Lexikální analýza, princip činnosti

Thursday, May 30, 2013 8:38 AM

Úkoly lexikálního analyzátoru

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

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 Name	popis (v podstatě pattern)	příklad lexémů	hodnota atributu
if	znaky i, f	if	-
else	zn. e, l, s, e	else	-
id	písmeno násl. písm./číslicí	foo, score, myId	pointer do tab. ~~Záznamů~~ symbolů
number	jakákoliv číselná konstanta	3.14, 10	pointer do tab. záznamů
literal	vše v uvozovkách	“hello world”	pointer do tab. záznamů
relop	<, <=, >, >=, =, <>	<, <=, >, >=, =, <>	LT, LE, GT, GE, EQ, NE

Token je tvořen dvěma částmi – názvem tokenu (token name) a hodnotou atributu (attribute value). Názvy tokenu jsou často abstraktní symboly, které jsou pak použity parserem pro syntaktickou analýzu. Jde např. o nějaké 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.

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).

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.

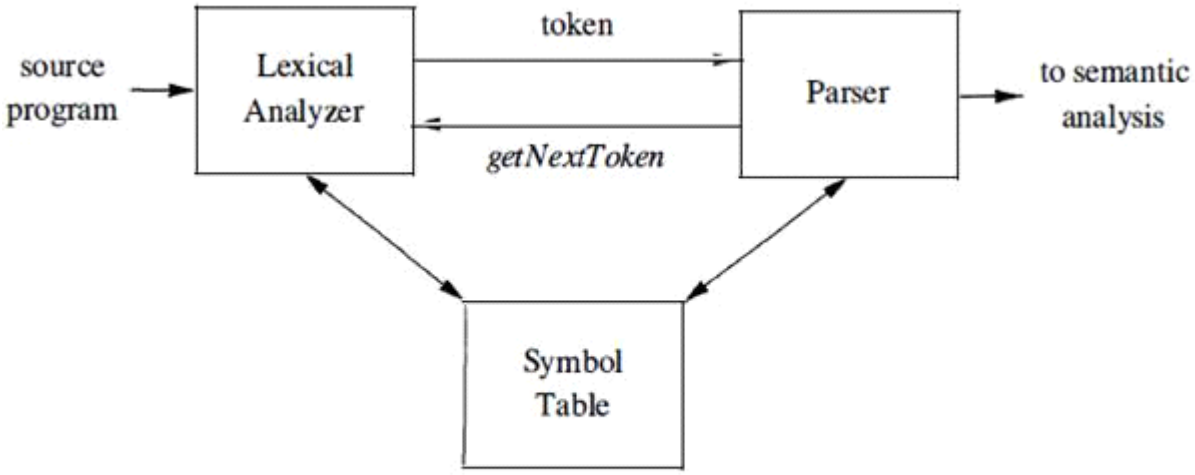

Figure 3.1: Interactions between the lexical analyzer and the parser

Často je potřeba dopředu skenovat vstup, aby se zjistilo, kde následující lexém končí. Proto lex. analyzátory typicky bufferují vstup.

From <https://d.docs.live.net/e3534876709763a3/Dokumenty/ZCU/Statnice/Statnice.docx>