SSZ 2016 SWI

Systémové programování

SP

Část: Formální jazyky a překladače

Seznam otázek

Všechny otázky se shodují s otázkami z minulých let :-)

1. Překladače – typy, struktura a princip činnosti.

Formálněje překladačzobrazením ze zdrojového jazyka do cílového jazyka.

Typy

Postup při tvorběcílového spustitelného programu:

Zdrojový program → `[preprocesor]` → upravený zdrojový program → `[kompilátor]` → cílový program v jazyce symbolických instrukcí → `[assembler]` → relokovatelný strojový kód &zvenku: knihovní soubory a další relokovatelné objektové soubory → `[linker/loader]` → cílový strojový kód

Preprocesory

Realizují vnořování částí programu do hostitelského jazyka. Např. expandují makra, přidají `include <něco.h>` apod. Jeho úkolem je posbírat zdrojový program.

Kompilátory

Generují zvyššího programovacího jazyka kód –strojový/symbolický/jiný jazyk (1. Fortran IBM, 50. léta). Assembly language (jazyk symbolických adres) je jednodušší vyplivnout jako výstup a je i jednodušší pro debugování.

Interprety

Namísto přeložení celého programu provádí příkaz za příkazem operace uvedené ve zdrojovém programu nad vstupními daty. Proto jsou interprety obecně pomalejší. Obvykle ale díky spouštění programu příkaz za příkazem lépe diagnostikují chyby než kompilátory.

Java -kombinace kompilace a interpretace

Zdrojový program je nejdříve zkompilován do bytecode a ten je pak interpretován virtuálním strojem.

Výhoda –bytecode může být zkompilován na jednom stroji a interpretován na jiném. Aby to bylo rychlejší, některé Java kompilátory (just-in-time kompilátory) převádí bytecode do strojového jazyka těsně předtím, než proběhne intermediate (vnitřní) program pro zpracování vstupních dat.

Zdrojový program → `[Translator]` → intermediate program + vstup→ `[Virtual Machine]` → výstup

Assemblery

Překládají z jazyka symbolických instrukcí (JSI) do strojového kódu. Přeložený strojový kód může být buďto **absolutní binární kód** nebo **přenositelný binární kód**.

Hlavní problémy, které řeší, je adresace symbolických jmen a makra.

Linker a loader

Větší programy jsou často kompilované po částech, takže vytvořený relokovatelný strojový kód (= lze jej umístit do libovolného místa v paměti) je potřeba spojit s dalšími relok. objektovými soubory a knihovními soubory. O to se stará **Linker**. Řeší adresy externí paměti, kde kód v jednom souboru může odkazovat na místo v jiném souboru. **Loader** pak narve všechny spustitelné objektové soubory do paměti pro spuštění.

Typy překladačů

Formátory textu

- jde o úpravu textu podle požadavkůuživatele
    - např. TeX → syntax highlighting, nebo přeformátování kódu (odsazení) apod.

- takovýpřekladač, který ze vstupního souboru definovaného určitým jazykem vygeneruje výstup.

Silikonový překladač

- Pro návrh integrovaných obvodů.
- Proměnné nereprezentují místo v paměti, ale logickou proměnnou obvodu.
- Výstupem je návrh obvodu.

Dávkový překladač

- dávkové zpracování

Inkrementální překladač

- je interaktivní a překládá po úsecích

Křížový překladač

- Překládá na jiném procesoru než na kterém se program (přeložený kód) spouští
    - např. zabudované (embedded) systémy

Kaskádní překladač

Máme již překlad z jazyka A do jazyka B, chceme ale AC. Pak vytvoříme kompilátor z jazyka B do jazyka C, pokud je to snazší než vytvořit kompilátor z jazyka A do jazyka C. Jazyk B je vnitřním jazykem, a pokud je to standardní všeobecně používaný jazyk, pak programy v jazyce A budou snadno přenositelné.

Nevýhodou je však, že oba překladače produkují chybové zprávy. Chybové zprávy druhého překladače jsou cizí pro uživatele jazyka A, protože jsou orientovány na jazyk B. Chybová hlášení výpočtu tak budou pomíchaná.

Paralelizující překladač

- Zjišťuje nezávislost úsekůprogramu

Optimalizující překladač

- Možnosti ovlivnění optimalizace času či paměti programátorem.

Konverzační překladač

- interaktivní

Struktura, principčinnosti

Překladače jsou dva druhy: kompilátory a interprety

Struktura Kompilátoru

- všechny příkazy překládá najednou, program lze spustit až po ukončení celého překladu (Pascal, C, Fortran, Ada, ...)
- ANALÝZA:
    - zdrojový program
    - **lexikální analýza** (lineární), programové symboly
    - **syntaktická analýza** (hierarchická)
    - derivační strom
- SYNTÉZA:
    - derivační strom
    - **zpracování sémantiky**, program ve vnitřní formě

    - **optimalizace** (příprava generování) upravený program ve vnitřní formě
    - **generování kódu**, cílový program

Všechny části spolupracují s pracovními tabulkami překladače. Základní tabulkou kompilátoru i interpretu je tabulka symbolů. Obsahuje záznamy o názvech proměnných, jejich typu, rozsahu, názvy procedur společněs věcmi jako počet a typy argumentů, metoda předání jednotlivých argumentů(odkazem nebo hodnotou) a návratový typ.

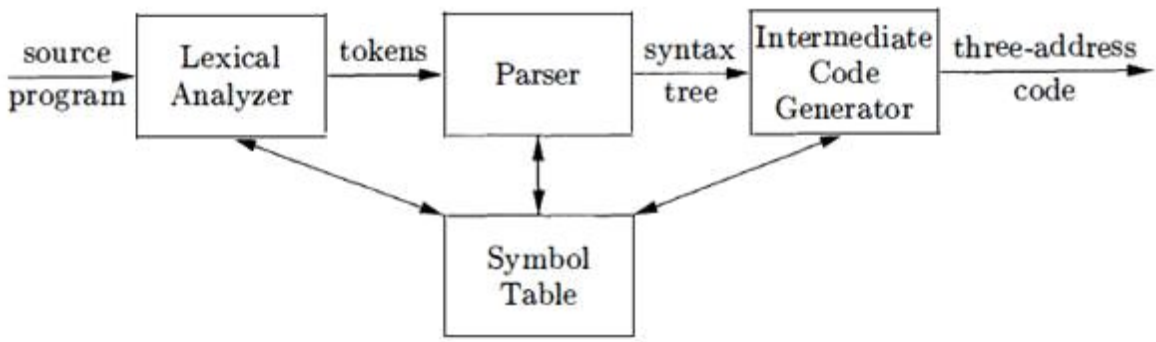
Výhodou kompilátoru je rychlá exekuce programu.

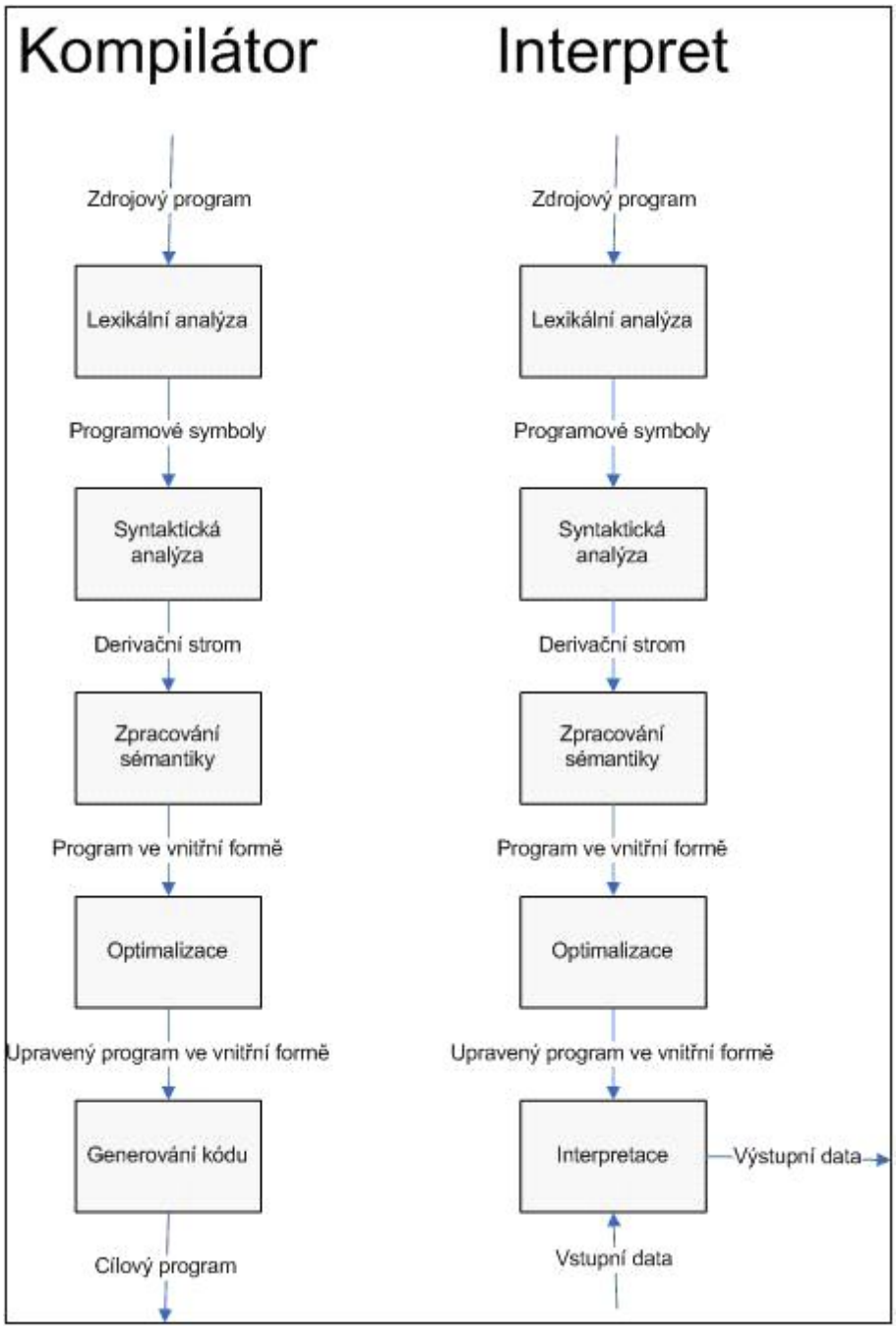
Struktura Interpreta

- zpracovává příkazy jednotlivě a každý provede okamžitě po jeho přeložení (Python, Perl, JavaScript, Ruby,...)
- ANALÝZA:
    - zdrojový program
    - **lexikální analýza** (lineární), programové symboly
    - **syntaktická analýza** (hierarchická)
    - derivační strom
- SYNTÉZA:
    - derivační strom
    - **zpracování sémantiky**,program ve vnitřní formě
    - **optimalizace** (příprava generování), upravený program ve vnitřní formě
    - **interpretace**, pracuje se vstupními daty, aby vygeneroval výstupní data

Výhodou interpretu je:

- Eliminace kroků cyklu „editace překlad sestavení exekuce“
- Snazší realizace ladících mechanismů(zachování původních jmen symbolů)



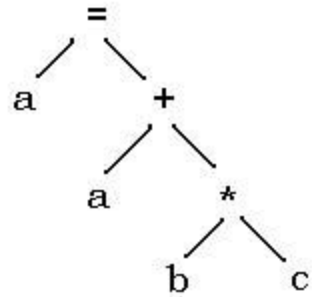


Lexikální analýza

- zdrojový kód vstupuje do procesu překladu jako posloupnost znaků
- tato posloupnost se čte lineárně zleva doprava a sestavují se z ní lexikální symboly jako konstanty, identifikátory, klíčová slova nebo operátory
- je založena na regulárních gramatikách
- výsledkem je posloupnost symbolů, např. je na vstupu rozeznáno klíčové slovo `begin` a do posloupnosti lexikálních symbolů bude zařazen nový symbol reprezentující právě toto klíčové slovo
- tyto symboly jsou programem snadno použitelné a dále zpracovatelné
- v této fázi se odstraňují veškeré komentáře

Syntaktická analýza

- z posloupnosti lexikálních symbolů se vytvářejí hierarchicky zanořené struktury (vnitřní jazyk překladače), které mají jako celek svůj vlastní význam
    - např. výrazy, příkazy, deklarace nebo program



- programy jsou psány většinou v infixové notaci (a=a+b*c) => analyzujeme a vytváříme hierarchické uspořádání derivačního stromu:

Notace vnitřního jazyka překladače:

- **Prefixová** (nemá závorky, operátory bezprostředně předchází operandy a pořadí operandů je zachováno)
- **Infixová**
- **Postfixová** (nemá závorky, operátory bezprostředně následují operandy a pořadí operandů je zachováno, vyhodnotitelná zásobníkem)

Sémantická analýza

Provádějí se některé kontroly, zajišťující správnost programu z hlediska vazeb, které nelze provádět v rámci syntaktické analýzy (některé konstrukty nejdou popsat bezkontextovou gramatikou, třeba např. kontrola deklarací, typová kontrola, kontrola, jestli index pole je integer apod.).

Typická reprezentace programu ve vnitřní formě(intermediate code) je sekvence trojic nebo čtveřic (3-nebo 4-adresových instrukcí)

Optimalizace

Optimalizátor kódu zajišťuje, aby se používalo co nejméně pomocných proměnných pro mezi výpočty, aby se v cyklu zbytečně několikrát nevyhodnocoval tentýž výraz, jestliže hodnota jeho prvků zůstává bez změny a vyhodnocení stačí provést jednou před cyklem, apod. Optimalizací prochází program obvykle v intermediálním tvaru –intermediální kód je již podobný cílovému programu, má však strukturu vhodnější pro optimalizaci. Může to být zápis podobný assembleru nebo třeba dynamická struktura (dynamický seznam stromů představujících jednotlivé příkazy).

Generování kódu

- poslední fází překladače je generování cílového kódu
- to je obvykle přemístitelný kód nebo program jazyka asembleru
- všem proměnným použitým v programu se přidělí místo v paměti
- potom se instrukce mezikódu překládají do posloupnosti strojových instrukcí, které provádějí stejnou činnost.

Vícefázový a víceprůchodový překladač

Fáze = logicky dekomponovaná část (může obsahovat více průchodů, např. optimalizace)

Průchod = čtení vstupního řetězce, zpracování, zápis výstupního řetězce –může obsahovat více fází

Jednoprůchodový překladač = všechny fáze probíhají v rámci jediného čtení zdrojového textu programu

- Omezená možnost kontextových kontrol
- Omezená možnost optimalizace
- Lepší možnosti zpracování chyb a ladění (tedy dobré pro výuku)

Na strukturu překladače mají vliv:

- Vlastnosti zdrojového a cílového jazyka
- Vlastnosti hostitelského počítače
- Rychlost/velikost překladače
- Rychlost/velikost cílového kódu
- Ladicí schopnosti (detekce chyb, zotavení)
- Velikost projektu, prostředky, termíny

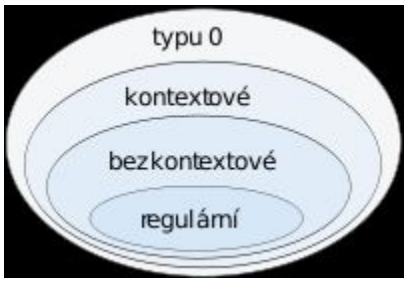
Testování a údržba překladače

- díky formální specifikaci jazyka je možné automatické provádění testů
- systematického testování lze dosáhnout regresními testy
    - sada testů doplňovaná o testy na odhalené chyby
    - po každé změně v překladači se provedou všechny testy a jejich výstupy se porovnají spředešlými

2. Regulární gramatiky, regulární výrazy a konečné automaty.

Obecně k jazykům a gramatikám

- ABECEDA = neprázdná množina, její prvky se nazývají PÍSMENA
- ŘETĚZEC = SLOVO = libovolná konečná posloupnost písmen abecedy
- JAZYK = libovolná množina řetězců (= slov) nad abecedou
- ABECEDA A={a,b,c} SLOVO (ŘETĚZEC) aaab, abc, cb,.. JAZYK L= {aaab, abc, cb}
- GRAMATIKA definuje jazyk jako množinu všech řětězců (slov), které lze v gramatice odvodit.
- Chomskeho hierarchie (místo „regulární“ má být asi „lineární“)



Regulární gramatiky

- Gramatika G je čtveřice (N,Σ,P,S), kde:
  - N je konečná množina neterminálních symbolů (neterminálů).
  - Σ je konečná množina terminálních symbolů tak, že žádný symbol nepatří do N a Σ zároveň (jsou disjunktní).
  - P je konečná množina odvozovacích pravidel. Každé pravidlo je tvaru

$$(\Sigma \cup N)^* \longrightarrow (\Sigma \cup N)^*$$

"cokoli poskládaný ze všech možnejch symbolů na cokoli"; S je prvek z N nazývaný počáteční symbol.

Větná forma

Řetězec α se nazývá větnou formou v gramatice G, s počátečním symbolem S, platí-li:

$S \Rightarrow^* \alpha,\ \text{kde}\ \alpha \in (N \cup T)^*$

Věta

Řetězec α se nazývá větou v gramatice G, s počátečním sysmbolem S, platí-li:

$S \Rightarrow^* \alpha,\ \text{kde}\ \alpha \in T^*$

Fráze

Nechť λ = α β γ je větná forma v gramatice G. Podřetězec β se nazývá frází větné formy λ vzhledem k neterminálnímu symbolu A, platí-li:

$S \Rightarrow^* \alpha A \gamma \quad a \quad A \Rightarrow^* \beta$

Jednoduchá fráze

větné formy α A γ vzhledek k neterm. A je podřetězec β, platí-li:

$S \Rightarrow^* \alpha A \gamma \quad a \quad A \Rightarrow \beta$

L-fráze

je nejlevější jednoduchou frází

Lineární gramatika

= bezkontextová gramatika, která má nanejvýš jeden neterminál na pravé straně.

Regulární gramatika je speciálním případem lineární gramatiky, kdy všechny neterminály jsou na levém konci/straně (levá lineární = levá regulární) nebo ekvivalentně pro pravou stranu.

Regulární gramatika

- je to gramatika typu 3 = lineární, navíc převedená do regulárního tvaru (podle Chomského hierarchie).

Pravidla těchto lineárních gramatik jsou omezena na jeden neterminál na levé straně. Pravá strana se u pravé regulární gramatiky skládá z jednoho terminálu (u lineární i z více), který může být následován jedním neterminálem, tedy:

$X \rightarrow wY$
$X \rightarrow w$

kde X, Y jsou neterminály a w je řetězcem terminálů. Regulární gramatiky se také nazývají **pravé lineární gramatiky**. Obdobně se definují i **levé regulární gramatiky**, které obsahují pravidla typu:

$X \rightarrow Yw$
$X \rightarrow w$

- pravé a levé gramatiky jsou ekvivalentní
- jazyky generované regulárními (=lineárními) gramatikami jsou právě jazyky rozpoznatelné konečným automatem.

Lineární gramatika = má na pravé straně právě jeden neterminál

Regulární gramatika = gramatika, která popisuje regulární jazyk, přesně definovaný tvar pravidel

- B → a, B→ aC, B -> e pro pravou regulární gramatiku

Regulární gramatika je tedy buď jen levá lineární gramatika nebo jen pravá lineární gramatika. Čistě lineární (levo-pravá) gramatika je pak taková gramatika, která sestává z pravých i levých pravidel současně.

Regulární výrazy

Regulární výrazy umožňují algebraické manipulace s regulárními množinami - umožňují **vyjádření regulárních množin**. Třída regulárních výrazů nad abecedou je definována takto:

- e a Ø jsou regulární výrazy
- každé písmeno (symbol - znak) $\sigma \in \Sigma$ je regulární výraz nad Σ
- jsou-li R1 a R2 regulární výrazy nad Σ, pak i (R1 + R2), (R1 . R2) a R1* jsou regulární výrazy nad Σ

Daná množina je regulární množina nad Σ, právě když může být popsána vhodným regulárním výrazem nad Σ. Každý regulární výraz **U** popisuje jistou množinu **Ũ** slov nad Σ: $\tilde{U} \subseteq \in^*$

Regulární množiny se vhodně charakterizují přechodovými grafy. Přechodový graf T nad abecedou Σ je konečný orientovaný graf, jehož každá hrana je pojmenována jistým slovem $w \in \Sigma^*$;alespoň jeden uzel je počáteční.

Množinu všech slov akceptovaných konečným automatem **A** označíme **Ã**. Množina je regulární nad Σ právě když je akceptována vhodným automatem nad Σ

Regulární výraz je řetězec popisující celou množinu řetězců (slov), konkrétně regulární jazyk. Používají se nejčastěji v počítačových programech a skriptovacích jazycích pro vyhledávání a úpravu textu. V případě, že uživatel chce v textu vyhledat nějaký řetězec, který nezná přesně nebo který může mít více variant, může zadat regulární výraz, který postihne všechny chtěné varianty. Program tak nalezne všechny části textu, které danému výrazu odpovídají.

Každý z regulárních výrazů označuje jistý regulární jazyk.

Konečné automaty

Základní popis

- konečný počet stavů
- konečný počet vstupů
- jednosznačně určený následující stav
- jednoznačně určený počáteční stav

Znázornit lze: Tabulkou, Stavovým diagramem (to jsou ty stavy a hrany), Stavovým stromem

3 typy konečných automatů

- Rozpoznávací (akceptuje / neakceptuje)
- Klasifikační
- S výstupní funkcí

Formální definice

- konečný automat je uspořádaná pětice (S,Σ,P,s,F) , kde:
  - **S** je konečná množina stavů.
  - **Σ** (velké sigma) je konečná množina vstupních symbolů nazývaná abeceda.
  - **P** je tzv. přechodová funkce (též přechodová tabulka), formálně zobrazení δ: S Σ S, popisující pravidla přechodů mezi stavy. Přechod je určen stavem ve kterém se automat nachází a symbolem, který přichází na vstup (nebo který je čten na vstupu)
  - **s** je počáteční stav (s naleží S)
  - **F** je množina koncových (přijímacích) stavů (F je podmnožinou S)

3. Ekvivalence konečných automatů a regulárních gramatik.

Lexikální symboly (lexémy, anglicky tokens) jsou regulární jazyk.

Regulární jazyk lze definovat gramatikou typu 3 nebo konečným automatem nebo regulárním výrazem.

Každou lineární gramatiku lze převést na regulární tvar.

Regulární gramatiky

- popisují všechny **regulární jazyky** a v tomto smyslu (ve schopnosti popisu jazyka) jsou ekvivalentní s konečnými automaty a regulárními výrazy
- regulární gramatika ja buď pravá regulární (neterminály jsou vpravo)
  nebo levá regulární (neterminály jsou vlevo).

Regulární jazyk

- je formální jazyk (množina (i nekonečná) slov složených z omezené abecedy), který:
- může být akceptován deterministickým/nedeterministickým konečným stavovým automatem
- lze popsat regulárním výrazem
- lze ho generovat regulární gramatikou

Příklady neregulárního jazyka

- je $a^n b^n$, kde n > 1 (alespoň jedno ***a*** následované stejným počtem ***b***)
- gramatika pro palindromy
- lze určit na základě *Nerodovy věty*, která se užívá v důkazech, že nějaký jazyk není rozpoznatelný konečným automatem

Každý regulární jazyk je rozpoznatelný konečným automatem.

⇕

Každý jazyk rozpoznatelný konečným automatem je regulární.

Kleenova věta

Libovolný jazyk je regulární, právě když je rozpoznatelný konečným automatem. Přechodový graf je T nad S je konečný orientovaný graf, jehož každá hrana je pojmenována jistým slovem ***wÎS****. Alespoň jeden z uzlů grafu je počáteční a některé uzly jsou koncové. Ke každému přechodovému grafu T nad abecedouSexistuje regulární výraz R nadStakový, že $\widehat{R} = \widehat{T}$ a ke každému regulárnímu výrazu R nadSexistuje konečný automat A takový, že $\widehat{A} = \widehat{R}$.

Postup převodu gramatiky na konečný automat

Potřebujeme získat gramatiku typu 3 ve standardní formě.

Regulární gramatika je ve **standardní formě**, jestliže obsahuje pouze pravidla tvaru

X → aY a ***X → a***, ***X → e*** kde:

X, ***Y*** jsou neterminály,
a je právě jeden terminál,
e je prázdný symbol.

Toho dosáhneme takto:

- Původní gramatika typu 3 (lineární): G = (N, T, S, P)
- Požadovaná regulární gramatika: G' = (N', T, S, P')
- Požadovaná gramatika G' bude mít stejné terminální symboly a stejný počáteční stav.
- Konstrukce přechodůP':
  - do ***P'*** zařadíme všechna pravidla z **P** ve tvaru ***X →aY*** a ***X →e***

- za každé pravidlo $X \rightarrow x_1 x_2 x_3 Y$ zařadíme do P'soustavu pravidel:
  - $X \rightarrow x_1 X_1$
  - $X_1 \rightarrow x_2 X_2$
  - $X_2 \rightarrow x_3 Y$
- za každé pravidlo $X \rightarrow z_1 z_2$, zařadíme do P'soustavu:
  - $X \rightarrow z_1 Z_1$
  - $Z_1 \rightarrow z_2 Z_2$
  - $Z_2 \rightarrow e$
- Místo pravidel tvaru $X \rightarrow Y$ musíme zajistit to, aby z každého stavu X pro který máme $X \rightarrow Y$, bylo možné odvodit všechny řetězce, které lze odvodit z Y
- N' vznikne obohacením N o všechny nově o vytvořené neterminální symboly

Zkonstruujeme automat z nově vytvořené gramatiky

- stavy budou odpovídat neterminálním symbolům
- vstupy budou odpovídat terminálním symbolům
- přechodovou funkci zkonstruujeme na základě analogií
  - $X \rightarrow aY \leftrightarrow$ přechod ze stavu X do stavu Y při vstupu symbolu a
- počáteční stav bude odpovídat počátečnímu symbolu
- množinu koncových stavů určíme z pravidel $X \rightarrow e$

*Tímto jsme získali **nedeterministický konečný automat**, který lze převést na **deterministický konečný automat**.*

Regulární atributované a překladové gramatiky

Atributovaná gramatika

AG = (G, Atributy, Sémantická pravidla)

Atributy jsou přiřazeny symbolům gramatiky a sémantická pravidla jednotlivým přepisovacím pravidlům. Při aplikaci přepisovacího pravidla se provedou příslušná sémantická pravidla a vypočtou hodnoty atributů.

Atributy vyhodnocované průchodem derivačním stromem zdola nahoru nazýváme syntetizované, shora dolů nazýváme dědičné.

Překladová gramatika

PG = (N, T u D, P, S) Obsahuje disjunktní množiny T a D, vstupních a výstupních terminálních symbol

4. Nedeterministický a deterministický konečný automat.

Deterministický konečný automat

Základní popis

- konečný počet stavů
- konečný počet vstupů
- jednosznačně určený následující stav
- jednoznačně určený počáteční stav

Formální definice

Deterministický konečný automat je uspořádaná pětice **A= (S, Σ, P, s, F)**, kde:

- **S** je konečná množina stavů.
- **Σ** je konečná množina vstupních symbolů nazývaná abeceda.
- **P** je tzv. přechodová funkce (též přechodová tabulka), popisující pravidla přechodů mezi stavy.
- **s** je počáteční stav (s naleží S)
- **F** je množina koncových stavů (F je podmnožinou S)

Nedeterministický konečný automat

Nedeterministickým konečným automatem (**NKA**) bez výstupu nazýváme každou pětici **A = (Q, Σ, δ, S, F)**, kde:

- **Q** je konečná, neprázdná, množina stavů
- **Σ** je konečná neprázdná množina vstupních symbolů (vstupní abeceda)
- **δ** (přechodová funkce) je zobrazení δ: QxΣ → P(Q). Kde P(Q) je potenční množina (množina všech podmnožin množiny Q včetně prázné množiny e)
- **S** je množina počátečních stavů (S náleží Q) - není jednoznačně určen počáteční stav
- **F** je množina koncových stavů (F náleží Q)

Oborem hodnot přechodové funkce jsou všechny podmnožiny množiny stavů.

Formálně je definován podobně jako DKA, ale obsahuje prvky nedeterminismu:

1. nejednoznačně určený počáteční stav (může jich být více)
2. nejednoznačné přechody (při přijetí stejného vstupu lze přejít do více stavů)
3. e - přechody (přechod do stavu bez přijetí vstupního symbolu)

- chování NKA lze popsat sekvencí pozic (množina stavů, ve kterých se automat může nacházet), z nichž každá jednoznačně definuje, zda je zpracovaný řetězec akceptován či zamítnut
- pozic je konečný počet
- přechody mezi pozicemi jsou jednoznačné
- nejdůležitější rozdíl mezi DKA a NKA je v tom, že výsledkem přechodové funkce není pouze jeden stav, ale množina stavů, která může být i prázdná
- to vše jsou vlastnosti DKA a proto **ke každému NKA existuje ekvivalentní DKA**

V případě nedeterministického konečného automatu (NKA) je vstupní slovo akceptováno (rozpoznáno,) pokud toto slovo může automat převést do některého z koncových stavů (množina F) z některého z počátečních stavů (množina S).

Převod NKA na DKA

1. Lineární gramatiku nejprve převedeme na regulární tvar (postup viz otázka [Ekvivalence konečných automatů a regulárních gramatik].
2. Pak zkonstruujeme nedeterministický konečný automat a z něho nakonec deterministický (jak viz dále).
    a. Hlavní myšlenka je taková, že **každý stav vytvořeného DFA odpovídá množině stavů NFA**.
    b. Nebo: Z nedeterministického automatu se vytváří **strom**, který již popisuje deterministický automat, popisující tentýž problém.

Postup převodu

- Samotný převod stojí na myšlence, že pokud lze ze vstupního uzlu ***S*** přejít do uzlu ***A*** a do uzlu ***B***, tak vytvoříme nový uzel, říkejme mu ***|A,B|***.
- Tento uzel bude mít stejné vstupy a výstupy jako sjednocení uzlů ***A*** a ***B*** .
- Nyní tabulka převedeného automatu obsahuje dva uzly ***{S, | A,B|}*** .
- Postup opakujeme pro uzel ***| A,B|***.
- Takto postupně projdeme všechny stavy nově vytvářeného deterministického automatu.
- Koncovými uzly převedeného deterministického automatu budou takové uzly, které jsou nadmnožinou koncových uzlů původního automatu
    - měl-li původně automat výstupní uzel ***A***, tak uzel ***|A,X|***, který vznikl jako sjednocení uzlu ***A*** a uzlu ***X***, bude také výstupní
- Tento postup zároveň eliminuje všechny stavy, do kterých se deterministická verze automatu nemůže vůbec dostat.
- Zároveň ale mohou vzniknout uzly, které mají totožné vlastnosti (vstupní a výstupní uzly, konečnost, vlastnost být počátečním uzlem).
    - Tyto uzly můžeme po doběhnutí algoritmu ztotožnit.

Příklad

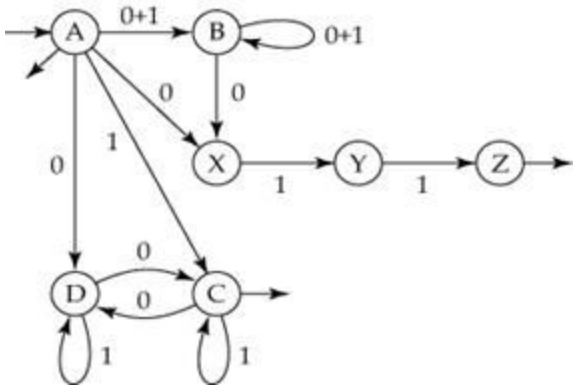
Zadaná pravá lineární gramatika

```
A --> B | C
B --> 0B | 1B | 011
C --> 0D | 1C | e
D --> 0C | 1D
```

Pravá regulární gramatika

```
A --> 0B | 1B | 0X | 0D | 1C | e
B --> 0B | 1B | 0X
X --> 1Y
Y --> 1Z
Z --> e
C --> 0D | 1C | e
D --> 0C | 1D
```

Nedeterministický konečný automat



Deterministický konečný automat

Přechodovou tabulku deterministického konečného automatu vytvoříme z přechodového diagramu nedeterministického kon. automatu takto:

- Do prvního řádku tabulky zapíšeme počáteční stav automatu a postupně zjistíme, do jakých množin stavů se nedeterministický automat může dostat z tohoto stavu přijmutím jednotlivých symbolů jeho vstupní abecedy.
- Z nalezených množin s více než jedním stavem vytvoříme tzv. kompozitní stavy det. automatu. Ty pak použijeme do přechodové tabulky det. automatu jako výstupy přechodové funkce pro počáteční stav a odpovídající vstupní symboly.
- Vzniklé kompozitní stavy (a případně i normální stavy) také využijeme v dalších řádcích přechodové tabulky a případně doplňujeme nové kompozitní stavy, do kterých se můžeme dostat z množin původních stavů každého kompozitního stavu přes vstupní symboly.
- Takto postupně vytvoříme celou přechodovou tabulku ekvivalentního deterministického automatu.
- Kompozitní stavy, zahrnující původní koncové stavy, můžeme označit také jako koncové.

	stav	0	1
↔	A	BXD	BC
	BXD	BXC	BYD
←	BC	BXD	BC
←	BXC	BXD	BYC
	BYD	BXC	BZD
←	BYC	BXD	BZC
←	BZD	BXC	BD
←	BZC	BXD	BC
	BD	BXC	BD

Nové stavy jsou A,BXD, BC, BXC, atd.

Přechody 0,1.

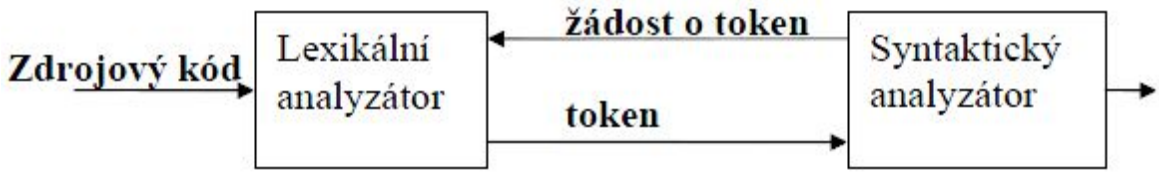
5. Lexikální analýza, princip činnosti.

Úkoly lexikálního analyzátoru

- Čtení zdrojového textu,
- Nalezení a rozpoznání lexikálních symbolů ve volném formátu textu, včetně případného rozlišení klíčových slov a identifikátorů. Vyžaduje spolupráci s SA.
- Vynechání mezer a komentářů,
- Interpretace direktiv překladače,
- Uchování informace pro hlášení chyb,
- Zobrazení protokolu o překladu.

Lexikální analyzátor

- je samostatnou částí pro jednodušší návrh překladače, zlepšení efektivity překladu a lepší přenositelnost.
- rozpoznává a zakóduje lexikální symboly jazyka = **lexémy** (anglicky *tokens*)



- **lexikální symboly jsou regulárním jazykem**

Lexikální analýza

- Je prováděna **lexikálním analyzátorem**, který je vstupní a nejjednodušší částí překladače.
- Čte znaky zdrojového programu, a jeho výstupem jsou **tokeny**.
- Vstupní posloupnost znaků - program - je slučována do lexikologicky smysluplných mnohoznakových jednotek, tzv. **lexémů** (např. if, foo123bar).
- Tokeny pak symbolicky reprezentují lexémy (např. if pro lexém klíčové slovo if, id pro identifikátor foo123bar) a lexémy jsou tak vlastně jejich instance.
- Podoba lexémů reprezentujících jednotlivé tokeny je vymezena **vzorem** (**pattern**), typicky regulárním výrazem.
- Kromě toho je jeho úkolem odstranění komentářů a eliminace přebytečných bílých znaků.

Token

- je tvořen dvěma částmi
  - **názvem tokenu** (token name)
  - **hodnotou atributu** (attribute value).
- názvy tokenu jsou často abstraktní symboly, které jsou pak použity parserem pro syntaktickou analýzu.
  - např. klíčové slovo nebo o soubor znaků představujících identifikátor
  - operátory, klíčová slova a další ve skutečnosti atributové hodnoty nepotřebují
  - pokud má token hodnotu atributu → jde o pointer do tabulky symbolů, která obsahuje dodatečné informace o tokenu, které nejsou součástí gramatiky

Parser

- Proud tokenů je předán parseru pro syntaktickou analýzu.
- Lexikální analyzátor také obvykle používá tabulku symbolů, do které ukládá objevené lexémy a ze kterých bere informace, aby mohl parseru podstrčit správný token.

- Jak název tokenu (typ - id, číslo,...), tak jeho atribut (číslo 0/1,...) ovlivňují rozhodování ve fázi parsování a pozdějších fázích.
- Parser proto potřebuje od analyzátoru dostat další token ke zpracování včetně informací z tabulky symbolů (viz obrázek níže).

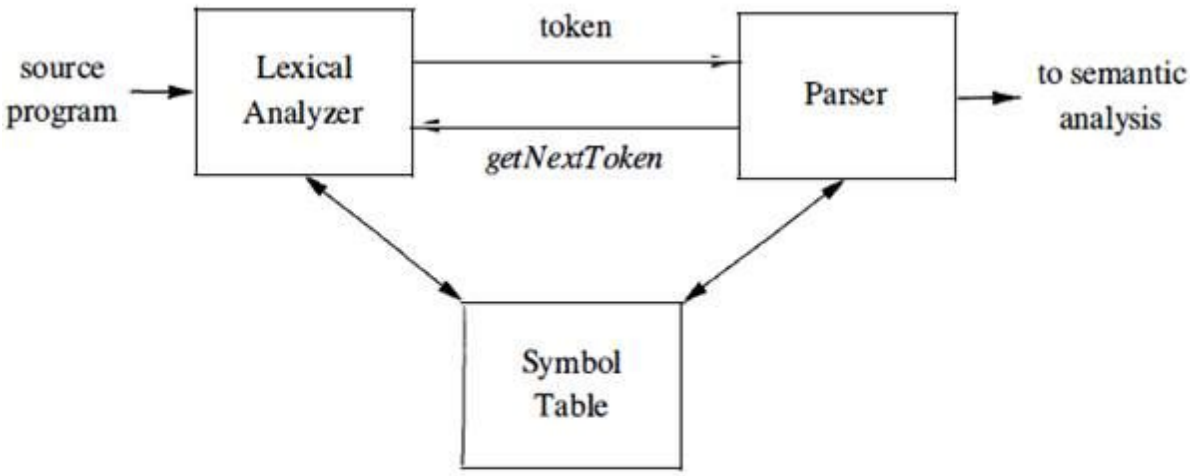


Figure 3.1: Interactions between the lexical analyzer and the parser

Princip činnosti

Princip lexikálního analyzátoru – nalezení a rozpoznání lexikálního symbolu

Třídy symbolů

- Identifikátory
- Klíčová slova (rezervované identifikátory)
- Celá čísla
- Jednoznakové omezovače
- Dvouznakové omezovače

Jiný postup z přednášky

- zpracování začíná prvým dosud nezpracovaným znakem ze vstupu
- zpracování končí, je-li automat v koncovém stavu a pro další vstupní znak již neexistuje žádný přechod
- pro každou kategorii předpokládáme samostatný koncový stav
- neohodnocená větev se vybere, pokud vstupujícímu znaku neodpovídá žádná z ohodnocených větví

Nejednoznačnost v lexikální analýze

- V lexikální analýze mohou nastat nejednoznačnosti, pokud je **jeden symbol prefixem jiného symbolu** (== apod.).
- Pak se hledá nejdelší symbol a je vyžadována nápověda od syntaktického analyzátoru.
- Často je potřeba dopředu skenovat vstup, aby se zjistilo, kde následující lexém končí.
    - Proto lexikální analyzátory typicky bufferují vstup.

6. Konstruktory lexikálních analyzátorů.

Lex

Program LEX (Lexical Analyzer Generator)

- Slouží k tvorbě jiných programů, které mají cosi udělat se vstupním (textovým) souborem za pomoci lexikální analýzy.
- Tím se myslí analýza struktur, které se dají zapsat lineárními gramatikami, konečnými automaty nebo regulárními výrazy.
- Typické použití LEXu je dvojí
  - vytvořený program pracuje samostatně
  - nebo slouží jako vstupní filtr pro jiný (syntaktický) analyzátor, např. **bison** či **yacc**.
- Lex umožňuje vytvořit lexikální analyzátor uvedením regulárních výrazů, které popisují vzory (**patterns**) pro tokeny.
- Vstupní notace pro Lex se nazývá **Lex language** a samotný nástroj je **Lex compiler**.
- Kompilátor Lexu transformuje vstupní vzory do přechodového diagramu (jádrem toho všeho je konečný automat.) a generuje kód do souboru **lex.yy.c**, ve kterém je simulován přechodový diagram.

Vstupem LEXu je soubor, obvykle s koncovkou .l, např. lex.l, je napsaný v jazyce Lexu a popisuje lexikální analyzátor k vygenerování (rozpoznávaná slova a akce, které se mají po jejich rozpoznání provést).

Slova (tokeny) se popisují regulárními výrazy, akce v cílovém programovacím jazyku.

Výstup LEXu je zdrojový kód hotového programu, který se potom musí běžným způsobem přeložit ⇒ pokud tedy používáme variantu LEXu, která generuje výstup v jazyku C, musí být i akce zapsané v jazyku C.

Lex kompilátor překlopí lex.l do programu v C, který je vždy uložen v souboru lex.yy.c. Pak je tento soubor zkompilován vždy do a.out. Výstupem je fungující lexikální analyzátor, který bere proud vstupních znaků a vytváří z nich proud tokenů.

Hodnoty atributů (= numerický kód/pointer do tabulky symbolů/nic) jsou umístěny v globální proměnné yylval, kterou sdílí lexikální analyzátor a parser.

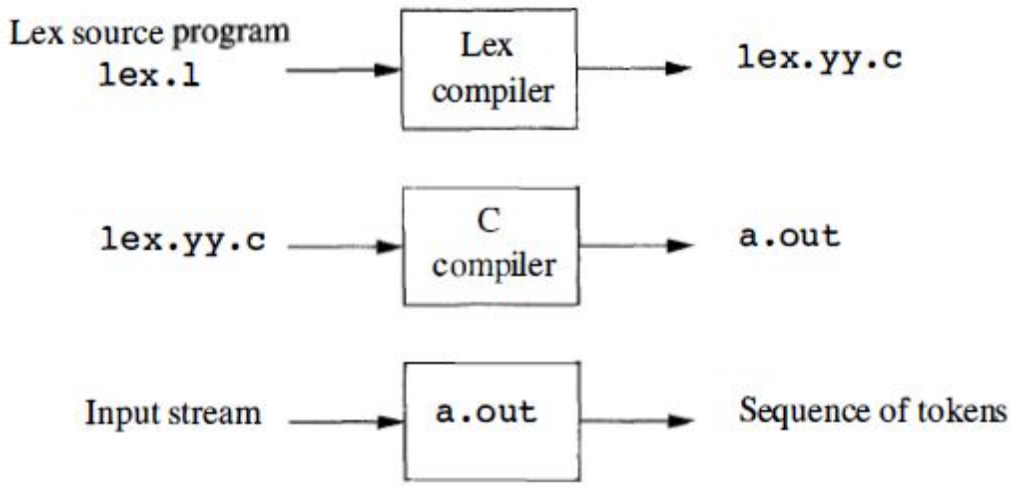


Figure 3.22: Creating a lexical analyzer with Lex

Struktura programu LEX

```
deklarace, definice
%%
popis slov a akcí
%%
další funkce (zapsané v cílovém jazyku)
```

Program, který ve vstupním souboru nahradí všechny identifikátory slovem "IDENTIFIKATOR"

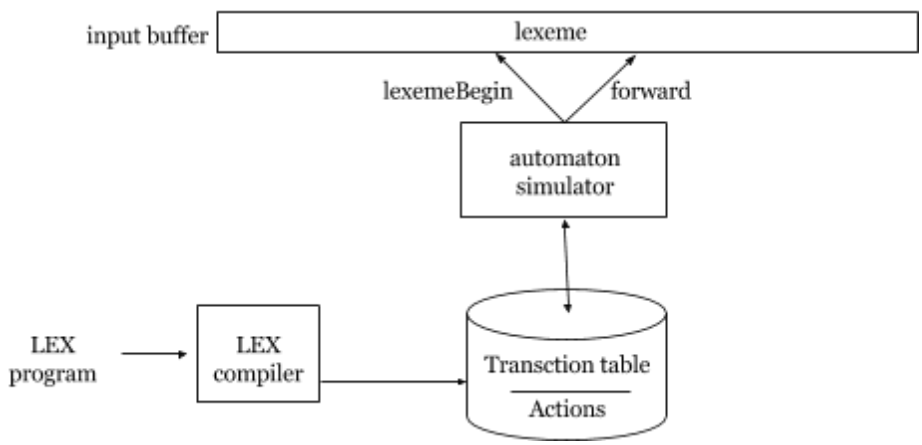
```
%%
[a-zA-Z_][0-9a-zA-Z_]*   printf("IDENTIFIKATOR");
%%
int main(void) {
   yylex();
   return 0;
}
```

Formát regulárních výrazů (popis slov)

x	znak "x"
[xy]	znak "x" nebo "y"
[x-z]	všechny znaky od "x" až k "z"
[^x]	jakýkoliv znak vyjma "x"
.	jakýkoliv znak, až na novou řádku
^x	znak "x", pokud se nachází na začátku řádky
x$	znak "x", pokud se nachází na konci řádky
x*	libovolný počet znaků "x"
x+	alespoň jeden znak "x"
x?	jeden nebo žádný znak "x"
x{m,n}	m až n výskytů znaku "x"
x|y	znak "x" nebo "y"
(x)	znak "x"
x/y	znak "x", je-li následován znakem "y"
{DEF}	doplnění definice z úvodní sekce
x	znak "x", je-li splněna podmínka y

Architektura lexikálního analyzátoru generovaného Lexem

Lex z definovaných regulárních výrazů ze vstupního souboru NKA DKA; pravidlo: v případě konfliktů přiřazuje lexém vzoru dle nejdelšího prefixu.



Další varianty LEXu

- **Flex** – volně dostupná implementace Lexu, pro C.
- **JLex** – volně dostupná implementace Lexu, pro Javu.
- **C# LEX** - varianta JLex pro C#.
- **PLY** - implementace Lexu v Pythonu

7. Bezkontextové gramatiky a zásobníkové automaty, formální popis, ekvivalence.

Bezkontextové gramatiky

- bezkontextové gramatiky (BKG) jsou gramatiky typu 2 podle Chomského hierarchie.
- gramatika je bezkontextová, **tvoří-li levou stranu všech přepisovacích pravidel právě jeden neterminální symbol**
- skládají se tedy z pravidel $A \rightarrow \gamma$ kde A je právě jeden neterminál a γ je řetězec terminálů a neterminálů
- pravidlo $S \rightarrow e$ je povoleno, pokud se S nevyskytuje na pravé straně žádného pravidla
- jazyky generované touto gramatikou jsou **rozpoznatelné nedeterministickým zásobníkovým automatem**

Formální definice

- *skládají se z terminalů, neterminálů, počátečního symbolu a přepisovacích (produkčních) pravidel*

G = (N,T,P,S)

- **T** = Terminály – jde o názvy tokenů (klíčová slova if, else, symboly „(“, „)“ atd.)
- **N** = Neterminály – syntaktické proměnné, pomáhají definovat jazyk generovaný gramatikou; zavádějí hierarchickou strukturu jazyka, která je klíčová pro syntaktickou analýzu
- **S** = Počáteční symbol – jeden z neterminálů
- **P** = přepisovací pravidla - ve tvaru $A \rightarrow \gamma$, kde $A \in N$ a $\gamma \in N \cup T$

Příklad

- jazyk $L = \{0^n1^n\}$, pro $n \geq 0$
- takovýto jazyk není rozpoznatelný konečným automatem
- zásobníkovým automatem ano („konečný automat neumí počítat“)
  - u programovacích jazyků by to třeba znamenalo, že není možné závorkovat (vnořovat kód) do libovolné úrovně.
- pro tento jazyk platí: N = {S}, T = {0,1}, P = {S → 0S1, S → e}, S = {S}

Zásobníkový automat

Formální definice

Zásobníkový automat je definován jako uspořádaná sedmice $(Q,T,G,\delta, q_o,z_o,F)$

- **Q** je konečná množina vnitřních stavů,
- **T** je konečná vstupní abeceda,
- **G** je konečná abeceda zásobníku,
- **δ** je tzv. přechodová funkce, popisující pravidla činnosti automatu (jeho program), je definováno jako zobrazení: $Qx(T \cup \{e\})xG^* \text{ do } QxG^*$
- $\mathbf{q_o}$ je počáteční stav,
- $\mathbf{z_o}$ popisuje symboly uložené na počátku v zásobníku,
- **F** je množina přijímajících stavů, $F \subseteq Q$

Je vidět, že zásobníkový automat se v podstatě skládá z konečného automatu, který má navíc k dispozici potenciálně nekonečné množství paměti ve formě zásobníku. Obsah tohoto zásobníku ovlivňuje činnost automatu tím, že vstupuje jako jeden z parametrů do přechodové funkce.

Zásobníkový automat se od konečného automatu liší ve dvou směrech

1. Využívá vršek zásobníku při rozhodování jaký přechod provést.
2. Může manipulovat se zásobníkem jako součást provádění přechodu.

Popis činnosti automatu

- Na počátku se automat nachází v definovaném počátečním stavu a zásobník obsahuje pouze počáteční symboly.
- Dále v každém kroku podle aktuálního stavu, symbolů na vrcholu zásobníku a symbolu na vstupu provede přechod, při kterém může vyjmout ze zásobníku několik symbolů, vložit místo nich jiné a na vstupu přečíst další symbol.
- Toto se opakuje.
- Po dokončení činnosti (po přečtení celého vstupu, pokud do té doby nedojde k chybě) je rozhodnuto, jestli automat vstupní řetězec přijal. K tomu mohou sloužit dvě kritéria:
    - stav, ve kterém se na konci automat nachází, patří do množiny přijímajících stavů
    - nebo zásobník je na konci prázdný

Konfigurace automatu

- lze popsat uspořádanou trojicí (q, w, α)
    - **q** - vnitřní stav
    - **w** - dosud nezpracovaná část vstupu
    - **α** - obsah zásobníku
- na počátku práce je automat v konfiguraci (q0, w, z0)

Příklad akceptace řetězce zásobníkovým automatem: viz cvičení 10 (LL gramatiky) na courseware FJP

Vztah bezkontextových gramatik a zásobníkových automatů

Zásobníkové automaty jsou ekvivalentní bezkontextovým gramatikám: pro každou bezkontextovou gramatiku existuje zásobníkový automat, který generuje (akceptuje) identický jazyk generovaný touto gramatikou a naopak.

Pro danou BKG gramatiku W=(N, T, P, S) můžeme sestrojit zásobníkový automat P takový, že L(W)=L(P). Jsou dvě varianty:

1. Konstrukce zásobníkového automatu, který je modelem **syntaktické analýzy shora dolů**:
    - Q = {q} (automat má jen jeden vnitřní stav),
    - T je shodná s množinou terminálních symbolů rozpoznávané gramatiky,
    - G = N+T, tj. v zásobníku se může vyskytnout jakýkoliv symbol rozpoznávané gramatiky,
    - δ je dáno rozkladovou tabulkou,
    - q0 = q, počáteční stav automatu je q, neboť automat jiné stavy nemá,
    - z0 = S, tj. na počátku je v zásobníku startovací symbol gramatiky
    - F = {}, což se interpretuje jako "automat akceptuje vyprázdněním zásobníku".
    - LL(k) gramatiky
2. **Analýza zdola nahoru** je obecnější a vyžaduje trochu složitější automat:
    - Q = {q, r}, stav q je "pracovní", stav r "akceptační",
    - T je shodná s množinou terminálních symbolů rozpoznávané gramatiky,
    - G je v nejjednodušším případě rovno N+T+{#}, tj. sjednocení symbolů gramatiky a speciálního symbolu "#"; deterministický automat může mít množinu G složitější
    - δ je dáno rozkladovou tabulkou,
    - q0 = q, z0 = #, F = {r}
    - gramatiky: LR, SLR, LALR

8. Nedeterministický syntaktický analyzátor.

Při syntaktické analýze konstruujeme derivační strom. Podle toho, jak je konstruován derivační strom věty, rozlišujeme dvě základní metody syntaktické analýzy:

1. **metoda shora dolů**
    - derivační strom konstruujeme od kořene k listům a zleva doprava (provádíme levou derivaci)
2. **metoda zdola nahoru**
    - postupujeme od listů směrem ke kořeni, ale také zleva doprava (provádíme pravou derivaci)

K syntaktické analýze se využívají zásobníkové automaty (ZA), které jsou obecně nedeterministické (nepoužitelné pro SA). Pro konstrukci SA lze použít buď:

- Deterministickou simulaci nedeterministického ZA = algoritmus syntaktické analýzy s návraty.
- Zdokonalit konstrukci ZA tak, aby byl pro určitou třídu BKG deterministický (pohled do zásobníku nebo dále do vstupního řetězce).

Obecný popis nedeterminismu a determinismu

Základem **nedeterminismu** je tedy vždy problém **výběr správného pravidla**, ať už analýzou shora dolů nebo zdola nahoru. Pokud se nemá analyzátor podle čeho rozhodnout, prostě **prochází prostor všech řešení** buď do šířky nebo do hloubky (backtracking) a hledá to správné řešení. Tato metoda je tedy značně **neefektivní**, protože v nejhorším případě může projít všechny možnosti a nenajít žádné správné řešení, tedy správnou množinu pravidel, jejichž expanzí/redukcí lze dosáhnout požadovaného výsledku.

V případě, že chceme analyzovat vstup **deterministicky**, musíme analyzátor **zdokonalit**. Ten musí mít přesnou informaci o tom, jaké pravidlo gramatiky v danou chvíli použít. Automat může využít informaci o **dosud provedené částečné derivaci** a také o **vstupu**, který ještě nebyl zpracován. Zásobníkovému automatu, který je abstraktním modelem syntaktického analyzátoru, tedy připravíme rozkladovou tabulku (lookup table), ve které bude určeno, jaké pravidlo má analyzátor použít podle vstupu a stavu zásobníku.

Metoda shora dolů

Derivační strom konstruujeme od kořene (ohodnoceného startovním symbolem) dolů k listům, zleva doprava podle levé derivace. Jedná se o zásobníkový automat LL. Počáteční konfigurace automatu se dá popsat uspořádanou trojicí **(q, w, α)**, kde:

- **q** =vnitřní stav
- **w** = dosud nezpracovaní část vstupu
- **α**= obsah zásobníku
- na počátku práce je automat v konfiguraci (q_0, w, z_0), napr. (q, abaaab, s).

Pokud jen generujeme větu v gramatice, můžeme v případě více pravidel se stejnou levou stranou náhodně vybírat. Naším úkolem však bývá spíše analýza již existující věty. Zde již náhoda nepřipadá v úvahu, protože posloupnost pravidel pro levou derivaci již nemusí být jednoznačná. Potřebujeme automat, který tuto analýzu provádí, a tento automat musí mít možnost jednoznačně vybírat mezi pravidly to správné.

Analýza s návratem (nedeterministická)

- postupně zkoušíme vhodná pravidla. Nejdřív první, pokračujeme dále ve výpočtu, a když se ukáže, že pravidlo nevyhovuje (dostaneme se do slepé uličky), vrátíme se zpátky a vyzkoušíme druhé pravidlo atd. Tato metoda je sice účinná. ale zbytečně pomalá.
- *rekurzivní sestup*

Deterministická analýza

- při výběru pravidla se řídíme dalšími informacemi. Může to být pohled do budoucnosti, kdy se díváme dále do vstupní posloupnosti symbolů a řídíme se tím, co později dostaneme na vstupu. Nebo například kontrolujeme obsah zásobníku (nestačí nám pouze vidět ten symbol, který ze zásobníku vyjímáme, ale i další, které jsou pod ním).
- *LL parsery*

Metoda zdola nahoru

- Konstruujeme derivační strom zdola od listů nahoru ke kořeni, přičemž postupujeme zleva doprava.
- Stejně jako u první metody i zde budeme používat lineární rozklad, tentokrát pro pravou derivaci - je to proces nalezení pravého rozkladu věty (LR gramatika).
- I zde musíme rozhodovat, která pravidla chceme použít. Tentokrát však nejde o pravidla se stejnou levou stranou (pro stejný neterminál), ale rozhodujeme se mezi pravidly, která mají podobnou pravou stranu a jsou proto použitelná pro tentýž podřetězec větné formy.

Analýza s návratem (nedeterministicky)

- Vybereme ve větné formě jeden podřetězec (jako první vybíráme ten, který začíná nejvíc nalevo, je co nejdelší a je shodný s pravou stranou některého pravidla), přepíšeme neterminálem na pravé straně pravidla a pokračujeme v konstrukci derivačního stromu.
- Pokud zjistíme, že tento krok nevede k úspěchu, vyzkoušíme jiný podřetězec atd. Tato metoda je příliš časově náročná.

dDeterministická analýza (LALR, SLR)

- Využíváme další informace získané při překladu, např. obsah nepřečtené části vstupního kódu nebo obsah zásobníku.

9. Derivace a derivační strom, víceznačnost gramatiky.

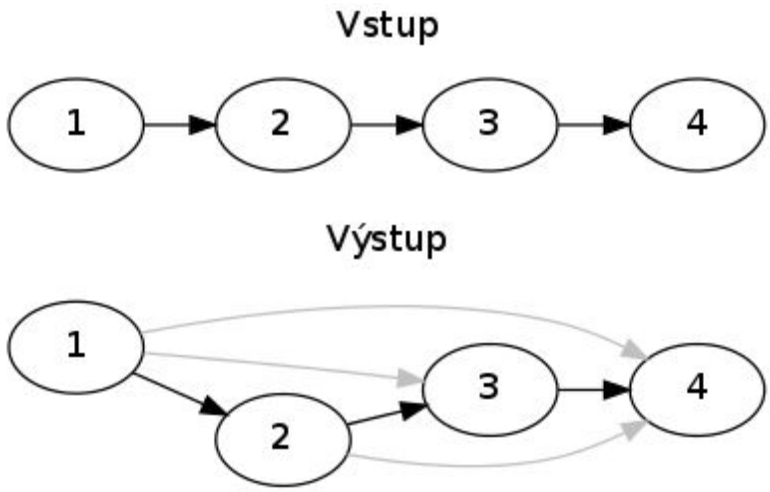
Derivace

- Posloupnost kroků odvození terminálu pomocí přepisovacích pravidel gramatiky
- Derivační pohled odpovídá konstrukci parsovacího (syntaktického) stromu shora dolů (top-down).
- Parsování zdola nahoru (bottom-up) je spjato s pravými derivacemi.
- Podle toho, který neterminál nahradit v každém kroku derivace, se rozlišují levá a pravá derivace.

DERIVACE řetězce **α** je posloupnost kroků odvození **α** pomocí přepisovacích pravidel gramatiky

$S = \alpha_1 \Rightarrow \alpha_2 \Rightarrow \ldots \Rightarrow \alpha_n = \alpha$ (totéž pro $S \Rightarrow^* \alpha$)

* je uzávěr relace (všechny přechody kam se dá transitivně dostat)



PŘÍMÁ DERIVACE: α A β => α γ β ,kde A -> γ ϵ P *(pozn.: P je množina pravidel, pomocí kterých lze odvodit jazyk)*

Derivační strom

Derivační strom (parse tree) je orientovaný acyklický graf a je grafickou reprezentací, která říká, v jakém pořadí byla přepisovací pravidla uplatňována na neterminály, tedy jak vznikla věta jazyka.

- Kořen stromu je označen startovacím symbolem gramatiky
- Každý vnitřní uzel je ohodnocen neterminálními symboly.
- Listy jsou ohodnoceny terminálními symboly.
- Listy se čtou zleva doprava a dávají větu (generovanou gramatikou).
- Jestliže uzly n1, n2, . . . , nk jsou bezprostřední následníci uzlu n, jsou ohodnoceny symboly A1, A2, . . . , Ak a uzel n je ohodnocen A, pak v množině pravidel gramatiky existuje pravidlo A A1A2 ... Ak.
- Není třeba značit orientaci hran.

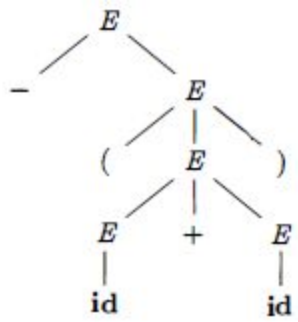
Jiná definice jazyka generovaného gramatikou je množina vět, které mohou být vytvořeny derivačním stromem. Proces hledání derivačního stromu pro danou větu (řetězec terminálů) se nazývá parsování tohoto řetězce

Derivační strom ignoruje variace v pořadí, v jakém jsou symboly přepisovány. Proto je mezi derivacemi a derivačními stromy vztah 1:N – např.:

$$E \Rightarrow -E \Rightarrow -(E) \Rightarrow -(E+E) \Rightarrow -(\mathbf{id}+E) \Rightarrow -(\mathbf{id}+\mathbf{id})$$

$$E \Rightarrow -E \Rightarrow -(E) \Rightarrow -(E+E) \Rightarrow -(E+\mathbf{id}) \Rightarrow -(\mathbf{id}+\mathbf{id})$$

- jsou různé derivace, jejich derivační strom ale vypadá stejně:



- Pro získání jednoznačného derivačního stromu pro derivaci se proto užívá buď pravá a nebo levá derivace.

Víceznačnost gramatik

Gramatika, která generuje větu, pro níž lze sestavit aspoň dva různé derivační stromy, je víceznačná. Jinak řečeno: je to taková gramatika, která produkuje více než jednu levou nebo víc než jednu pravou derivaci pro tutéž větu.

Příklad

Pro gramatiku $E \rightarrow E + E \mid E * E \mid (E) \mid \mathbf{id}$ umožňuje vytvořit dvě levé derivace pro id + id * id (násobení není upřednostněno před sčítáním):

$$
\begin{array}{lll}
E \Rightarrow E+E & \qquad & E \Rightarrow E*E \\
\Rightarrow \mathbf{id}+E & & \Rightarrow E+E*E \\
\Rightarrow \mathbf{id}+E*E & & \Rightarrow \mathbf{id}+E*E \\
\Rightarrow \mathbf{id}+\mathbf{id}*E & & \Rightarrow \mathbf{id}+\mathbf{id}*E \\
\Rightarrow \mathbf{id}+\mathbf{id}*\mathbf{id} & & \Rightarrow \mathbf{id}+\mathbf{id}*\mathbf{id}
\end{array}
$$

- Nutnou podmínkou jednoznačnosti gramatiky je, aby pro žádný neterminální symbol neexistovalo jak pravidlo rekurzivní zprava, tak i pravidlo rekurzivní zleva
- Problém nejednoznačnosti bezkontextových jazyků je algoritmicky nerozhodnutelný.
- Je potřeba buďto vytvořit jednoznačné gramatiky pro kompilaci aplikací, nebo u nejednoznačných gramatik zavést dodatečná pravidla, která řeší případné nejednoznačnosti.

Odstranění levé rekurze

- Levorekurzivní gramatiku nelze použít k analýze shora dolů
- Odstranění pravidla rekurzivního zleva:
  - Nechť je dána BKG G = (N, T, P, S), ve které,
    - A → Aα1 | Aα2 | ... | Aαm | β1 | β2 | ... | βn
      jsou všechna A pravidla v P a žádné z β nezačíná A.
  - Pak G' = (N ∪ { A'}, T, P', S), kde P' obsahuje místo uvedených pravidel pravidla:
    - A → β1 | β2 | ... | βn | β1 A'| β2 A'| ... | βn A'
  - A' → α1 | α2 | ... | αm | α1 A' | α2 A'| ...| αm A'

10. Deterministická syntaktická analýza.

Je k ní zapotřebí **prediktivní parse**r = rekurzivně sestupný parser, který nevyžaduje zpětné kroky. Je ho možné vytvořit jen pro LL(k) gramatiky, což jsou bezkontextové gramatiky, pro které existuje kladné k, které umožní rekurzivně sestupnému parseru rozhodnout se, které přepisovací pravidlo použít na základě k dalších načtených symbolů. LL(k) gramatiky vylučují mnohoznačnost a levou rekurzi. Jakákoliv bezkontextová gramatika může být transformována na ekvivalentní nelevorekurzivní gramatiku, ale odstranění levé rekurze ne vždy vede k LL(k) gramatice. Prediktivní parser běží v lineárním čase.

Deterministická syntaktická analýza využívá další informace získané při překladu – obsah zásobníku a obsah nepřečtených vstupních tokenů. Na základě těchto informací umožňují následující funkce vhodný výběr přepisovacích pravidel, a tím prediktivní parsování:

FIRST(A)

= množina terminálů, kterými mohou začínat řetězce odvozené z A **(terminály na začátku A)**

- funkce first zjišťuje, co vznikne přepsáním jednotlivých neterminálů na levé straně všech pravidel
- $A \Rightarrow bxayB \rightarrow b \in FIRST(A)$

Algoritmus výpočtu

- FIRST(A) pro A = terminál nebo ***e***: je terminál/e
- Když je A neterminál:
  - Je to první terminál ve všech přepisovacích pravidlech s A na levé straně
  - Pokud jsou v přepisovacím pravidle na P straně jen neterminály, hledá se FIRST prvního neterminálu na pravé straně
  - Pokud lze nějaký z těchto neterminálů přepsat na prázdný řetězec e, je potřeba se podívat na first neterminálu následujícího po něm v některém z přepisovacích pravidel

FOLLOW(A)

= množina terminálů, které mohou následovat za A v některé větné formě v derivacích (terminály hned za A)

- $S \Rightarrow bxCAyZ \rightarrow y \in FOLLOW(A)$

Algoritmus výpočtu

1. Polož FOLLOW (A) = θ
2. Je-li A počáteční symbol G, přidej e do FOLLOW(A)
3. Pro všechny pravé strany pravidel z G tvaru α A β přidej FIRST (β) do FOLLOW (A), nepřidávej ale e.
4. Je-li v G pravidlo L → α A nebo L → α A β , kde FIRST (β) obsahuje e, pak přidej do FOLLOW (A) množinu FOLLOW (L)

- ***Vytváříme vždy pro všechny neterminály zároveň!***

FIRSTk(A), FOLLOWk(A) = zobecnění na množiny terminálních řetězců o délce nejvýše **k,** tyto funkce slouží k vytvoření **rozkladové tabulky**, která nahrazuje přechodovou funkci. V první řádce jsou uvedeny všechny možné vstupy, v prvním sloupci všechny možné stavy vrcholu zásobníku (vč. dna zásobníku #). Má stavy:

- **srovnání (pop)** – na vstupu i na vrcholu zásobníku jsou stejné hodnoty
- **přijetí (accept)** – bylo dosaženo dna zásobníku a přijímá se prázdný symbol e
- **expanze (expand)** – aplikace přepisovacího pravidla, které je uvedené v buňce tabulky určené vstupem (který sloupec) a vrcholem zásobníku (který řádek)
- **chyba (error)** – pokud pro vstup a hodnotu na vrcholu zásobníku je buňka v tabulce prázdná vstupní řetězec není větou jazyka

11. Rekurzivní sestup.

Rekurzivní sestup (rekurzivní sestupný parser)

- postupuje shora dolů a je sestaven ze vzájemně se volajících procedur
- každá taková procedura obvykle implementuje jedno přepisovací pravidlo gramatiky
    - kromě něj do top-down parserů patří prediktivní parsery založené na LL(k) gramatikách

Prediktivní parser

- rekurzivně sestupný parser, který nevyžaduje zpětné kroky
- je ho možné vytvořit jen pro LL(k) gramatiky, což jsou bezkontextové gramatiky, pro které existuje kladné k, které umožní rekurzivně sestupnému parseru rozhodnout se, které přepisovací pravidlo použít na základě k dalších načtených symbolů
- LL(k) gramatiky vylučují mnohoznačnost a levou rekurzi
- jakákoliv bezkontextová gramatika může být transformována na ekvivalentní nelevorekurzivní gramatiku, ale odstranění levé rekurze ne vždy vede k LL(k) gramatice
- prediktivní parser běží v lineárním čase.

Rekurzivní sestup s návratem

- technika určování použitého produkčního pravidla zkoušením všech pravidel
- není limitován na LL(k) gramatiky, ale nemá zaručeno skončit, pokud gramatika není LL(k)
- může vyžadovat exponenciální čas pro svůj běh

Princip

Hlavní myšlenka je taková, že pro každý neterminál gramatiky je implementována příslušná fce v programu.

- každému neterminálnímu symbolu A odpovídá procedura A
- tělo procedur je dáno pravými stranami pravidel pro A
- pravé strany musí být rozlišitelné na základě symbolů vstupního řetězce
- je-li rozpoznána pravá strana, pak v případě neterminálního symbolu vyvolá A proceduru pro rozpoznaný neterminální symbol, v případě terminálního symbolu, ověří A jeho přítomnost ve vstupním řetězci a zajistí přečtení dalšího znaku ze vstupu
- rozpoznané pravidlo analyzátor oznámí (např. jeho číslo)
- chybnou strukturu vstupního řetězce oznámí chybovým hlášením

Terminál na pravé straně je porovnán s dalším vstupním symbolem. Pokud se shodují, přejde se na další vstupní symbol a na další symbol na pravé straně. V opačném případě je nahlášena chyba.

O neterminál na pravé straně je postaráno voláním příslušné funkce. Po jejím vykonání se pokračuje dalším symbolem na pravé straně.

Pokud na pravé straně už nejsou žádné symboly, funkce končí (function returns).

Takto se postupně volají všechny funkce, až se nakonec opět ocitneme ve funkci pro startovací symbol, která byla zavolána jako první. Ta také oznamuje úspěšný průběh, pokud se prošel celý vstupní řetězec.

```
void A() {
1)      Choose an A-production, A → X1X2 ··· Xk;
2)      for ( i = 1 to k ) {
3)          if ( Xi is a nonterminal )
4)              call procedure Xi();
5)          else if ( Xi equals the current input symbol a )
6)              advance the input to the next symbol;
7)          else /* an error has occurred */;
        }
}
```

Tento pseudokód je nedeterministický, protože začíná volbou A-přepisovacího pravidla, které se používá blíže nepopsaným způsobem.

Obecně může rekurzivní sestup potřebovat backtracking, tzn. někdy je třeba se vrátit a opakovaně číst vstup. Backtracking je však potřeba zřídkakdy. Kód výše neumožňuje backtracking, bylo by je nutno modifikovat – v řádce 7 se pak vrátit na řádku 1 a zvolit jiné pravidlo, popř. nahlásit chybu, když už žádné další nejde použít.

Sémantické zpracování

Při rekurzivním sestupu se může provádět také sémantické zpracování. Sémantické zpracování zahrnuje vyhodnocení atributů symbolů v derivačním stromu.

Atributy = vlastnosti gramatických symbolů nesoucí sémantickou informaci (hodnota, adresa, typ, scope, spojitost mezi formálními a skutečnými parametry apod.).

Způsoby vyhodnocení

1. procházením stromem od listů ke kořenu = syntetizované atributy
2. procházením stromem od rodiče k potomkovi, od staršího bratra k mladšímu = dědičné atributy (např vícenásobné deklarace v C – int x,y,z)

Je nutné doplnit procedury lex. analýzy (LA) i syntakt. analýzy (SA) takto:

- LA bude předávat s přečteným vstupním symbolem i jeho atributy.
- procedury SA pro neterminály doplnit o:
    - vstupní parametry odpovídající dědičným atributům
    - výstupní parametry odpovídající syntetizovaným atributům
    - zavést lokální proměnné pro uložení atributů pravostranných symbolů
    - před vyvoláním procedury korespondujícího neterminálu z pravé strany vypočítat hodnoty jeho dědičných atributů
    - na konec procedury popisující pravou stranu pravidla zařadit příkazy vyhodnocující syntetizované atributy

Vlastnosti

- pro metodu rekurzivního sestupu, tj. analýza shora dolů, se používají LL gramatiky
- jednoduchá LL gramatika je taková gramatika, kde levou stranu tvoří právě jeden neterminální symbol a kde každá pravá strana začíná terminálním symbolem
- navíc musí platit, že např. pro pravidla A->... jsou počáteční symboly různé
- obecná LL gramatika nemá omezení, ale musí pro ni existovat rozkladová tabulka

12. Principy a podmínky LL analýzy.

LL-gramatika

LL gramatika je jakákoliv gramatika, z níž se dá udělat rozkladová tabulka pro LL parser.

LL(k) parser se kouká při parsování věty na následujících k tokenů, aby věděl, co dál. Pokud takový parser může být použit pro nějakou gramatiku, aniž by se musel použít backtracking, jedná se o **LL(k) gramatiku**.

Aby se ze vstupní gramatiky dala udělat LL(1) gramatika – eliminace levé rekurze, levá faktorizace (eliminace překrývajících se množin FIRST

Například: `STAT => if EXP then STAT | if EXP then STAT else  STAT => if EXP then STAT ElsePart;  ElsePart => else STAT | e)`

Podmínky

- Nesmí být přítomna levá rekurze.
- Nesmí dojít k first-follow (u neterminálu, který se přepisuje na "e") kolizi, first-first kolizi

Třídy jazyků LL(k)

L = Left to right -> vstupní text (soubor) čteme zleva doprava
L = Left parse -> vytváříme levý rozklad
K = při rozhodování mezi pravidly potřebujeme vidět nejvýše k znaků z nepřečtené části vstupu

Tzn.: **LL(k)** gramatika provádí deterministický rozbor *čtením textu z **L**eva doprava, s použitím **L**evé derivace a prohlédnutí **k** dalších symbolů vstupního textu.*

- gramatika je typu **LL(k)**, jestliže ji lze použít pro deterministickou syntaktickou analýzu metodou shora dolů (tj. vytváříme levý rozklad) a při rozhodování mezi pravidly potřebujeme znát nejvýše **k** symbolů ze vstupu.
- jazyk je typu **LL(k)**, pokud je generován některou **LL(k)** gramatikou

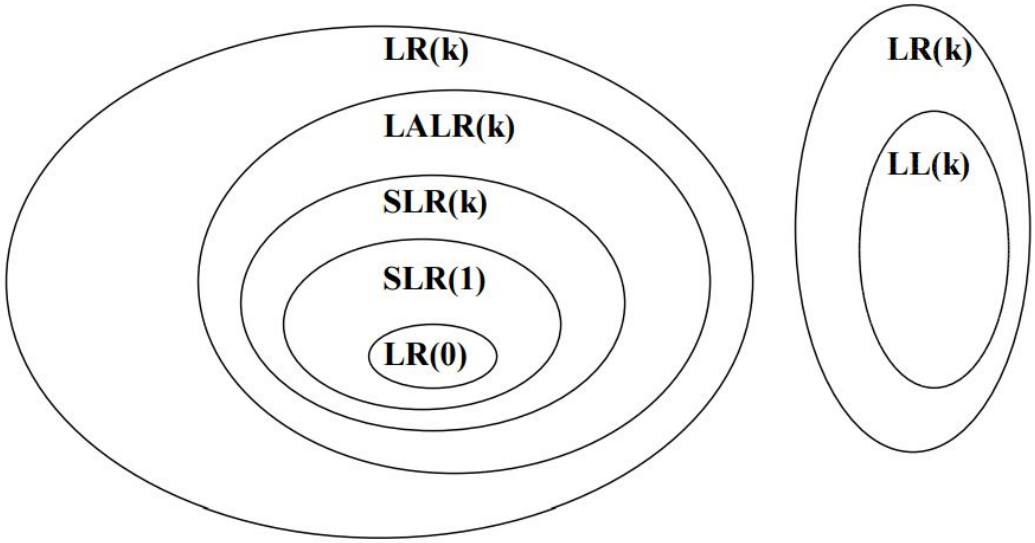
LL(0) gramatika

- lze určit správné pravidlo aniž bychom předem potřebovali vidět nějaký znak na vstupu
- každý neterminál musí mít jen jednu jedinou pravou stranu (jen jedno přepisovací pravidlo)
- neumožňuje rekurzi
- prostě jen určují, jestli sekvence patří do jazyka nebo ne, žádné rozhodování není potřeba.
- LL(0) gramatiky jsou nevhodné pro popis programovacích jazyků, protože zde není možná rekurze a pro každý neterminál existuje právě jedno pravidlo (důsledek faktu, že gramatika se nemůže rozhodnout podle následujícího vstupního symbolu), tudíž mohou generovat jen jazyk s jediným slovem.

LL(1) gramatika

- pro danou gramatiku G se vystačí při rozhodování o výběru pravidla pro expanzi s informací o dopředu prohlíženém řetězci délky 1 -> proto LL(1) je **gramatika silná**
- **jednoduchá LL (1) gramatika** je taková bezkontextová gramatika jestliže platí:
  - pravá strana každého pravidla začíná terminálním symbolem např. A -> aB
  - pokud mají 2 pravidla stejnou levou stranu, pak pravé strany začínají různými terminálními symboly např. A -> aB, A->bB
  - to znamená, že v každém políčku rozkladové tabulky bude právě jeden element
- **Obecná LL(1):**
  - gramatika nemá omezení, ale musí pro ni existovat rozkladová tabulka

Mohutnosti gramatik



Typy analýzy

- shora (top-down)
- sdola (bottom-up) (vyžadují LR gramatiku, takže se netýkají této otázky)

Analýza shora = analýza top-down

- Při hledání derivace začínáme počátečním symbolem a snažíme se dostat k hledanému slovu
- LL analýza: hledáme levou derivaci, vstupní slovo analyzujeme zleva
  - Přesně určuje volbu pravidel při analýze a umožňuje jednoznačný postup při odvození
  - Gramatika, která je jednoznačná a lze ji takto analyzovat: LL gramatika
  - Využívá se zásobníkový automat
- LL(k) označení gramatiky pro LL analýzu, číslo k určuje, kolik následujících symbolů na vstupu je nutné znát pro analýzu slova
- LL(1): nejpoužívanější gramatika, stačí znát jeden následující symbol
- LL(0): umožňuje jen jazyky s konečným počtem slov
- LL gramatiky s k>1 lze převést na LL gramatiky s k = 1
  - Existují přesné popisy, jak jednotlivá pravidla nahrazovat (přidávají se neterminály a pravidla se upravují, aby při analýze stačilo znát jeden další symbol)

LL parsery = parsery s analýzou top-down

LL parsery používají parsing shora dolů, zpracovávají vstup zleva doprava a konstruují nejlevější derivaci. Proto se také nazývá L (left-to-right) L(leftmost derivation). Občas se setkáváme s označením LL(k), kde k značí počet tokenů, které potřebujeme znát při rozhodování o průběhu další analýzy bez toho, aby bylo třeba používat backtracking (= prediktivní parser). Také se v této souvislosti používá pojem look-ahead. Prakticky do nedávné doby se tyto gramatiky příliš nepoužívaly, ovšem na počátku 90. let minulého století došlo ke změně přístupu.

Syntaktická analýza LL gramatik

Budeme se zabývat algoritmem syntaktické analýzy, který vytváří derivační strom analyzovaného řetězce směrem shora dolů. Základní princip syntaktické analýzy můžeme v tomto případě formulovat takto:

Je dána bezkontextová gramatika $G=(N,T,P,S)$ a řetězec $w = a_1\ a_2\ \dots\ a_n$, který je větou z L(G). Pak existuje levá derivace $S = \gamma_1 \Rightarrow \gamma_2 \Rightarrow \dots \Rightarrow \gamma_n = w$.

Vzhledem k tomu, že derivace je levá, má každá větná forma γ_i tvar: $\boldsymbol{\gamma_i = a_1\, a_2 \ldots a_j\, A_i\, \beta_i}$, kde $\boldsymbol{a_1, a_2 \ldots, a_j}$ jsou terminální symboly, $\boldsymbol{A_i}$ je neterminální symbol, $\boldsymbol{\beta_i}$ je řetězec terminálních a neterminálních symbolů. Přitom řetězec $\boldsymbol{a_1 a_2 \ldots a_j}$ je předponou věty $\boldsymbol{w}$, $j \geq 0$.

Podmínky LL analýzy

Předpokládejme, že A → α1 | α2 | ... | αn jsou všechna pravidla v P s neterminálním symbolem A na levé straně. Pak základní problém syntaktické analýzy metodou shora dolů spočívá v nalezení toho pravidla A → αk, jehož aplikací dostaneme z větné formy γi větnou formu γi + 1.

Pro výběr pravidla A → αk, je možno použít:

1. informaci o dosavadním průběhu (historii) analýzy,
2. informaci o dosud nepřečtené části vstupního řetězce (dopředu prohlíženém řetězci omezené délky).

Pokud tyto informace vždy stačí k jednoznačnému výběru pravidla A → αk, pak se gramatika G nazývá LL gramatika. Název je odvozen od toho, že při čtení vstupního řetězce zleva je vytvářen levý rozklad. Při syntaktické analýze LL gramatik jsou do zásobníku ukládány řetězce, které odpovídají levým větným formám nebo takovým jejich příponám, které vzniknou odejmutím předpony tvořené řetězcem terminálních symbolů.

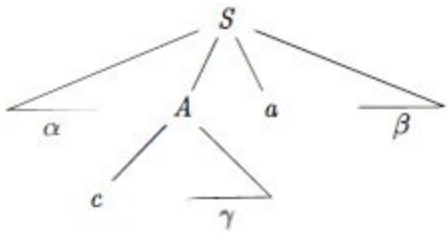
Základními operacemi syntaktického analyzátoru pro LL gramatiky (LL analyzátoru)

- **Expanze** – neterminální symbol na vrcholu zásobníku je nahrazen pravou stranou vybraného pravidla
- **Srovnání** – terminální symbol na vrcholu zásobníku se ze zásobníku vyloučí, jestliže je shodný se symbolem, který byl ze vstupního řetězce přečten.
- **Přijetí** – vstupní řetězec je přečten a zásobník je prázdný.
- **Chyba** – ve všech ostatních případech.

Pokud pro danou gramatiku G vystačíme při rozhodování o výběru pravidla pro expanzi s informací o dopředu prohlíženém řetězci délky nejvýše k, pak se gramatika G nazývá silná LL(k) gramatika. Při analýze silných LL(k) gramatik jsou do zásobníku ukládány přímo symboly gramatiky a syntaktický analyzátor je řízen rozkladovou tabulkou.

Funkce FIRST a FOLLOW

- **top-down** i **bottom-up parsery** používají funkce *FIRST* a *FOLLOW*, spojené s gramatikou G
- při parsování shora dolů nám *FIRST* a *FOLLOW* říkají, které přepisovací pravidlo uplatnit v závislosti na dalším vstupním symbolu.
- během zotavení z chyby při panic módu mohou být množiny tokenů získané pomocí *FOLLOW* použity jako synchronizační tokeny



Terminal c is in FIRST(A) and a is in FOLLOW(A)

Algoritmus

Výpočet funkce FOLLOW

Vstup: Bezkontextová gramatika $G=(N,T,P,S)$ a neterminální symbol A
Výstup: FOLLOW(A).
Metoda:

1. Vytvoříme množinu $Ne = \{ B : B \Rightarrow^* e, B \in N\}$, tj. neterminálních symbolů, ze kterých je možno generovat prázdné řetězce.

2. Vytvoříme množinu F takto:
   a) Vytvoříme fiktivní pravidlo $A \rightarrow A$ a $F := \{ A \rightarrow A.\}$.
   b) Jestliže v množině F je položka, ve které je tečka na konci pravidla, tj. položka $B \rightarrow \gamma.$, vložíme do F nové položky vytvořené tak, že vezmeme všechna pravidla z P, ve kterých se na pravých stranách vyskytuje symbol B a tečku v nich umístíme právě za tento symbol B:
   $F := F \cup \{ C \rightarrow \alpha B. \beta : B \rightarrow \gamma. \in F, C \rightarrow \alpha B \beta \in P \}$.
   c) Jestliže v množině F je prvek, ve kterém je bezprostředně za tečkou neterminální symbol, který patří do množiny Ne, přidáme do F další položku, kterou vytvoříme z uvažované položky posunutím tečky o jeden symbol doprava:
   $F := F \cup \{ A \rightarrow \alpha B. \beta : A \rightarrow \alpha .B \beta \in F, B \in Ne \}$.
   d) Kroky b) a c) opakujeme tak dlouho, dokud je možno do F přidávat další prvky.
   e) Jestliže v množině F je prvek, ve kterém je bezprostředně za tečkou neterminální symbol B, přidáme do množiny F všechna pravidla z P se symbolem B na levé straně a tečku umístíme před první symbol pravé strany:
   $F := F \cup \{ B \rightarrow . \alpha : C \rightarrow \gamma. B \beta \in F, B \in N \ B \rightarrow \alpha \in P \}$.
   f) Jestliže v množině F je prvek, ve kterém je bezprostředně za tečkou neterminální symbol, který patří do množiny Ne, přidáme do F další položku, kterou vytvoříme z uvažované položky posunutím tečky o jeden symbol doprava:
   $F := F \cup \{ A \rightarrow \alpha B. \beta : A \rightarrow \alpha. B \beta \in F, B \in Ne \}$.
   g) Kroky e) a f) opakujeme tak dlouho, dokud je možno do F přidávat další prvky.

3. Množinu FOLLOW(A) vytvoříme tak, že do ní vložíme všechny terminální symboly, které se vyskytují bezprostředně za tečkou v některém prvku množiny F. Jestliže je v množině F prvek, ve kterém se vyskytuje tečka na konci pravidla a na levé straně je symbol S (tj. počáteční symbol gramatiky), přidáme do FOLLOW(A) prázdný řetězec:
   FOLLOW(A) := $\{a : a \in T, B \rightarrow \alpha . a \beta \in F\} \cup \{e : S \rightarrow \alpha . \in F\}$.

13. Vnitřní jazyky překladačů – druhy, použití v jednotlivých fázích překladu, překlad jednoduchých jazykových konstrukcí.

Po ukončení syntaktické a sémantické analýzy generují některé překladače explicitní intermediální reprezentaci zdrojového programu (mezikód).

Intermediální reprezentaci můžeme považovat za program pro nějaký abstraktní počítač. Tato reprezentace by měla mít dvě důležité vlastnosti: měla by být jednoduchá pro vytváření a jednoduchá pro překlad do tvaru cílového programu.

Intermediální kód slouží obvykle jako podklad pro optimalizaci a generování cílového kódu. Může však být také konečným produktem překladu v interpretačním překladači, který vygenerovaný mezikód přímo provádí. Intermediální reprezentace mohou mít různé formy.

Postfixová notace

- operátory následují ihned za operandy
- `A B C * D + - => A - (B * C + D)`
- efektivní zpracování pomocí zásobníku, musíme vědět prioritu operátorů

Prefixová notace

- operátory a pak operandy
- `+ A B + C D => (A + B) * (C + D)`

Tříadresový kód

- Abstraktní forma mezikódu sestávající ze sekvence příkazů ve tvaru **x := y op z**
  - ***x***, ***y*** a ***z*** jsou jména, konstanty nebo dočasné proměnné
  - ***op*** je nějaký operátor.
  - na levé straně je **adresa**, na pravé **instrukce**
  - adresou může být název (ze zdrojového programu, je pak nahrazen pointerem do jeho tabulky symbolů), konstanta, kompilátorem generovaná dočasná proměnná (užitečné pro optimalizaci)
- Jde o linearizovanou podobu syntaktického stromu.

Překlad výrazu x+y*z na tříadresový kód

```
t1 := y * z
t2 := x + t1
```

Další formy tříadresových instrukcí

- s unárním operátorem (x = -y)
- copy instrukce (x = y)
- indexované copy instrukce (x = y[o])
- nepodmíněný skok (goto L)
- podmíněný skok (if x goto L)
- volání procedur (call p, n; předtím uvedeno n parameterů)

Trojice a čtveřice

Implementací tříadresového kódu jsou záznamy se třemi nebo čtyřmi poli: trojice resp. čtveřice. Následující příklady budou ukázány na výrazu: `a := b * (-c) + d [ b ]`

Čtveřice

Záznam má čtyři položky nazývané **op, arg1, arg2 a res**. Tříadresový příkaz ve tvaru ***x := y op z*** je reprezentován umístěním

- op do op
- y do arg1
- z do arg2
- x do res
- *některé tříadresové příkazy nepotřebují všechny položky (např. x := y)*

Výhoda čtveřic oproti trojicím – v optimalizaci kompilátoru, kdy jsou instrukce často přemísťovány; přesun čtveřic je ok (nová pozice se dá hned určit podle dočasných proměnných), u trojic je při posunu třeba změnit reference na výsledky, protože jsou určeny svou pozicí.

Trojice

Jestliže se chceme vyhnout generování dočasných proměnných, je možné použít formu trojic. Trojice obsahuje op, arg1 a arg2. Místo dočasných proměnných jsou indexy do pole trojic (jejich pozice).

Příklady

Ukažme si tříadresový kód, čtveřice a trojice na příkladě výrazu: `a := b * (- c) + d [ b ]`

Tříadresový kód

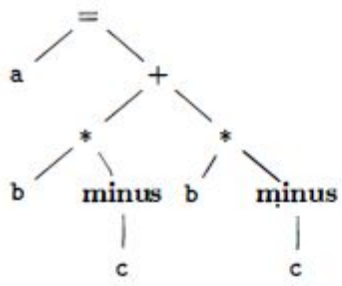
```
t1 := - c
t2 := b * t1
t3 := d [ b ]
t4 := t2 + t3
a := t4
```

Čtveřice

	op	arg1	arg2	res
(1)	uminus	c		t1
(2)	*	b	t1	t2
(3)	loadidx	d	b	t3
(4)	+	t2	t3	t4
(5)	:=	t4		a

Trojice

	op	arg1	arg2
(1)	uminus	c	
(2)	*	b	(1)
(3)	loadidx	d	b
(4)	+	(2)	(3)
(5)	:=	a	(4)



(a) Syntax tree

	op	arg_1	arg_2
0	minus	c	
1	*	b	(0)
2	minus	c	
3	*	b	(2)
4	+	(1)	(3)
5	=	a	(4)
	...		

(b) Triples

Figure 6.11: Representations of $a + a * (b - c) + (b - c) * d$

14. Tabulka symbolů – obsah, způsob manipulace při vytváření a využívání při překladu.

Jakmile syntaktický analyzátor najde určitou konstrukci symbolů, tedy frázi, je třeba této konstrukci přiřadit význam. Součástí syntaktického analyzátoru bývá procedura (nebo více procedur či funkci), která je postupně pro každou frázi volaná a jejím úkolem je doplnit údaje do tabulky symbolů nebo do interního kódu.

OBSAH

Do tabulky symbolů (tabulky objektů) **ukládáme postupně všechny objekty** - pojmenované **identifikátory** (které nejsou klíčovými slovy), **proměnné** nebo **konstanty**, **uživatelské datové typy**, **funkce**, **procedury**, návěští apod., na které v kódu narazíme. Pojem objekt zde budeme chápat obecněji než je obvyklé v teorii programování, bude to **prostě jakýkoliv identifikátor, který není klíčovým slovem** a lexikální analýza ho proto odlišila od jiných identifikátorů.

Zapisujeme zde obvykle název, typ, adresu, případně počáteční hodnotu objektu, počet a typ parametrů funkce a další informace potřebné při dalším překladu, ale také při provádění programu.

Název	Typ	Délka	Deklarováno	Adresa	Použito
`delky`	integer array 10	40 B	A	.....	N
`I`	byte	1 B	A	.....	A
`pocet`	integer	4 B	A	.....	N
`x1`	real	6 B	A	.....	N
`z1`	*nedefinováno*	0	N	0	A

Tabulka symbolů

V tabulce vidíme objekty délky (pole o délce 10 prvků, prvky jsou celá čísla), I, pocet a x1, které již byly deklarovány a objekt I také použit. Objekt z1 ještě nebyl deklarován, ale už je v kódu použit. V jazyce, který umožňuje pracovat pouze s deklarovanými proměnnými, se jedná o sémantickou chybu.

U každého typu objektu potřebujeme uchovávat různé druhy informací. Například u proměnné je to název, adresa, datový typ, velikost potřebné paměti apod., u funkce název, adresa, návratový typ, počet a typ jednotlivých parametrů, příp. zda jsou volány hodnotou nebo odkazem (jestliže jsou volány odkazem, musí sémantický analyzátor navíc ošetřit, aby ve volání funkce byly jako skutečné parametry použity pouze názvy proměnných a nikoli například výrazy nebo konstantní hodnoty), u dalších typů objektů to budou opět jiné údaje. Řádky tabulky mohou být navzájem závislé (jeden uživatelský datový typ může využívat deklaraci již dříve uvedeného, popř. proměnná je typu deklarovaného dříve, . . .), nesmí se však jednat o kruhovou závislost.

Tato tabulka nám slouží k mnoha účelům. **Využívá ji zejména sémantický analyzátor** (kontroluje, zda proměnná použitá v kódu je deklarovaná a zda její datový typ odpovídá jejímu použití, jestli u funkce souhlasí počet a typ argumentů, atd.), **používá se také u generování cílového kódu** (překladač musí vědět, kolik místa v paměti má vyhradit pro jednotlivé symboly).

Při interpretaci obvykle není nutné uchovávat informaci o adrese, samotná tabulka symbolů může sloužit jako úschovna symbolů, se kterou pak neustále pracujeme.

Způsob manipulace při vytváření a využívání při překladu

Tabulka symbolů může být **vytvářena již lexikálním analyzátorem**, ten však má **omezené možnosti** při zjišťování některých údajů, proto je v mnoha případech vhodnější přenechat tuto práci syntaktickému nebo sémantickému analyzátoru. Často používaný postup je vytváření tabulky lexikálním analyzátorem (kdykoliv narazí

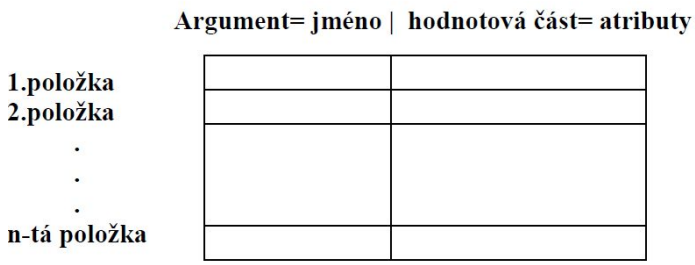
na identifikátor, který není klíčovým slovem, uloží ho do tabulky) s tím, že **další části překladače doplňují zbývající informace o vlastnostech uloženého identifikátoru**.

Otázkou je, **jak** vlastně **řadit** jednotlivé objekty v tabulce. Důležitým kritériem je rychlost vyhledávání, protože k tabulce symbolů přistupuje zejména sémantický analyzátor velmi často. U jednodušších jazyků je možné tabulku automaticky řadit **podle abecedy**, u složitějších jazyků řešíme **indexací**, kdy zároveň s tabulkou vytváříme indexový seznam (příp. soubor), ve kterém jsou odkazy na objekty seřazené podle abecedy.

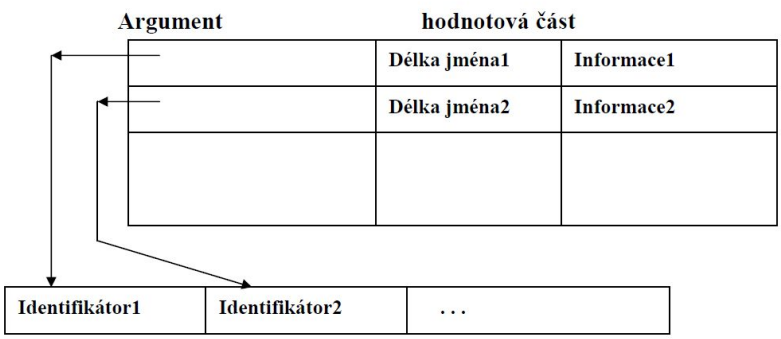
Speciální implementaci vyžaduje tabulka symbolů **pro jazyk s blokovou strukturou**, jako je třeba Pascal. Rozlišují se zde **lokální a globální objekty** a přístupnost lokálních je omezena. Každá proměnná je viditelná v tom bloku, ve kterém je deklarovaná, a také ve všech blocích vnořených. Když v určitém bloku použijeme proměnnou, hledáme informace o ní nejdřív v tom bloku, ve kterém se nacházíme. Při neúspěchu se posouváme do nadřízeného bloku a tak postupujeme, dokud ji nenajdeme. Pokud neuspějeme ani v hlavním bloku, znamená to, že byla použita proměnná, která není deklarovaná, jde o sémantickou chybu. **Každý blok má svoji vlastní tabulku**. S celou strukturou **se pracuje jako s klasickým zásobníkem**. Každá z tabulek má svou vlastní organizaci a je z ní přístupná nadřízená tabulka. „Aktivní tabulka je na vrcholu zásobníku, kde také začínáme prohledávat. Při vyhodnocení konce bloku se aktivní tabulka ze zásobníku odstraní. Po jejím odstranění se sem přesune nadřízená tabulka. Tabulka hlavního bloku zůstává v zásobníku až do konce vyhodnocování programu, je odstraněna až jako poslední po vyhodnocení celého programu.

Tabulka symbolů

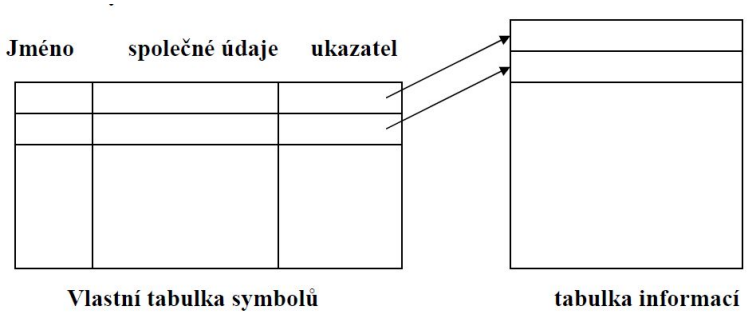
- uchovává informace o objektech
- umožňuje kontextové kontroly
- umožňuje operace
  - inicializaci informace pro standardní jména
  - vyhledání jména
  - doplnění informace ke jménu
  - přidání položky pro nové jméno
  - vypuštění položky či skupiny položek



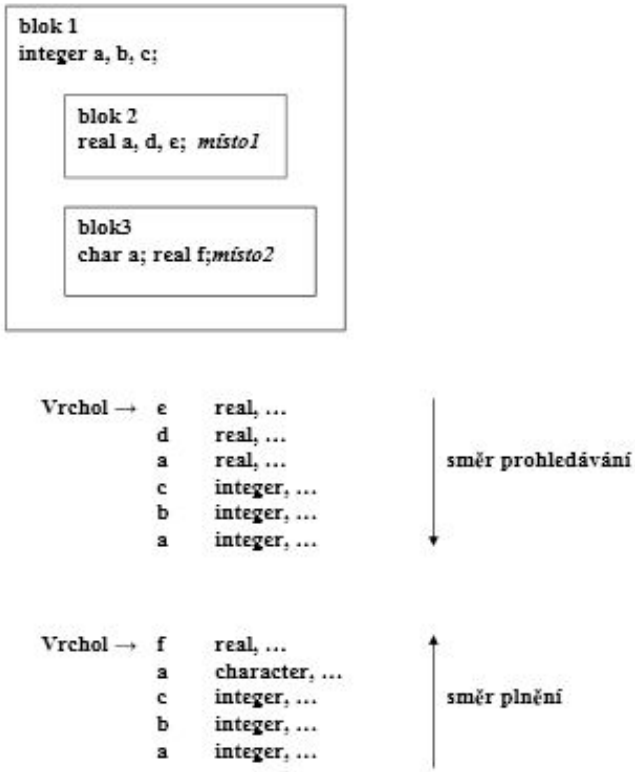
s jednoduchou strukturou



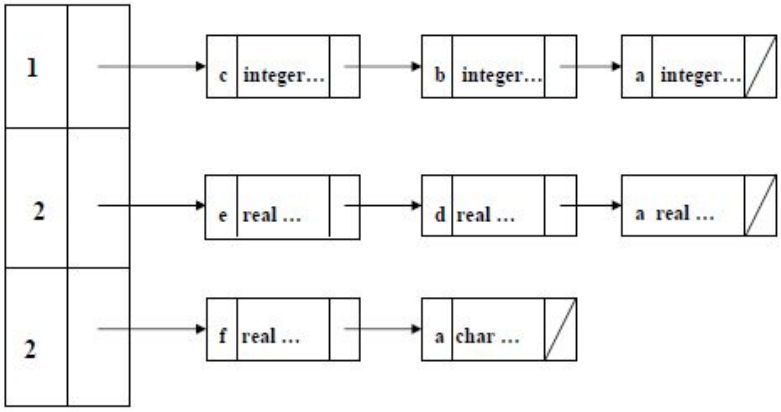
s oddělenou tabulkou identifikátorů



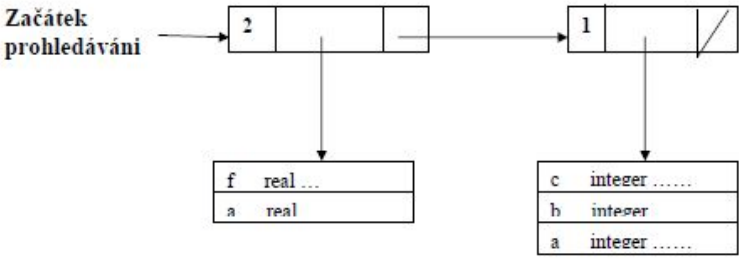
s oddělenou tabulkou informací



uspořádané do podoby zásobníku



Překládá-li se uvnitř bloku 3:



s blokovou strukturou

Implementace tabulky symbolů

- Vyhledávací netříděné tabulky (jen pro krátké programy)
  - prostá struktura
  - lineární seznam
- Vyhledávací setříděné tabulky
  - průběžné setřiďování
  - setřídění po zaplnění
- Frekvenčně uspořádané tabulky
- Binární vyhledávací stromy
- Tabulky s rozptýlenými položkami

Ukládání polí a struktur

Pole i struktury mají pevnou adresu začátku pole a pro přístup k jednotlivým prvkům se výsledná adresa dopočítává. Pole mohou být v paměti uložena buď po řádcích nebo po sloupcích. Tomu musí odpovídat mapovací funkce, která vypočítává relativní adresu prvků. K této adrese musí být připočtena adresa začátku pole.

15. Principy přidělování paměti překladačem.

Přeložený program dostane od operačního počítače k dispozici blok paměti, který obecně může být rozdělen na následující části:

- Vygenerovaný cílový kód
- Statická data
- Řídící zásobník
- Hromada

Základní způsoby přidělování:

- **Statické** (přidělení paměti v čase překladu)
- **Dynamické** (přiděleno v run time) - **v zásobíku** nebo **na haldě**

Velikost vygenerovaného kódu je známa již v době překladu, takže jej může překladač umístit **do staticky definované oblasti** (Code/cílový kód programu), obvykle na začátek přiděleného paměťového prostoru.

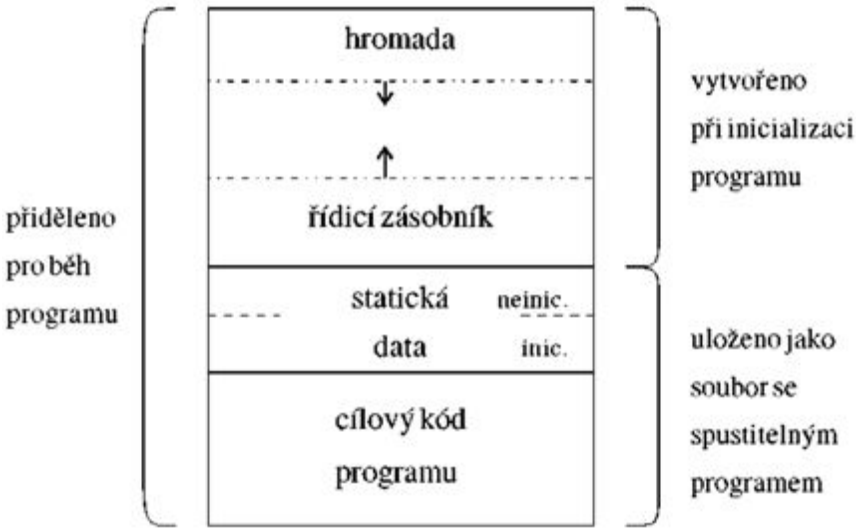
Rovněž velikost **statických datových objektů** může být známa již v době překladu a překladač je může **umístit za program** nebo uložit dokonce jako součást programu (to lze pouze u těch programovacích jazyků, které neumožňují rekurzivní volání procedur – Fortran).

Jazyky umožňující rekurzi (Pascal, C, ...) využívají pro aktivace podprogramů řídicího **zásobníku (stack)**, do kterého se ukládají jednotlivé ***aktivační záznamy*** (AZ jsou generovány při voláních procedur).

Pro účely **dynamického přidělování paměti** (explicitně vyžadovaného voláním příslušných funkcí nebo implicitně při přidělování paměti například pro pole s dynamickými rozměry) se používá zvláštní část paměti zvaná **hromada (heap)**.

Vzhledem k tomu, že se velikosti použité části paměti pro zásobník a hromadu v průběhu činnosti programu mohou značně měnit, je výhodné pro obě části využít opačné konce společné části paměti – viz obrázek.

Zásobník roste směrem k nižším adresám, hromada směrem k vyšším.
Nedostatek paměti se rozpozná tehdy, jestliže ukazatel konce některé oblasti překročí hodnotu ukazatele konce druhé oblasti.



Pro zmíněné datové oblasti se používají následující hlavní metody přidělování paměti:

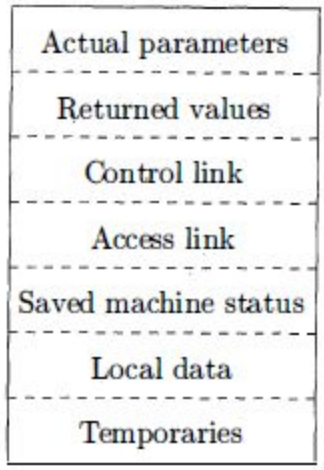
- **Statické** přidělování paměti v době překladu
- **Dynamické** přidělování paměti za běhu programu:
  - Přidělování paměti na zásobníku
  - Přidělování paměti z hromady

Statické přidělování paměti v době překladu

Při statickém přidělování paměti jsou všem objektům v programu **přiděleny adresy již v době překladu**. Při kterémkoliv volání podprogramu jsou jeho lokální proměnné vždy na stejném místě, což umožňuje zachovávat hodnoty lokálních proměnných nezměněné mezi různými aktivacemi podprogramu. Statická alokace proměnných však klade na zdrojový jazyk určitá omezení. Údaje o velikosti a počtu všech datových objektů musejí být známy již v době překladu, **rekurzivní podprogramy mají velmi omezené možnosti**, neboť všechny aktivace podprogramu sdílejí tytéž proměnné, a konečně nelze vytvářet dynamické datové struktury.

Přidělování na zásobníku

Přidělování paměti pro aktivační záznamy na zásobníku se používá běžně u jazyků, které **umožňují rekurzivní volání podprogramů** nebo které **používají staticky do sebe zanořené podprogramy**.



Paměť pro lokální proměnné je přidělena při aktivaci podprogramu vždy na vrcholu zásobníku a při návratu je opět uvolněna.

To ale zároveň znamená, že hodnoty lokálních proměnných se **mezi dvěma aktivacemi podprogramu nezachovávají**.

Aktivace procedur při běhu programu jde znázornit ***aktivačním stromem***. Co uzel, to jedna aktivace, kořen je aktivací hlavní procedury, která je volána po spuštění programu; potomci uzlu p = volání procedur z procedury p. Kořen aktivačního stromu je na dně zásobníku, poslední aktivace má svůj záznam na vrcholu zásobníku.

Každá *„živá“* aktivace má **aktivační záznam**. Obsah aktivačního záznamu se liší podle implementovaného jazyka.

1. **dočasné hodnoty** – vypadnou po vyhodnocení výrazů, jsou tu, pokud nemohou být udržovány v registrech
2. **lokální data** – patřící k dané proceduře s příslušným aktivačním záznamem
3. **uložený strojový status** – info o stavu stroje před voláním procedury. Typicky jde o:
   - návratovou adresu (= hodnota program counteru, kam se má pak procedura vrátit) a o
   - obsah registrů použitých procedurou (musí být obnoveny po návratu z procedury)
4. **access link** = statický ukazatel – pro lokaci dat, která procedura potřebuje, ale která se nachází v jiném aktivačním záznamu
5. **control link** – ukazuje na aktivační záznam volajícího (caller)
6. **vrácené hodnoty** – prostor pro návratovou hodnotu volané funkce (kvůli rychlosti lepší dávat do registru)
7. **vlastní parametry** – parametry použité volající procedurou; pokud je to možné, jsou umístěny radši v registrech kvůli výkonnosti.

Při implementaci přidělování paměti na zásobníku bývá **jeden registr vyhrazen jako ukazatel na začátek aktivačního záznamu na vrcholu zásobníku**. Relativně k tomuto registru se pak počítají všechny adresy datových objektů, které jsou umístěny v aktivačním záznamu. Naplnění registru a přidělení nového aktivačního záznamu je součástí volací posloupnosti, obnovení stavu před voláním se provádí během návratové posloupnosti.

Volací (a návratové) posloupnosti se od sebe v různých implementacích liší. Jejich činnost bývá rozdělena mezi volající a volaný program. Obvykle volající program určí adresu začátku nového aktivačního záznamu (k tomu potřebuje znát velikost záznamu vlastního), přesune do něj předávané argumenty a spustí volaný podprogram zároveň s uložením návratové adresy do určitého registru nebo na známé místo v paměti. Volaný podprogram nejprve uschová do svého aktivačního záznamu stavovou informaci (obsahy registrů, stavové slovo procesoru, návratovou adresu), inicializuje svá lokální data a pokračuje zpracováním svého těla. Při návratu opět volaný

podprogram uloží hodnotu výsledku do registru nebo do paměti, obnoví uschovanou stavovou informaci a provede návrat do volajícího programu. Ten si převezme návratovou hodnotu a tím je volání podprogramu ukončeno.

Přístup k nelokálním proměnným při statickém =lexikálním rozsahu platnosti jmen. To řeší tzv. řetězec statických ukazatelů (access links). Pro zrychlení přístupu k nelokálním proměnným se zavádí vektor ukazatelů – displej. Zamezí se tak průchod aktivačními záznamy pro hluboko zanořené podprogramy.

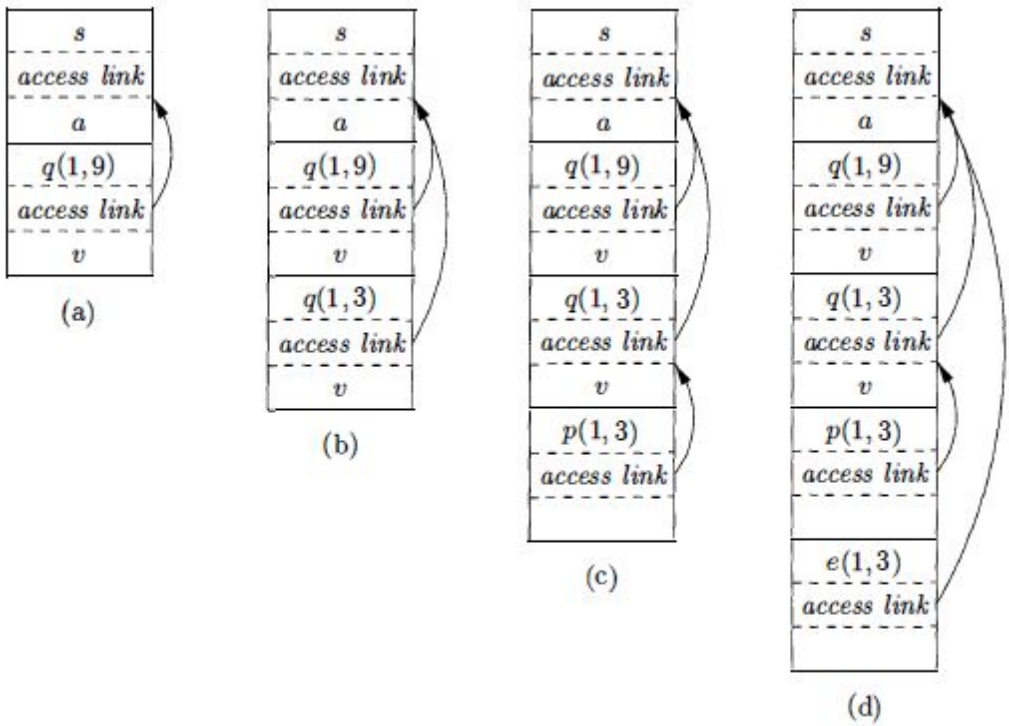


Figure 7.11: Access links for finding nonlocal data

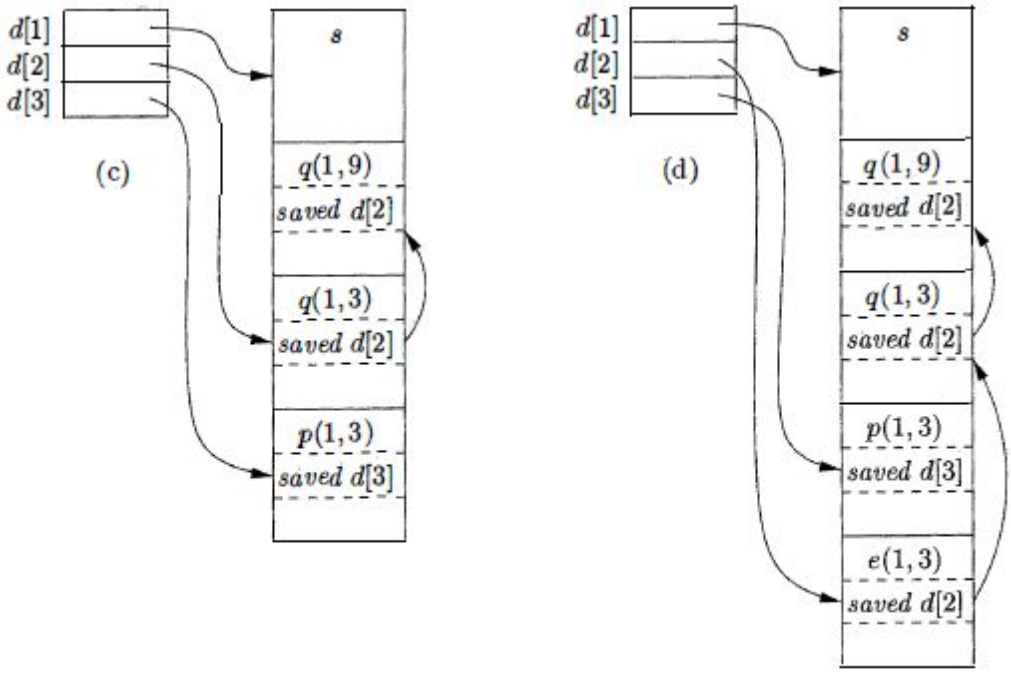


Figure 7.14: Maintaining the display

Přidělování z hromady

Strategie přidělování na zásobníku je nepoužitelná, pokud mohou hodnoty lokálních proměnných přetrvávat i po ukončení aktivace, případně pokud aktivace volaného podprogramu může přežít aktivaci

volajícího. V těchto případech přidělování a uvolňování aktivačních záznamů se mohou překrývat, takže nemůžeme paměť organizovat jako zásobník. Aktivační záznamy se mohou v těchto nejobecnějších situacích přidělovat z volné oblasti paměti (hromady), která se jinak používá pro dynamické datové struktury vytvářené uživatelem. Přidělené aktivační záznamy se uvolňují až tehdy, pokud se ukončí aktivace příslušného podprogramu nebo pokud už nejsou lokální data potřebná.

Při použití této strategie se pro vlastní přidělování a uvolňování paměti používají stejné techniky jako pro dynamické proměnné.

Správce paměti (memory manager)

- alokuje a dealokuje místo na heapu. V důsledku toho může dojít k fragmentaci heapu (vznik malých, nesouvislých míst = děr). Strategie best fit = alokuj nejmenší vhodnou a dostupnou díru.

Garbage collection

- hledá místo na heapu, které se už nepoužívá a může být proto realokované pro uchovávání dalších dat (Java, C#). Automaticky uvolňuje již nepoužívané objekty z paměti.

16. Vlastnosti jazykových konstrukcí pro statický a pro dynamický způsob přidělování paměti.

Důležitá hlediska jazykových konstrukcí

- Dynamické typy
- Dynamické proměnné
- Rekurze
- Konstrukce pro paralelní výpočty

Podstatný je rovněž způsob

- Omezování existence entit v programu
- Předávání parametrů funkce (hodnotou, odkazem)
- Určování přístupu k nelokálním entitám
    - na základě statického vnořování rozsahových jednotek,
    - na základě dynamického vnoření rozsahových jednotek.

Statické přidělování paměti

- Globální proměnné
- Statické proměnné
- Proměnné jazyka bez rekurze (i s blokovou strukturou) (možno staticky na zásobníku)

Důležitá hlediska jazykových konstrukcí

- Dynamické typy
- Dynamické proměnné
- Rekurze
- Konstrukce pro paralelní výpočty
- Podstatný je rovněž způsob:
- Omezování existence entit v programu

Předávání parametrů funkce (hodnotou, odkazem)

- Určování přístupu k nelokálním entitám
- na základě statického vnořování rozsahových jednotek,
- na základě dynamického vnoření rozsahových jednotek.
- Statické přidělování paměti:
- Globální proměnné
- Statické proměnné
- Proměnné jazyka bez rekurze (i s blokovou strukturou) (možno staticky na zásobníku)

Dynamické přidělování v zásobníku

Aktivační záznam obsahuje místo pro

- Lokální proměnné
- Parametry
- Návratovou adresu
- Funkční hodnotu (je-li podpr. funkcí)
- Pomocné proměnné (pro mezivýsledky)
- Další informace potřebné k uspořádání aktivačních záznamů

Statická typová kontrola

Referenční prostředí podprogramů je definováno staticky, tj. při překladu zdrojového programu. Pro každou deklaraci je staticky vymezen rozsah platnosti, tj. část zdrojového kódu, ve kterém lze deklarované jméno použít. V podprogramu pak kromě lokálních jmen lze použít ta nelokání jména, do jejichž rozsahu platnosti je definice podprogramu vnořena. Statické referenční prostředí je často založeno na blokové struktuře programu.

Statické referenční prostředí je při překladu reprezentováno tabulkou symbolů. Překlad deklarace znamená rozšíření tabulky symbolů o nový záznam, při dosažení místa konce platnosti deklarace se záznam odstraní nebo skryje. Při překladu těla podprogramu překladač na základě tabulky symbolů pro každé jméno rozhodne, zda je či není v daném místě použitelné a jaký datový objekt (nebo jiný programový prvek) označuje. Tím se dosáhne vyšší bezpečnosti programu (nepoužitelné jméno je odhaleno již při překladu) i vyšší efektivity cílového programu.

Dynamická typová kontrola

Např. Lisp, referenční prostředí podprogramů je definováno dynamicky. Dynamicky definované referenční prostředí **se nevytváří ani nekontroluje při překladu, ale až při provádění programu**. Při spuštění programu se vytvoří referenční prostředí tvořené vazbami jmen definovaných jazykem. Při každém vstupu do podprogramu se referenční prostředí rozšíří o vazby lokálních jmen podprogramu, při návratu z podprogramu se jeho lokální prostředí odstraní. Při provádění příkazů se pro každé jméno hledá jeho vazba.

Dyn. definované ref. prostředí snižuje bezpečnost programu i jeho efektivitu (význam jména se hledá při provádění programu). Jeho výhodou je jednoduchá sémantika – nenajde-li se jméno v lokálním prostředí podprogramu A, hledá se v lokálním prostředí podprogramu B, ze kterého byl podprogram A vyvolán, případně v lokálním prostředí podprogramu C, z čehož byl vyvolán podprogram B apod.