



Diplomová práce

Návrh jazyka odvozeného z C a implementace nástrojů pro překlad

Studijní program:

B0613A140005 – Informační technologie

Studijní obor:

Aplikovaná informatika

Autor práce:

Maxim Osolotkin

Vedoucí práce:

Ing. Lenka Koskova Třísková Ph.D.

Liberec 2025

Tento list nahradte
originálem zadání.

Prohlášení

Prohlašuji, že svou diplomovou práci jsem vypracoval samostatně jako původní dílo s použitím uvedené literatury a na základě konzultací s vedoucím mé diplomové práce a konzultantem.

Jsem si vědom toho, že na mou diplomovou práci se plně vztahuje zákon č. 121/2000 Sb., o právu autorském, zejména § 60 – školní dílo.

Beru na vědomí, že Technická univerzita v Liberci nezasahuje do mých autorských práv užitím mé diplomové práce pro vnitřní potřebu Technické univerzity v Liberci.

Užiji-li diplomovou práci nebo poskytnu-li licenci k jejímu využití, jsem si vědom povinnosti informovat o této skutečnosti Technickou univerzitu v Liberci; v tomto případě má Technická univerzita v Liberci právo ode mne požadovat úhradu nákladů, které vynaložila na vytvoření díla, až do jejich skutečné výše.

Současně čestně prohlašuji, že text elektronické podoby práce vložený do IS STAG se shoduje s textem tištěné podoby práce.

Beru na vědomí, že má diplomová práce bude zveřejněna Technickou univerzitou v Liberci v souladu s § 47b zákona č. 111/1998 Sb., o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách), ve znění pozdějších předpisů.

Jsem si vědom následků, které podle zákona o vysokých školách mohou vyplývat z porušení tohoto prohlášení.

26. 3. 2025

Maxim Osolotkin

Návrh jazyka odvozeného z C a implementace nástrojů pro překlad

Abstrakt

Klíčová slova: programovací jazyk, překladač

Design of a C-derived language and compiler tools implementation

Abstract

Keywords: programming language, compiler

Poděkování

Obsah

Úvod

Dnes, v době, kdy člověk se spíš zeptá, umí-li to JavaScript, než, běží-li na tom Doom, C je stále C.

I když mně jazyk C imponuje, málokdy jsem se v něm našel dělat projekty. Ve většině případů jsem se uchýlil k používání C++, protože mi chyběly některé triviální moderní vlastnosti, které jsou dnes součástí mnoha jazyků (např. i pouhá namespace). Ovšem programování v C++ bylo vždy spojeno s utrpením. Tak jsem si položil otázku, zda existuje alternativa, jestli je tu něco, co by bylo jako C, ale mělo tu tak potřebnou špetku současnosti.

Odpovědí byl Odin a, popřípadě, Zig, které ve výsledku řešili můj problém, i když z části nepřímou. Ovšem, nebyl jsem v obou případech spokojen s přístupem k syntaxi, která, na rozdíl od C, šla cestou implicitnosti, ve stopách Go. Kdežto pro mě jednou z hlavních imponujících vlastností C byla i explicitní syntaxe, která i dělala dojem jazyka, kde to — co se přečte — se i vykoná.

Protože jsem ve výsledku nebyl úplně spokojen s existujícími řešeními a obecně mi přišlo, že je spíše řešena problematika náhrady C++ se zaměřením na bezpečnost, než na tvorbu jazyka, ve kterém by se dalo příjemně trávit čas při psaní vlastní aplikace, dospěl jsem k myšlence návrhu vlastního jazyka, a v důsledku psaní této práce.

Na úvod se dotknu teorie ohledně kompilátorů a programovacích jazyků. Dále představím možné nástroje/způsoby zlehčující práci s vlastním jazykem. A ve druhé polovině práce se budu věnovat samotnému návrhu jazyka, kde mimo zdůvodnění, proč je něco tak či onak, se budu odkazovat na jiné jazyky a diskutovat jejich přístup. Následně se dotknu i konkrétní implementace kompilátoru.

Nyní vás opustím a předám trpnému rodu.

1 Kompilátor

Kompilátorem nazveme program, který převádí vstupní text do výstupního textu zachovávající význam, kde oba texty jsou zapsané nějakým jazykem. Samotný proces převodu se nazývá kompilací nebo také překladem. V kontextu programovacích jazyků se jedná o převod zdrojového kódu konkrétního programovacího jazyka do jiného programovacího jazyka, nebo přímo strojového kódu.

Existence kompilátoru pro libovolný programovací jazyk je zásadní, protože z podstaty věci finálním cílem je dostat program reprezentující zdrojový kód běžící na nějakém stroji, či v nějakém virtuálním prostředí.

Za cíl se také může klást i návrh jazyka čistě pro zápis programů. Ovšem, pokud neexistuje nástroj pro překlad tohoto zápisu do jazyka, který ve výsledku je schopen být přeložen do spustitelné podoby, onen zápis nemá žádnou technickou relevanci.

Často tedy dochází k případům, kdy pojmy kompilátor a jazyk splynou nebo se zaměňují. Kdy se při použití názvu jazyka implicitně bere i na mysl konkrétní kompilátor, např. Go. Nebo kdy se naopak místo názvu jazyka používá název kompilátoru.

Protože kompilátor je jen program jako každý jiný, může být napsán v jakémkoliv již existujícím programovacím jazyce a zkompileován příslušným libovolným kompilátorem. Dokonce může být napsán v jazyce, který sám kompiluje a přeložen sam sebou. Takovýto jev se nazývá bootstrapping. To vše vede na problém o kuřeti a vejci. Zde ovšem máme jasné řešení, jelikož ve výsledku existuje stroj schopný vykonání nějakého souboru instrukcí. One instrukce vlastně tvoří jazyk, který je spustitelný, a tudíž se dá vnímat jako nejtriviálnější kompilátor pro daný stroj.

1.1 Přechodná reprezentace

Programovací jazyk slouží jako abstrakce semantiky programu a jeho skutečné podoby na konkrétním hardwaru a, po případě, operačním systému. Je zřejmé, že takto lze proložit chtěné množství vrstev abstrakcí před překladem do strojového kódu. Ovšem obecně má smysl jen jedna taková další abstrakce, kdy se jazyk přeloží nejprve do tzv. přechodné reprezentace, IR (intermediate representation), a až ona do kódu konkrétního hardwaru. Smyslem je vytvořit rozhraní pro výrobce hardwaru a jazyku. Část určená pro překlad do IR se označuje jako front-end a část převádějící IR do spustitelného kódu back-end.

Je nutno podotknout, že jak back-end, tak i front-end jsou samostatné celky, které jsou implementované pro problémy/potřeby, které chtějí řešit/naplnit. Proto jejich samotné implementace mohou také obsahovat svoje front-endy a back-endy. A tedy jak výrobce jazyků, tak i hardwaru, nemusí přímo implementovat práci IR, ale třeba využije nějakého rozhraní poskytnutého již existujícím obecným back-endem či front-endem.

Samotná IR může být reprezentována jak rozhraním a objekty či strukturami v programovacím jazyce, nebo přímo jako jazyk, tzv. mezijazyk, IL (intermediate language).

Dále se specifikuje pár ukázek IR s krátkým popisem a ukázkou reprezentace následující C funkce.

```
unsigned add1(unsigned a, unsigned b) {  
    return a+b;  
}
```

LLVM IR Forma, která se dá využívat napříč LLVM nástroji, hlavně tedy pro optimalizaci a kompilaci. Jedná se o jazyk, který je někde na pomezí C a assembleru. Může být jak v normální textové formě, nebo i přímo implementován v paměti.

```
define i32 @add1(i32 %a, i32 %b) {  
entry:  
    %tmp1 = add i32 %a, %b  
    ret i32 %tmp1  
}
```

GCC GIMPLE Jedna z mezi frémy využívaných GCC, která je využívána při optimalizacích. Výrazy převádí do tříadresného formátu.

```
unsigned int add1(unsigned int a, unsigned int b) {  
    unsigned int _tmp;  
    _tmp = a + b;  
    return _tmp;  
}
```

Java bytecode Jedná se o instrukční sadu JVM (Java Virtual Machine). Jméno vyplývá z faktu, že každá instrukce je reprezentována jedním bytem. Bytecode je využíván JVM k JIT (viz []) kompilaci, lze ho tedy spustit, pokud existuje příslušné JVM.

```
.method public static add1(II)I  
.limit stack 2  
.limit locals 2  
iload_0  
iload_1  
iadd  
ireturn  
.end method
```

CIL Zkráceno z Common Intermediate Language. Jedná se o obdobu Java bytecode vyvinutou Microsoftem. Ke spuštění CIL je potřeba platforma, která podporuje nějakou implementaci tzv. Common Language Infrastructure, zkráceně CLI, jako je .NET.

```
.method uint32 add1(uint32 a, uint32 b) cil managed {  
    .maxstack 2  
    ldarg.0  
    ldarg.1  
    add  
    ret  
}
```

C I samotné Céčko může sloužit jako IR, i když nebylo přímo pro to navrženo. Jedná se o jazyk blízký k assembleru a využívaný v některých různých operačních systémech. Existuje pro něj jak velký výběr kompilátorů pro různé platformy, tak i jiných vývojových nástrojů.

1.2 Konstrukce kompilátoru

Samotný překlad se může rozdělit do pár základních kroků. Nejprve je provedena lexikální a syntaktická analýza a čirá sled textu je převedena na abstraktní reprezentaci. Následně je tato reprezentace zvalidována dle příslušných semantických pravidel. Validní reprezentace pak převedena do vybraného IR či přímo do spustitelné podoby. Viz obr[].

Samozřejmě, překlad může obsahovat mnohem více kroků, např. po validaci může následovat optimalizace. Ovšem, vytknuté tři kroky jsou nezbytné pro jakýkoliv překlad.

1.2.1 Lexikální a syntaktická analýza

Cílem je převést zdrojový kód dle gramatiky jazyka do nějaké abstraktní datové struktury v paměti kompilátoru. Taková struktura je často reprezentována stromem, jelikož je to nej přirozenější vyjádření gramatiky, a nese ustálený název AST (abstract syntax tree).

Zde se nabízí zřejmá a 'dobrá' abstrakce lexikální a syntaktické analýzy. Kde modul lexikální analýzy se stará o zpracování vstupního textu a převádí slova na reprezentaci v paměti kompilátoru, která se nazývá token. Syntaktická analýza pak bude již se slovy pracovat jako s abstraktními celky. Lexikální část se často nazývá lexer a syntaktická parser.

Zároveň se smysluplná abstrakce nabízí i mezi samotnou gramatikou a lexémem-parsrem. Možnost vyjádření jazyka za pomoci jistého standardu gramatiky (viz.[]) umožňuje i existenci příslušných nástrojů napomáhajících při generaci AST.

Mezi takové nástroje patří třeba YACC a ANTLR.

YACC Neboli Yet Another Compiler-Compiler. Nástroj umožňující parsing na úrovni gramatiky jazyka (pro bližší představení samotného formátu gramatiky viz. []). Vše probíhá v jakémsi dialektu C, kde se jednotlivým syntaktickým celkům dají přiřadit funkce, jež budou vykonány při jejich rozpoznání, jednotlivým za pomoci asociovaných pravidel. Jako lexer YACC využívá uživatelem definovanou funkci, standardně se využívá nástroj Lex. YACC ve výsledku generuje C kód (hlavně `yyparse` funkci), který se již používá v samotném kompilátoru.

ANTLR Neboli Another Tool for Language Recognition. Jedná se o nástroj umožňující generaci parseru z gramatiky. Kromě generace samotného parseru ANTLR vygeneruje i tzv. procházeče stromu, které umožní aplikaci vykonávat vlastní kód. Základním jazykem, pro který ANTLR generuje parser, je Java, ale umožňuje generaci i do jiných jazyků, jako C#, Python, Go atd.

1.2.2 Validace a linkování AST

Cílem je provést semantickou kontrolu AST. Kontrola se bude lišit od jazyka k jazyku v závislosti na striktnosti jeho pravidel. Může se zde provést ověření existence příslušných deklarací vyskytujících se proměnných v příslušných jmenných prostorech; kontrola datových typů proměnných a výrazů; nalezení vhodné funkce v případě function-overloadingu; a podobně. Kromě validace se zároveň vyskytujícím se symbolům spojí příslušné definice, je-li to třeba z hlediska navrženého AST. Tedy například uzlu reprezentujícímu proměnnou se přiřadí reference na její definici.

1.2.3 Vykreslení

Na konec se AST má vykreslit do finální podoby. Tedy, obecně by pro každý node měla existovat posloupnost instrukcí, které by to činily. Nejprirozenější způsob je existence funkce pro každý typ uzlu AST, kdy by se volala vždy příslušná funkce při procházení stromu. Ovšem, je to ve výsledku jen obyčejný program, takže implementace může být vždy přizpůsobena konkrétnímu problému.

Vykreslení lze rozdělit v závislosti na finálním produktu.

Kod Výsledkem je kód v jiném jazyce, tedy buď v IL, nebo přímo strojový kód. Zde buď kompilátor končí a očekává se, že výsledný kód se přeloží již jiným nástrojem do spustitelné podoby. Může to být i v podobě skriptu, který je pak součástí většího celku, jako je game engine. Nebo se může jednat i o tzv. transkripci, jako v případě TypeScriptu.

Interpretace Za místo získání nějakého výsledného kódu můžeme rovnou každý uzel interpretovat. Tedy, za místo napsání funkce, jejíž výstupem by byl text v jiném jazyce, lze v ní implementovat rovnou chování uzlu, jeho logiku. Takovéto kompilátory se z pravidla označují za interprety.

JIT I když se formálně jedná o první kategorii, tak samotná koncepce je významná a stojí za samostatnou zmínku. JIT stojí za Just In Time. Jedná se o způsob kompilace, kdy je generovaná IR reprezentace, která se pak předá programu

tzv. JIT kompilátoru, který už přeloží IR do konkrétního strojového kódu mašiny, na které běží. Důležitým bla bla bla..

1.3 Cross-Compilation

Někdy je vhodné přeložit program do strojového kódu jiného hardwaru, než na kterém běží kompilátor. Tomuto procesu se říká cross-compilation. Může to být v případech, kdy se program vyvíjí na vysoce výkonném stroji obsahujícím všechny potřebné nástroje pro rychlou a pohodlnou práci, a výsledný software má být určen jinému stroji neobsahujícím takovou infrastrukturu. Příčinou může být operační systém, nebo i samotný hardware stroje.

Také se jedná o případy, kdy se program kompiluje i pro jiné operační systémy, než na kterém je vyvíjen. Například, Doom byl vyvíjen na NeXT počítači s operačním systémem NeXTSTEP, zatímco byl spuštěn pod MSDOS.

Je zřejmé, že o cross-compilaci má smysl mluvit pouze v případě kompilace do strojového kódu. V jiných případech se jedná o kód, který je mezivýsledkem a jeho spuštění závisí na jiném nástroji, který sám musí být přeložen pro odpovídající stroj.

2 Gramatiky

Při návrhu programovacího jazyka hraje důležitou roli samotná syntaxe, která ho z velké části definuje. Syntaxe je totiž jakýmsi rozhraním mezi člověkem a jazykem, obzvlášť v případě programovacích jazyků, kde se v textových editorech či vývojových prostředích běžně různé části syntaxe různě zvýrazňují. Je tedy vhodné mít možnost ji nějakým způsobem formálně popsat, jak z teoretického hlediska, tak i z praktického, kdy definice gramatiky jazyka se může používat v různých nástrojích, např. jak již bylo zmíněno, syntaktické zvýrazňování.

K definici syntaxe jazyka slouží tzv. formální gramatika. Formální gramatiku můžeme definovat následovně:

Definice 2.1. Formální gramatika G je čtveřice (σ, V, S, P) , kde:

- σ je konečná neprázdná množina terminálních symbolů, tzv. terminálů.
- V je konečná neprázdná množina neterminálních symbolů, tzv. neterminálů.
- S je počáteční neterminál.
- P je konečná množina pravidel.

Terminály jsou dále nedělitelné symboly jazyka. Jsou to například klíčová slova nebo jednotlivá písmena sloužící pro definici proměnných. Neterminaly jsou pak jakési proměnné, které se dále dělí na další terminály nebo neterminaly. Neterminál může například představovat binární operátor, který pak bude definován jako množina již terminálních symbolů představujících jednotlivé binární operátory. Prázdný symbol se označuje jako ϵ .

Obecně pravidlo gramatiky můžeme vyjádřit jako zobrazení: ¹

$$\alpha \rightarrow \beta, \text{ kde } \alpha \in (\Sigma \cup V)^* V (\Sigma \cup V)^*, \text{ a } \beta \in (\Sigma \cup V)^*$$

Tedy vzorem je posloupnost terminálů a neterminálů obsahující alespoň jeden neterminál. Obrazem pak je libovolná posloupnost terminálů a neterminálů.

Gramatiky lze členit na základě striktnosti pravidel dle tzv. Chomského hierarchie.

Definice 2.2. Necht $G = (\sigma, V, S, P)$ je gramatika, pak:

¹Hvězdíčka (*) představuje symbol libovolného opakování výrazu, a to i žádného.

- G je gramatika typu 0 nebo take neomezená gramatika právě tehdy, když ...
- G je typu 1 nebo take kontextová gramatika právě tehdy, kde pro každé pravidlo $\alpha \rightarrow \beta$ z P platí $|\beta| \geq |\alpha|$ a zároveň pravidlo $S \rightarrow \epsilon$ se nevyskytuje na pravé straně.
- S je typu 2 nebo take bezkontextová gramatika právě tehdy, když pro každé pravidlo $\alpha \rightarrow \beta$ z P platí $|\alpha| = 1$. Neboli, že α je pouze neterminal.
- P je typu 3 nebo take regulární gramatika právě tehdy, když každé pravidlo z P je v jedné z forem:

$$A \rightarrow cB, A \rightarrow c, A \rightarrow \epsilon,$$

kde A, B jsou libovolné neterminaly a c je terminal.

Z hlediska programovacích jazyků prakticky se lze omezit na gramatiky bezkontextové.

2.1 Bezkontextová gramatika

Bezkontextovou gramatiku, kromě definice uvedené výše, lze také definovat z hlediska samotných pravidel, což bude názornější pro navazující text.

Definice 2.3. Gramatika je bezkontextová právě tehdy, když pro každé pravidlo z P platí buď

$$S \rightarrow \epsilon,$$

nebo

$$\alpha A \beta \rightarrow \alpha \gamma \beta,$$

kde

$$A \in N, \alpha, \beta \in (N \cup \Sigma \setminus \{S\})^* \text{ a } \gamma \in (N \cup \Sigma \setminus \{S\})^+$$

Například, pravidlo pro sestavení goto výrazu v jazyce C může být vyjádřeno ve volné formě třeba následovně

$$\text{goto} \rightarrow \text{'goto' identifier ';'}$$

Pravidlo definuje nonterminal `goto` jako možné slovo počínající goto, čířym textem následujícím nonterminalem `identifier`, který je definován v nějakém jiném pravidle, představujícím identifikátor. To vše je zakončeno terminálem představujícím symbol středníku.

Definuje-li se pak třeba `identifier` za pomoci regulárního výrazu následovně

$$\text{identifire} \rightarrow [\text{a-zA-Z}]^+$$

`goto` pravidlo bude třeba generovat slova jako

$$\text{goto FooLabel ;}$$


```
goto Me ;  
goto UnhandledException ;
```

Je zřejmé, že zápis pravidel může být různorodý. Pro sjednocení zápisu existují různé standardy. Může se vytknout několik relevantních a stručně předvést na příkladu s `goto` příkazem.

BNF zkraceno od Backus–Naur forma.

```
<goto-stmt> ::= "goto" <identifier> ";"  
<identifier> ::= <letter> <letters>  
<letters> ::= <letter> <letters> | \epsilon  
<letter> ::= "a" | "b" | ... | "Z"
```

EBNF rozěříená (Extended) Backus-Naur forma. Existuje několik verzí a má ISO Standard.

```
goto-stmt = "goto", identifier, ";"  
identifier = letter, { letter }  
letter = "a".."z" | "A".."Z";
```

YACC notace bližší seznámení s YACC'em může být naleznuto zde [\[\]](#).

```
goto_stmt : KW_GOTO identifier ';' { .. };
```

Složené závorky představují místo, kam se umísťuje C kód, který se má provést při parsingu oných elementů. Neterminaly z příslušných pravidel jsou přístupné za použití symbolu \$. Například \$ označuje proměnnou samotného pravidla, \$1 proměnnou prvního terminálu či neterminálu (KW_GOTO v případě pravidla `goto`), \$2 respektive druhého atd.

Definice `identifier` je pak součástí jiného programu zvaného Lex, na výstup kterého YACC spoléhá.

```
identifier : [A-Za-z]+
```

ANTLR notace bližší seznámení s ANTLR'em může být naleznuto zde [\[\]](#).

```
goto_stmt : 'goto' identifier ';' ;  
identifier : [a-zA-Z]+ ;
```

3 Vývojové nástroje

Po mimo samotného překladače se k práci s programovacím jazykem běžně využívají různé nástroje usnadňující práci.

Základem je samotný textový editor, bez kterého by nebylo možné samotný kód v celku psát. Samotné editory pak, většinou za pomoci pluginů, umožňují přidávat podporu různých jazyků. Mezi takové populární editory patří například VS Code nebo Vim/Nvim. Lze jich tedy využít jako platformu pro tvorbu jakéhosi IDE pro vlastní jazyk. Tento text se omezí pouze na VS Code a Nvim.

Pluginy, nebo také Extensions, se ve VS Code dají standardně psát za pomoci TypeScriptu či JavaScriptu. Jako v prohlížečích, je zde i možnost využití WebAssembly. Lze tedy využít i jiný jazyk, který by šel do WebAssembly zkompileovat, jako třeba Rust nebo C++. Ke komunikaci s editorem je zde VS Code API, které umožňuje přístup k elementům uživatelského rozhraní editoru, poslouchání různých eventů, přístup k debuggeru atd. Všechny pluginy se pak dají nahrát do jednotného oficiálního marketplace, kde budou dostupné uživatelům a umožní automatické aktualizace.

V případě Nvim je zde kromě klasických možností využití Vimscriptu, jak v případě Vim, dostupná možnost skriptování za pomoci integrovaného Lua script engine. Celé Vim API je pak dostupné skrz Lua. Lze tedy přímo přistupovat k bufferům a měnit rozhraní celého editoru. Pluginy jsou ve své podstatě jen zdrojové kódy, které jsou načtené v konfigu. Většinou za pomoci nějakého packer-manageru, který umožní načtení složky s pluginem jedním řádkem, a to i třeba z git repozitáře. Klasickým způsobem distribuce pluginů je pak git repozitář se samotným kódem pluginu, link na který uživatel předá packer-manageru. Takto uživatel bude moci i stáhnout updaty, jestli bude chtít.

3.1 Zvýraznění kódu

Za základní potřebnou vlastnost se může klást zvýraznění kódu, která je prakticky zřejmostí.

K definici vlastního zvýrazňování se v případě VS Code využívá TextMate, který umožňuje definovat vlastní gramatiku v JSON souboru za využití regulárních výrazů. V případě Vim se používá vlastní formát, který také umožňuje využití regex.

Oba tyto přístupy využívají tak či onak prohledávání a parsování zdrojového kódu

pro zvýraznění. Existuje však i jiný přístup, který je v praxi rychlejší a přesnější, a to za využití LSP. Většinou totiž i tak máme aktivní LSP, které nám zajišťuje např. doplňování slov, a tudíž už máme informaci o všech symbolech a jejich roli v jazyce. Oba vybrané editory mají vestavěnou podporu LSP.

3.1.1 Dokumentace

Po mimo zvýraznění kódu v editorech je někdy třeba mít možnost zvýraznění kódu ve statických dokumentech. Například jako dokumentace, která je nezbytná pro popis semantiky jazyka uživateli.

V takovémto případě lze využít například nástroje Shiki. Jedná se o JavaScript knihovnu, která využívá TextMate gramatiky k generaci zvýrazněného výstupu. V základu Shiki umí generovat výstup jako HTML. Klasické užití je pak napsání drobného skriptu v NodeJS, který by procházel HTML dokument a nahrazoval vybrané elementy, např. code, výstupem z Shiki. Pak výsledný HTML dokument neobsahuje žádný JS run-time kód. Obdobným způsobem je generována dokumentace obsažená v příloze.

V případě Vim syntaxe je zde možnost využití jeho samotného ke generování HTML z kódu. Je zde opět nutnost napsání nějakého skriptu, který by automaticky procházel HTML soubor a přepisoval ho.

```
vim -c 'syntax_on' -c 'T0html' -c 'wq' myfile.html
```

Bohužel, zde nejsou výrazné nástroje, které by umožnily využití Vim syntaxe pro generování zvýrazněného výstupu, jako v případě TextMate gramatiky. Pro využití v HTML dokumentech je zde jen opce využití vim.js, tedy portu Vim pro prohlížeče, který by v run-time mohl zvýrazňovat kód. Ovšem využití tohoto řešení jen pro zvýraznění kódu je zbytečně náročné.

3.2 Language server protocol

Language server protocol, zkráceně LSP. Jedná se o protokol využívaný pro komunikaci language serveru a klienta, kterým je třeba IDE nebo textový editor, kde language server předává informaci klientovi o textu z pohledu jazyka. Smyslem je nabídnout rozhraní mezi programem nabízejícím syntaktickou a semantickou informaci o kódu a vývojovým nástrojem.

V základu k implementaci lze použít část kódu ze samotné implementace kompilátoru, či dokonce celé moduly. Protože práce serveru je od části shodná až do fáze vykreslení. Ovšem, v případě kompilátoru je možnost ukončení kompilace při první chybě. V případě LSP by měl program umět kompilovat i neúplně správný syntaktický kód, a to i semanticky, a dávat výsledky o tom, co se podařilo převést do AST. Navíc, LSP by měl fungovat v reálném čase obnovujíc informaci o kódu po každém inputu uživatele. Tvorba LSP tedy není triviálním úkolem i za podmínky existence

kompilátoru, jelikož jak kompilátor, tak i LSP by měly fungovat co nejrychleji, ovšem jejich potřeby se protirečí.

bla bla bla

3.3 Debugger

Finalním nástrojem při tvorbě programů je debugger. Zde opět lze využít vybraných textových editorů jako platformy. Ovšem psaní vlastního debuggeru není zcela žádoucí, jelikož je to další aplikace, která se bude muset s jazykem vyvíjet a udržovat. Je výhodnější využít již existujících řešení skrze nějaký dostupný interface.

3.3.1 Debug informace

Debug informace je veškerá informace, která není obsažená ve spustitelném souboru, ale je napomocná debuggeru k propojení zdrojového kódu a konečných instrukcí. Debugger pak může umět například krokovat zkompilovaný program ve zdrojovém kódu, zobrazovat hodnoty proměnných atd.

Uvažujeme-li jazyk, který se bude kompilovat do strojového kódu, tak stačí mít program jako spustitelný soubor a k němu vygenerovanou debug informaci. První problém řeší samotný kompilátor, a tedy zbývá vyřešit otázku generace debug informace.

V zásadě existují dva hlavní formáty využívané moderními debuggery, a to PDB a DWARF.

PDB zkraceno z Program Database. Jedná se o soubory převážně využívané Microsoftem, například pro debugování ve Visual Studio. PDB vnitřně, pro definici samotných debug symbolů, využívá formátu CodeView. V rámci Windowsu existuje API, které umožňuje získání informací z PDB souboru bez znalosti formátu.

DWARF zkraceno z Debugging With Arbitrary Record Formats. Formát využívaný například GDB a LLDB. Převažně pro programy na Linux a macOS. Často se používá v rámci ELF souborů. Je standardně vestavěn do spustitelného souboru.

Samou přímočarou možností je vlastnoruční generace těchto souborů. Naštěstí některé backendy umožňují generaci oněch symbolů.

V případě LLVM IR lze třeba definovat podrobnější informace o původním kódu za pomoci maker `#dbg_value`, `#dbg_declare` a `#dbg_assign`. Může to vypadat následovně

```
%i.addr = alloca i32, align 4
#dbg_declare(ptr %i.addr, !1, !DIExpression(), !2)
; ...
!1 = !DILocalVariable(name: "i", ...) ; int i
!2 = !DILocation(...)
```

Kde první řádek představuje deklaraci proměnné `i` typu `int32_t`. Následující pak přidává oně deklaraci metadata a na dalších řádcích se některá metadata specifikují. Lze to použít jak pro generaci PDB, tak i DWARF.

Nebo, například při použití C jako IL, lze využít vybraného kompilátoru umožňujícího generaci potřebného formátu. Pro mapování zdrojového kódu na C kód pak lze využít direktivy `#line`, která umožňuje specifikovat číslo řádku a název souboru. Direktiva však funguje jen na bezprostředně následující řádek kódu, což lze řešit generací kódu obecně bez nových řádků a přidáváním je vždy s použitím oně direktivy.

3.4 Řešení

Pokud je možné generovat debug informaci se spustitelným souborem, lze využít libovolného již existujícího debuggeru podporujícího formát oně informace.

Protože VS-Code má standardně implementované rozhraní pro debuggery, lze vytvořit vlastní konfiguraci pro již existující debugger plugin, kde se zamění příkaz kompilace. A po případě se to dá zabalit i do samostatné extension.

Nvim nemá standardní interface pro debugger, takže v případě každého debugger pluginu je konfigurace individuální, jestli je vůbec v konkrétním případě dostupná. Vždy ale lze udělat fork ...

4 Návrh jazyka

Prvně bych vytvořil nějakou představu o vizi jazyka. Protože v zásadě cílem je přijít s moderní obdobou C, tak by bylo k věci jim i začít.

V jazyce C se mi v zásadě líbí jeho jednoduchost, co se v zásadě napíše, to i platí, neprovádí se v pozadí tučt instrukcí (no hidden control flow); explicitní syntaxe, která se dá číst; možnost přímé a volné práce s pamětí, která je ve většině jazyku potlačena kvůli bezpečnosti.

Tedy, chtěl bych proceduralní systemový general purpose jazyk umožňující robustní explicitní compile-time vyhodnocení, který by umožňovat jednoduchou a neomezenou manipulaci s pamětí, byl by čitelný sam o sobě i na ukor osvědčeným postupem. Interpretace kodu po přečtení by měla co nejvíce odpovídat skutečnosti. Tedy, například, deklarace proměnné by neměla být ve vychozím případě konstatní, protože po přečtení kodu, který nespecifikuje vlastností deklarce je přirozenější se domnívat, že žádných vlastností nenabyva, než že jsou nějaké standardní skryté. Syntaxe by měla být srozumitelná intuitivně i bez znalostí gramatiky. Tedy, v zásadě, akce spíš interpretovaný slovy než symboly.

Jazyk by neměl přímo plnit nějaký cíl, nebo být určen speciální sféře. Jedná se o tvorbu nástroje pro moje vlastní užití, jazyk, ve kterém bych mohl psát vlastní projekty.

Existují jazyky v tom či jiném smyslu řešící muj problem. Některé z mého hlediska jsou vskutku skvělými a zajímavými variantami. Ovšem, nenašel jsem se v nich na tolik, abych pak něco nevyčítal. Přejde mi tedy vhodným je představit a vytknout věci, které jsou odlišné od moji idelaní představy.

C++ Pesprostřední nasledník C, který je i často s ním nerozlučně poejn jako C/C++. C++ ponechává v zakladu C s drobnými změnami a staví na tom v zakladu za pomoci standardní knihovny. C++ tedy dědí i špatné vlastností C, jako jsou makra a system importu, a nové vlastností často nejsou částí jazyka, ale objekty právě std knihovny, jako třeba i pole.

Abstrahujic se od neřešené C částí, jazyk nabízí mn

D

Zig

Odin

Dále bych se prošel po zásadním celku a předvedl vlastní

4.1 Vstupní bod programu

Obvykle vstupním bodem programu ve vyšším programovacím jazyce je nějaká tzv. `main` funkce. Taková funkce může mít za úkol předání vstupních argumentů programu a oddělení globálního scope.

Samotný koncept mi přijde obskurtním:

- Prvně, `main` funkce zvyšuje indenci kódu a zesložituje strukturu programu bez možnosti se tomu vyhnout. Když, například, uživatel bude chtít začít v lokálním scope, protože je to to, co se mu líbí na `main` funkci, tak to může udělat přímo za pomoci odpovídajícího syntaktického konstruktů. Dokonce, když to udělá, tak jasně dá čtenáři znát svou myšlenku.
- A za druhý, ruší chápání pořadí vykonání instrukcí. Instrukce se totiž běžně mohou objevit i v globálním scope. Ovšem, protože z `main` funkce nelze nijak skočit do globálního scope, ale stále se jedná o místo, kde by měl program začít své konání, tak není jasné jak, a jestli vůbec, se provedou globální instrukce.

Jako vstupní bod programu jsem tedy zvolil počátek souboru, obdobně jako v Pythonu, nebo, když mám vybírat z C-like jazyku, jak v HolyC.

4.1.1 Alokace

Dynamickou alokaci bych nevnímal jako funkci, operator, nebo výraz, ale jako samostatný celek, který by sloužil alternativou při přiřazení. Tedy, že by přiřazení buď bylo alokací, nebo výřezem. Smyslem je vždy zaručit, že dynamicky alocovaná paměť bude vždy přiřazena proměnné.

To sice nevyřeší ...

Syntaxi bych volil následující

```
int^ ptr = alloc 8; // alokuje 8 bytu
int^ ptr = alloc int[8]; // alokuje 8 * sizeof int
```

Navíc bych umožnil při deklaraci vynacházet datový typ na pravé straně, jestli má být schodný s tím na levé. Formálně je to možné, protože `alloc` je alternativní `rvalue`, oproti výrazu `new` v C++ či D, který by měl splňovat pravidla výrazu.

Následující řádky by tedy vyjadřovali to samé.

```
int^ ptr = alloc int[8];
int^ ptr = alloc [8];
```

4.1.2 Komentáře

V C, nebo například i v C++ a D nelze vnořovat blokové komentáře `/**/`. Není to vyznáčným nedostatek, ale stále nepříjemný, pokud se komentuje něco, co už obsahuje komentář. Navíc se je to nekonzistentní s řádkovými komentáři, které se mohou vnořovat.

Volil bych následující syntaxi.

```
\{  
    \{ comment \}  
\}
```

4.1.3 Datové typy

4.2 Pole

Jedna se o nejzákladnější a nejužitečnější datovou strukturu, která se v programování vyskytuje. V C se ovšem pole dají používat jen ve statických případech, kdy velikost lze stanovit ještě při kompilaci. Ovšem, i v případech, kde je to dostačující, tak při snaze využít pole jako argument funkce, tak buď se ona funkce musí definovat pro konkrétní velikost pole, nebo využít ukazatele.

První případ je použitelný jen zřídka, jelikož nepřináší abstrakci, která se intuitivně pojí s vodbou datového typu, muselá by se vytvořit pro různé velikosti poli samostatná funkce. To se řeší předáním pole přes pointer a, po případě, ukončením pole nějakým specifickým symbolem, nebo předáním doplňujícího parametru délky.

V takovém to případě se však ztrácí jakákoliv výhoda vybraného datového typu, a vlastně i konceptuální smysl onoho. Kod je ve výsledku méně explicitní, a navíc více nadolný na chyby, jelikož compile-time informaci, která se pojí jen k jedné proměnné, rozvadíme do dvou run-time.

Stanovil bych tedy některé základní požadavky pro pole. Mělo by být využitelné ve funkcích bez ztráty identity a při tom být implicitním ukazatelem na počátek svých dat při přiřazení do pointeru. Navíc, rozšířit typ i na dynamické pole a dynamické pole variabilní délky.

4.2.1 Délka pole

Následně bych zavedl získatelnou délku pole.

```
int[8] arr;  
arr.length; // vrati delku pole, tedy 8
```

Při předání pole do funkce by se tedy třeba předal ukazatel na data a jako skrytý parametr velikost pole. Pro případy, kdy není potřeba předávat velikost, se může použít pointer a implicitní přetypování.

4.2.2 Typy poli

Po mimo klasického rozdělení poli na statická a dynamická, bych chtěl umožnit jejich dělení v závislosti na variabilnosti délky. To by umožnilo vytvářet více specifické rozhraní pro využití poli ve funkcích. Například by `int[const]` by vyznačovalo konstantní délku.

Navíc bych chtěl integrovat array list do jazyka v rámci poli, jelikož je to často využívaná struktura. Array list bych viděl jako automaticky rozšiřovatelné pole tak, aby vždy šlo zapsat na zvolený index.

Pole konstantní compile-time známé délky

Jedna se o pole analogické tomu, co je v C. Tedy

```
int[8] arr;
```

By vytvořilo pole o delce 8. Spočtená délka by se vždy dosazovala compile-time, realná proměnná by se pro její uchování negenerovala.

Pole konstantní run-time známé délky

Jedna se o analogii alokace konstantního ukazatele v C, který by byl využíván jako pole.

```
int* const arr = malloc(sizeof(int) * 8);
```

Tedy

```
int[const] arr = alloc int[8];
```

By alokovalo pole o delce 8 na heapu a vygenerovalo by příslušnou proměnnou pro uložení délky někde v paměti programu. Pole by, samozřejmě, nešlo realokovat, jelikož délka pole je obecně run-time známa, a tedy není možnosti ověřit při kompilaci její neměnnost.

Pole variabilní run-time známé délky

Analogie využití ukazatele jako pole v C.

```
int* arr = malloc(sizeof(int) * 8);
```

Tedy

```
int[dynamic] arr = alloc int[8];
```

By alokovalo pole o delce 8 na heapu a vygenerovalo by příslušnou proměnnou pro uložení délky někde v paměti programu. Pole by šlo realokovat.

Array List

Šel by vytvořit následovně

```
int[] arr;
```

nebo se specifickou počáteční delkou, v tomto případě 8.

```
int[] arr = alloc int[8];
```

Volba kvalifikátoru

Lze také vnímat dynamické pole za výchozí, a ne array list. Pak by se využil kvalifikátor při inicializaci arraylistu, pole variabilní run-time známé délky by se inicializovalo bez kvalifikátoru. Takovým kvalifikátorem by mohl být třeba `auton` od slova `autonomus`.

Taková to varianta se protířečí s základní představou o jazyce viz. Ale na druhou stranu se zbavuje skrytého flow, kdy se akce, která nese v sobě implicitní alokace není jako výchozí, ale musí se konkrétně zvolit.

Příklady jiných jazyků

C++ ponechává vše jak v C. Řeší vše za pomoci kontejneru definovaných v std knihovně. Nema však analog dynamického pole, ale nabízí `std::span`, který může sloužit jako interface pro souvislý kontejner, nemůže však sloužit přímo jako kontejner.

```
std::array<int, 8> arr;  
std::vector<> arr;
```

D obohacuje jak statické, tak i dynamické pole o délku.

```
int[8] arr;  
int[] arr = new int[8];  
// v obou případech délka dostupná jako  
arr.length;
```

Odin má k dispozici statická pole, slice, které se využívají obdobně jako `std::span`, ale umožňují alokaci, a tzv dynamická pole, které je analogem array listu pro jehož deklaraci se využívá kvalifikátoru na místě délky. Všechny mají dostupnou funkci `len` vracející jejich délku a dynamické pole může využít funkci `cap` k získání reálné alokované délky.

```
arr: [5]int // static  
arr: make([]int, 8) // slice  
arr: [dynamic]int  
len(arr);  
cap(arr);
```

4.2.3 Práce s polem

Protože cílem je vytknout identitu pole oproti pointeru, tak bych zduraznil rozdíl mezi nimi i z pohledu práce s daty. Pole je na rozdíl od pointeru kus paměti obsahující stejně elementy za sebou. Tedy bych pole vnímal jako proměnnou definující stejné prvky. Akce prováděné se samotnou proměnnou jako takovou (bez indexace), by se týkaly všech prvků v poli.

I když C fakticky rozděluje mezi ukazatelem a polem, tak prakticky struggles identity. Při následné deklaraci pole

```
int arr[8];
```

se vytvoří kontejner pro osm proměnných: `arr[0] .. arr[7]` o datovém typu `int`, který se specifikuje jen jednou. Použije-li se kvalifikátor, tak se taky aplikuje na všechny prvky.

```
const int arr[8];
```

Aritmetické operace se však chovají k proměnné jako k pointeru a nejsou aplikované na všechny prvky.

```
int* ptr = arr + 1;
```

V důsledku dvojakosti následující přiřazení není dovoleno,

```
int arr1[8];
int arr2[8];
arr1 = arr2;
```

jelikož pak není jasné, zda se jedná o přiřazení všech prvků `arr2` do `arr1`, nebo přepsání ukazatele `arr1` na ukazatel `arr2`. Což není z přetčení kodu zřejmé, jelikož standardně se využívá kvalifikátoru `const` k zakázání přiřazení do proměnné. Navíc, syntaxe řešící onen problém již v jazyce je, což přináší ještě špetku zmatku.

```
const int* const ptr;
```

Jasně rozdělení mezi ukazatelem a polem na úrovni typu umožňuje se vyhnout podobným problémům. Navrhoval bych, že kvalifikatory jsou vždy vztaženy ke všem proměnným v poli, že přiřazení pole do pole je definováno jako přiřazení jednotlivých prvků pole, že libovolné operace jsou vztaženy vždy na všechny prvky pole. Vyjimkou by bylo přiřazení pole do ukazatele, kde se jedná o cast, a operace concatínace viz[].

Tedy, například by pak bylo možné inicializovat všechny prvky pole na 0 následně.

```
int[8] arr1 = 0;
```

Nebo sečíst dva pole

```
int[3] arr1 = [ 1, 2, 3 ];
int[3] arr2 = arr1 + arr1; // [ 2, 4, 6]
```

atd.

4.3 String

V C string literaly jsou pouze hezčí verzí zapsání pole konstantních znaků. Vcelku, i když je to primitivní, tak zcela postačující. Problemem je zde 'absence' identity pole jako typu, jak již bylo zmíněno viz[]. Tedy vlastně se s každým stringem pracuje jako s pointerem.

Jelikož já definuji pole jinak, tak lze rozvinout možnosti pole, aby umožňovali ve výsledku lehčí práci i se stringy. Samotný typ pro string existovat nebude, ale bude jen podpora string literalu, který se při kompilaci rozloží na pole.

4.3.1 UTF-8

Bylo by vhodné rozšířit podporu literalu z ASCII na jiné kodování, které by umožnilo jednoduchou manipulaci se složitějšími symboly. Jako takové kodování bych volil utf8, protože je kompatibilní s ascii, blokem je byte, tedy není závislé na edianech a je velmi rozšířené.

Protože symboly v utf8 jsou variabilní délky, viděl bych jako nejlepší možnost compile-time vyhodnocení největší délky potřebné pro uložení jednoho symbolu příslušného literalu a následnou konverzi na pole intů o patřičné velikosti, kde každý element bude samostatným symbolem zakódovaným v utf8.

```
int16 str = "čau";  
str[0]; \\ 'č'  
str[1]; \\ 'a'  
str[1]; \\ 'u'
```

To umožní pracovat se symboly samostatně, využívat všechny výhody pole a mít pro ASCII text stejně velké pole jako v C. Tedy, například, levá strana může definovat libovolnou velikost k uložení symbolu.

```
int32 str = "čau";
```

Aby bylo umožněno pracovat se string literaly jako prostě s kusem paměti, tak bych přidal možnost definovat tzv. raw string literal následujícím způsobem.

```
int^ str = "Hello"R;
```

Jiné jazyky

V D jsou string literaly standardně v utf8 formátu jako immutable pole znaku, ale za pomoci postfixu u string literalu lze je konvertovat do wchar(utf16), či dchar(utf32).

```
"hello"w;  
r"ab\n" // Wysiwyg ("what you see is what you get") quoted strings can be d
```

V Odinu jsou stringy taky utf8, má ale koncepci tzv. rune umožňující operovat v takových stringech po samostatných symbolech, kde právě rune represents a Unicode code point. Má také speciální datový typ reprezentující null-terminated stringy s C, cstring.

```
str := "čau"  
for r in str { fmt.print(r, '.') }  
// vypíše "č .a .u ."
```

V Zig string literal je faktický jen null-terminated pole bytu, které lze zapsat pomocí utf8 symbolu. Další rozšíření práce Prace je skrz standardní knihovnu.

```
const arr = "čau";  
print("{d}\n", .{arr[0]}); // první byte 'č': 196  
print("{d}\n", .{arr[2]}); // byte reprezentující a: 97
```

4.3.2 Operace

Jako jediné konvenční operace nad stringy které by se měly integrovat do syntaxe, bych viděl zřetězení a slice. Ostatní operace by už měly být obsažené v standardní knihovně.

Zřetězení

Níc nového bych nevymyslel, a použil operator `..` jako v Lua.

```
u8[] str1 = "Hello";
u8[] str1 = "World";
u8[] str3 = str1 .. " " .. str2;
```

Jelikož delká libovolného pole je získatelná, lze to zobecnit na jaký koliv typ pole. Je však důležité, aby implementace neobsahovala žádnou alokaci paměti, bylo by to zavádějící. Tedy případné výsledné pole by se muselo samostatně standardně alokovat prostředky jazyka.

Pro dynamické pole by to mohlo vypadadat následně

```
u8[const] arrC3 = alloc [] : arrA .. arrB;
```

Slice

Se tzv. slice se dá setkat v mnoha jazycích v té nebo jiné podobě. Jedná se o reprezentaci jen libovolné souvislé části pole, která sama o sobě neobsahuje data, ale odkazuje se na původní pole. Tedy v C by se slice dal definovat třeba následujícím způsobem

```
struct Slice {
    int* dataPtr;
    int len;
};
```

kde `dataPtr` by ukazoval na nějaký element v původním poli, a `len` by specifikovalo délku. Odin a Zig například implementují slice právě jako pointer na data a délku.

Odin, Zig a D například vnímají slice jako datové typy. Využívají je jako nějaký interface pro dynamická pole.

Já bych nevnímal slice jako datový typ, ale spíš jako operaci, protože ve své podstatě je to to same. ...

4.4 Namespace

Zručný nástroj k organizaci kódu. Umožňuje zhlukovat proměnné pod jedním společným názvem, který je rozlišitelný parsrem. Na rozdíl od použití identifikujících prefixu / postfixu v názvech je strukturním celkem z hlediska nástrojů operujících s kódem (např LSP).

Umožňuje také při kompilaci hromadně pracovat s vevnitř definovanými proměnnými, a tedy se dá dobře využívat i pro import a export částí kódů. Například v Python

```
import foo;
import from foo x;
```

To, po mimo jiné také umožňuje řešit kolizi názvu při importování knihoven (pokud nejste C++).

Namespace bych viděl jednoduše jako pojmenovaný scope.

```
namespace Foo {
    int x;
}
```

Nebo možná prostě

```
Foo {
    int x;
}
```

K přístupu prvku bych využil se syntaxe z C++ `Foo::x;`.

4.5 Import

Klasificoval bych hlavičkové soubory a s nimi související systém importu jako nejhorší část C. Hlavní nevýhodou kterých je duplicita definic. Slouží však k dobrému úmyslů, k izolaci implementace a definici rozhraní. Cílem tedy bude tuto myšlenku ponechat, a vyhnout se jak preprocesoru, tak i duplicitě.

Základním celkem bude soubor, jelikož jeto to co se ve výsledku předá překladači. Překladač dostane jen jeden vstupní soubor, který následně již za pomoci prostředku jazyka umožní načíst další soubory. Všechny importy však budou probíhat v rámci AST, každý soubor by tedy měl být samostatně parsovatelným celkem.

Intitivně se nabízí možnost přímého importu souboru následující syntaxi.

```
import filename;
```

Ověšem, této možností bych se vzdal. Přejde mi, že by jen vybízelo k "nesprávnému" přístupu při importování viz [namespaces], a nepřenašelo nic, co by nešlo řešit jinak.

Zavedl bych tedy variantu, která by vždy zabalovala soubor ze strany uživatele importu.

```
import filename as namespace Foo;
```

To by vytvořilo namespace `Foo` a překopirovalo kořen rozparsovaného souboru `filename` do něj.

Syntakticky se specialně specifikuje namespace, protože by onen konstrukt mohl být využit k implementaci jiných způsobu zabalení souboru.

Např.

```
import filename as scope;  
import filename as fcn foo;
```

apod.

Tento konstrukt lze rozšířit a zavést import jen vybraných symbolů ze souboru.

```
import from filename foo, boo as namespace Foo;
```

V zásadě tohle umožní robustní import, a více prostředků není potřeba. Zbývá zohlednit viditelnost jednotlivých identifikátorů.

Lze buď vycházet z toho, že vše je viditelné, a my omezujeme viditelnost, nebo naopak, vše je nepřístupné, a my rozšiřujeme přístup. Druhý přístup je víc praktický, ale ztrácí na explicitnosti, protože, když importujeme soubor, tak intuitivně očekáváme, že se nám tam naimportuje všechno (nebo alespoň něco), než nic.

Podstatnější je otázka viditelnosti vnořených importů. Tedy, importuje-li soubor identifikátory z jiného souboru, budou-li viditelné také. Zřejmé je, že pokud jsou přístupné při importu, tak by měly být přístupné i pro další importy, jelikož jsou na stejné úrovni jako kód souboru, a nekladli jsme žádným způsobem omezení.

Tedy, navrhol bych umožnit omezit viditelnost importu, než omezovat viditelnost samostatných identifikátorů. Pak by byla decentní explicitní možnost omezení viditelnosti symbolu, aniž by se to muselo řešit poprvakově, a navíc by jsme stále měli možnost vytvoření případného rozhraní z dostupných symbolů, které by se umísitili do jednoho souboru a zbyte by se importovali lokálně.

K označení lokálních importů bych použil slovo `local`

```
import filename as local namespace Foo
```

Samotná klasifikace proměnných dle viditelnosti by se mohla kdyžtak řešit za pomoci direktiv. Například

```
#private  
fcn foo();
```

4.6 Function Overloading

I když se jedná o implicitní konstrukt, který skryje od čtenáře pravou identitu volané funkce, tak přináší, z mého hlediska, jednu zásadní věc, zjednotěnou jména funkcí. Tedy, namísto třeba vepisování datového typu do jména funkce, k rozlišení funkce, lze jen uvést její činnost.

To zjednodušuje vnímání samotného programu, jelikož při práci s vlastními datovými typy, které definují složité objekty, jména funkcí budou už znatelnou zátíží, oproti např. `maxi`, `lmaxf`, `lmaxu`, kde lze přibližně vydedukovat typ očekávané proměnné.

Navíc jména samotných funkcí s použitím postfixu/prefixu, které si zvolíme pro identifikaci, nebudou samostatnými celky z hlediska nástroje pracujících s kódem, tedy v základu samotným kompilátorem a např LSP.

Samotná abstrakce nad konkrétní volanou funkcí pro četnáře není nikterak zavádějící. Nebo spíš, je stejně zavádějící jako for loop, který za místo abstraktní instrukce for provádí skoky a sem a tam. Smysl čtenář získá ze samotného názvu funkce a vstupních proměnných, a vnímá konkrétní funkci jako černou skříňku. I když ona funkce dostává int, tak nemůže vědět, že ten int není hned první instrukci přetypován do floatu. Tedy jediné co to ovlivní je rychlost nalezení správné funkce při potřebě se podívat na její kód, což, bez užití LSP, bude zřejmě delší, ovšem, neřekl bych, že se jedná o něco závažného.

Z mého hlediska, je lepší ho mít, než nemít. Zbývá tedy rozhodnout, zda povolit implicitní overloading, jestli může platit

```
int foo(int x);  
foo(1.0);
```

nebo pro jiný datový typ musí být explicitní cast

```
foo((int)1.0);
```

Explicitním vepisem datového typu se identifikuje volaná funkce. Ovšem, existuje-li potřebná varianta se dá dozvědět při psání kódu jen z LSP. V takovém to případě je cast, jen z hlediska informace, navíc (porovnávali-li s implicitním overloadingem). Nebo až po kompilaci.

Tedy, faktické využití explicitního overloadingu je jako assert, kdy se explicitně vyžaduje konkrétní varianta funkce a v případě absencí se očekává chyba při kompilaci. Ovšem, cast se bude muset specifikovat u každého volání overloaded funkce, což se přeci s tím, že overloading zavádím hlavně z důvodu zjednodušené jmenové stopy. A to nemluvě o tom, že vlastně tutež informaci existuje dvakrát v kódu, jednou při definici, po druhý při volání.

Bylo by tedy vhodné mít implicitní overloading, ale s opcí v jistých případech specifikovat konkrétní požadovaný datový typ. Zavedl bych tedy příslušnou symboliku

```
foo!();
```

Využití prapodivného symbolu v tomto případě není zavádějící, jelikož očekávané intuitivní chování výrazu se nemění. Jedná se stále o function call, který nijak nemění výsledky volání ani jeho vstupy, z hlediska čtenáře je prakticky irelevantní.

4.6.1 Implementace

V C++ implementuji následovně bla bla bla ... https://en.cppreference.com/w/cpp/language/overload_resolution

My budeme postupovat obdobně.

Pro jmeno volané funkce najdeme všechny funkce se stejnými jmeny a v odpovídajícím scope. Uložíme do pole, kde v každém chlivečku bude struktura odkazující se na funkci a doplňující případné informace popisující schodu. Pro zatím, neuvažujeme-li polymorfismus, genericitu atd... si postačíme jen s jednou jedinou proměnnou typu `int` určující podobnost funkce našemu vzoru z volání.

Budeme procházet ono pole postupně funkci po funkci a buď je vyřazovat, nebo sestavovat skóre podobnosti. Nakonci vybereme funkci s největším skóre. Chyba nastane pokud budou dvě a více stejná maximální score, nebo nezbude žádná funkce se skórem.

4.6.2 Přístup jiných jazyků

V Odin je pouze explicitní, jelikož jazyk umožňuje definovat vnořené funkce ve funkcích, a tudíž rozlišení konkrétní funkce, která se má zavolat není triviální.

Zig nema function overloading, ale podobného chování lze docílit při kompilaci za pomoci tzv. duck typing.

D a C++ mají implicitní function overloading.

4.7 Správa chyb

Uvažuje-li se C, tak jazyk nenabízí přímý způsob správy chyb. Chyby se mohou řešit například navratovou hodnotou, nějakým specifickým stavem očekávané vystupní proměnné předané přes ukazatel (většinou `NULL`), speciální funkci, která vrací poslední chybu atd. V zásadě je to na programátorovi, aby vytvořil nějaký systém pro správu chyb, a jestli vůbec.

Při práci s libovolným kódem je pak nutné číst komentáře k funkcím, externí dokumentaci apod. To opět vede na problém, kdy důležitá informace není součástí strukturních elementů kódu, ke kterým by měly různé nástroje přistup. Take to postráda jednotnost, kde různé knihovny mohou řešit zprávu chyb vždy jinak, a ve výsledném programu se bude muset řešit zbytečný problém, jak s tím naložit.

To vše mně ve výsledku vede k myšlence o přidání standardního systému pro správu chyb.

Z metod řešení jiných programovacích jazyků standartizovanou správu chyb lze v zásadě vyčlenit dva přístupy.

Navratová hodnota Chyba je vracena jako navratová proměnná nebo její součást. Obvykle je to spojeno s možností návratu několika proměnných, kde se vyděluje jedno, např. poslední místo, pro případnou chybu (Odin), nebo je přímo speciální doplňující navratová hodnota vydělena jen pro chybu (Go). Nebo, třeba, se může vracet struktura obsahující jak případnou chybu, tak i navratovou hodnotu (Rust).

Tenhle přístup je přímočarý a explicitní a dává svobodu programátorovi jak a kde s chybou naložit. Zpracování chyby je pak přípmou součástí code-flow. Tedy chybový stav je prakticky jen další stav programu.

Try-Catch Využívá se system tzv. exceptions, kde případně chybové místo je zabaleno do try bloku, a případná chyby odchycena v catch bloku. To umožňuje např. nezatěžovat kód správou chyb, a psát ho v try bloku tak, jako kdyby žádná chyba nastat nemohla, a následně jakoukoliv chybu zohlednit v catch bloku.

S try-catch se většinou pojí i tzv. throw mechanismus umožňující označit případné chyby, které může kód nějaké funkce vyvolat, a propagovat jejich ošetření do bloku, jež onu funkci volal.

<https://www.youtube.com/watch?v=uoIutDC5iBE>

4.7.1 Definice požadavku

Neprve bych si definoval požadavky na chybový system:

- Jednotný datový typ.
- Umožnit vytvoření množin chyb, které by se mohly kompozičně skladat do nových množin. Např. můžeme vytvořit množinu chyb pro načtení souboru a množinu chyb pro zápis do souboru. Pak, budeme-li chtít vytvořit funkci, která čte a zapisuje do souboru, tak by jsme měli mít možnost spojit oně dvě množiny do jedné.
- Definice funkce musí specifikovat, které chyby mohou být při jejím volání vráceny.
- Umožnit jednoduchou propagaci chyby stakem funkcí dal. Tedy zjednodušit obdobný konstrukt `err = foo(); if err != null : return err;`, který je relativně frekventní.

4.7.2 Implementace

Protože nahlížet na chybu jako jen na další stav programu, i když, řekněme, speciální, je z mého hlediska přirozenější a implicitní přístup, tak se vydáme cestou navratové proměnné.

Jelikož funkce mají k dispozici jen jednu navratovou proměnnou, chyba se bude vracet samostatným kanálem. Ovšem, nechtěl bych vnímat chybu jako přímo navratovou hodnotu určenou jen pro chybu, jak je tomu např. v Go. Protože pak se pro každé volání funkce musí řešit dvě vystupní proměnné. To ve výsledku povede k vytvoření buď implicitních pravidel, nebo k mnohomluvné (verbose) syntaxi.

K bližší předstě uvedu následující příklad v Go, symbol `:=` vyjadřuje, obdobně jako v Pascalu, definici s inicializací.

```

func foo() (int, error) {
    return 42, nil;
}

val1, err := foo();
if err != nil { ... }

val2, err := foo();
if err != nil { ... }

```

Zde není zcela zřejmé, co se má dít. Prvně provádíme definici `val1` a `err`, a následně, v týmtýž scope provádíme definici `val2` a opět `err`.

Samozřejmě, je to zohledněné pravidly jazyka, kód je kompilovatelný a nová definice `err` se neprovede. Ovšem, dochází ke sporu syntaxe a semantiky, kde ze syntaktického hlediska se `err` chová jen jako druhá návratová hodnota, ale ze semantického se implicitně provádí 'vyjimky' v pravidlech, jen protože je to chybová hodnota.

Navíc se to kompiluje přidáním kvalifikátoru. Bude-li se chtít označit `val1` jako `const` ale ne `err`, nebo naopak, budeme-li chtít mít jedno `embed` a druhý `const`, atd. To vše lze řešit na úkor upravené syntaxe, budeme-li chtít být explicitní, nebo přidáním implicitních pravidel. Proto se pokusím najít jiné řešení.

K návratu chyby bych tedy využil pravé strany příkazu. To umožní oddělit syntaktický samotný příkaz a ošetření chyby. Navíc to může do budoucna umožnit odchycení chyby nejen z jednoho volání funkce, ale i z libovolného výrazu, který by mohl obsahovat několik volání funkcí.

Navrhol bych následující syntaxi.

```

error err;
int x = foo() catch err;

```

Kde se případná chyba uloží do proměnné `err`.

Zde bych stanovil, že nechci zbytečně obohacovat datový typ chyby o implicitní chování, nebo konstrukty pro tvorbu chyb. Chyba by byla vždy datového typu `error` a chovála by se vždy stejně.

Kuriozně se lze v takovém to případě dopustit jedné vyjimky – pominutí samotné definice chyby před odchycením – jelikož je redundantní, místo odchytu totiž může přiřazovat jen pracovat s datovým typem `error`.

```

int x = foo() catch err;

```

Množiny chyb

Samotná chyba by měla být jednoduše identifikovatelná přes své jméno, aby ji bylo možné používat pro určení stavu. Např.

```

if err == ErrName : foo();

```

Chyby by měly být shlukovány do uživatelem definovaných skupin, které by pak sloužily pro určení chybového rozhraní funkcí. Skupiny by měly být shlukovatelné, jelikož funkce by měla mít možnost navracet i chyby užívaných funkcí, které mohou být definované samostatně, aniž by se pro ní redundantně definovaly nové chyby.

Tedy, řekněme, že budeme moct definovat jakési množiny chyb, a jen je. Použijeme následující syntaxi.

```
error ErrorSetA {
    ErrorA;
    ErrorB;
};
error ErrorSetB {
    ErrorSetA;
    ErrorB;
};
```

Pak `ErrorSetA` je množina obsahující prázdné množiny `ErrorA` a `ErrorB` a `ErrorSetB` obsahuje množinu `ErrorSetA` a prázdnou množinu `ErrorB`. Libovolná s těchto množin by měla být identifikovatelná svým jménem a být přiřazena do datového typu `error`.

```
error err = ErrorSetB::ErrorB;
```

K definici chybového rozhraní funkce se pak použije následující syntaxi.

```
fcn foo() using ErrorSetB -> int {...}
```

Funkce `foo` pak může vracet chyby definované v `ErrorSetB`.

Protože oné množiny mají smysl jen při definici samotných funkcí a definice funkce ve funkci není umožněná, tak jejich definice uvnitř funkcí je zavádějící, a tudíž zakazána. A tedy můžeme vnímat oné množiny jako nadstavbu nad namespace pro chyby, a tedy k jejich diferencí používat stejný symbol `::`, jak již bylo naznačeno viz.[...].

Toto řešení je jednoduché a relativně všestranné. Umožňuje například rozvíjet některou prázdnou množinu na plnohodnotnou, aniž by se rozbil kód využívající onu množinu. Ovšem, má jeden základní nedostatek – vracíme pouze stav. Tedy nemůžeme vrátit informaci o chybě. Teoreticky je to řešitelné přidáním nových stavů, ovšem to zdaleka není praktické.

Prakticky to omezuje jen při logování chyby, protože jinak vždy popisujeme stav programu, který je nezbytný z hlediska jeho činnosti. Tudíž přidání v takovýchto případech chybového stavu je vlastně nezbytné (uvažuje-li se, že tento stav je vhodné vnímat jako chybový, obecně to lze řešit normální cestou).

Ve výsledku je to jen něco, co slouží jako doplnění systému. Něco, co je využíváno přímo při zpracování samotné chyby, a tedy neruší samotnou standartizaci, která se kladla za cíl, protože popis samotné chyby už není obecně standartizovatelný, a tak či onak se jedná o konkrétní záležitost.

Pokud by se navrhnutý model zobecnil definici chyby za pomoci struktury nebo unie, tak vlastně dojde k rozporu se standartizací. System se totiž zobecní natolik, že bude moci být využíván i pro jiné věci, a mnohy způsoby. A tudíž se vlastně postavený problém nevyřeší, jen se přesuneme jinam.

Mohl bych povolit přiřazení chybám konkrétních hodnot, což by mohl být postačující kompromis. To umožní pak indexovat pole hodnotami chyb, což je ve výsledku velmi silný nástroj.

Navrat chyby

Možnost v chybovém stavu vrátit i normální hodnotu z funkce je zručná záležitost. Muže posloužit jako například doplňující informace o chybovém stavu. Navíc, je to dokonce nutná záležitost, jelikož vnímáme chybu jen jako další stav, a ne jako něco zvláštního.

Volil bych následující intuitivní syntaxi:

```
return value, err;
```

Kde `value` představuje proměnnou s navratovou hodnotou, a `err` navratovou chybu.

Pak `navrat` jen hodnoty je nezměnný, ale otázkou je `navrat` jen chyby. Lze k tomu přistoupit tak, že vlastně takovýto případ existovat nebude, tedy vždy spolu s chybou se vrátí i hodnota. To také zaručí, že proměnná, do které se запиše navratová hodnota, nebudeme mít neurčenou hodnotu. I když je to skvělé chování, nelze ho použít. Ma-li být jazyk nízkourovňový, musí také dát programátorovi i kontrolu. Nelze jen tak zbytečně vnucovat instrukci. Tedy, přidal bych symbol, např. `_` definující přeskočení proměnné, a skončil s následujícím kódem:

```
return _, err;
```

Implementace v jiných jazycích

4.8 Kontext

Koncept, který umožňuje měnit chování kódu dle svých potřeb a tím i znovuvyžít kód pro konkrétní potřeby. Ovšem, když je kontext explicitní, tedy je předáván jako parametr funkce. Pak je to jen na konkrétní knihovně nabídnout-li vůbec kontext a když ano, tak jaký.

Implicitní kontext, který integrován do jazyka pak umožňuje řešit onen problém. Obzvlášť je to výhodné v jazycích s manuální kontrolou paměti, kde za pomoci kontextu se dá řešit problém použitím vlastních alokátorů. Takovými jazyky jsou třeba Odin a Jai.

`print`, `mem alloc`, etc.

4.8.1 custom alloc

4.9 Compile-time exekuce

V rámci optimalizace kompilatory mohou provádět vypočty některých výrazů pokud je dostatek informace. Tedy například:

```
int x = 5 + 3 * 9 - 2;
```

Výraz definující `x` se může předpočítat a za běhu programu se nebude muset nic vypočítávat.

Ovšem, to vše je prováděno implicitně. Podstatnou možností je ale mít kontrolu nad compile-time exekucí ze stranky jazyka. Koncepce je zpravidla obsažená v nějaké formě v nízkoúrovňových jazycích, jako například C++, D, Odin, Zig, Rust, atd.

K deklaraci compile-time proměnné bych využil klasifikátoru `embed` od slova `embedded` (vestavený), protože slovo odraží smysl, a vzniká obdobně jako `const`. Navíc má stejnou delku, což by při následných deklaracích vypadalo dobře:

```
const int x;  
embed int y;
```

Implementačně by proměnná neexistovala, ale vždy by se využívala její spočtená hodnota.

Takový to jednoduchý klasifikátor pak umožní provádět velmi složité compile-time vypočty. Protože ze své podstaty `embed` specifikuje, že proměnná má být spočtená při kompilaci, fakticky tedy jako koliv nebyla pravá strana výrazu, kompilátor se ji pokusí vypočíst, a buď uspět, nebo ukončit kompilaci s chybou.

what should happen at compile-time, does happen at compile-time.

Lze se tedy například pokusit vypočíst hodnotu libovolné funkce při kompilaci přičemž funkce se nemusí speciálně předeklarovat jako compile-time jako v C++:

```
fn add(i32 x, i32 y) -> i32 { .. }  
embed int ans = add(4, 2);
```

Obdobně to funguje v D, kde se da využít enumeratoru k definici compile-time proměnné a spočíst pravou stranu:

```
int add(int a, int b) { .. }  
enum ans = add(4, 2);
```

Nebo třeba v Zig parametry funkce `comptime` umožňuje parciálně `comptime-run-time` funkce.

```
fn add(a: u32, b: u32) u32 { .. }  
const ans = comptime add(4, 2);
```

V C++

```
constexpr int add(int a, int b) { .. }  
constexpr int ans = add(4, 2);
```

5 Implementace kompilátoru

K implementaci kompilátoru jsem použil jazyk C++ využívající standardu C++20. Jako cílovou a vyvojovou platformu jsem volil Windows, kde jsem využil Visual Studio kompilátoru (cl) a Visual Studio 2022 k debugování. Jako IL jsem vybral jazyk C, protože jsem s ním a s nástroji pro práci s ním dobře znám. Pro kompilaci vygenerovaného C kodu jsem zvolil Tiny C Compiler (TCC), a konkrétně jeho knihovnu libtcc, která umožňuje vestavenou kompilaci C kodu.

Jedná se o konzolovou aplikaci, která pracuje se soubory na disku. Tedy, závislost na operačním systému není tak velká a Unix verze nebude složita k realizaci. Jedná se třeba o rozdíly v získání `getExePath` cesty, která nejde získat skrz `std::filesystem`. Navíc, je zde rozdíl v edianech, který mohl nebyť někde v kodě zohledněn, protože se často operovalo s byty. Proto, i když samotný program se kompiluje pod linux a i funguje, tak nemůžu zaručit jeho funkční bez řádného testování.

Strukturá aplikace je následující. Nejprve je zpracován uživatelský vstup ve formě argumentu z příkazové řádky. Ten je převeden do konfigurace modulu Compiler. Modul compiler je reprezentován namespacem, a tedy je statický. Modul provádí v zásadě čtyři věci postupně. Nejprve běží parser, který inicializuje AST a převede zdrojový kod do jeho uzlu. Dále běží validator, který již provede semantickou kontrolu AST. Následně je kod přeložen instancí modulu translator. V závislosti na uživatelském vstupu výstup překladače může být převeden do binární podoby a popřípadě hned spuštěn.

Cílem bylo dospět k robustné implementaci, která by dokázala vytvořit základnu pro další, řekněme, ideální, implementaci kompilátoru, který by byl specifický pro daný jazyk. Tedy, hlavním úkolem bylo přijít s AST, které by se dalo využít jak pro překlad do libovolného IL, tak i které by bylo možné využít ve vestaveném interpretu.

Zatím se za cíl nekladla rychlost řešení, ale nabídnout proof of concept navrženému jazyku, který by se dal využít...

5.1 Uživatelské rozhraní

Pro komunikaci s uživatelem se využívá argumentu příkazové řádky. Uživatel má v základu tři hlavní příkazy následující jménem vstupního souboru, na který mají být

aplikovane:

build Příkaz sestaví program do spustitelné podoby.

run Příkaz sestaví program do spustitelné podoby a bezprostředně ho spustí v terminálu.

translate Příkaz pouze přeloží do vybraného IL.

Následně mohou být uvedené doplňující možnosti dostupné pro každý příkaz:

-ol (Output Language) Vybere, který IL použít, prozatím jen C.

-of (Output File) Jméno výstupního spustitelného souboru bez rozšíření.

-od (Output Directory) Určí složku, do které se uloží výstupní přeložené IL soubory.

-gd (Generate Debug information) Jestli se má vygenerovat debug informace.

-h (Help) Vypíše nápovědu ohledně uživatelského rozhraní.

Samotná implementace samotného rozhraní byla triviální, pouze za pomoci for loopu a strcmp. Žádné hashování se neprovádělo. Jedinou zajímavou částí byla implementace příkazu run.

Příkaz run potřeboval umožnit spustit zkompileovaný program po kompilaci. Lze třeba použít std::system, ale protože Lze použít napřímo systémových funkcí, v případě windows se jedná o CreateProcessA, která umožní vytvořit process a specifikovat různé možnosti, například jako, jestli se má process spustit v nové konzoli. Pro případ emulace, lze použít WaitForSingleObject, což umožní počkat na vykonání kódu programu. Tedy ze strany uživatele se to bude chovat jako jeden program.

Já jsem volil cestu ukončení hlavního programu a

5.2 AST

Centrum programu je statické AST jehož uzel se reprezentuje za pomoci struktury SyntaxNode. Kořen stromu je statickým prvkem SyntaxNode, což dělá zápis. SyntaxNode obsahuje i další statické prvky, které odkazují na kontejnery s cachovanou informací, která by byla dostupná přímo, aniž by se musel procházet strom, jako odkazy na proměnné, na definice, funkce atd. Samotná definice pak vypadala obdobně:

```
struct SyntaxNode {
    static Scope* root;
    ...
    NodeType type;
    Scope* scope;
    Location* loc;
    ...
};
```


Každý uzel teda ve své podstatě nese informaci o svém typu, v jakém scope se nachází a lokaci ve zdrojovém kodě.

Každý konkrétní uzel pak dědí SyntaxNode, obdobně třeba vypadá uzel reprezentující while loop:

```
struct WhileLoop : SyntaxNode {
    Scope* bodyScope;
    Variable* expression;

    WhileLoop() : SyntaxNode(NT_WHILE_LOOP) {};
};
```

I přesto, že SyntaxNode představuje uzel stromu, tak neslouží k vyjádření všech syntaktických prvků. Samostatnou reprezentaci mají výrazy. Je to dané tím, že vnitřně výrazy představují vyjádření hodnoty proměnné. Tedy v AST je vždy výraz zastoupen proměnnou.

Obecný výraz je definován velmi jednoduše:

```
struct Expression {
    ExpressionType type;
};
```

Pak konkrétní výraz vypadá například následovně:

```
struct OperatorExpression : Expression {
    OperatorEnum operType;
};

struct BinaryExpression : OperatorExpression {
    BinaryExpression() { type = EXT_BINARY; };

    Variable* operandA;
    Variable* operandB;
};
```

Vyjádření hodnoty proměnné je pak představeno následovně:

```
struct Operand : SyntaxNode {
    VariableDefinition* def;

    Value cvalue; // c as compiler
    Value ivalue; // i as interpreter

    std::vector<Value> istack;

    Expression* expression;
    ...
}
```

Jedná se o vyjádření obecného operandu výrazu, proměnná se pak jen definuje jako pojmenovaný operand. Hodnota je tedy buď vyjádřená odkazem na definici, přímo hodnotou, nebo výrazem. Samotná hodnota se využívá i pro určení datového typu operandu a vypadá následovně:

```

struct Value {
    DataTypeEnum dtypeEnum;
    int hasValue = 0;
    union {
        int32_t      i32;
        int64_t      i64;
        ...
        void*         any;
    };
};

```

Jak lze usoudit z atributu `ivalue` a `istack` v definici operandu, doplňující abstrakce nad hodnotou ve formě `Value` je především z hlediska interpreteru, který s onym celkém pracuje. Podrobnějc oné atributy budou představené v příslušné sekci [].

5.3 Parsing

Rozhodl jsem se nejit cestou generatoru jako je YACC, protože, ze zkušeností, bych stejně musel napsat veškerý kod pro sestavení stromu, neměl bych jednoduchý způsob vypisování chyb způsobem, kterým bych chtěl, měl bych omezeny přístup k buffrum souboru, neměl bych kontrolu nad paměti atd. Navíc bych nevyužil výhod při prototypování nad syntaxi, protože mam konkrétní typ jazyka, který chci implementovat, tedy bych jen mohl implementovat nutnou abstrakci pro zaměnu syntaktických celku.

Při implementaci parsru jsem se rozhodl nepoužít abstrakci ve tvaři lexru, jelikož mně jen zajímalo jaky to je, protože ve většině publikaci se lexer používá a chtěl jsem si udělat názor jinou implementaci.

5.3.1 Importy

Zvnějšku parser je jen jedná funkce přijímající na vstup jméno souboru, který oná vníma jako main soubor programu. Vnitřně oná již využívá funkce k parsingu samotných souboru. Každý soubor se parsuje do předem dodané instance `Scope`. Při parsování souboru se případné importy zařazují postupně do fronty. Prvním rozparsovaným souborem je main. Každý případný importovaný soubor je následně parsován a výsledek převáděn do potřebné podoby a zařazen do stromu.

Aby se předešlo parsování 2x téhož souboru, každý rozparsovaný soubor se označí unikátním id. Id je ve formě:

Protož

Nevyhodou je

Je nutné podotknout, že importovat symboly dopředu scope je relativně drahé, protože se prvky v každém array-list budou muset posunout. V případě třeba funkci nezáleží na pořadí, protože neosahují sam o sobě spustitelný kod a jejich definice

nemusí předcházet použití. Tedy je lze prostě zařadit nakonec. V případech, kdy to nejde, krom zřejmé změny kontejneru na třeba list, lze alespoň posunout prvky jen jednou za soubor udělav posun až po zpracování všech importu.

5.3.2 Samotný parsing

Parser jsem implementoval proceduralně, kde každá funkce relativně odpovídala syntaktickému celku, který má za úkol rozparsovat. Každá taková funkce dostává pointer na buffer s textem a pointer na Location, které definuje lokaci, tedy hladvně index, který slouží jako přímý index do bufferu a číslo řádku. Povinnosti funkce je tedy updatovat správně řádek

Ovšem prakticky to není tak hezké, protože rozhodnutí, kterou větví gramatiky se vydat se rozhodují vždy individualně a využívajíc napřímo bufferu textu. Kdežto v případě lexru bych mohl využít tokenu a buď indexace pole, nebo switch-casu v každém případě, což by bylo víc čitelné.

5.4 Validace AST

5.5 Generace C kódu

Překladač AST byl řešen na rozdíl od ostatních modulu abstraktně. Má totiž smysl umět překladat strom do různých IR, na rozdíl třeba od parsru, kterých nemá mít smysl několikero, jelikož pracuje s jedním konkrétním jazykem.

Překladačem je vlastně instance následující jednoduché struktury, která definuje interface:

```
struct Translator {
    FILE* mainFile;
    int debugInfo;

    void (*init)                (char* const dirName);
    void (*printNode)           (FILE* file, int level, SyntaxNode* const node);
    void (*printExpression)     (FILE* file, int level, Expression* const node);
    void (*printForeignCode)    ();
    void (*exit)                ();
};
```

Zasadními funkcemi jsou printNode a printExpression

5.6 Vestavená kompilace C kódu

Protože jsem chtěl, aby kompilator obsahoval konvenční možnost generace spustitelného souboru, tak jsem dospěl k integraci TCC kompilátoru. Obecně bych mohl použít třeba `std::system("gcc build my code pls")`, tak jsem jak nechtěl být závislý na něčem, co už má uživatel předinstalovaný.

Opci by bylo buď distribuovat gcc, nebo i jiný překladač, spolu s kompilátorem, což není šikovný, protože program pak nejde distribuovat jako zdrojový kód. Lepší cestou by bylo integrovat kompilátor do kodu přes příslušnou knihovnu.

GCC

LLVM

TCC jsem vybral proto, že je to malý kus softwaru, který vypadal vhodně pro potřebu vestavení, jelikož by nezabral moc místa.

Protože jsem nenašel moc rozumnou informaci o správném použití knihovny pro moje potřeby, tak popíšu použitý postup trochu detailněji.

Samotná instalace knihovny je standardní, tcc obsahuje include a libtcc složky, které se musí předat kompilátoru jako include path a lib path respektivě a jeden dll soubor. Knihovně potřebné pro samotnou kompilaci C kodu v run-timu jsou pak dostupné ve složce lib.

Nejprve je nutné vytvořit instanci TCCState, předat potřebné cmd argumenty a nastavit výstupní program. Záleží na pořadí!

```
TCCState *state = tcc_new();
tcc_set_options(state, "-ggdb");
tcc_set_output_type(state, TCC_OUTPUT_EXE);

tcc_add_library_path(state, libPath.c_str());
tcc_add_include_path(state, tccIncPath.c_str());

tcc_add_library(state, "gdi32");
...
tcc_add_library(state, "msvcrt");

tcc_add_file(state, "main.c");

tcc_output_file(state, Compiler::outFile);
```

Většina funkcí vrací chybový stav zohlednění kterého se v ukazce pominulo. K uvolnění resursu alokovaných *tcc_new* použijte funkci *tcc_delete*.

Ke kompilaci C kodu

5.7 Správa chyb a logování

Protože chyby a varování jsou fakticky jediným komunikačním prostředkem kompilátoru s uživatelem a jejich kvalita je zcela zásadní, potřeboval jsem si vytvořit jednotný a robustní systém chyb a definovat pravidla jednotná pravidla pro správu a logování chyb.

Aby se docílilo jednoznačnosti, definoval jsem si pro sebe pravidlo – chyba musí být logována přímo ve funkci vyskytu. Jinak se musí řešit jestli použitá funkce vrací chybu už provedla logování nebo ne. Navíc to umožňuje klasickou propagaci chyby

až do mainu, kde se v každé funkci, buď díky lenosti, nebo nerozhodností, se chyba neošetří a jen se předá dál.

Problemem při logování v místě chyby může být ne vždy úplná znalost kontextu, ovšem, protože postupně parsujeme AST, tak lze případně retrospektivně podrobnější informací získat. Navíc, vnější funkce může v případě nutnosti provést svůj doplňující log, pokud to z jejího kontextu přijde vhodné.

Chybu jsem tedy reprezentoval jednoduše jako negativní integralní hodnotu, kde bezchybový explicitní bezchybový stav je 0. Každá funkce by tedy měla mít návratový typ `int` a vrátet případnou chybu. Pro každou chybu jsem definoval take standartní chybovou hlášku `printf` syntaxi. Hlášky byli umístěné v pole a indexovatelné absolutní hodnotou příslušné chyby.

Logovací systém se skládá z pár funkcí a statických hodnot pod namespacem `Logger`. Za pomoci bitových hodnot lze filtrovat typy hlášek, například potlačit informace a varování. Logovací funkce pak umožňovaly po mimo hezčího vypisu samotné hlášky take vypsát konkrétní místo ve zdrojovém kodě v řádkovém formátě a podtrhnout nutnou část viz obr[].

Protože někdy chyba funkce nemusí znamenat kompletní chybu parsingu, někdy se může parser vydat jednou cestou, zjistit, že to nejde rozparsovát a zkusit jinou cestu, tak je nutné umět se vyhnout logování. Pro takový to případ je v rámci `Loggeru` dostupná proměnná `mute`, kde každé vlákno může nastavit vlastní bit v závislosti na svém id. Prozatím ale funguje jen jako `bool` hodnotá, protože multithreading v aplikaci nebyl řešen.

Ovšem, protože chyba samotného parsru vlastně vede ke konci kompilace, tak generace chybové zpravy může být obecně složitá, a tedy i když řešení skrz samotný `Logger` je čisté, tak není optimální. Lepší by bylo předávat potřebnou informaci skrz vstupní proměnné funkce, v ideále zavést nějaký context a předavat obecně potřebné proměnné skrz něho.

6 Závěr