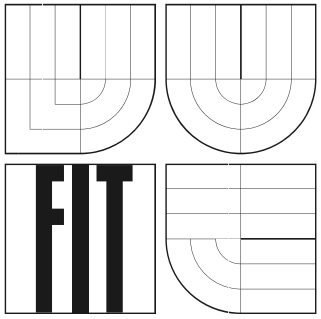
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ



Syntaktická analýza založená na bezkontextových gramatikách nad volnou grupou

Semestrální projekt

2007 **Zdeněk Melkes**

Syntaktická analýza založená na bezkontextových gramatikách nad volnou grupou

Prohlášení

Prohlašuji, že jsem tuto semestrální práci vypracoval samostatně pod vedením ing. Petra Blatného.
Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

……
Zdeněk Melkes
3. 1. 2007

Abstrakt

Tento dokument se zabývá možnostmi syntaktické analýzy založené na bezkontextových gramatikách nad volnou grupou. Práce obsahuje úvod do problematiky gramatik, jazyků a volných grup a stručný přehled způsobů syntaktické analýzy.

Klíčová slova

Bezkontextové gramatiky, gramatiky typu 2, gramatiky bez omezení, gramatiky typu 0, volná grupa, bezkontextové gramatiky nad volnou grupou, syntatická analýza shora dolů, syntaktická analýza zdola nahoru

Poděkování

Velmi rád bych poděkoval svému vedoucímu, ing. Petru Blatnému, za materiály a především za mimořádnou ochotu při spolupráci, díky které jsem měl možnost práci dokončit.

Abstract

This work handle with possibilities of syntax analysis based on context – free grammars over free group. The work contains introduction to problem of grammars, languages and free groups and short view of syntax analysis.

Keywords

Context – free grammars, grammars type 2, unrestricted grammars, grammars type 0, free group, context – free grammars over free group, sytax analysis up top down, syntax analysis bottom up

Obsah

1 Úvod

Cílem této práce je stručně uvést do problematiky formálních jazyků a především definovat vhodnou strukturu, která by byla schopna generovat celou třídu rekurzivně vyčíslitelných jazyků. Základem použitým pro tuto strukturu je bezkontextová gramatika, rozšířená o algebraickou strukturou, kterou je volná grupa nad totální abecedou této gramatiky. Tato práce má sloužit především jako teoretický základ pro pochopení dané problematiky a pro další rozšíření a implementaci daného problematiky.

2 Základy teorie jazyků

Pro pochopení pozdějšího výkladu je třeba pro začátek definovat přinejmenším základní pojmy teorie jazyků, kterými jsou abeceda, řetězec, jazyk nad abecedou a gramatika.

2.1 Abeceda

Abecedou se rozumí libovolná konečná množina Σ, jejíž prvky se nazývají znaky (případně také symboly nebo písmena) abecedy.

2.2 Řetězec nad abecedou

Řetězec (nebo také slovo) nad abecedou Σ je libovolná konečná posloupnost znaků této abecedy. Prázdné posloupnosti znaků odpovídá tzv. prázdný řetězec, označovaný jako ε.

Délkou řetězce rozumíme celé nezáporné číslo, které udává počet symbolů v řetězci. Délku řetězce x značíme jako | x |. Je-li $x = a_1 \ldots a_n$, kde $a_i \in \Sigma$ pro i=1,2 … n, potom | x | = n. Délka prázdného řetězce je rovna nule.

2.3 Jazyk nad abecedou

Množina všech řetězců nad abecedou Σ se značí jako Σ^*, množina všech neprázdných řetězců pak jako Σ^+, tedy Σ^* je $\Sigma^+ \cup \{\varepsilon\}$. Jazykem L nad abecedou Σ je libovolná množina slov nad Σ. Pro množinu L platí, že $L \subseteq \Sigma^*$.

Teď jsou uvedeny a vysvětleny nejzákladnější pojmy, potřebné pro jazyk nad abecedou. Dále je zapotřebí konečným způsobem reprezentovat jazyk. Nejjednodušším možným způsobem by byl v případě konečného jazyka pouhý přímý výčet všech slov daného jazyka. Tento způsob je ale problematický i u popisu rozsáhlejších konečných jazyků a zcela selhává při popisování nekonečných jazyků. Proto se nejčastěji jako konečná reprezentace konečných i nekonečných jazyků používá gramatika.

3 Gramatiky

Gramatika je konečná reprezentace jazyků. Gramatika je čtveřice $G = (N, \Sigma, P, S)$, kde

- N je neprázdná konečná množina neterminálních symbolů.
- Σ je konečná množina terminálních symbolů taková, že $N \cap \Sigma = \varnothing$. $N \cup \Sigma$ je množina všech symbolů gramatiky, která je obvykle označována jako V.
- $P \subseteq (N \cup \Sigma)^* N (N \cup \Sigma)^* \times (N \cup \Sigma)^*$ je konečná množina přepisovacích pravidel. Přepisovací pravidlo (α, β) se obvykle zapisuje ve tvaru $\alpha \rightarrow \beta$. Řetězec α se nazývá levou, řetězec β pravou stranou přepisovacího pravidla. Podmínka kladená na tvar pravidel $\alpha \rightarrow \beta$ pouze požaduje, aby α obsahovala alespoň jeden neterminál. Na β nejsou obecně kladeny žádné podmínky.
- $S \in N$ je speciální výchozí (počáteční) symbol gramatiky, nazývaný také kořen gramatiky.

3.1 Derivace v gramatice

Nechť $G = (N, \Sigma, P, S)$ je gramatika a nechť λ , μ jsou řetězce z $(N \cap \Sigma)^*$. Mezi λ a μ platí binární relace $\Rightarrow_G$, zvaná přímá derivace, můžeme-li řetězce λ a μ vyjádřit ve tvaru

$\lambda = \gamma \alpha \delta$

$\mu = \gamma \beta \delta$

$\gamma, \delta \subseteq (N \cup \Sigma)^*$ a $\alpha \rightarrow \beta$ je některé z přepisovacích pravidel z P. Pokud mezi řetězci λ a μ platí relace přímé derivace, pak píšeme $\lambda \Rightarrow \mu$ a říkáme, že řetězec μ lze přímo generovat z řetězce λ v uvažované gramatice G.

Nechť $G = (N, \Sigma, P, S)$ je gramatika a $\Rightarrow$ relace přímé derivace na $(N \cup \Sigma)^*$. Relace $\Rightarrow^+$ pak označuje tranzitivní uzávěr relace $\Rightarrow$ a nazývá se relací derivace. Platí-li $\lambda \Rightarrow^+ \mu$, pak existuje posloupnost

$\lambda \Rightarrow v_0 \Rightarrow v_1 \ldots \Rightarrow v_n = \mu, n \geq 1$

která se nazývá derivací délky n.

Relace $\Rightarrow^*$ označuje reflexivní a tranzitivní uzávěr relace $\Rightarrow$:

$\lambda \Rightarrow^* \mu \quad \Rightarrow \quad \lambda \Rightarrow^+ \mu, \quad$ nebo $\quad \lambda = \mu$

3.2 Jazyk nad gramatikou

Nechť $G = (N, \Sigma, P, S)$ je gramatika. Řetězec α je podmnožinou $(N \cup \Sigma)^*$ se nazývá větnou formou, právě tehdy, když platí $S \Rightarrow^* \alpha$, tzn. řetězec α je generovatelný z počátečního symbolu S. Větná forma, která obsahuje pouze terminální symboly se nazývá věta. Jazyk L(G), generovaný gramatikou G, je definován množinou všech vět

$$L(G) = \{ w \mid S \Rightarrow^* w, \text{ kde } w \in \Sigma^* \}$$

generovatelných z výchozího symbolu S gramatiky G.

3.3 Chomského hierarchie gramatik

Chomského hierarchie gramatik (a jazyků) dělí gramatiky do čtyř skupin (typů) na základě různých omezení, kladených na tvar přepisovacích pravidel. Tyto typy jsou označeny jako typ 0, typ 1, typ 2 a typ 3. Dále popíšeme jen typy, kterými se v této práci budeme zabývat.

Typ 0 – obecné gramatiky

Na gramatiky typu 0 nejsou kladena žádná omezení (vyjma výše uvedeného požadavku na nejméně jeden neterminální symbol na levé straně přepisovacího pravidla). Tyto gramatiky se označují také jako gramatiky bez omezení nebo frázové gramatiky.

Typ 2 – bezkontextové gramatiky

Gramatika je typu 2 v případě, jestliže každé její pravidlo je ve tvaru $A \rightarrow \alpha$, kde $| \alpha | \geq 1$ s eventuální výjimkou pravidla $S \rightarrow \varepsilon$, pokud se S nevyskytuje na pravé straně žádného přepisovacího pravidla.

3.4 Penttonenova normální forma

Gramatiky typu 0 mají ve své základní podobě velké množství různých tvarů přepisovacích pravidel. Proto je vhodné zavést pro větší přehlednost gramatiky a usnadnění práce s ní jednotný omezený tvar přepisovacích pravidel. Tím je právě Penttonenova normální forma.

Gramatika G = (N, Σ, P, S) je v Penttonenově normální formě, pokud její množina přepisovacích pravidel P obsahuje pouze pravidla tvaru:

- $AB \rightarrow AC$
- $A \rightarrow BC$
- $A \rightarrow a$
- $A \rightarrow \varepsilon$

kde A, B, C ∈ N, a ∈ T a ε je prázdný řetězec. Každou gramatiku G typu 0 je možné transformovat na ekvivalentní gramatiku H v Pettonenově normální formě tak, že L(G) = L(H). Důkaz správnosti transformace viz [3].

3.5 Chomského normální forma

Podobně jako gramatiky typu 0 mohou i gramatiky typu 2 nabývat množství různých tvarů. Pro větší přehlednost a pro pozdější využití zavedeme jednotný omezený tvar přepisovacích pravidel pro bezkontextové gramatiky. Tím je právě Chomského normální forma.

Gramatika G = (N, Σ, P, S) je v Chomského normální formě, pokud její množina přepisovacích pravidel P obsahuje pouze pravidla tvaru:

- $A \rightarrow BC$
- $A \rightarrow a$
- $S \rightarrow \varepsilon$ v případě, že ε ∈ L(G)

kde A, B, C ∈ N, a ∈ T a ε je prázdný řetězec. Každou gramatiku G typu 2 je možné transformovat na ekvivalentní gramatiku H v Chomského normální formě tak, že L(G) = L(H). Důkaz správnosti transformace viz [1].

4 Základní algebraické struktury

Pro pochopení později definovaného pojmu volná grupa je zapotřebí uvést alespoň stručný přehled a popis některých základních algebraických struktur, se kterými budeme nadále pracovat.

4.1 Finitární algebraické operace

Základním pojmem v algebraických strukturách je pojem algebraické operace. Mějme dáno celé nezáporné číslo n a množinu M. Řekneme, že na množině M je definována n−ární algebraická operace •, když každé uspořádané n−tici prvku $a_1, a_2, \ldots, a_n \in M$ je přiřazen jednoznačně určený

prvek z M. Tento prvek nazveme výsledkem operace • na dané prvky a budeme jej označovat symbolem $a_1\ a_2\ \ldots\ a_n\bullet$. Číslo n nazýváme aritou operace •.

- Nulární operací na množině M nazýváme zobrazení $\bullet : M^0 \to M$, jehož výsledkem je jeden z prvku množiny M, který není závislý na volbě prvků z M. Tuto operaci můžeme chápat jako vyčlenění jednoho význačného prvku z množiny M.
- Unární operací na množině M nazýváme zobrazení $\bullet : M \to M$, které každému prvku $a \in M$ přiřazuje právě jeden prvek $a\bullet \in M$.
- Binární operací na množině M nazýváme zobrazení $\bullet : M^2 \to M$. Prvek $a\bullet b$, kde $a, b, a\bullet b \in M$ nazýváme kompozicí prvku a, b vzhledem k binární operaci na množině M.

Podle základních požadavku kladených na binární operace na dané množině M dostáváme objekty s různými algebraickými vlastnostmi. Tyto objekty nazýváme algebraické struktury. Binární operace na určité množině, které splňují axiómy příslušné algebraické struktury, nazýváme operacemi této struktury.

4.2 Axiomy binárních operací

Každý axiom budeme chápat jako požadavek, který na danou binární operaci klademe. Uvažujme binární operaci • a množinu M. Nyní uvedeme základní axiomy, které pro nás budou v souvislosti s naší problematikou vyznamné.

- Axiom A_0 - uzavřenost operace: ke každé dvojici prvku $a, b \in M$ přiřazujeme prvek $c = a \bullet b$. Platí $c \in M$.
- Axiom A_1 - asociativní zákon: Operace • splňuje asociativní zákon, pokud $(a \bullet b) \bullet c = a \bullet (b \bullet c)$, kde $a, b, c \in M$.
- Axiom A_2 - existence neutrálního prvku: existuje takový prvek $\varepsilon \in M$, pro který platí $\varepsilon\bullet a = a\bullet\varepsilon = a$ pro všechna $a \in M$.
- Axiom A_3 - existence inverzního prvku: ke každému prvku $a \in M$ existuje tzv. inverzní prvek $\bar{a} \in M$ takový, že platí $a \bullet \bar{a} = \bar{a} \bullet a = \varepsilon$, kde $\varepsilon \in M$ je neutrální prvek operace • na množině M podle axiomu A_2.

Je zapotřebí poznamenat, že existují další axiomy definující další vlastnosti kladené a operace. V této práci však není jejich definice nutná, protože si v tomto textu vystačíme s výše uvedenými a další axiomy proto nebudeme uvádět.

4.3 Grupoidy, pologrupy, monoidy a grupy

Než si definujeme volnou grupu, která nás ze všech algebraických struktur bude v tomto výkladu zajímat nejvíc, musíme si definovat i struktury, ze kterých bude definice volné grupy vycházet.

- Algebraická struktura $(M, \bullet)$ definována binární operací $\bullet$ na množině M se nazývá grupoid, pokud operace $\bullet$ splňuje axiom A_0 uzavřenosti operace.
- Algebraická struktura $(M, \bullet)$ jejíž operace splňuje axiom A_0 uzavřenosti operace a asociativní zákon podle axiomu A_1 se nazývá pologrupa. Ekvivalentní název je též asociativní grupoid.
- Algebraická struktura $(M, \bullet, \varepsilon)$ jejíž operace splňuje axiomy A_0, A_1 a A_2 se nazývá monoid s neutrálním prvkem ε. Ekvivalentní název je též pologrupa s neutrálním prvkem. Výběr neutrálního prvku můžeme chápat jako definici určité nulární operace na množině M, která dává jako svůj výsledek právě prvek $\varepsilon \in M$.
- Algebraická struktura $(M, \bullet, \varepsilon, {}^{-1})$ která splňuje axiomy A_0, A_1, A_2 a A_3 se nazývá grupa. Grupu můžeme chápat jako monoid obsahující ke každému prvku $a \in M$ prvek inverzní $\bar{a} \in M$, kde jeho přiřazení každému prvku můžeme provést definicí unární operace ${}^{-1}: M \to M$.

Opět je zapotřebí poznamenat, že existují další algebraické struktury, které se liší počtem binárních operací, popřípadě dalšími axiomy, které musí jejich operace splňovat. Operace mohou být v obecném případě n−ární a jejich arity mohou být navzájem různé. Tyto struktury však v této prácí nejsou potřebné a proto by bylo jejich další rozepisování zbytečné.

5 Volné grupy

Nyní se dostáváme k pro nás nejdůležitější algebraické struktuře, kterou je volná grupa. Na základě její definice budeme dále stavět. Nejprve ale ukážeme volnou variantu jednodušší algebraické struktury, volného monoidu.

5.1 Volný monoid

Volný monoid na množině V je monoid, jehož prvky jsou všechny konečné řetězce složené z prvku množiny V za pomoci binární operace konkatenace včetně prázdného řetězce ε, který je neutrálním prvkem. Tyto řetězce často nazýváme slova a množinu všech slov nad množinou V značíme V^*.

Definujme nyní výše uvedené formálně. Množina V se nazývá volná báze monoidu G, jestliže V generuje G a každé zobrazení $f : V \rightarrow H$, kde H je monoid, lze rozšířit na homomorfismus $g : G \rightarrow H$. Monoid G se nazývá volný, jestliže má alespoň jednu volnou bázi. Volný monoid generovaný množinou generátoru V pomocí operace konkatenace je v teorii formálních jazyku značen jednoduše symbolem V^*.

5.2 Volná grupa

Podobně jako v případě volného monoidu, definujme si pojem volné grupy nejprve neformálně. Volná grupa na množině M je grupa, jejíž prvky jsou všechny konečné řetězce složené z prvku množiny M za pomoci binární operace konkatenace a to včetně prázdného řetězce ε, který je neutrálním prvkem. Připomeňme, že aby se jednalo o grupu, musí množina M ke každému svému prvku obsahovat prvek inverzní (výjimku tvoří pouze neutrální prvek, který je inverzní sám k sobe).

Uveďme si formální definici. Množina V se nazývá volná báze grupy G, jestliže V generuje G a každé zobrazení $f : V \rightarrow H$, kde H je grupa, lze rozšířit na homomorfismus $g : G \rightarrow H$. Grupa G se nazývá volná, jestliže má (alespoň jednu) volnou bázi.

Aby nedošlo k záměně označení volného monoidu a volné grupy, stanovme si zde, že volnou grupu generovanou množinou generátoru V pomocí operace konkatenace budeme značit jednoduše symbolem $V^{\bullet}$. Protože množina V obsahuje ke každému prvku i prvek inverzní, mohou se všechny prvky z V vyskytovat v jednotlivých slovech množiny $V^{\bullet}$. Jisté problémy činí případ, kdy bychom měli slovo

. . . . . $x\bar{x}$ . . . . . .

nebo

. . . . . $\bar{x}x$ . . . . . .

Těchto dvojic vzájemně inverzních symbolu by se ve slově samozřejmě mohlo vyskytovat více. V tomto případě by se však nejednalo o volnou grupu, protože by pro $x\bar{x}$ případně pro $\bar{x}x$ nebyl splněn axiom inverzních prvku. Každé slovo které obsahuje výše zmíněné dvojice však lze dále redukovat až k jejich úplnému odstranění. Slovo, které neobsahuje žádné výše uvedené dvojice prvku vzájemně inverzních se potom nazývá redukované. Lze dokázat, že každému slovu odpovídá pouze jediné slovo redukované a to bez ohledu na poradí redukcí jednotlivých dvojic vzájemně inverzních symbolu. Důkaz je možné najít např. v [4].

V pozdějším spojení bezkontextové gramatiky a volné grupy však budeme možnost redukce dvojic vzájemně inverzních symbolu často využívat.

6 Bezkontextové gramatiky nad volnými grupami

Nyní již máme definované všechny nezbytné pojmy a můžeme si konečně zavést strukturu, která je pro tuto práci mimořádně významná. Výše jsme si uvedli matematické a algebraické základy formálně, zde zavedeme určité zjednodušení. Uvažujme libovolnou abecedu V . Znovu zdůrazněme, že symbolem $V^{\bullet}$ budeme označovat volnou grupu generovanou množinou generátoru V operací konkatenace. Přitom platí, že pro každý symbol $a \in V$ existuje právě jeden inverzní symbol $\bar{a} \in V$. Prázdný řetězec $\varepsilon \in V^{\bullet}$ je neutrálním prvkem. Spojením bezkontextové gramatiky a volné grupy generované totální abecedou této gramatiky získáme bezkontextovou gramatiku nad volnou grupou, kterou si nyní definujeme.

Nechť G = (N, T, P, S) je bezkontextová gramatika s totální abecedou $V = N \cup T$ takovou, že pro každé $x \in V$ existuje právě jeden prvek $\bar{x} \in V$ takový, že platí $x\bar{x} = \bar{x}x = \varepsilon$. Dvojici $\Gamma = (G, V^{\bullet})$ potom nazveme bezkontextovou gramatikou nad volnou grupou.

Pro účely této práce je zajímavá především schopnost struktury generovat jazyky. Naším cílem je samozřejmě co nejvyšší síla — ideálně schopnost generovat celou třídu jazyku typu 0. Před dalším definováním struktury a jejími možnosti je ale třeba definovat relaci přímé derivace $\Rightarrow_{\Gamma}$.

6.1 Derivace v bezkontextové gramatice nad volnou grupou

Nechť $\Gamma = (G, V^{\bullet})$ je bezkontextová gramatika nad volnou grupou, kde G = (N, T, P, S), $V = N \cup T$. Mezi řetězci λ a μ platí relace $\Rightarrow_{\Gamma}$ zvaná přímá derivace, pokud oba můžeme vyjádřit ve tvaru

$\lambda = \gamma\ \alpha\ \delta$

$\mu = \gamma\ \beta\ \delta$

a zároveň $p = \alpha \rightarrow \beta \in P$, kde $\gamma, \delta \in V^{\bullet}$ Potom píšeme $\lambda \Rightarrow_{\Gamma} \mu\ [p]$ a říkáme, že řetězec λ přímo derivuje řetězec μ podle pravidla p v bezkontextové gramatice nad volnou grupou Γ.

V případě bezprostředního výskytu dvojice $x\bar{x}$ nebo $\bar{x}x$, kde $x, \bar{x} \in V$, ve větné formě se tato okamžitě redukuje podle pravidel popsaných výše. Protože je výsledná redukovaná větná forma nezávislá na poradí případných redukcí jednotlivých dvojic vzájemně inverzních symbolu, nečiní tento jev ani žádné problémy při generování vět jazyka.

Tranzitivní uzávěr $\Rightarrow_\Gamma^+$, reflexivní a tranzitivní uzávěr $\Rightarrow_\Gamma^*$ a derivace délky n $\Rightarrow_\Gamma^n$ jsou definovány zcela přirozeně stejným způsobem jako v případě běžných gramatik. Stejně tak symbol Γ jako dolní index symbolu zavedené relace se dále nebude uvádět, pokud bude z kontextu zřejmé, o kterou bezkontextovou gramatiku nad volnou grupou se jedná.

6.2 Jazyk generovaný bezkontextovou gramatikou nad volnou grupou

Nechť $\Gamma = (G, V^\bullet)$ je bezkontextová gramatika nad volnou grupou, kde $G = (N, T, P, S)$, $V = N \cup T$. Řetězec $\alpha \in V^\bullet$ nazýváme větnou formou, jestliže platí $S \Rightarrow^* \alpha$, tj. řetězec α je generovatelný z výchozího symbolu S. Větná forma, která obsahuje pouze terminální symboly, se nazývá věta. Jazyk L(G), generovaný bezkontextovou gramatikou nad volnou grupou Γ, je definován množinou všech vět

$$L(G) = \{w \mid S \Rightarrow^* w \wedge w \in T^*\}$$

generovatelných z výchozího symbolu této gramatiky.

6.3 Generativní schopnosti bezkontextových gramatik nad volnou grupou

Nyní se budeme zabývat generativními schopnostmi naší nove zavedené struktury. Naším cílem je ukázat, že pro každou gramatiku H typu 0 existuje ekvivalentní bezkontextová gramatika nad volnou grupou Γ taková, že L(H) = L(Γ) a především ukázat konkrétní algoritmus transformace gramatik typu 0 na bezkontextové gramatiky nad volnými grupami.

Z [2] je převzata následující pomocná věta. Pro každou gramatiku $H = (N, T, P, S_H)$ typu 0 existuje ekvivalentní gramatika $G = (N \cup \{X, Y\}, T, P, S_G)$ typu 0, kde $\{X, Y, S_G\} \cap \{N \cup T\} = \varnothing$ taková, že její množina přepisovacích pravidel obsahuje pouze pravidla tvaru

- $AB \rightarrow CD$

- $A \rightarrow BC$
- $A \rightarrow a$
- $XY \rightarrow Z$

kde $A,B,C,D \in N$, $a \in T$ a $Z \in \{X, \varepsilon\}$.

Z důkazu nám postačí pouze uvedení konstrukce, kterou je též možné najít v [2]. Uvažujme gramatiku $H = (N', T, P', S')$. Bez ztráty obecnosti můžeme předpokládat, že H je v Penttonenově normální formě, tedy množina přepisovacích pravidel P' obsahuje pouze pravidla tvaru

- $AB \rightarrow AC$
- $A \rightarrow BC$
- $A \rightarrow a$
- $A \rightarrow \varepsilon$

kde $A, B, C \in N'$, $a \in T$ a ε je prázdný řetězec. Navíc předpokládejme, že $\{S_G, X, Y\} \cap (N' \cup T) = \varnothing$. Definujme gramatiku $G = (N, T, P, S_G)$, kde množina $N = N' \cup \{X, Y\}$. Množinu přepisovacích pravidel P zkonstruujeme následovně:

1. každé pravidlo $p \in P'$, kde p obsahuje na své pravé straně dva neterminální symboly, vlož do P
2. pro každé $A \rightarrow \varepsilon \in P'$, přidej $A \rightarrow Y$ do P
3. pro každé $A \in N'$, přidej $AY \rightarrow YA$ do P
4. přidej $S_G \rightarrow XS_H$, $XY \rightarrow X$ a $X \rightarrow \varepsilon$ do P

Tím je konstrukce hotova.

V tuto chvíli je nutné upozornit na několik důležitých vlastností takto zavedené gramatiky. Každá větná forma této gramatiky neobsahuje neterminální symbol X vůbec (poté, co byl odstraněn pomocí pravidla $X \rightarrow \varepsilon$), nebo obsahuje právě jeden jeho výskyt a tento výskyt je současně prvním symbolem dané větné formy. Jako další věc je nutno připomenout, že pokud odstraníme symbol X z větné formy, ale tato větná forma ještě někde obsahuje jeden nebo více symbolu Y , případně budou v dalších derivačních krocích další symboly Y vyderivovány, nevygenerujeme již nikdy z této větné formy větu tvořenou pouze terminálními symboly.

Z výše napsaného vyplývá, že symbol X by mel být z větné formy odstraněn až v okamžiku, kdy je jisté, že žádný další symbol Y již nebude vyderivován a zároveň XY je prefix aktuální větné formy, na který bude aplikováno pravidlo $X \rightarrow \varepsilon$. To potom zaručí korektní odstranění symbolu X a posledního výskytu symbolu Y.

Podle zavedených konvencí z teorie formálních jazyku je třída všech jazyku typu 0 značena symbolem RE. Označme jako FG(2) třídu všech jazyků generovaných bezkontextovými gramatikami nad volnými grupami. Pak platí, že

FG(2)=RE

Jinými slovy říkáme, že pro každou gramatiku G typu 0 existuje ekvivalentní bezkontextová gramatika nad volnou grupou Γ taková, že L(G) = L(Γ). Poznamenejme, že symbolem Γ budeme značit bezkontextovou gramatiku nad volnou grupou včetně struktury volné grupy. Symbolem G_Γ potom pouze bezkontextovou gramatiku patřící ke struktuře volné grupy (bez nutnosti tuto gramatiku uvažovat včetně struktury volné grupy v daném kontextu).

V této práci není zapotřebí uvedení celého důkazu tvrzení FG(2)=RE. Proto se zaměříme pouze na část, popisující vytvoření bezkontextové gramatiky nad volnou grupou odpovídající gramatice typu 0. Celý důkaz platnosti tvrzení FG(2)=RE lze nalézt v [5].

Předpokládejme, že G = (N, T, P, S) je gramatika typu 0 s výše uvedenými modifikacemi. Vytvoříme bezkontextovou gramatiku nad volnou grupou $G_\Gamma = (N_\Gamma, T \cup \{\overline{a} \mid a \in T\}, P_\Gamma, S)$, kde definujeme $N_\Gamma = N \cup N_{CS} \cup N$ tak, že

$N = \{\overline{A} \mid A \in N\}$

$N_{CS} = \{<ABCD> \mid AB \to CD \in P\} \cup \{<XYX>, <XY\varepsilon>\}$.

Množina terminálních symbolu této gramatiky je vytvořena jako sjednocení

$T_\Gamma = T \cup \{\overline{a} \mid a \in T\}$.

Množina N obsahuje inverzní symboly k symbolům množiny N. Množina N_{CS} obsahuje nezbytně nutné neterminály a jejich inverzní protějšky, které jsou potřeba pro další kroky transformace a které se týkají kontextových přepisovacích pravidel. Konečně potom množina terminálních symbolu je konstruována sjednocením všech terminálních symbolu v množině T a jejich inverzních protějšků. Existence inverzních symbolu ke každému prvku totální abecedy gramatiky je vyžadována obecnou definicí grupy. My však budeme z inverzních symbolu využívat pouze ty, které se nachází v množině N_{CS}.

Nyní konstruujme množinu přepisovacích pravidel P_Γ následovně:

1. každé $A \to BC \in P$ a každé $A \to a$ 2 P, kde $A,B,C \in N$, $a \in T$ vlož do P_Γ
2. pro každé $AB \to CD \in P$ přidej $A \to C<ABCD>$ a $B \to <\overline{ABCD}>D$ do P_Γ
3. místo $XY \to X \in P$ přidej $X \to X<XYX>$ a $Y \to <\overline{XYX}>$ do P_Γ
4. místo $XY \to \varepsilon \in P$ přidej $X \to <XY\ \varepsilon>$ a $Y \to <\overline{XY\ \varepsilon}>$ do P_Γ

Tím je konstrukce G_{Γ} dokončena. Podle výše uvedeného platí, že $L(G) = L(G_{\Gamma})$. Na tuto konstrukci navazuje pokračování důkazu z [5].

7 Syntaktická analýza

Syntaktickou analýzou rozumíme především proces analyzování sekvence symbolů s cílem určit, zda je ona sekvence platná pro danou gramatiku. Výsledkem syntaktické analýzy je tedy jednoznačná odpověď, zda zkoumaný řetězec patří nebo nepatří do dané gramatiky, případně může být výstupem i syntaktický strom nebo odvozená posloupnost. Obecně existují dva základní přístupy k syntaktické analýze, kterými jsou syntaktická analýza shora dolů a zdola nahoru. Oba přístupy zde pro úplnost stručně popíšeme.

7.1 Syntaktická analýza shora dolů

Tento přístup vychází při analýze z myšlenky postupného dopředného odvozování na zásobníku od počátečního neterminálu a srovnávání analyzovaného sekvence symbolů (slova) po terminálních symbolech, které se objeví na zásobníku. V momentě, kdy dojdeme do situace, ve které nám nezbude již nic ke srovnání a to jak v původním slově, tak v přepisovaných řetězcích od počátečního symbolu na zásobníku je analýza shora dolů skončena a lze říci, zda slovo patří nebo nepatří do jazyka, generovaného zadanou gramatikou Tyto kroky, kdy přepisujeme od počátečního neterminálu nazýváme expanze (jelikož neterminály rozšiřujeme na řetězce podle pravidel).

7.2 Syntaktická analýza zdola nahoru

Druhým základním principem syntaktické analýzy je metoda zdola nahoru. Tento přístup vychází z myšlenky postupného zpětného odvozovaní. Při této analýze postupně vkládáme symboly na zásobník a v případě, že se v zásobníku vyskytne řetězec, který se vyskytuje u některého z přepisovacích pravidel na pravé straně, tak provedeme zpětně odvození na daný neterminál. Princip je tedy oproti analýze shora dolů zcela opačný – řetězce zjednodušujeme na neterminály. Tomuto opaku expanze se říká redukce. Sekvence symbolů je přijata, pokud dojdeme k situaci, že je celá sekvence přečtena a na zásobníku zbyl pouze počáteční neterminál.

7.3 Algoritmus implementace

Pro další rozvoj projektu je zapotřebí si uvést a rozepsat algoritmus, pomocí kterého budeme celou problematiku implementovat. Základní, zjednodušeně zapsaný algoritmus pro implementaci bude vypadat následovně:

```
Vstup: gramatika, testovaný  řetězec
1. převod gramatiky na Penttonenovu normální formu
2. úprava  gramatiky  v Pentonnenově  normální  formě  podle
   pravidel, uvedených v kapitole 6.3 a převod na bezkontextovou
   gramatiku nad volnou grupou
3. syntaktická  analýza  testovaného  řetězce  nad  získanou
   bezkontextovou gramatiku nad volnou grupou
Výstup: testovaný řetězec patří / nepatří do zadané gramatiky
```

Pro provedení syntaktické analýzy je zapotřebí zvolit jeden ze dvou výše zmíněných přístupů, buď metodu shora dolů nebo metodu zdola nahoru. Vzhledem k principům bezkontextových gramatik nad volnou grupou se jako lepší metoda jeví princip shora dolů. Pokud bychom vybrali metodu zdola nahoru, došlo by při analýze k značnému nedeterminismu, protože na místě každého prázdného řetězce ε by bylo nutné rozhodnout, zda se nejedná o některou dvojici vzájemně inverzních neterminálů, která byla na redukována.

Z možných variant analyzátorů, využívajících metody shora dolů, jsme vybrali Earlyho algoritmus. Je založen na tzv. tečkové notaci, kdy tečka značí místo, které chceme prozkoumat (např. pravidlo $A \rightarrow B.CD$ odpovídá situaci, ve které B již bylo analyzováno a zbývá ještě analyzovat CD).

8 Závěr

V této práci jsme zavedli novou strukturu, která generuje rekurzivně vyčíslitelné jazyky. Díky přidání volné grupy nad totální abecedou gramatiky získáváme výpočetní sílu Turingova stroje, čímž výrazně zvyšujeme schopnosti původní bezkontextové gramatiky. Tato struktura má i nevýhody, především velký nárůst neterminálních symbolů.

Dále jsme popsali a ukázali algoritmus, pomocí kterého budeme v dalším rozvoji projektu implementovat a analyzovat.

Literatura

[1] Češka, M.: Přednášky z předmětu Teoretická informatika, Brno, FIT VUT 2006

[2] Meduna, A., Kolář, D.: Homogenous Grammars with a Reduced Number of Non-Context-Free Productions. In: Information Processing Letters, roc. 2002, c. 81, Amsterdam, NL, p. 253-257, ISSN 0020-0190.

[3] Penttonen, M.: One-Sided and Two-Sided Context in Formal Grammars. Inf. Control 25, pp. 371-392, 1974.

[4] Milne, J.S.: Group Theory. Course Notes, 2003, URL http://www.jmilne.org/math/CourseNotes/math594g.pdf

[5] Bidlo, R.: Bezkontextové gramatiky nad volnými grupami, Proceedings of 8th International Conference ISIM'05 Information System Implementation and Modeling, Ostrava, CZ, MARQ, 2005, s. 95-100