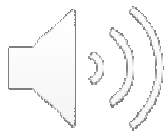
FORMÁLNÍ LINGVISTIKA



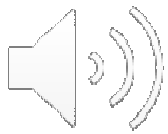
ZÁKLADNÍ POJMY

Abeceda je konečná množina symbolů. Např. $\{a, b\}$.

Slovo je libovolná konečná posloupnost prvků Σ, např. $aabab$.

Délka slova je velikost této posloupnosti, např. $|aabab| = 5$.

Prázdné slovo je slovo nulové délky, značíme je ϵ.



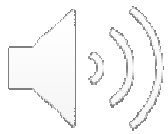
ZÁKLADNÍ POJMY

Σ* je množina všech slov nad abecedou Σ, např. $\{a, b\}^* = \{\epsilon, a, b, aa, bb, ab, ba, aab, abb, ...\}$.

Operace zřetězení slov, značíme tečkou (.), je pro dvě slova u a v definována jako $u \,.\, v = uv$, např. $aab \,.\, ab = aabab$.

Mocnina slova je pro slovo u a přirozené číslo i značena ui a je definována induktivně: $u0 = \epsilon$, $ui{+}1 = u.ui$ Např. $(ab)3 = ababab$.

Jazyk je množina některých slov nad danou abecedou, tedy pro každý jazyk L platí $L \subseteq \Sigma^*$



FORMÁLNÍ GRAMATIKA

Formální gramatika přepisovací systém, jímž lze vygenerovat slova jazyka.

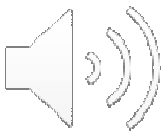
Formálně, gramatika je čtveřice (N, Σ, P, S), kde:

N je množina neterminálů (značíme nejčastěji velkými písmeny)

Σ je množina terminálů (symbolů abecedy, značíme malými písmeny), je disjunktní s množinou N a $N \cup \Sigma$ označujeme jako V (množina všech symbolů)

P je množina pravidel, tj. množina dvojic, kde prvním prvkem je řetězec obsahující alespoň jeden neterminál a druhým prvkem je libovolný řetězec.

S je počáteční symbol gramatiky



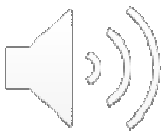
GRAMATIKA

Pravidla gramatiky jako dvojice řetězců (slov nad množinou V) (α, β) zapisujeme jako $\alpha \rightarrow \beta$. α nazýváme levou stranou pravidla a musí obsahovat alespoň jeden neterminál. β nazýváme pravou stranou pravidla.

Gramatika je modelem, kterým lze generovat slova jazyka:

začneme z počátečního symbolu gramatiky S a pravidla gramatiky používáme jako přepisovací systém, to znamená, že v jednom kroku přepisu můžeme nahradit některý řetězec terminálů a neterminálů, který je současně na levé straně nějakého pravidla, pravou stranou tohoto pravidla. Tento postup opakujeme, dokud nedostaneme řetězec terminálních symbolů (čili slovo nad Σ).

Tomuto procesu říkáme odvození slova z gramatiky.



GRAMATIKA

Řekneme, že gramatika G generuje jazyk L, pokud existuje odvození každého slova jazyka L z gramatiky G. Jazyk generovany gramatikou G značíme většinou $L(G)$.

Příklad!

Mějme gramatiku (N, Σ, P, S), kde

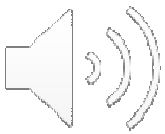
$\Sigma = \{a, b\}$

$N = \{S, A\}$

$P = \{ S \rightarrow A,\ A \rightarrow AA,\ A \rightarrow a \}$

$S \Rightarrow A \Rightarrow a$

$S \Rightarrow A \Rightarrow AA \Rightarrow aA \Rightarrow aAA \Rightarrow aaA \Rightarrow aaa$



CHOMSKÉHO HIERARCHIE JAZYKŮ

Gramatika typu 0 neklade žádná omezení na množinu pravidel, libovolná gramatika je gramatikou typu 0.

Gramatika typu 1 neboli kontextová gramatika klade na všechna pravidla $\alpha \rightarrow \beta$ podmínku $|\alpha| \leq |\beta|$, tedy levá strana každého pravidla musí být kratší než jeho pravá strana. Výjimkou z tohoto omezení je pravidlo $S \rightarrow \epsilon$, které může být přítomno.

Gramatika typu 2 neboli bezkontextová gramatika má všechna pravidla ve tvaru $A \rightarrow \beta$ (tak, že $A \in N$), tedy na levé straně je vždy právě jeden neterminál a β je neprázdné. Výjimkou z tohoto omezení je pravidlo $S \rightarrow \epsilon$, které může být přítomno.

Gramatika typu 3 neboli regulární gramatika má všechna pravidla ve tvaru $A \rightarrow aB$ nebo $A \rightarrow a$, kde A,B jsou neterminály a a je terminál. Výjimkou z tohoto omezení je pravidlo $S \rightarrow \epsilon$, které může být přítomno.