

Dokumentace k projektu pro předměty IFJ a IAL

Implementace interpretu imperativního jazyka

IFJ14

Tým 113, varianta a/1/I

Vedoucí týmu: Antonín Marko

14. prosince 2014

Řešitelé:	3BIT	Antonín Marko,	xmarko07@stud.fit.vutbr.cz
	1BIT	Tomáš Pružina,	xpruzi01@stud.fit.vutbr.cz
	3BIT	Martin Kubíček,	xkubic34@stud.fit.vutbr.cz
	3BIT	Martin Juřík,	xjurik08@stud.fit.vutbr.cz
	3BIT	Petr David	xdavid15@stud.fit.vutbr.cz

Fakulta Informačních Technologií
Vysoké Učení Technické v Brně

Obsah

1	Úvod	1
1.1	Rozdělení bodů	1
1.2	Rozšíření	1
2	Moduly interpretu jazyka	2
2.1	Lexikální analyzátor	2
2.2	Syntaktický analyzátor	2
2.2.1	Shunting-yard algoritmus	3
2.3	Interpret	3
2.3.1	Tabulka symbolů	4
3	Postup při implementaci řešení	4
3.1	Funkce <code>sort</code>	4
3.2	Funkce <code>find</code>	5
3.3	Funkce <code>length</code>	5
3.4	Funkce <code>copy</code>	5
3.5	Shunting-yard algoritmus	5
4	LL-tabulka	6
5	Závěr	8
A	Metriky kódu	8

1 Úvod

Implementace překladače imperativního jazyka je netriviální úloha, proto je vhodné ji rozdělit na podúlohy (2), které jsou vhodně rozdělené mezi členy týmu, který je vedené a kontrolován týmovým vedoucím.

Tento dokument sa skládá z N částí, které popisují jednotlivé moduly interpretu imperativního jazyka IFJ14, a to jsou lexikální analyzátor (2.1), syntaktický analyzátor (2.2) využívající rekursivní sestup (), který je pro potřeby syntaktické analýzy výrazů rozšířený o Shunting-Yard (2.2.1) algoritmus¹. Kontrola syntaxe a sémantiky probíhá z velké části při jednom průchodu, následné kontroly sémantiky probíhají až za běhu.

Jelikož je projekt založen na týmové spolupráci, rozhodli jsme se využít možnosti používat verzovacího systému, jmenovitě **bitbucket**. Vzhledem k tomu, že práce byla rozdělena mezi jednotlivé členy a probíhaly kontroly v době schůzek jevil se tento systém jako nejlepší, do jisté míry totiž umožňuje porovnávání verzí a tím se dá kontrolovat zdali daný člen splnil svou část úkolu či nikoliv.

Práce v teamu byla rozdělena následovně: Interpret dostal na starosti **Tomáš Pružina**. Parser a tedy syntaktickou a sémantickou část projektu vytvářel **Antonín Marko.Martin Juřík**, měl na starost zpracování lexikálního analyzátoru a jeho funkcí otestování. Celkové testování, zpracování LL gramatiky a dokumentaci měl na starost **Martin Kubíček**. **Petr David** uprava knihovních funkcí a vestavěné funkce do předmětu **Algoritmy**.

Za účelem testování a ověření plné funkčnosti projektu byla napsána sada testů, obsahující i vlastní testovací skript, čítající zhruba **X** odvětví s celkovým počtem **XXX** testů. Úkolem testování, bylo především schopnost odchyťovat všechny možné chyby, které by mohly nastat a vrácení správného chybového kódu. Ovšem ne vždy bylo určení správného chybového kódu jednoduché jelikož jsme narazili na pár sporných momentů, které jsme se snažili vyřešit důkladnějším studiem dokumentace k projektu nebo společnou debatou v týmu. Dále jsme testy ověřovaly do jaké míry jsme schopni správně interpretovat validní data.

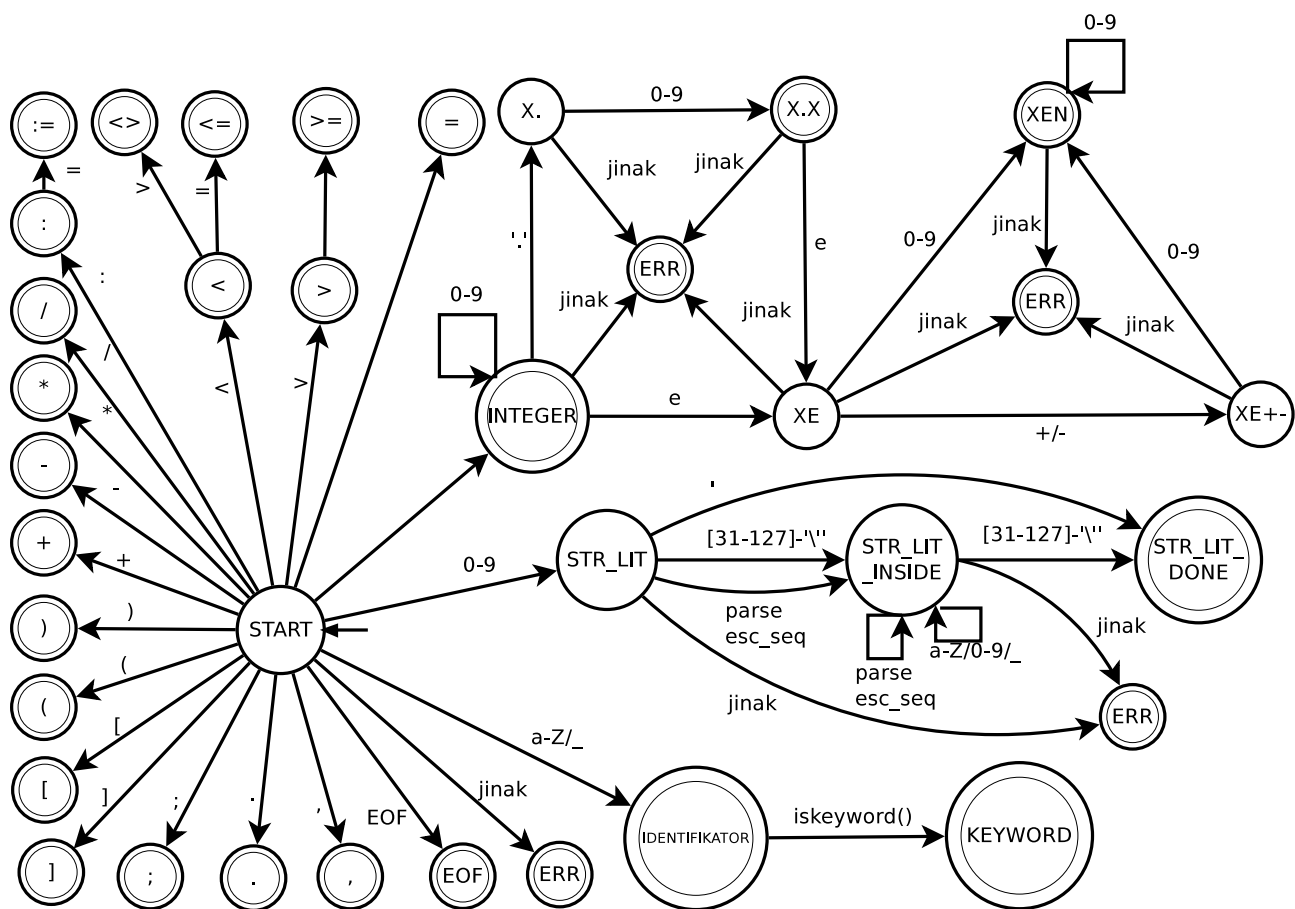
1.1 Rozdělení bodů

xmarko07:20
xpruzi01:20
xkubic34:20
xhjurik08:20
xdavid15:20

1.2 Rozšíření

ARRAY
REPEAT
ELSEIF
BOOLOP
FOR

¹Byla nám dovolena výjimka, implementovat a náležitě zdokumentovat tento algoritmus oproti předepsané precedenční analýze



Obrázek 1: Konečný automat lexikálního analyzátoru

2 Moduly interpretu jazyka

2.1 Lexikální analyzátor

Syntaktický analyzátor potřebuje ke své činnosti lexikální analyzátor. Ten předává na žádost syntaktického analyzátoru tzv. tokeny, které získává postupným čtením vstupního souboru, který obsahuje zdrojový kód napsaný v jazyce IFJ14. Samotný výsledný token je reprezentací lexému (identifikátor, klíčové slovo, příkaz přiřazení atp.).

Pro úspěšné rozlišení typu lexému se v naší implementaci používá konečný automat, jehož struktura je znázorněna na obrázku (2.1).

Jestliže se konečný automat v době zpracovávání řetězce zdrojového souboru `document` do chybového stavu `lexerref`, řetězec je nepřijatý a jedná se o lexikální chybu. Samotná implementace lexikálního analyzátoru patří k jednodušším částem projektu, avšak jeho návrh a testování zabralo nemálo času.

2.2 Syntaktický analyzátor

Jak již bylo zmíněno, syntaktický analyzátor žádá lexikální analyzátor o tokeny a na základě sekvence po sobě jdoucích tokenů sestavuje abstraktní syntaktický strom (dále jen AST). Tato sekvence musí odpovídat pravidlům syntaxe jazyka IFJ14, v opačném případě se jedná o syntaktickou chybu.

Jazyk IFJ14 je staticky typovaný jazyk, tedy typové kontroly probíhají během překladač, v našem případě právě při průchodu syntaktickým analyzátořem. Je tedy nutné zachytávat chybové vstupy a přímo ukončovat průchod a tím i celou aplikaci.

Jako vnitřní kód programu jsme si zvolili AST, tedy binární strom a rozšířili jsme jeho strukturu o potřebné položky pro práci interpretu. Jak již bylo zmíněno výše, jedná se o binární strom, má tedy dva synovské uzly. Avšak pro určité typy uzlů je mít nemusí. Například uzel pro operaci NOT má pouze levý synovský podstrom.

Pro převod příkazů obsahujících tělo na AST uzly byla v týmu domluvena konvence, že primárním uzlem pro tělo bude levý podstrom. Tedy například cyklus `while` má svoje tělo uložené v levém podstromu. Zde je potřeba zmínit, že příkaz `if` má v podstatě dvě sekvence příkazů a to pravdivou a nepravdivou. Je tedy potřeba zajistit, aby se obě uchovaly v jednom uzlu AST.

Řešení je jednoduché a více než samozřejmé, uložení pravdivé sekce do levého podzlu stromu a nepravdivou sekci do uzlu pravého. U příkazů s podmíněným chováním (`if`, `while`, `repeat-until`, atp.) bylo nutné uchovat podmínku, tedy výraz, který určuje, jestli je příkaz proveditelný či není. Primárně pro tento účel slouží dodatečná položka uzlů AST, která svým typem `void*` umožňuje mnohostranné využití.

Zvláštní případ použití dodatečné položky je například u definice či volání funkce. V tomto případě obsahuje strukturu `String` se jménem funkce. Levý podstrom opět obsahuje tělo funkce a v pravém podstromu je uložen uzel se strukturou `varspars`, která se složená z dvou front, jedné pro lokální proměnné funkce a druhá pro parametry funkce. Tyto fronty jsou naplněny při definici funkce, při deklaraci (použití `forward` deklarace) je naplněna pouze fronta parametrů. Více o funkcích v ??.

2.2.1 Shunting-yard algoritmus

Pro zpracování výrazu mělo být původně v projektu užito precedenční analýzy. My jsme ovšem toto zpracování řešili pomocí Shunting-yard algoritmu². Algoritmus zpracování výrazů je založen na prioritě operátorů, které buď vkládá nebo odebírá ze zásobníku a vkládá do výstupní sekvence. Tento algoritmus primárně slouží k převodu infixového zápisu výrazů do postfixového, tedy lépe zpracovatelného a převeditelného do AST.

Tento algoritmus byl pro potřeby projektu upraven, aby přímo generoval AST místo výstupního postfixového zápisu a zároveň u tvorby jednotlivých uzlů zastupujících operace kontroloval, zda-li datový typ vrcholových uzlů synovských uzlů odpovídá sémantickým pravidlům zpracovávání výrazů. Což nám poskytuje možnost pouze jediného průchodu, který tak stačí na zpracování.

Toto řešení ovšem není tak přínosné jak se zdálo, jelikož při tvorbě výrazů je nutné kontrolovat spoustu aspektů validního výrazu. To jsou například datové typy jednotlivých operandů v případě binárních operací.

2.3 Interpret

Interpret je závěrečnou částí projektu. Rekurzivně zpracovává AST reprezentovaný binárním stromem a postupně vykonává operace nad ním definované.

Kořenem tohoto AST je vždy typ převedené operace, přičemž obsahuje předem definované nepovinné parametry (příkladem je podstrom volání funkce, který obsahuje předávání

²Byla nám dovolena výjimka, implementovat a náležitě zdokumentovat tento algoritmus oproti předepsané precedenční analýze

proměnné). Tedy každý uzel AST představuje jednu atomickou operaci, kterou je řízený samotný běh programu.

Interpret samozřejmě nepracuje jen se samotným syntaktickým stromem, ale taktéž s tabulkou symbolů, která obsahuje data potřebná na to, aby interpret mohl vykonávat nějakou užitečnou práci definovanou v těle programu.

Interpret tedy spravuje tabulky symbolů, dynamicky je za běhu vykonávaného programu vytváří a (podla potřeby) ruší.

V neposlední řadě interpret vykonává sémantickou kontrolu u operacích, které syntaktický analyzátor v době zpracování vstupního programu a generování abstraktního syntaktického stromu neřešil.

Příkladem takovéto kontroly je používání (čtení) proměnné před její řádnou definicí (přiřazením hodnoty) anebo dělení nulou v aritmetickém výrazu.

Takovéto běhové chyby jsou v interpretě řešeny ukončením interpretace a navrácením chybového kódu podle zadání imperativního jazyka IFJ14.

2.3.1 Tabulka symbolů

Tabulka symbolů je implementována pomocí binárního vyhledávacího stromu, přičemž se rozlišuje několik úrovní tabulky symbolů.

Každý program obsahuje globální tabulku symbolů (ve které jsou uloženy ukazatele na funkce) a lokální tabulku symbolů, která obsahuje běhové proměnné vykonávaného programu.

Z implemetačního hlediska jsou lokální tabulky symbolů (stromy) ukládány na zásobník, přičemž platí, že při volání funkce se vytvoří nová (lokální) tabulka symbolů, do které se skopírují příslušné parametry volané funkce z nižší vrstvy lokální tabulky (popř. z globální tabulky symbolů).

Taková lokální tabulka taktéž obsahuje návratovou hodnotu vykonávané funkce, která umožňuje rekurzivní volání funkcí a samozřejmě umožňuje vykonávat tělo funkce bez nutnosti jakkoliv zasahovat do abstraktního syntaktického stromu.

3 Postup při implementaci řešení

3.1 Funkce sort

K implementaci funkce sort jsme podle zadání použili algoritmus quicksort. Kromě funkce quicksort jsme potřebovali ještě pomocnou funkci **partition**, která rozděljuje řazené pole na dvě části. V jedné takto vzniklé části jsou prvky s menší hodnotou než je prvek rozdělující tyto části (pseudomedian, dále jen PM) a v druhé části prvky s hodnotou větší než PM. Index PM určíme jako $(l + p)/2$, kde i je index nejlevějšího prvku řazeného úseku pole a j index nejpravějšího prvku tohoto pole. Všechny prvky s hodnotou nižší než je PM poté přesuneme na levou stranu od PM a prvky s vyšší hodnotou na pravou stranu. Poté aplikujeme algoritmus quicksort na obě poloviny pole.

Konkrétní implementace algoritmu je převzata z přednášek předmětu Algoritmy (IAL2014 [2] – 10. přednáška 33. slide). Pro snazší konverzi do jazyka C a pro pochopení algoritmu samotného jsme vytvořili vývojový graf a podle grafu jsme sepsali samotný kód v jazyce C.

Modifikace: Algoritmus popsáný v přednáškách obsahoval prohození prvku se sebou samým, takže jsme provedli drobnou úpravu, která tomuto zabránila.

3.2 Funkce find

Pro funkci `find` jsme dle zadání použili KMP (Knuth-Morris-Prattův algoritmus). Jedná se o algoritmus využívající konečný automat. Tento konečný automat využívá hrany ANO/NE. Nejprve bylo nutné vytvořit si pole `FAIL`, do kterého se uloží zjištěné hodnoty posunů. Poté přijde na řadu samotné vyhledávání řetězce. Nejprve zarovnáme podřetězec a řetězec na první index. Pak se jen porovnávají znaky, pokud jsou shodné, zvyšuje se index o 1 v podřetězci, který hledáme (`pattern`) a zároveň i v řetězci, ve kterém vyhledáváme (`text`). Pokud nikoliv, tak se vracím po hraně "N" (NE hrana) neboli `FAIL` vektor. Je-li pak hodnota indexu v hledaném řetězci rovna délce hledaného podřetězce, tak byl nalezen vzorek na indexu $TInd - PL$, kde $TInd$ je index řetězce a PL je délka hledaného podřetězce, jinak nic nenašel a vrací se 0.

Algoritmus byl převzán a přepsán do jazyka C z přednášek předmětu Algoritmy (IAL2014 [2] – 11. přednáška 18. slide).

Modifikace: Pro návrat 0 v případě neúspěchu byla provedena modifikace algoritmu z přednášek, kde byl vracen poslední dosažený index, tedy délka řetězce.

3.3 Funkce length

Funkce vrací velikost stringu pro zadanou hodnotu. Více zde není nutno implementovat.

3.4 Funkce copy

Umožňuje kopírovat hodnotu z zadaného stringu, do nového. K vytvoření nového stringu složí funkce `makeNewString`, která je podrobně popsána v knihovně `string.c`, který zajišťuje práci s řetězci. Avšak před samotným uložením se provedou sady kontrol. Jako například zda předávaná hodnota či celá struktura nejsou nulové, jestli počáteční index odpovídá hodnotě 0 tedy standardu C-like a v poslední řadě, zda délka není menší či rovna 0. Po samotné alokaci paměti pro daný string se upraví indexování aby odpovídalo C-like standardu. Jelikož, je použit implementační jazyk C, který neumí pracovat se stringy je zde string interpretován jako poslounost za sebou jdoucích znaků typu `char` uložených do pole.

3.5 Shunting-yard algoritmus

Jak už bylo výše zmíněno algoritmus funguje na bázi priority, kde v našem případě jsou priority stanoveny takto (bráno od nejvyšší po nejnižší, je zda zahrnuto i rozšíření): `[not]`, `[Logický and, násobení, dělení]`, `[sčítání, odčítání, or, xor]`, `[porovnávací operátory]` pokud se zde objeví něco jiného je nahlášena chyba. Konkrétně je algoritmus implementován jako funkce `parseExpression` v `parser.c`. Obsahuje jeden nekonečný cyklus, který se ukončí pouze v případě chyby či ukončení výrazu. Zpracování obsluhuje jeden obsáhlý switch, který se snaží pokrýt všechny možné situace, poté provede příslušné úkony dle toho jaká část výrazu právě zpracovává. Největším problémem zde bylo odchyťávání specifických chyb. Např.: `(3(+8))`, která je řešena tak, že pokud načteme levou závorku uschováme si ji a pokud za ní následuje operátor je vypsána příslušná chyba.

4 LL-tabulka

PROGRAM	-> VARS FUNC begin CMD_LIST enddot
VARs	-> var VAR_DEF VAR_DEFS
VARs	-> eps
VAR_DEFS	-> VAR_DEF VAR_DEFS
VAR_DEFS	-> eps
VAR_DEF	-> id : DT_TYPE ;
DT_TYPE	-> integer
DT_TYPE	-> real
DT_TYPE	-> boolean
DT_TYPE	-> string
FUNC	-> eps
FUNC	-> HEADER
HEADER	-> function id (DEF_PARAMS) : DT_TYPE ; AFTER_HEADER
AFTER_HEADER	-> forward ;
AFTER_HEADER	-> VARS BODY ;
DEF_PARAMS	-> eps
DEF_PARAMS	-> DEF_PAR DEF_PAR_LIST
DEF_PAR_LIST	-> eps
DEF_PAR_LIST	-> ; DEF_PAR
DEF_PAR	-> id : DT_TYPE
BODYN	-> begin CMD_LIST_N end
CMD_LIST_N	-> eps
CMD_LIST_N	-> CMD CMD_FOLLOW
CMD_FOLLOW	-> eps
CMD_FOLLOW	-> ; CMD CMD_FOLLOW
BODY	-> begin CMD_LIST end
CMD_LIST	-> CMD CMD_FOLLOW
CMD	-> ASSIGN
CMD	-> BODYN
CMD	-> IF
CMD	-> WHILE
CMD	-> WRITE
CMD	-> READLN

ASSIGN	-> id := AFTER_ASSIGN
AFTER_ASSIGN	-> EXPR
AFTER_ASSIGN	-> CALL
IF	-> if EXPR then BODYN IF_ELSE
IF_ELSE	-> eps
IF_ELSE	-> else BODYN
WHILE	-> while EXPR do BODYN
CALL	-> id (TERM_LIST)
CALL	-> sort (dt_str)
CALL	-> find (dt_str , dt_str)
CALL	-> length (dt_str)
CALL	-> copy (dt_str , dt_str , dt_int)
READLN	-> readln (id)
WRIT	-> write (TERM_LIST)
TERM_LIST	-> TERM
TERM_FOLLOW	-> , TERM TERM_FOLLOW
TERM_FOLLOW	-> eps
TERM	-> id
TERM	-> dt_int
TERM	-> dt_real
TERM	-> dt_bool
TERM	-> dt_str
EXPR	-> vyraz

5 Závěr

A Metriky kódu

Reference

- [1] HONZÍK J. M.: *Studijní opora pro předmět Algoritmy*. Elektronický text. FIT VUT v Brně
- [2] HONZÍK J. M.: *Přednášky k předmětu Algoritmy*. Soubor elektronických textů. FIT VUT v Brně
- [3] MEDUNA A., LUKÁŠ R., *Podklady k přednáškám*. Elektronický text. FIT VUT v Brně
- [4] MEDUNA A., LUKÁŠ R., *Studijní opora pro předmět Formální jazyky a překladače*. Elektronický text. FIT VUT v Brně