



Projektová dokumentace
Implementace překladače imperativního jazyka IFJ17
Tým 104, varianta II

29. listopadu 2017

Dominik Harmim	(xharmi00)	25 %
Vojtěch Hertl	(xhertl04)	25 %
Timotej Halás	(xhalas10)	25 %
Matej Karas	(xkaras34)	25 %

Obsah

1	Úvod	1
2	Návrh a implementace	1
2.1	Lexikální analýza	1
2.2	Syntaktická analýza	1
2.2.1	Syntaktická analýza založená na LL – gramatice využívající metodu rekurzivního sestupu	1
2.2.2	Zpracování výrazů pomocí precedenční syntaktické analýzy	1
2.3	Sémantická analýza	1
2.4	Generování cílového kódu	1
2.4.1	Rozhraní generátoru kódu	1
2.4.2	Začátek generování	2
2.4.3	Generování funkcí	2
2.4.4	Generování výrazů	2
2.4.5	Generování návěští	2
2.5	Překladačový systém	2
2.5.1	CMake	2
2.5.2	GNU Make	2
3	Speciální algoritmy a datové struktury	3
3.1	Tabulka s rozptýlenými položkami	3
3.2	Dynamický řetězec	3
3.3	Zásobník symbolů pro precedenční syntaktickou analýzu	4
4	Práce v týmu	4
4.1	Způsob práce v týmu	4
4.1.1	Verzovací systém	4
4.1.2	Komunikace	4
4.2	Rozdělení práce mezi členy týmu	4
5	Závěr	4
A	Diagram konečného automatu specifikující lexikální analyzátor	7
B	LL – gramatika	8
C	LL – tabulka	8
D	Precedenční tabulka	8

1 Úvod

Cílem projektu bylo vytvořit program v jazyce C, který načte zdrojový kód zapsaný ve zdrojovém jazyce IFJ17, jenž je zjednodušenou podmnožinou jazyka FreeBASIC a přeloží jej do cílového jazyka IFJcode17 (mezikód).

Program funguje jako konzolová aplikace, které načítá zdrojový program ze standardního vstupu a generuje výsledný mezikód na standardní výstup nebo v případě chyby vrací odpovídající chybový kód.

2 Návrh a implementace

Projekt jsme sestavili z několika námi implementovaných dílčích částí, které jsou představeny v této kapitole. Je zde také uvedeno, jakým způsobem spolu jednotlivé dílčí části spolupracují.

2.1 Lexikální analýza

TODO

2.2 Syntaktická analýza

TODO

2.2.1 Syntaktická analýza založená na LL – gramatice využívající metodu rekurzivního sestupu

TODO

2.2.2 Zpracování výrazů pomocí precedenční syntaktické analýzy

TODO

2.3 Sémantická analýza

TODO

2.4 Generování cílového kódu

Generování cílového kódu pro nás znamená generování mezikódu IFJcode17. Kód je generován na standardní výstup po dokončení všech analýz, přičemž jednotlivé dílčí části jsou generovány v průběhu analýz do interní paměti programu. Při generování mezikódu jsou generovány i komentáře pro jeho zpřehlednění.

2.4.1 Rozhraní generátoru kódu

Generování kódu je implementováno jako samostatný modul v souboru `code_generator.c`, jehož rozhraní se nachází v souboru `code_generator.h`. Rozhraní nabízí 3 základní funkce. Jedná se o funkci, která připraví generování kódu, tato funkce se volá na začátku syntaktické analýzy. Funkce, která uvolní z paměti všechny rezervované prostředky generátorem kódu se volá po dokončení syntaktické analýzy. Poslední důležitá funkce generuje vytvořený kód na patřičný výstup. Rozhraní nabízí spoustu dalších funkcí pro generování jednotlivých částí programu, např. generování začátku funkce, volání funkce, deklarace proměnné, příkazu `input`, funkcí pro zpracování výrazů, podmínek, cyklů, aj. Všechny tyto funkce se volají ve správnou chvíli a se správnými parametry v průběhu syntaktické analýzy a syntaktické analýzy výrazů.

2.4.2 Začátek generování

Na začátku generování jsou inicializovány potřebné datové struktury (které jsou na závěr opět uvolněny), vygenerována hlavička mezikódu, která zahrnuje potřebné náležitosti pro korektní interpretaci mezikódu, definici globálních proměnných na globálním rámci a skok do hlavního těla programu. Poté jsou vygenerovány vestavěné funkce, které jsou zapsány přímo v jazyce IFJcode17.

2.4.3 Generování funkcí

Každá funkce je tvořena návěští ve tvaru `$identifikator_funkce` a svým lokálním rámcem. Pro funkce se vždy generuje výchozí návratová hodnota podle jejího návratového datového typu, která se ukládá na lokální rámec, který je po odchodu z funkce dostupný jako dočasný rámec. Před zavoláním funkce jsou hodnoty parametrů uloženy do dočasného rámce s číslem, které odpovídá pořadí daného parametru a po vstupu do funkce jsou všechny předané parametry uloženy na lokální rámec se svým názvem. Při generování návratu z funkce je předán výsledek zpracovaného výrazu do proměnné s návratovou hodnotou funkce a proveden skok na konec funkce.

2.4.4 Generování výrazů

Všechny výrazy jsou během syntaktické analýzy výrazů ukládány na datový zásobník a v pravou chvíli jsou provedeny příslušné operace s hodnotami na vrcholu zásobníku. Hodnota zpracovaného výrazu se ukládá do globální proměnné na globálním rámci.

2.4.5 Generování návěští

Všechna návěští, pokud neuvažujeme návěští pro funkce a jiná speciální návěští, jsou generovány ve tvaru `$identifikator_funkce%label_deep%label_index`, kde `identifikator_funkce` zajistí unikátnost návěští mezi jednotlivými funkcemi, `label_deep` je hloubka zanoření návěští (například u podmínek a cyklů) a `label_index` je číselná hodnota, která se inkrementuje s každým dalším návěští uvnitř dané funkce.

2.5 Překladačový systém

Projekt je možné přeložit dvěma způsoby, buď nástrojem CMake nebo nástrojem GNU Make. Oba způsoby vytvoří spustitelný soubor `ifj17_compiler`.

2.5.1 CMake

Při vývoji, ladění a testování projektu někteří z nás používaly CMake, protože překlad tímto nástrojem je integrován ve vývojovém prostředí CLion, které většina členů týmu používala a protože nastavení pravidel pro překlad tímto nástrojem je jednoduché.

V souboru `CMakeLists.txt` jsou nastavena pravidla pro překlad. Je zde nastaven překladač `gcc` a všechny potřebné parametry pro překlad. Dále je zde uvedeno, že se pro překlad mají používat všechny soubory s příponou `.c` a `.h` a taky je zde uveden název spustitelného souboru.

Soubory pro překlad je možné vygenerovat příkazem `cmake .` a překlad provést příkazem `cmake --build .` nebo pro toto využít nástroje vývojového prostředí.

2.5.2 GNU Make

Jedním z požadavků pro odevzdání projektu bylo přiložit k odevzdanému projektu i soubor `Makefile` a přeložení projektu příkazem `make`. Z tohoto důvodu jsme vytvořili překladačový systém i pomocí nástroje GNU Make.

V souboru `Makefile` jsou nastavena pravidla pro překlad. Je zde nastaven překladač `gcc` a všechny potřebné parametry pro překlad. Vytvořili jsme pravidla, které nejdříve vytvoří objektové soubory ze všech souborů s příponou `.c` v daném adresáři za pomoci vytvořených automatických pravidel a překladače `gcc`, který dokáže automaticky generovat tyto pravidla, více viz *Dependency Generation* v kapitole *C and C++* knihy [2]. Následně je ze všech objektových souborů vytvořen jeden spustitelný soubor.

Nástroj GNU Make jsme mimo překlad využili i pro automatické zabalení projektu pro odezvdání, generování dokumentace, spouštění automatických testů a čištění dočasných souborů.

3 Speciální algoritmy a datové struktury

Při tvorbě překladače jsme implementovali několik speciálních datových struktur, které jsou v této kapitole představeny.

3.1 Tabulka s rozptýlenými položkami

Implementovali jsme tabulku s rozptýlenými položkami, která slouží jako tabulka symbolů, což byl mimo jiné i požadavek vyplývající ze zadání, který souvisí s předmětem IAL. Tabulku jsme implementovali jako tabulku s explicitním zřetěžením synonym, protože jsme potřebovali, aby počet položek, který do ní můžeme uložit byl dynamický a teoreticky neomezený. Pro zřetězení synonym jsme použili lineárně vážené seznamy. Velikost mapovacího pole jsme dymenzovali tak, aby byla rovna prvočíslu a její naplnění nepřesáhlo 75 %. Jako optimální velikost jsme tedy zvolili číslo 27 457.

Jako mapovací funkci jsme použili GNU ELF Hash použitou v algoritmu, který používá UNIX ELF. Tato funkce je variantou PJW Hash. Použití této mapovací funkce jsme vyhodnotili jako nejvhodnější, více viz [3]. Funkci jsme upravili tak, aby vracela datový typ a hodnotu, které je přípustná pro naše mapovací pole.

Každý položka tabulky obsahuje unikátní klíč v podobě řetězce, kterým je identifikátor funkce či proměnné. Dále obsahuje ukazatel na další prvek (synonymum) a datovou část. Datová část obsahuje informaci o tom, jakého datového typu je daný symbol (v případě funkce se jedná o datový typ návratové hodnoty), pravdivostní hodnoty, které indikují, jestli byl daný symbol již definován a zda je symbol globální. Posledním prvkem datové části, který se používá pouze v případě, že se jedná o funkci, je počet a datové typy formálních parametrů funkce, který je ve formátu řetězce (Pokud např. daná funkce má dva parametry, první typu `Double` a druhý typu `Integer`, uloží se následující řetězec `di.`).

Implementovali jsme i několik potřebných funkcí, které tvoří rozhraní pro práci s tabulkou. Jedná se o následující funkce: inicializace, přidání nové položky, přidání určitého parametru pro konkrétní položku tabulky (v případě, že se jedná o funkci), vyhledání položky, odstranění položky a uvolnění tabulky z paměti. Při implementaci jsme se inspirovali studijní literaturou předmětu IAL [1].

3.2 Dynamický řetězec

Vytvořili jsme strukturu `struct dynamic_string` pro práci s řetězcem dynamické délky, kterou používáme např. pro ukládání řetězce nebo identifikátoru u atributu tokenu v lexikální analýze, protože dopředu nevíme, jak bude tento řetězec dlouhý.

Daná struktura a rozhraní operací nad ní je popsáno v souboru `dynamic_string.h` a operace jsou implementovány v souboru `dynamic_string.c`.

Struktura v sobě uchovává ukazatel na řetězec, velikost řetězce a informaci o tom, kolik paměti je pro řetězec vyhrazeno. Implementovali jsme operace pro inicializaci struktury, uvolnění vyhrazených dat, přidání znaku na konec řetězce, porovnávání řetězců, kopie celých řetězců a další. Při inicializaci struktury se vyhradí paměťový prostor pouze pro určitý počet znaků (určeno konstantou), až při nedostatku vyhrazené paměti se velikost paměti zvýší.

3.3 Zásobník symbolů pro precedenční syntaktickou analýzu

TODO

4 Práce v týmu

4.1 Způsob práce v týmu

Na projektu jsme začali pracovat koncem října. Práci jsme si dělili postupně, tj. neměli jsme od začátku stanovený kompletní plán rozdělení práce. Na dílčích částech projektu pracovali většinou jednotlivci nebo dvojice členů týmu. Nejprve jsme si stanovili strukturu projektu a vytvořili překladový systém.

4.1.1 Verzovací systém

Pro správu souborů projektu jsme používali verzovací systém Git. Jako vzdálený repositář jsme používali GitHub.

Git nám umožnil pracovat na více úkolech na projektu současně v tzv. větvích. Většinu úkolů jsme nejdříve připravili do větve a až po otestování a schválení úprav ostatními členy týmu jsme tyto úpravy začlenili do hlavní vývojové větve.

4.1.2 Komunikace

Komunikace mezi členy týmů probíhala převážně osobně nebo prostřednictvím aplikace Slack, kde jsme si buď psali přímo mezi sebou nebo si vytvářeli různé skupinové konverzace podle tématu.

V průběhu řešení projektu jsme měli i několik osobních setkání, kde jsem probírali a řešili problémy týkající se různých částí projektu.

4.2 Rozdělení práce mezi členy týmu

Práci na projektu jsme si rozdělili rovnoměrně s ohledem na její složitost a časovou náročnost. Každý tedy dostal procentuální hodnocení 25 %. Tabulka 1 shrnuje rozdělení práce v týmu mezi jednotlivými členy.

Člen týmu	Přidělená práce
Dominik Harmim	vedení týmu, organizace práce, dohlížení na provádění práce, konzultace, kontrola, testování, dokumentace, struktura projektu, generování cílového kódu
Vojtěch Hertl	lexikální analýza, testování, dokumentace, prezentace
Timotej Halás	syntaktická analýza, syntaktická analýza výrazů, sémantická analýza, testování, dokumentace
Matej Karas	syntaktická analýza, sémantická analýza, testování, dokumentace

Tabulka 1: Rozdělení práce v týmu mezi jednotlivými členy

5 Závěr

Projekt nás zprvu trochu zaskočil svým rozsahem a složitostí. Postupem času, až jsme získali dostatek znalostí o tvorbě překladačů na přednáškách IFJ, jsme projekt začali řešit.

Náš tým jsme měli sestaven velmi brzy, byly jsme již předem domluveni na komunikačních kanálech, osobních schůzkách a na používání verzovacího systému, tudíž jsme s týmovou prací neměli žádný problém a pracovalo se nám společně velmi dobře.

Na projektu jsme začali pracovat trochu později, takže jsme neměli časovou rezervu, ale nakonec jsme všechno bez problému stihli. Jednotlivé části projektu jsme řešili většinou individuálně za použití znalostí z přednášek nebo materiálů do předmětů IFJ a IAL.

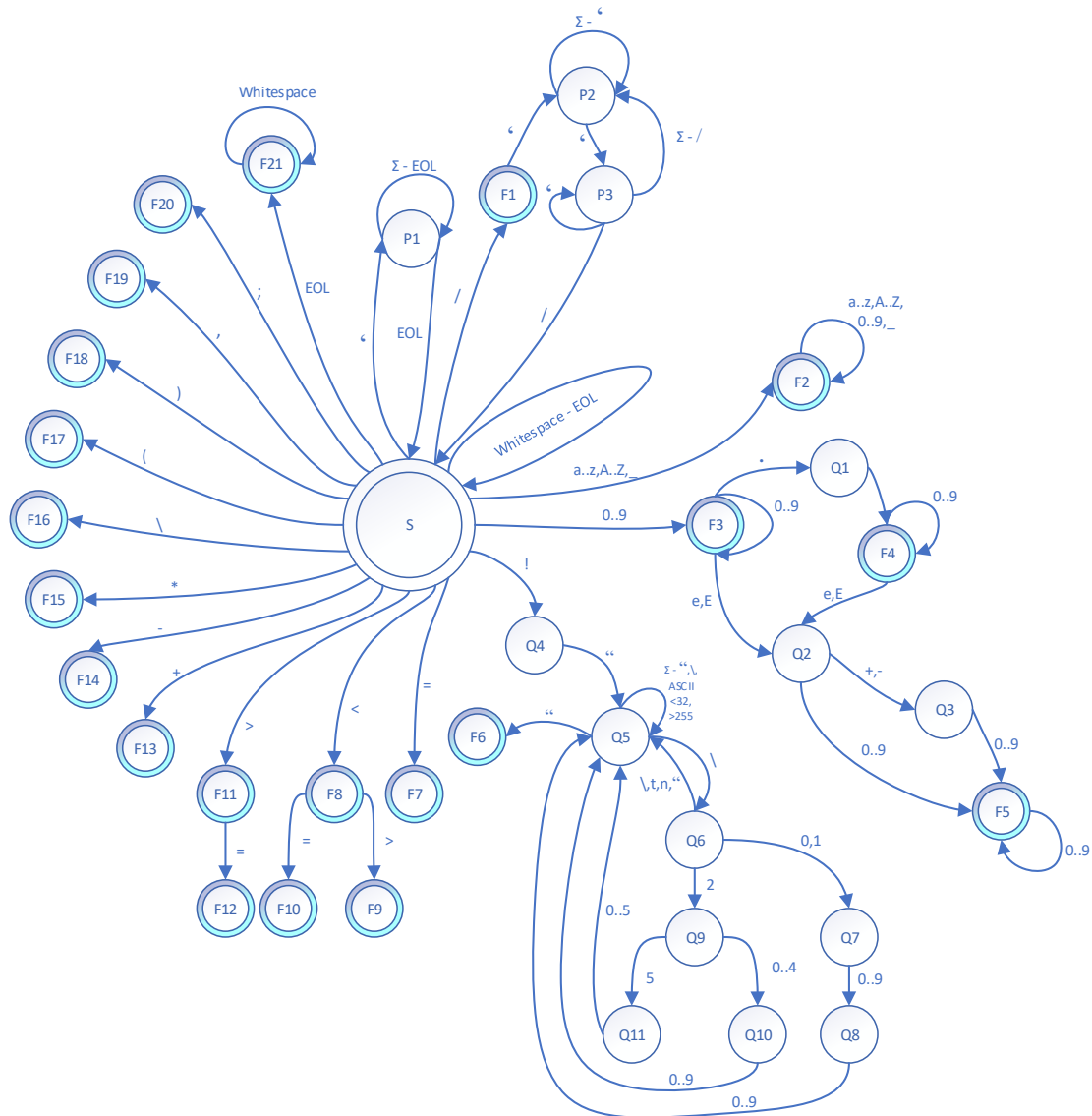
V průběhu vývoje jsme se potýkali s menšími problémy týkajícími se nejasností v zadání, ale tyto jsme vyřešili díky fóru k projektu. Správnost řešení jsme si ověřili automatickými testy a pokusným odevzdáním, díky čemuž jsme byli schopni projekt ještě více odladit.

Tento projekt nám celkově přinesl spoustu znalostí ohledně fungování překladačů, prakticky nám objasnil probíranou látku v předmětech IFJ a IAL a přinesl nám zkušenosti s projekty tohoto rozsahu.

Literatura

- [1] Honzík, J. M.; Hruška, T.; Máčel, M.: *Vybrané kapitoly z programovacích technik*. Brno: VUT v Brně, třetí vydání, 1991, ISBN 80-214-0345-4.
- [2] Mecklenburg, R. W.; Oram, A.: *Managing projects with GNU make*. Cambridge [Mass.]: O'Reilly, třetí vydání, 2005, ISBN 05-9600-610-1.
- [3] Wikipedia: PJW hash function. [online], rev. 5. srpen 2017, [vid. 2017-11-29]. Dostupné z: https://en.wikipedia.org/wiki/PJW_hash_function

A Diagram konečného automatu specifikující lexikální analyzátor



Legenda:

S	START
P1	COMMENTARY
F1	SLASH
P2	BLOCK_COMMENTARY
P3	BLOCK_COMMENTARY_LEAVE
F2	IDENTIFIER_OR_KEYWORD
F3	NUMBER
Q1	NUMBER_POINT
F4	NUMBER_DOUBLE
Q2	NUMBER_EXPONENT
Q3	NUMBER_EXPONENT_SIGN
F5	NUMBER_EXPONENT_FINAL
Q4	STRING_START
Q5	STRING
Q6	STRING_ESCAPE
Q7	STRING_ESCAPE_ZEROONE
Q8	STRING_ESCAPE_ZEROONE_ZERONINE
Q9	STRING_ESCAPE_TWO
Q10	STRING_ESCAPE_TWO_ZEROFOUR
Q11	STRING_ESCAPE_TWO_FIVE

F7	EQUALS
F8	LESS_THAN
F9	NOT_EQUAL
F10	LESS_OR_EQUAL
F11	MORE_THAN
F12	MORE_OR_EQUAL
F13	PLUS
F14	MINUS
F15	TIMES
F16	DIV
F17	LEFT_BRACKET
F18	RIGHT_BRACKET
F19	COMMA
F20	SEMICOLON
F21	EOL

Obrázek 1: Diagram konečného automatu specifikující lexikální analyzátor

B LL – gramatika

TODO

C LL – tabulka

TODO

D Precedenční tabulka

TODO