}, \{a,b\}, S, P)$
  - přepisovací pravidla

    $S \xrightarrow{1} aA$

    $A \xrightarrow{0.7} bB$

    $A \xrightarrow{0.3} a$

    $B \xrightarrow{0.4} b$

    $B \xrightarrow{0.6} aS$

  - pravděpodobnost odvození řetězce *abaabb*

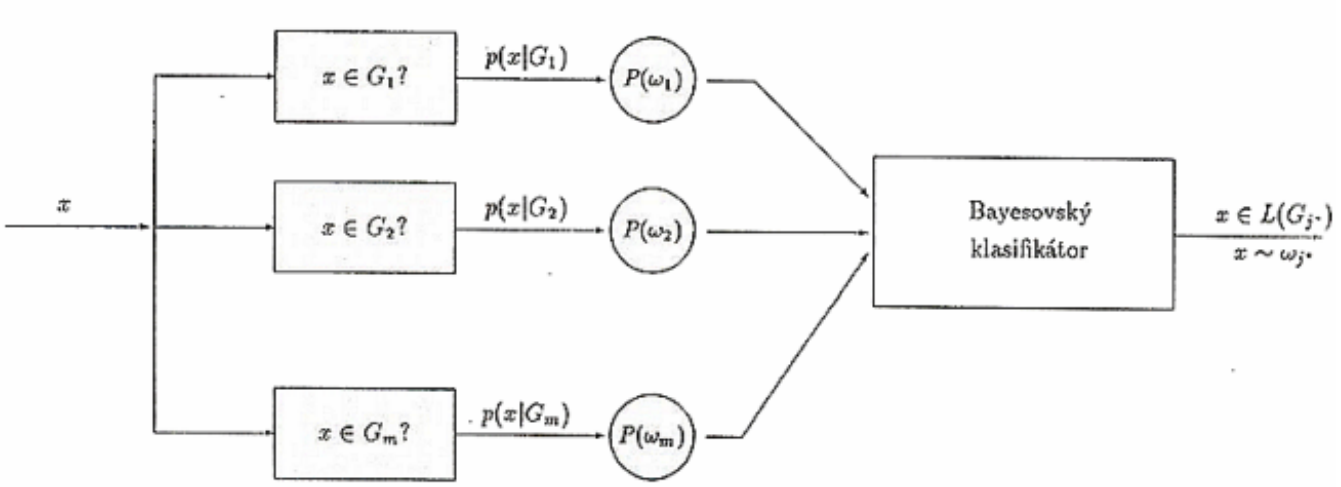
    $S \Rightarrow aA \Rightarrow abB \Rightarrow abaS \Rightarrow abaaA \Rightarrow abaabB \Rightarrow abaabb$

    $p(abaabb) = 1 \cdot 0{,}7 \cdot 0{,}6 \cdot 1 \cdot 0{,}7 \cdot 0{,}4 = 0{,}1176$

Klasifikace pomocí stochastických gramatik

- klasifikace
  - v určité třídě se vzory vyskytují častěji
    - nechtěné řetězce → malé pravděpodobnosti
  - řešení problému nejednoznačnosti
    - vybrán derivační strom s největší pravděpodobností

- stochastický klasifikátor



Stanovení pravděpodobností pravidel

- vstup
  - trénovací množina vzorů $S_t = \{ (x_1,f_1), \ldots, (x_t,f_t) \}$
- výstup
  - (odhady) pravděpodobnosti jednotlivých pravidel

f_L ... pravď. výskytu vzoru x_L ve třídě

- postup pro bezkontextovou gramatiku
  - bezkontextová gramatika G s pravidly $A_i \xrightarrow{p_{ik}} \gamma_{ik}$

$A_i \in V_N$ a $\gamma_k \in (V_N \cup V_T)^+$

Krok 1 – syntaktická analýza všech vzorů x_L z S_t

Krok 2 – pro vzor x_L a pravidlo $A_i \rightarrow \gamma_k$ zjištění absolutní četnosti $N_{ik}(x_L)$ pravidla při derivaci x_L

Krok 3 – očekávaná absolutní četnost výskytu pravidla $A_i \rightarrow \gamma_k$ při analýze celé S_t

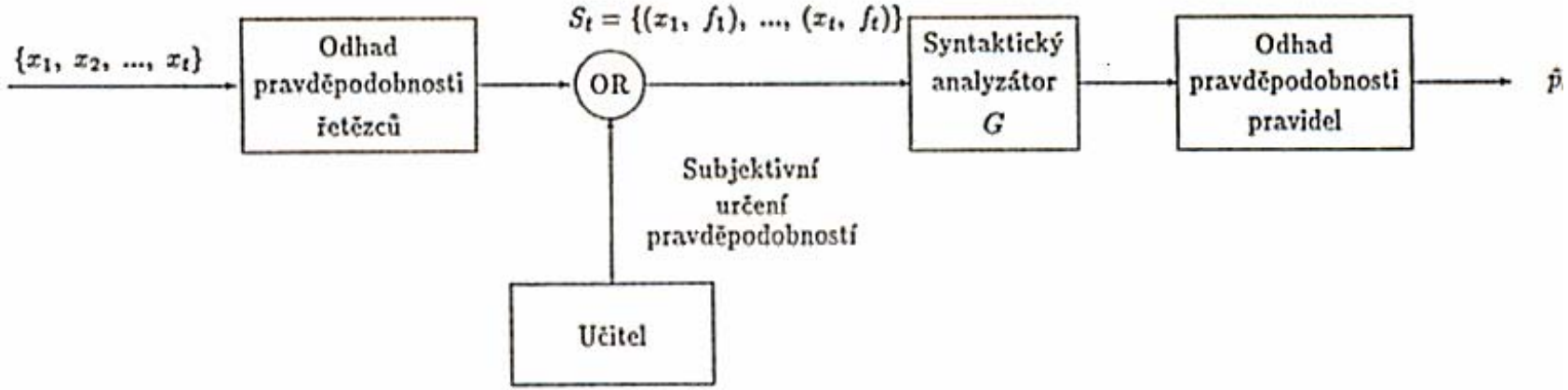
$$n_{ik} = \sum_{x_k \in S_t} f_k \cdot N_{ik}(x_k)$$

Krok 4 – odhad pravděpodobnosti p_{ik} pravidla $A_i \rightarrow \gamma_k$

$$\hat{p}_{ik} = \frac{n_{ik}}{\sum_j n_{ij}}$$

Stanovení pravděpodobností pravidel

- systém pro odvození pravděpodobností jednotlivých pravidel



- platí

$$\hat{p}_{ik} \rightarrow p_{ik} \quad \text{pro} \quad S_t \rightarrow L(G)$$

Stanovení pravděpodobností pravidel – příklad (1)

- určit pravděpodobnosti stochastické gramatiky
  - *G = ({S,X}, {a,b,c,d}, S, P)*
  - přepisovací pravidla

    $S \xrightarrow{p11} aXS$

    $S \xrightarrow{p12} cX$

    $X \xrightarrow{p21} d$

    $X \xrightarrow{p22} bX$
  - trénovací data – 100 vzorů vygenerovaných gramatikou

Řetězec x_k	Absolutní četnost	Relativní četnost
adcd	9	0,09
cd	77	0,77
adcbd	2	0,02
cbd	6	0,06
abbdadcbd	1	0,01
abdcd	2	0,02
abdadadcd	1	0,01
adadcd	2	0,02

Stanovení pravděpodobností pravidel – příklad (2)

- Krok 1
  - syntaktická analýza všech vzorů

- Krok 2
  - absolutní četnosti $N_{ik}(x_L)$ výskytu pravidel při odvození řetězců

Řetězec	Počet výskytů pravidla				Četnost řetězce
	$S \rightarrow aXS$	$S \rightarrow cX$	$X \rightarrow d$	$X \rightarrow bX$	
adcd	1	1	2	0	9
cd	0	1	1	0	77
adcbd	1	1	2	1	2
cbd	0	1	1	1	6
abbdadcbd	2	1	3	3	1
abdcd	1	1	2	1	2
abdadadcd	3	1	4	1	1
adadcd	2	1	3	0	2

Stanovení pravděpodobnosti pravidel – příklad (3)

- Krok 3
  - absolutní četnosti výskytu pravidel při generování celé S_t
  - př. $S \rightarrow aXS$

    $1 \cdot 9 + 0 \cdot 77 + 1 \cdot 2 + 0 \cdot 6 + 2 \cdot 1 + 1 \cdot 2 + 3 \cdot 1 + 2 \cdot 2 = 22$

Řetězec	Počet výskytů pravidla				Četnost řetězce
	$S \rightarrow aXS$	$S \rightarrow cX$	$X \rightarrow d$	$X \rightarrow bX$	
adcd	1	1	2	0	9
cd	0	1	1	0	77
adcbd	1	1	2	1	2
cbd	0	1	1	1	6
abbdadcbd	2	1	3	3	1
abdcd	1	1	2	1	2
abdadadcd	3	1	4	1	1
adadcd	2	1	3	0	2

  - výsledné četnosti pravidel

Pravidlo	$S \rightarrow aXS$	$S \rightarrow cX$	$X \rightarrow d$	$X \rightarrow bX$
Použito celkem	22	100	122	14

- Krok 4
  - odhady pravděpodobností

$$\hat{p}_{11} = \frac{22}{100+22} = 0.18 \quad \hat{p}_{12} = \frac{100}{100+22} = 0.82$$

$$\hat{p}_{21} = \frac{122}{122+14} = 0.9 \quad \hat{p}_{12} = \frac{14}{122+14} = 0.1$$

Stochastická syntaktická analýza

- stochastická syntaktická analýza
  - možnost volby z více pravidel
    → volba pravidla s největší pravděpodobností
    → snížení pravděpodobnosti chybné klasifikace
  - existují speciální algoritmy na urychlení procesu stochastické syntaktické analýzy

Vliv syntaktických deformací

Řešení vlivu syntaktických deformací

- deformace ve struktuře vzorů
  - vliv nepřesností/poruch

    → vzor x reprezentován vzorem x' $(x \neq x')$

  - deformace řetězců
    - vložení symbolu (terminálu)
    - vynechání symbolu (terminálu)
    - záměna symbolu

    → míra podobnosti x a x' → editační vzdálenost

- problém
  - při syntaktické analýze odmítnuty deformované vzory
    - i když strukturálně podobné trénovacím vzorům

- řešení
  - syntaktická analýza s opravou chyb

Syntaktická analýza s opravou chyb

- vstup
  - gramatika *G* třídy
  - deformovaný vzor *y*
    - deformace: vložení/vypuštění/záměna symbolu
- výstup
  - nalezení nejpravděpodobnější reprezentace (opravy) deformovaného vzoru *y*

    → nalezení $x \in L(G)$ kde $edit(x,y) = min\ \{edit(z,y), z \in L(G)\}$

- Krok 1
  - rozšíření gramatiky *G* o deformační pravidla → *G'*
    - *G'* generuje *L(G)* + deformované řetězce
- Krok 2
  - syntaktický analyzátor s opravou chyb pracuje s *G'*
    - hledání derivace deformovaného vzoru *y* s nejmenším počtem deformačních pravidel
- Krok 3
  - řetězec *x* (= nejlepší oprava *y*) se získá z derivace *y* vypuštěním deformačních pravidel

Konstrukce rozšířené gramatiky

- konstrukce rozšířené gramatiky *G'*
  - vstup: bezkontextová $G = (V_N, V_T, S, P)$
  - výstup: rozšířená $G' = (V'_N, V'_T, S', P')$

- Krok 1

  $V'_N = V_N \cup \{S'\} \cup \{E_b \text{ pro } b \in V_T\}$

  $V'_T = V_T$

  $\alpha_i \in V_N^*$ $b_i \in V_T$

- Krok 2
  - v P pravidlo $A \rightarrow \alpha_0 b_1 \alpha_1 \ldots b_m \alpha_m \Rightarrow$ do P' se přidá pravidlo $A \rightarrow \alpha_0 E_{b1} \alpha_1 \ldots E_{bm} \alpha_m$
- Krok 3
  - do P' se přidají nová pravidla (D … deformační pravidlo)

	$S' \rightarrow S$	
(D)	$S' \rightarrow Sa$	pro $a \in V_T$
	$E_a \rightarrow a$	pro $a \in V_T$
(D)	$E_a \rightarrow b$	pro $a \in V_T$, $b \in V_T$, $b \neq a$
(D)	$E_a \rightarrow \lambda$	pro $a \in V_T$
(D)	$E_a \rightarrow bE_a$	pro $a \in V_T$, $b \in V_T$

Syntaktická analýza s opravou chyb

- syntaktická analýza podle rozšířené gramatiky *G‘*
  - více pravidel
  - nejednoznačná
    - deformovaný řetězec lze odvodit více způsoby

→ klasické algoritmy syntaktické analýzy typicky neefektivní

Syntaktická analýza s opravou chyb – praxe

- velká a dostatečně reprezentovaná trénovací množina
  - trénovací množina musí obsahovat i zašuměná data

  → stochastická gramatika dostatečně spolehlivě reprezentuje vzory

  → syntaktická analýza bez opravy chyb

- trénovací množina málo reprezentativní
  - gramatika necharakterizuje i deformované vzory

  → syntaktická analýza bez opravy chyb nepřijme deformované vzory

  → nutná syntaktická analýza s opravou chyb

- syntaktická analýza s opravou chyb
  - rozpozná i zašuměné vzory (+)
  - výrazně delší doba na zpracování (-)
  - vylepšení
    - paralelizace
    - speciální algoritmy stochastické syntaktické analýzy