

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Dokumentácia – IFJ 2018

Tým 40, varianta II

Rozšírenie: BASE

Adam Hostin	xhosti02	25 %
Sabína Gregušová	xgregu02	25 %
Dominik Peza	xpezad00	25 %
Adrián Tulušák	xtulus00	25 %

Obsah

1	Úvod	2
2	Lexikálna analýza	2
2.1	Štruktúra Token.t	2
2.2	Spracovanie reťazcov	2
3	Syntaktická analýza	2
3.1	Precedenčná analýza výrazov	3
3.2	Sématická analýza	3
4	Generovanie inštrukcií	3
5	Práca v tíme	4
5.1	Rozdelenie práce	4
5.2	Komunikácia	4
5.3	Verzovanie	4
5.4	Testovanie	4
5.5	Hodnotenie	4
A	Deterministický konečný automat	6
B	LL-gramatika	8
C	Tabuľka LL-gramatiky	8
D	Precedenčná tabuľka	9

1 Úvod

Naším cieľom je implementovať prekladač imperatívneho jazyka IFJ18 do predmetov IFJ a IAL v jazyku C. Hlavnou náplňou práce bola implementácia: lexikálneho analyzátoru, parsera (syntaktická a sématická analýza) a generátora inštrukcií.

2 Lexikálna analýza

Na začiatku sme implementovali lexikálny analyzátor v súbore `scanner.c`, ktorého základom je deterministický konečný automat (ďalej iba DKA). Hlavnou funkciou v tomto súbore je `get_next_token`, ktorá číta jednotlivé znaky a pomocou príkazu `switch` prechádza do nasledujúcich stavov podľa DKA až kým nevyhodnotí lexikálne správny token, inak vracia `ER_LEX`. Lexikálny analyzátor musí mať nadstavený vstupný súbor, ktorý obsahuje program napísaný v jazyku IFJ18.

Pri úspešnom vyhodnotení tokenu sa správne uvoľní všetká alokovaná pamäť. Matematické a relačné operátory sú vyhodnotené vcelku rýchlo a jednoducho, identifikátory, reťazce a čísla vyžadujú viacej prechodov a používajú dynamický reťazec, o ktorom ďalej pojednáva sekcia Spracovanie reťazcov. Pri identifikátore sa kontrolujú povolené znaky na základe pozície v reťazci a na záver sa identifikátor porovná so všetkými kľúčovými slovami; ak sa nájde zhoda, je to kľúčové slovo, inak je to identifikátor. Reťazce sú ohraňované dvojitémi úvodzovkami (") a môžu obsahovať escape sequence. Pre tento prípad existuje špeciálny stav `STATE_BACKSLASH_LITERAL`, do ktorého sa prechádza pri prečítaní znaku spätné lomítko, (\) za ktorým môže nasledovať escape sekvencia. V prípade, že bol tento znak prečítaný ale nasleduje znak ktorý nepatrí do známej escape sekvencie, je spätné lomítko ignorované.

2.1 Štruktúra `Token_t`

Pre jednoduchšiu prácu s tokenmi sme použili štruktúru `Token_t`, ktorá obsahovala:

- `union Token_attr`
- `struct Token_type`

Union `Token_attr` obsahuje možné atribúty tokenu, konkrétne to sú: string, integer, flt a keyword.

Struct `Token_type` obsahuje typy tokenov, konkrétne to sú: EOF, EOL, identifikátor, kľúčové slovo, relačné a matematické operátory, ľava a pravá zátvorka, čiarka, komentár, int, float a string.

2.2 Spracovanie reťazcov

Pre jednoduchšie spracovanie reťazcov sme sa rozhodli implementovať súbor `dynamic_string.c`. Jeho súčasťou je aj štruktúra `string_t`, ktorá obsahuje samotný ukazateľ na dynamický reťazec, súčasnú veľkosť reťazca a celkovú veľkosť bufferu. Na začiatku je alokovaný reťazec s veľkosťou 10 bytov a pri každom pridaní znaku sa kontroluje, či je ešte v reťazci miesto.

Keď sa blížíme k zaplneniu reťazca, funkcia `check_empty_bites` zväčší veľkosť bufferu o 5, čím zaistí dostatočnú veľkosť pre reťazec. Všetky alokácie pamäte sú kontrolované a ich zlyhanie je adekvátne ošetrené vrátením internej chyby `ER_INTERNAL`.

3 Syntaktická analýza

Syntaktická analýza je hlavnou časťou programu. Je založená na LL-gramatike a podľa pravidiel LL-tabuľky prechádza zdrojovým súborom rekurzívnym zostupom. Funkcie v syntaktickom analyzátore predstavujú jednotlivé časti LL-gramatiky. Toto platí pre celú syntaktickú analýzu s výnimkou spracovania výrazov. V rámci syntactickej analýzy je volaná funkcia lexikálneho analyzátoru `get_next_token`, ktorá naplní štruktúru `Data_t` aktuálne získaným tokenom. Ak je daný token komentár, je funkcia volaná v cykle až kým daný token nie je komentár. Na základe tohto tokenu sa syntaktický analyzátor rozhoduje, ktoré pravidlo z LL-tabuľky uplatniť a ktorú z funkcií následne zavolať.

Štruktúra `Data_t` obsahuje okrem typu tokenu a jeho atribútov aj pomocné súčasti ako `in_while_or_if`, `in_definition`, `in_declare`, ktoré pomáhajú rozlíšiť situácie, v ktorých by bolo možné s jedným typom tokenu uplatniť viacero pravidiel. Vždy po analyzovaní tokenov potrebných na presné určenie o aký prípad sa jedná, je volaná príslušná funkcia na generovanie kódu. V prípade, že lexikálny analyzátor vráti ako token identifikátor, sú volané funkcie sémantickej analýzy, ktoré skontrolujú platnosť identifikátoru.

3.1 Precedenčná analýza výrazov

Syntaktická analýza výrazov je implementovaná v súbore `expression.c` pomocou precedenčnej tabuľky. Hlavné telo `expression.c` pozostáva z funkcie `handle_expression`, ktorá postupne spracováva jednotlivé tokeny alebo symboly a vyhodnocuje ich syntaktickú správnosť. Keďže nie je vždy úplne jednoznačne možné určiť, či nasledujúci token bude súčasťou výrazu alebo nie, rozhodli sme sa implementovať jednosmerne viazaný zoznam s pracovným názvom `buffer`, kam sa postupne ukladajú tokeny vždy na koniec zoznamu.

Ak parser vyhodnotí, že dané tokeny nie sú súčasťou výrazu, vyčistí `buffer` a syntaktická analýza pokračuje v parseri. V opačnom prípade je zavolaná funkcia `handle_expression`, ktorá pri spracovaní používa tokeny v `bufferi`, až kým nie je `buffer` prázdny a ďalej si žiada tokeny pomocou funkcie `get_next_token`. Z každého tokenu si vytvoríme symbol na základe jeho typu, ktorý si ďalej ukladáme na zásobník. Spracovávanie symbolov je naprogramované na základe algoritmu uvedeného v prezentácii [1]. Symboly majú svoj status, ktorý sa kopíruje aj pri redukcii danej časti výrazu. Časti výrazu sa postupne dávajú na zásobník generátora a akonáhle už raz bol daný výraz daný do generátora, jeho status sa mení na `ON_GENERATOR_STACK` a pri ďalšej redukcii už nebude znova pridaný na zásobník generátora. Ak prebehne syntaktická analýza výrazu bez problémov, výsledok, ktorý je na vrchole zásobníka sa popne a adekvátne spracuje.

Pri kontrole jednotlivých symbolov na zásobníku sa zároveň kontroluje aj `flag errorov`, a jeho nadstavenie na čokoľvek iné ako `EXPRESSION_OK` vedie na ukončenie spracovávania výrazov s adekvátnym návratovým kódom.

3.2 Sématická analýza

Sématickú analýzu sme implementovali ako tabuľku s rozptýlenými položkami v súbore `symtable.c`. Synonymá sú v tabuľke zoradené explicitne, čo zabezpečuje toereticky neobmedzený počet položiek uchovávateľných tabuľkou symbolov. Synonymá sú zreťazené v jednosmerne viazaných zoznamoch. Veľkosť mapovacieho poľa sme vyberali tak, aby bola rovná prvočíslu. Naša tabuľka má veľkosť 6 421. Očakávame, že naplnenie tabuľky nepresiahne 75%. Mapovaciu funkciu sme prevzali z druhej domácej úlohy z predmetu `IAL`, lebo nám pripadala efektívna a ľahko pochopiteľná.

Funkcia spočítava `ASCII` hodnotu jednotlivých znakov kľúča a nakoniec vracia modulo veľkosti tabuľky z daného súčtu. Každá položka tabuľky obsahuje svoj vlastný unikátny kľúč, ktorý sa ukladá v podobe reťazca. Kľúče značia identifikátory funkcií a premenných. Každá položka taktiež obsahuje svoj typ, značiaci či sa jedná o premennú alebo funkciu, boolovskú hodnotu, určujúcu či bola položka definovaná, ukazateľ na ďalší prvok v zozname a integer značiaci počet parametrov v prípade, že sa jedná o funkciu. Položky sa ukladajú do 2 identických tabuliek s rozptýlenými položkami v závislosti od toho, či sa jedná o lokálne alebo globálne premenné.

Všetky funkcie sa ukladajú do globálnej tabuľky symbolov. Implementovali sme taktiež niekoľko funkcií zabezpečujúce pohodlnú prácu s tabuľkou. Funkcie zabezpečujú inicializáciu, pridávanie prvkov, vyhľadávanie v tabuľke, kontrolu jednotlivých atribútov položiek a čistenie tabuľky.

Keďže nie je možné skontrolovať všetkú sématicku počas prekladu, mnohé sématické kontroly prebehnú až pri interpretácii kódu.

4 Generovanie inštrukcií

Generovanie inštrukcií je realizované v súbore `instructions.c`. Na začiatku činnosti parseru sa inicializuje pamäť na ukladanie inštrukcií a vygeneruje sa hlavička medzikódu, ktorá obsahuje vstavané funkcie a interné funkcie určené na typovú kontrolu.

Po skontrolovaní syntaktickej správnosti časti kódu, volá parser príslušnú funkciu realizujúcu generovanie inštrukcií. Podobne ako pri lexikálnej analýze aj tu sa využíva štruktúra dynamického reťazca, ktorý je používaný na "lepenie"

inštrukcií. Ak syntaktický analyzátor vyhodnotí kód ako preložiteľný, obsah dynamického reťazca sa vypíše na štandardný výstup.

Pri testovaní inštrukcií sme narazili na problém s inštrukciou `DEFVAR` v prípade cyklu `while` alebo `if`-e. Tento problém sme vyriešili použitím pomocného dynamického stringu, kam sa ukladali všetky inštrukcie okrem `DEFVAR`-ov a keď sa skončil cyklus `while` alebo `if`, obsah pomocného dynamického stringu sa prilepil na koniec nášho hlavného dynamického stringu.

5 Práca v tíme

Náš tím sme si zostavili pomerne skoro. Po rozdelení práce na menšie celky sme začali každý pracovať na časti pridelenej vedúcim tímu. Približne 3 týždne pred pokusným odovzdaním sme začali jednotlivé časti spájať do celku.

Pokusné odovzdávanie sme využili, no výsledok nás nemilo prekvapil. Zistili sme, že hoci všetko fungovalo pomerne správne, návratové hodnoty sme vždy “natvrdo” vracali buď ako 0 alebo 1, čo sa aj prejavilo na celkovom percentuálnom hodnotení. Pri tejto príležitosti sme spravili riadnu revíziu kódu a poopravovali čo najviac chýb a nedostatkov.

5.1 Rozdelenie práce

Prácu sme si rozdelili na základe osobitých požiadaviek členov a ich skúseností nasledovne:

Meno	Povinnosti
Dominik Peza	vedenie tímu, pridelenie úloh, kontrola práce, konzultácie, generovanie inštrukcií
Sabína Gregušová	lexikálna analýza, precedenčná analýza výrazov, dokumentácia, testovanie
Adam Hostin	sématická analýza, dokumentácia, testovanie, diagramy
Adrián Tulušák	syntaktická analýza, sématická analýza, testovanie, dokumentácia

Tabuľka 1: Rozdelenie práce

5.2 Komunikácia

Už na začiatku sme sa dohodli na pravidelných týždenných stretnutiach, kde sme diskutovali o našej ďalšej práci na nadchádzajúci týždeň. Komunikovali sme najmä cez facebook a skype a každú nedeľu sme sa stretli a prebrali výsledky našej práce za posledný pracovný týždeň.

5.3 Verzovanie

Pre správu projektu sme používali verzovací systém Git a zdieľaný repozitár GitHub. Tento spôsob správy projektu nám umožňoval pracovať na viacerých častiach projektu súčasne. Po otestovaní jednotlivých podčastí sme súbory začali spájať do celku a každý mal možnosť testovať projekt už ako celok.

5.4 Testovanie

Keďže bolo pomerne náročné odladiť tento projekt, rozhodli sme sa rozdeliť si úlohy nasledovne - 2 ľudia písali testy a chystali testovací skript, ktorý spustil všetky testy naraz aj s vyhodnotením, a zvyšní 2 riešili výsledky týchto testov a hľadali bugy. Tento spôsob sa nám veľmi overil, a podarilo sa nám relatívne dobre odladiť veľké množstvo chýb.

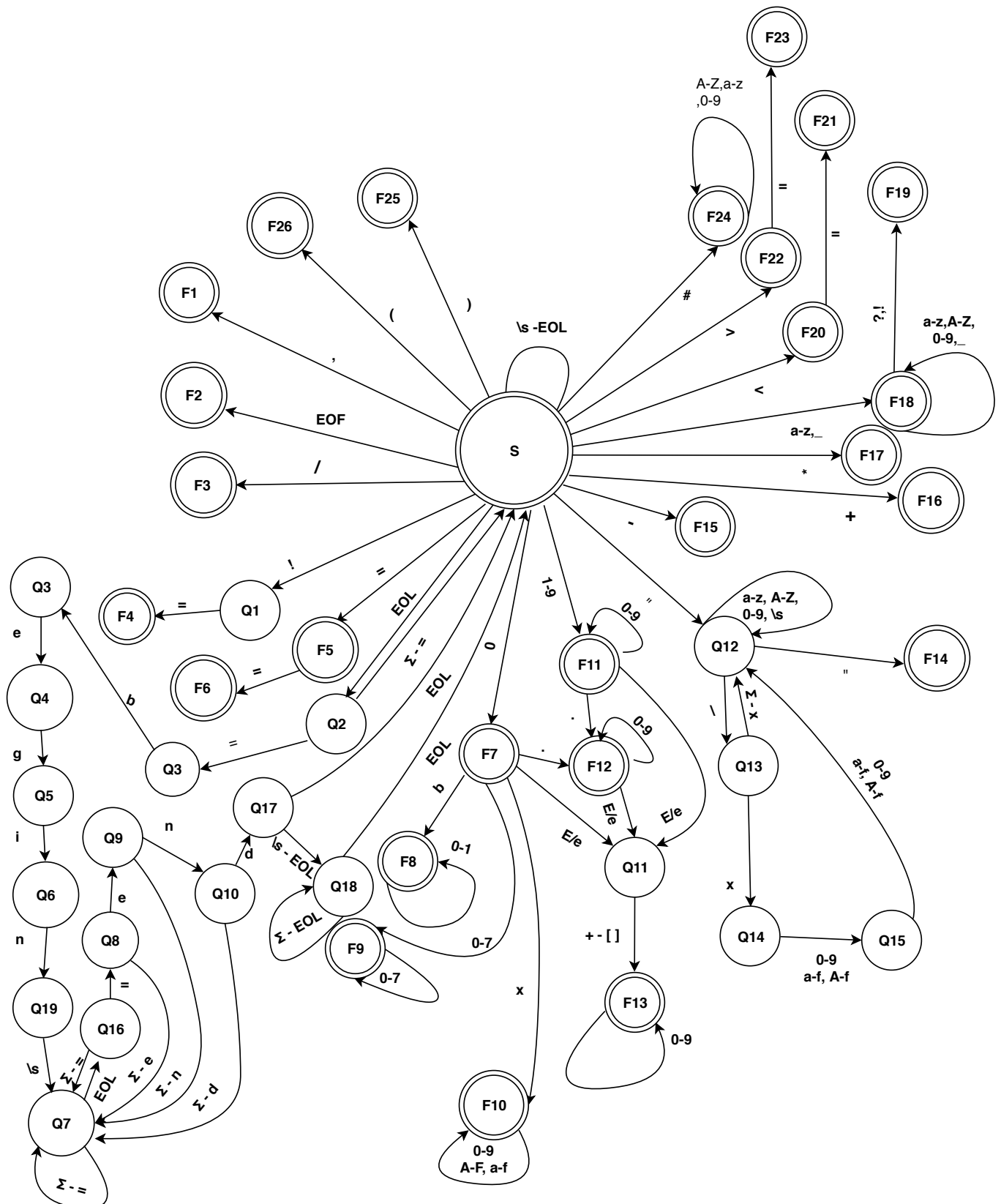
5.5 Hodnotenie

Celkovo hodnotíme tento projekt kladne, hoci na začiatku vyzeral pomerne zložito, postupne sme mali nápady ako dané problémy vyriešiť a nakoniec sme všetko stihli v časovom predstihu, takže sme mali možnosť odstraňovať chyby a testovať správnosť a určite tento projekt prispel k zlepšeniu našich programátorských schopností.

Použitá literatúra

- [1] MEDUNA, A.; LUKÁŠ, R.: Kapitola VIII. Syntaktická analýza zdola nahoru, rev. 1. septembra 2018, [vid. 3. novembra 2018].

A Deterministický konečný automat



Legenda:

S	start
F1	čiarka
F2	EOF
F3	div
F4	nerovná sa
F5	priradenie
F6	rovná sa
F7	nula
F8	binárne číslo
F9	číslo v osmičkovej sústave
F10	číslo v hexadecimálnej sústave
F11	celé číslo v desiatkovej sústave
F12	desatinné číslo
F13	číslo s exponentom
F14	reťazec
F15	mínus
F16	plus
F17	krát
F18	identifikátor premennej
F19	identifikátor funkcie
F20	ostro menší
F21	menší alebo rovný
F22	ostro väčší
F23	väčší rovný
F24	jednoriadkový komentár
F25	pravá zátvorka
F26	ľavá zátvorka

B LL-gramatika

1. $\langle \text{prog} \rangle \rightarrow \text{DEF ID_FUNC (} \langle \text{params} \rangle \text{) EOL } \langle \text{statement} \rangle \text{ END } \langle \text{prog} \rangle$
2. $\langle \text{prog} \rangle \rightarrow \text{EOL } \langle \text{prog} \rangle$
3. $\langle \text{prog} \rangle \rightarrow \text{EOF}$
4. $\langle \text{prog} \rangle \rightarrow \langle \text{statement} \rangle \langle \text{prog} \rangle$
5. $\langle \text{statement} \rangle \rightarrow \text{IF } \langle \text{expression} \rangle \text{ THEN EOL } \langle \text{statement} \rangle \text{ ELSE EOL } \langle \text{statement} \rangle \text{ END}$
6. $\langle \text{statement} \rangle \rightarrow \text{WHILE } \langle \text{expression} \rangle \text{ DO EOL } \langle \text{statement} \rangle \text{ END EOL}$
7. $\langle \text{statement} \rangle \rightarrow \langle \text{function} \rangle \text{ EOL}$
8. $\langle \text{statement} \rangle \rightarrow \text{ID } \langle \text{declare} \rangle \text{ EOL}$
9. $\langle \text{statement} \rangle \rightarrow \langle \text{expression} \rangle \text{ EOL}$
10. $\langle \text{declare} \rangle \rightarrow = \langle \text{expression} \rangle$
11. $\langle \text{declare} \rangle \rightarrow = \langle \text{function} \rangle$
12. $\langle \text{declare} \rangle \rightarrow \epsilon$
13. $\langle \text{params} \rangle \rightarrow \text{ID } \langle \text{param} \rangle$
14. $\langle \text{params} \rangle \rightarrow \epsilon$
15. $\langle \text{param} \rangle , \text{ID } \langle \text{param} \rangle$
16. $\langle \text{param} \rangle \rightarrow \epsilon$
17. $\langle \text{argvs} \rangle \rightarrow \langle \text{value} \rangle \langle \text{arg} \rangle$
18. $\langle \text{argvs} \rangle \rightarrow \epsilon$
19. $\langle \text{arg} \rangle \rightarrow , \langle \text{value} \rangle \langle \text{arg} \rangle$
20. $\langle \text{arg} \rangle \rightarrow \epsilon$
21. $\langle \text{value} \rangle \rightarrow \text{INT}$
22. $\langle \text{value} \rangle \rightarrow \text{FLOAT}$
23. $\langle \text{value} \rangle \rightarrow \text{STRING}$
24. $\langle \text{value} \rangle \rightarrow \text{ID}$
25. $\langle \text{function} \rangle \rightarrow \text{PRINT (} \langle \text{argvs} \rangle \text{)}$
26. $\langle \text{function} \rangle \rightarrow \text{LENGTH (} \langle \text{argvs} \rangle \text{)}$
27. $\langle \text{function} \rangle \rightarrow \text{SUBSTR (} \langle \text{argvs} \rangle \text{)}$
28. $\langle \text{function} \rangle \rightarrow \text{ORD (} \langle \text{argvs} \rangle \text{)}$
29. $\langle \text{function} \rangle \rightarrow \text{CHR (} \langle \text{argvs} \rangle \text{)}$
30. $\langle \text{function} \rangle \rightarrow \text{INPUTS}$
31. $\langle \text{function} \rangle \rightarrow \text{INPUTI}$
32. $\langle \text{function} \rangle \rightarrow \text{INPUTF}$
33. $\langle \text{function} \rangle \rightarrow \text{ID_FUNC (} \langle \text{argvs} \rangle \text{)}$

C Tabulka LL-gramatiky

	INT	FLOAT	STRING	PRINT	LENGTH	SUBSTR	ORD	CHR	INPUTS	INPUTI	INPUTF	ID_FUNC	EOL	EOF	IF	WHILE	DEF	ID	\$
$\langle \text{prog} \rangle$	4	4	4	4	4	4	4	4	4	4	4	4	2	3	4	4	1	4	
$\langle \text{statement} \rangle$				7	7	7	7	7	7	7	7	7			5	6		8	
$\langle \text{declare} \rangle$				11	11	11	11	11	11	11	11	11							12
$\langle \text{params} \rangle$																		13	14
$\langle \text{param} \rangle$																		15	16
$\langle \text{argvs} \rangle$	17	17	17															17	18
$\langle \text{arg} \rangle$	19	19	19															19	20
$\langle \text{value} \rangle$	21	22	23															24	
$\langle \text{function} \rangle$				25	26	27	28	29	30	31	32	33							

Tabulka 2: LL-gramatika

D Precedenční tabulka

	+	-	*	/	()	i	R	\$
+	>	<	<	<	<	>	<	>	>
-	>	>	<	<	<	>	<	>	>
*	>	>	>	>	<	>	<	>	>
/	<	>	>	>	<	>	<	>	>
(<	<	<	<	<	=	<	<	
)	>	>	>	>		>		>	>
i	>	>	>	>		>		>	>
R	<	<	<	<	<	>	<		>
\$	<	<	<	<	<		<	<	

Tabulka 3: Precedenční analýza výrazov