Implementace překladače imperativního jazyka IFJ24

Tým xvacla37, varianta výškově vyvážený binární strom

5. prosince 2024

Vedoucí: **Dominik Václavík (xvacla37)** – 25% Marek Slaný (xslany03) – 25% František Sedláček (xsedla2d) – 25% Radim Dvořák (xdvorar00) – 25%

Rozšíření:

ORELSE, UNREACHABLE, WHILE, FUNEXP

Obsah

1	Úvo	od		2										
2	Zpr	acován	า์	3										
	2.1	Návrh		. 3										
	2.2		překladače											
		2.2.1	Lexikální analýza											
		2.2.2	Syntaktická analýza											
		2.2.3	Sémantická analýza											
		2.2.4	Generátor kódu											
		2.2.5	Struktura generátoru kódu											
		2.2.6	Napojení částí											
3	Roz	zdělení	práce	6										
4	Kor	nečný a	automat	8										
5	Návrh syntaktického analyzátoru													
	5.1 LL gramatika													
	5.2	LL-ta	bulka	. 11										
	5.3	Drogo	donění tabulka	19										

1. **Úvod**

Dokumentace popisuje návrh a implementaci překladače imperativního jazyka IFJ24, který je zjednodušenou podmnožinou jazyka Zig. Pro náš projekt vedoucí týmu vybral úkol implementace tabulky symbolů ve výškově vyváženém binárním stromě. Implementace se nachází v souboru symtable.c.

2. Zpracování

2.1 Návrh

Při návrhu překladače bylo rozhodnuto za dúvodu odlehčení míry pokročilosti na rozdělení překladače na 4 části a rozhraní mezi těmito moduly.

2.2 Části překladače

- Lexikální analýza
- Rekurzivní syntaktická analýza
- Sémantická analýza
- Generace kódu

2.2.1 Lexikální analýza

Struktura tokenu je definována v souboru scanner.h. Token se sestává ze dvou atributů:

- typ tokenu daný výčtem.
- nepovinným obsahem, kde jsou ukládány hodnoty literálů a identifikátorů.

Implementace lexikální analýzy jako deterministický konečný automat je obsažena v souboru scanner.c ve funkci get_token(). Funkce je volána syntatickým analyzátorem při požadavku na další token. Pro ukládání řetezců určených jako číselné literály, řetezcové literály nebo identifikátory byla použita knihovna pro generický vektor definována v vector.h.

Při čtení nového tokenu je vstupní soubor procházén znak po znaku a na základě přečteného znaku dochází k přechodu do jiného stavu. Pokud v nějakém stavu dojde k přečtení znaku, který už neodpovídá danému tokenu, ale token už je platný, tak dochází k vrácení znaku na vstup a k vrácení platného tokenu syntatické analýze, jinak dojde k lexikální chybě, která je pak propagována dále. U tokenu, u kterého je potřeba si uchovat jeho obsah tak je znak pushnut do buffer, který je datového typu vector, a při úspěsném načtení výsledného tokenu je obsah buffer překopírován a u číselných literálů překonvertován do obsahu tokenu.

Vstupní soubor je načítán dokud se nenačte symbol EOF nebo pokud nedojde k lexikální nebo syntatické chybě.

2.2.2 Syntaktická analýza

Syntaktická analýza probíhá postupně rekurzivním sestupem pomocí LL-gramatiky a LL-pravidel. při sestupu je postupně generovaný derivační strom, který je použitý pro sémantickou analýzu a naslednou generaci mezi-kódu. Implementace se nachází v souborech parser.c a parser.h. Pro načítání tokenů je volána funkce get_token, která vrací nově načtený token nebo chybný kód, který je propagován. Rozhraní pro generaci derivačního stromu se nachází v souboru ast.h a působí jako rozhraní pro práci s moduly semantic a code_gen Pro řešení výrazů je také použitý rekurzivní sestup z důvodu vytvaření derivačního stromu. Derivační strom který obsahuje výraz v jazyce IFJ24 je binární. Dle priority operátorů se provede sestup až k operandu, který může být jeden z následujících:

- Proměnná
- Funkce
- Literál
- Závorky
- Ternární operátor

Závorky je považovány jako *pod-exprese* a stejným způsobem se řeší jako exprese (provede se rekurze).

2.2.3 Sémantická analýza

Soubor semantic.c. Před samotnou analýzou je inicializován zásobník tabulek symbolů – symtable_stack a jsou na něj vloženy vestavěné funkce.

Analýza je prováděna rekurzivním voláním funkce semantically_analyse(), zahájena je voláním této funkce na kořen AST, potom je volána na všechny děti, děti těchto dětí atd. Zpravidla probíhá rekurzivní volání před analýzou konkrétního uzlu, vyjímkou je pokud je potřeba vejít do nového scopu, v tom případě je prvně na symtable_stack vložena nová tabulka symbolů, po analýze daného uzlu je tato vrchní tabulka odstraněna. Další vyjímkou je kořenový uzel ve kterém se prioritně nadefinují signatury funkcí a je zkontrolována přítomnost a korektnost funkce main.

Při samotné kontrole jsou podle typu uzlu volány jednotlivé funkce na kontrolu daného uzlu. V rámci těchto funkcí mimo kontrolu také nastavuje/propaguje návratové typy a přiřazuje identifikátorům scope_id, které slouží pro jednoznačnou identifikaci symbolu v rámci funkce později při generaci kódu. Identifikátory jsou kontrolovány tak že při definici jsou vloženy do tabulky symbolů na vrchu symtable_stack. U výskytu symbolů se poté prochází postupně odshora všechny tabulky symbolů. Při odchodu ze scopu je také kontrolováno použití proměnných které se nachází v odstraňované tabulce.

2.2.4 Generátor kódu

Generátor kódu je klíčovou částí překladače jazyka IFJ24, která transformuje syntakticky a sémanticky správný abstraktní syntaktický strom (AST) na cílový mezikód ve specifickém pseudo assemblerovém formátu.

2.2.5 Struktura generátoru kódu

Generátor kódu se skládá z několika komponent, které kooperují při vytváření mezikódu:

- code_gen.c: Obsahuje všeobecný kod pro generaci
- **code_gen_help.c**: Obsahuje implementaci pomocných funkcí, například pro formátování escape sequencích
- func_look_up.c: Implementuje vyhledávající tabulku pro vestavěné funkce pomocí binárního stromu

Generátor začíná alokací paměti pro pomocné datové struktury (vyhledávací tabulka pro vestavěné fukce a halda pro identifikaci navěští), následně je vygenerováná hlavička programu

```
.IFJcode24
DEFVAR GF@null
CALL main
EXIT int@0
```

která se skládá ze návěští **.IFJcode24**, globální proměnou **GF@null** sloužící pro vyprázdnění zásobníku při volání funkce s throw away, zavolání main fukce a následní ukončení programu s ukončovacím kodem 0. Následně jsou vygenerováný postupně všechny funkce programu.

```
LABEL nazevFunkce
CREATEFRAME
DEFVAR TF@argument2
POPS TF@argument2
DEFVAR TF@argument1
POPS TF@argument1
DEFVAR TF@promnenaFukce
PUSHFRAME
```

Kod fuknce

POPFRAME RETURN

Kostra fuknce se sklada z LABELu fukce, vytvoření nového pamětového rámce, deklarací argumentů funkce a jejich naplnění, deklarací ostatních proměných funkce a následné pushnutí rámce na lokální zásobník. Při volání návratu z fukce se popne momentání pamětový rámec a zavolá se return. Následně generátor proiteruje všemy instrukcemi kodu a zpracuje je. If a While jsou zpracovány rekurzivně identifikaci návěští se používá zásobník návěští. Všechny exprese jsou řešeny pomocí rekuze a využívají zásobníkové instrukce.

2.2.6 Napojení částí

V main.c je main() funkce která volá syntaktický analyzátor (parser.c) přes funkci parse(). Ten vnitřně volá lexikální analyzátor(scanner.c) přes funkci get_token(). Syntaktický analyzátor sestaví syntakticky správný strom, který vrací. Strom je následně funkcí semantically_analyse() předán sémantickému analyzéru (semantic.c), ten ověří sémantickou správnost a doplní do stromu některé dodatečné informace. Strom je nakonec předán generátoru kódu (codegen.c) funkcí generate_code().

3. Rozdělení práce

Dominik Václavík – generace kódu

- util.c, util.h pomocná funkce d_string()
- code_gen.c, code_gen.h generace kódu
- code_gen_help.c, code_gen_help.h pomocné funkce pro generaci kódu
- symtable.c, symtable.h implementace tabulky symbolů pomocí výškově vyváženého binárního stromu
- func_look_up.c, func_look_up.h implementace stromu pro vyhledávání builting funkcí.

Marek Slaný – sémantická analýza

- ast.c, ast.h implementace abstraktního syntaktického stromu
- data_types.c, data_types.h definice datových typů, pomocných pseudotypů a pomocných funkcí pro práci s datovými typy
- scope_stack.c, scope_stack.h implementace zásobníku identifikátorů scopu
- semantic.c, semantic.h implementace sémantické analýzy a pomocných funkcí pro ni použitých
- util.c, util.h pomocná funkce has_decimals()
- vector.c, vector.h implementace generického vectoru a funkcí pro práci s ním
- testy

František Sedláček – syntaktická analýza

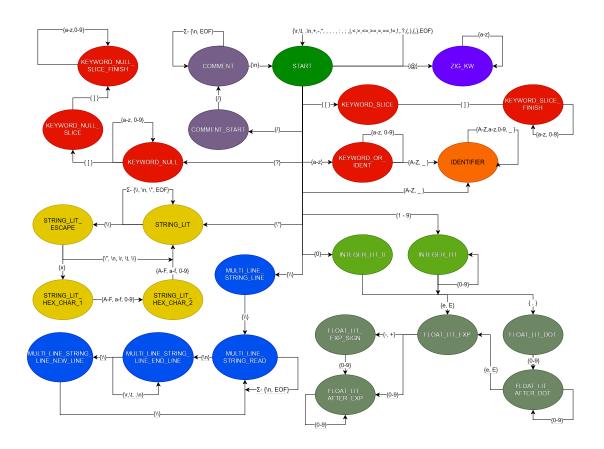
• error.c, error.h – definice chybových kódu a debugovacích funkcí

- parser.c, parser.h implementace syntaktické analýzy
- tvorba LL gramatiky

Radim Dvořák – lexikální analýza

- scanner.c, scanner.h implementace lexikální analýzy
- návrh stavového automatu pro lexikální analýzu

4. Konečný automat



5. Návrh syntaktického analyzátoru

5.1 LL gramatika

```
START \longrightarrow PROLOG FUNC LIST EOF
PROLOG \longrightarrow const \ ifj = @import("ifj24.zig");
FUNC LIST \longrightarrow \varepsilon
{\tt FUNC\_LIST} \longrightarrow {\tt FUNC\_DEF} \ {\tt FUNC\_LIST}
FUNC\_DEF \longrightarrow pub \text{ fn id ( PARAM\_LIST ) RET\_TYPE STMT\_LIST}
PARAM\_LIST \longrightarrow PARAM
PARAM_LIST \longrightarrow \varepsilon
PARAM \longrightarrow id : D TYPE P SEP
P SEP \longrightarrow PARAM
P SEP \longrightarrow \varepsilon
\operatorname{RET} \_\operatorname{TYPE} \longrightarrow \operatorname{void}
RET\_TYPE \longrightarrow D\_TYPE
{\tt STMT\_LIST} \longrightarrow {\tt STMT\ STMT\_LIST}
{\rm STMT\_LIST} \longrightarrow \varepsilon
STMT \longrightarrow VAR DEF;
STMT \longrightarrow id ASSIGN OR CALL OR LABEL STMT // ACL STMT short
\mathrm{STMT} \longrightarrow \mathrm{IF} \ \mathrm{STMT}
STMT \longrightarrow WHILE\_STMT
STMT \longrightarrow RET\_STMT;
\mathrm{STMT} \longrightarrow \mathrm{FOR}\_\mathrm{STMT}
VAR\_DEF \longrightarrow VAR\_CON \ id \ TYPE\_AND\_ASSIGN
VAR\_CON \longrightarrow const
VAR\_CON \longrightarrow var
{\tt TYPE\_AND\_ASSIGN} \longrightarrow {\tt OPT\_TYPE} = {\tt expr}
OPT TYPE \longrightarrow : D TYPE
\mathsf{OPT}\mathsf{\_TYPE}\longrightarrow\varepsilon
D TYPE \longrightarrow i32
D TYPE \longrightarrow f64
D_TYPE \longrightarrow ?i32
D TYPE \longrightarrow ?f64
D TYPE \longrightarrow []u8
D_TYPE \longrightarrow ?[]u8
ASSIGN_OR_CALL_LABEL_STMT \longrightarrow ASSIGN_STMT;
ASSIGN_OR_CALL_LABEL_STMT \longrightarrow FUNC_CALL_STMT;
ASSIGN OR CALL LABEL STMT \longrightarrow WHILE STMT
ASSIGN\_STMT \longrightarrow = expr
```

```
IF\_STMT \longrightarrow if COND STMT\_LIST ELSE\_BLOCK
COND \longrightarrow (expr) NULL\_COND
NULL COND \longrightarrow \varepsilon
NULL\_COND \longrightarrow | id |
\texttt{ELSE\_BLOCK} \longrightarrow \varepsilon
\texttt{ELSE\_BLOCK} \longrightarrow \texttt{else} \ \texttt{ELSE\_PART}
ELSE\_PART \longrightarrow IF\_STMT
ELSE\_PART \longrightarrow STMT\_LIST
FUNC\_CALL\_STMT \longrightarrow (INPUT\_PARAM\_LIST)
INPUT PARAM LIST \longrightarrow INPUT PARAM
INPUT PARAM LIST \longrightarrow \varepsilon
{\tt INPUT\_PARAM} \longrightarrow {\tt expr} \ {\tt IP\_SEP}
IP SEP \longrightarrow \varepsilon
IP\_SEP \longrightarrow , INPUT\_PARAM\_LIST
RET\_STMT \longrightarrow return OPT\_EXPR
OPT\_EXPR \longrightarrow \varepsilon
OPT\_EXPR \longrightarrow expr
FOR\_STMT \longrightarrow for (expr) \mid id \mid STMT\_LIST // null condition is a must.
WHILE STMT \longrightarrow WHILE_NAME while ( expr ) NULL_COND SPEC_WHILE_CNT_STMT_LIST
SPEC WHILE ELSE
WHILE_NAME \longrightarrow \varepsilon
WHILE_NAME \longrightarrow:
{\tt SPEC\_WHILE\_CNT} \longrightarrow \varepsilon
SPEC\_WHILE\_CNT \longrightarrow : (CNT\_BLOCK)
CNT BLOCK \longrightarrow STMT
\text{CNT} BLOCK \longrightarrow STMT LIST
{\tt SPEC\_WHILE\_ELSE} \longrightarrow \varepsilon
{\tt SPEC\_WHILE\_ELSE} \longrightarrow {\tt else} \ {\tt STMT\_LIST}
STMT \longrightarrow CONTINUE STMT;
STMT \longrightarrow BREAK\_STMT;
CONTINUE STMT \longrightarrow continue OPT LABEL
BREAK STMT \longrightarrow break OPT LABEL
OPT\_LABEL \longrightarrow \varepsilon
OPT\_LABEL \longrightarrow id
expr - expression (expression, math, function call)
id – identification (of a variable, function name, label)
```

5.2 LL–tabulka

	M	major			Tpih 1		"weld"	Tear?	55	100178417802178451545116451645	***	181	"else"	10	717	1111	" for "	"white"	"continue"	"break"	"return"
CART				START -> PROLOG PUNC USE end																	
				PROLOG -> semi-lity - (limper(CR2Lety))		_			_				_	_	_	_			_		_
4000					FUNCTION - NUMBER OF	_	-												_		_
UNC,SRT			PUNC LIST ++		FUNC LET FUNC DEF-+ puls in all																
PUNC DEF					(NAME LET) SET, TYPE (STMT LET)																
					(sterjust)	_			_	-			_	_	_	_		-	_		_
PARKING LIST	PARAMICLET - PARAMI PARAMICLET		PRISON USE + 1																		
PARKE	PARKED - M. D. TYPE P SEP																				
- 107	F 107 ***					P MP															
BET TOPE	7,001.1						SET TYPE -			BET TYPE D TYPE											_
	STMT UST ->			arner unt +		_	-	STREET LIST +	_	MAIL THE TRUITS	_	and lat + and	_	_	_	_	STREET LIST ->		STMT USE -	STREET LIST ->	STMT LET -
mer,use	smerkmer part		STMT_SST → *	arner armer paer				and the par				amer Lat					amer anner pase	energian → sterramegian	merane uar	amer's mery part	STMT STMT US
STAT	STRET - WACL STRET			STREET VAN DREF				armer year per				amer - p amer						STRET WHILE STRET	STMT -> CONTINUE STMT	arner	STMT +
LIMI .	STATE OF THE PARTY NAMED IN					_	_	ELINE TO CHANGE	_		_	LINE TP, LINE	_	_	_		EIRC TOR EIRC	and wanted that	CORTINUE DIRECT	mana, steet	part and the
NAS DEP				HAN DEF VMX CON H TVPE, AND, ASSON																	
NAK CON				MAK CON - House				VMI_CON → sar							_						_
			TYPE, AND, ASSIGN →																		
TYPE, AND ASSIGN OPT, TYPE			OPT TYPE - maps			_				TYPE AND ASSOCIATION TYPE - maps			_		_				_		_
D TYPE						_	_			D_TVPE -+ tope			_	_	_	_			_		_
ACL STMT									ACL STMT → WHILE STMT		ACL STREET ->			ACL, STHET →							
						_	_		WHILE STAT	-			_	PUNE CALL STREET	_	_			_	_	_
ALBICH, STMT	_										- sain	IF STATT - IF COND			_	_					_
P STMT												But BOOK									
						_	_	_	_		_	HAR PLACE		COND -> (map)		_			_		_
COND	_					_							EUR EUR Andre	MMT COMB.	_				_		_
MULL_COMD	NULL_COND + +												ELM PLOCE -1 shar ELM PART			NULL COND + W					
BLBE, BLOCK	ELIE_BLOCK++					_	-		_		_		_						_	_	_
BAR,PART												ELIE PART - IF STMT			ELSE PART -> (ETS-IT_LEST)						
														FUNC, CALL, STREET ->							
PUNC_CALL_STMT		REPORT PROMISE LIST +				_				-			_	DAPUT PRISM LIST	_	_			_		_
INPUT, PARAM, LIST		INPUT FAMILY	DIPUT, PARKIN, LIST ++																		
NPUT PARAM		PUT NAME - repr				F 107 1															
P, MP			P. MP ++			NPUT PRISON															
BET STMT																					RET, ETMT ->
CPT_EXPE		OFT DOTE - repr	OPT_EXPE ++																		
																	POR STMT -> har Comp 1(16) CSTMT LIST I				
POR_STMT																					
																		MALL COND IPEC WHEE CAT STAT LIST SPEC WHEE ELSE			
WHEE THE			MINELE NAME +						WHILE NAME +:									I BAC MHT LTIS			
mma, note		_	perma mode + s		_	_	_			-			_	_	_	_		_	_		+
SPEC MINUS CNT			SPEC WHILE CHT ++						SPEC WHELE CHT												
ONT BLOCK	CMT BLOCK → STMT			ONT BLOCK + STMT				OUT BLOCK + STREET				OR BLOCK + STHE			CHT/BLDCK -+			ONT BLOCK STMT	CNF, BLOCK +	CNT_BLOCK +	ONT BLOCK ->
	No. BOOK + ITMT			DATE BOOK + ETMT		_	_	NAME AND ASSOCIATION OF THE PARTY OF T	_		_	AND MAKES + ETSET	SPEC WHEE ELSE ->	_	CALIFORNIA LIET	_	DATE BROCK + ETHER	CHI MILES THE	2000	2.00	
IPIC,WHILE, ILIE			SPEC_WHILE_ELSE ++			_	_						ete (Street List)	_	_				CONTINUE STREET		_
CONTINUE STHEFT			1												1					RETURN STMT -+	1
BREAK STIMT						_				1				_	_	_		1			_
OPT_LABEL	OPT LASES W		OPT LAMEL -> x																		

5.3 Precedenční tabulka

Priorita	Operátory	Asociativita
1	Postfix ".?"	bez
2	" <u>- !</u> "	pravá
3	"* /"	levá
4	"+ -"	levá
5	"orelse"	levá
6	"==!= < > <= >="	levá
7	"and"	levá
8	"or"	levá