

UNIVERZITET U BEOGRADU
MATEMATIČKI FAKULTET



Ivan Ristović

**JEZIČKI INVARIJANTNA PROVERA
SEMANTIČKE EKVIVALENTNOSTI
STRUKTURNO SLIČNIH SEGMENTA
IMPERATIVNOG KODA**

master rad

Beograd, 2020.

Mentor:

doc. dr Milena VUJOŠEVIĆ-JANIČIĆ
Univerzitet u Beogradu, Matematički fakultet

Članovi komisije:

prof. dr Filip MARIĆ
Univerzitet u Beogradu, Matematički fakultet

doc. dr Milan BANKOVIĆ
Univerzitet u Beogradu, Matematički fakultet

Datum odbrane: _____

TODO zahvalnica

Naslov master rada: Jezički invarijantna provera semantičke ekvivalentnosti strukturno sličnih segmenata imperativnog koda

Rezime: Apstraktno sintaksičko stablo (engl. abstract syntax tree, skraćeno AST) nastaje kao rezultat parsiranja ulaznog programa i predstavlja osnovu za semantičku analizu koda. U okviru semantičke analize koda veoma su važne analize koje omogućavaju obezbeđivanje visokog kvaliteta softvera, bilo kroz analize koje uključuju statičko otkrivanje grešaka u kodu ili kroz analize koje imaju za cilj automatsku transformaciju koda u semantički ekvivalentan oblik sa nekom željenom karakteristikom (na primer, u okviru procesa refaktorisanja koda). Međutim, apstraktno sintaksičko stablo je specifično za konkretan viši programski jezik i na osnovu njega se ne mogu porediti karakteristike programa napisanih u različitim programskim jezicima, što je posebno važno u okviru procesa migracije na nove tehnologije. Osnovni cilj rada je kreiranje i implementacija opšte AST reprezentacije za imperativne programske jezike sa ciljem njene primene u analizi semantičke ekvivalentnosti implementacija algoritama u različitim programskim jezicima. Implementacija će dozvoljavati kreiranje apstrakcije na osnovu zadate gramatike imperativnog programskog jezika, pri čemu će koncept biti prikazan na primeru jednostavnog pseudokoda, kao i na programskim jezicima C i Lua. Implementacija će biti urađena u programskom jeziku C# koristeći ANTLR4 generator parsera.

Ključne reči: TODO

Sadržaj

1	Uvod	1
2	Generisanje i korišćenje apstraktnih sintaksičkih stabala	3
2.1	Apstraktna sintaksička stabla — AST	3
2.2	Parsiranje gramatika programskih jezika	10
2.3	Korišćenje generisanih stabala	23
2.4	Programske paradigme i gramatičke razlike programskih jezika . . .	26
3	Simboličko izračunavanje	35
4	Opis opštih AST apstrakcija za imperativne jezike	39
4.1	Čvorovi deklaracija	42
4.2	Čvorovi operatora	44
4.3	Čvorovi izraza	47
4.4	Čvorovi naredbi	49
5	Semantičko poređenje opštih AST	53
5.1	Upoređivač blokova naredbi	55
6	Implementacija	58
6.1	Implementacija apstrakcije	60
6.2	Implementacija upoređivača	64
6.3	Implementacija vizualnog prikaza AST	64
6.4	Implementacija korisničkog interfejsa	65
6.5	Implementacija testova	66
7	Zaključak	70
	Literatura	72

Glava 1

Uvod

Apstraktno sintaksičko stablo (engl. *abstract syntax tree*, skr. *AST*) programa ima značajnu ulogu u procesu kreiranja izvršivog programa od izvornog koda. AST nastaje parsiranjem izvornog koda, kao rezultat apstrahovanja stabla parsiranja koje generiše *parser*. Parser čita izvorni kôd i pokušava da u njemu pronade primene određenih pravila jezika čiji je kôd dizajniran da parsira. Svaki programski jezik ima specifična sintaksna pravila pa su stoga i skupovi pravila (tzv. *gramatike*) programskih jezika raznorodni, što se zatim prenosi i na generisana stabla parsiranja. Stablo parsiranja se apstrahuje tako što se iz njega izvuku samo bitne sintaksne a uklone neke tehničke informacije.

Ovakva apstrakcija se najpre koristi u semantičkoj analizi programa koju vrši prevodilac nakon faze parsiranja i provere sintaksne ispravnosti koda. Ukoliko program prođe semantičke provere, prelazi se na prevođenje istog u međureprezentaciju i fazu optimizacije. Nakon faze optimizacije sledi generisanje asemblerskog koda koji se zatim prevodi u mašinski kôd.

AST, zbog svoje uloge u semantičkoj analizi, može poslužiti i za analizu programa pre samog prevođenja, kroz proces poznat pod nazivom *statička analiza*. Posmatranje programa kroz AST pruža mogućnost za poređenje dva programa na apstraktnom nivou. Jedna primena ove ideje u okviru statičke analize može biti provera semantičke ekvivalentnosti. Provera semantičke ekvivalentnosti dva programa je neodlučiv problem u opštem slučaju, međutim pod određenim pretpostavkama koje pojednostavljuju problem moguće je dizajnirati algoritme koji daju smislene rezultate u praksi. Jedna od često korišćenih pretpostavki je pretpostavka sličnosti strukture dva programa. Interesantno, provera da li dva programa zadovoljavaju ovu premisu se može proveriti posmatranjem izgleda i sličnosti u

strukturi na apstraktnom nivou — problem koji se može rešiti primenom algoritama za rad sa stablima jer je u pitanju AST (ali i grafovima uopšte, jer je stablo specijalizacija grafa).

Iako veoma konceptualno moćan alat, AST je ipak specifičan za konkretni programski jezik s obzirom da nastaje od stabla parsiranja koje je usko vezano za gramatiku konkretnog programskog jezika. Motivacija za ovaj rad dolazi od nepostojanja opštih apstrakcija sintaksičkih stabala koje bi se mogle koristiti za analizu programa napisanih u različitim programskim jezicima. Iako je broj programskih jezika danas veoma veliki, u okviru iste programske paradigme jezici moraju implementirati koncepte koji su potrebni da bi se programiralo u toj paradigmi i ta zajednička svojstva se mogu iskoristiti za formiranje opšte apstrakcije na AST nivou.

U ovom radu će biti predstavljena opšta AST apstrakcija za imperativne programske jezike, sa ciljem da se omogući zajednička apstraktna reprezentacija za što više imperativnih jezika, pa čak i onih koji pripadaju skript paradigmi. Njena upotreba će biti demonstrirana na problemu semantičke ekvivalentnosti dobijenih apstrakcija kroz naivni algoritam poređenja simboličkih promenljivih. Štaviše, na apstraktnom nivou nije važno od kog se programskog jezika dobio AST, što može imati primenu u procesu migracije na nove tehnologije.

Naravno, semantička ekvivalentnost se ne mora zasnivati na apstrahovanju programa, već se takođe često rešava spuštanjem na nivo međukoda između višeg programskog jezika i assemblera. U nekim slučajevima se može ići i do assemblera pa i mašinskog jezika. Ukoliko bi se posmatrali assemblyski ili mašinski kod, vršilo bi se poređenje kodova prilagođenih određenoj arhitekturi procesora. U ovom radu je odabran AST-zasnovan pristup, s obzirom na važnosti i značaj apstraktnih sintaksičkih stabala, ali i zbog nedostatka opštih apstrakcija.

U poglavlju 2 će biti opisani relevantni pojmovi potrebni za razumevanje rada uz akcenat na apstraktnim sintaksičkim stablima i procesu njihovog dobijanja. U poglavlju 3 će biti opisano simboličko izračunavanje koje se koristi prilikom provere semantičke ekvivalentnosti apstraktnih sintaksičkih stabala. Opšta AST apstrakcija za imperativne jezike biće opisana u poglavlju 4, a njena upotreba u problemu odlučivanja semantičke ekvivalentnosti kao i sam algoritam za poređenje opštih apstrakcija biće opisani u poglavlju 5. Implementacija apstrakcije i algoritma semantičkog poređenja će biti opisana u poglavlju 6. Na kraju, biće dati glavni zaključci ovog rada kao i moguća unapređenja i budući koraci.

Glava 2

Generisanje i korišćenje apstraktnih sintaksičkih stabala

U ovom poglavlju će biti opisani koncepti i alati čije je razumevanje potrebno kako bi se razumeo proces kreiranja i poređenja opštih apstrakcija, kao i implementacije programa za njihovo kreiranje i poređenje. Umesto analize samog sadržaja izvornog koda analizira se *apstraktno sintaksičko stablo*, opisano u odeljku 2.1. Kako bi se od izvornog koda došlo do stabla parsiranja a potom i do apstraktnog sintaksičkog stabla, koriste se *lekseri* i *parseri*. S obzirom da je cilj kreirati univerzalnu reprezentaciju, neophodno je kreirati leksera i parsere za proizvoljne programske jezike. Više reči o samom procesu dobijanja stabla parsiranja od izvornog koda i alatima koji mogu da generišu leksera i parsere biće u odeljku 2.2, sa akcentom na alat *Another Tool For Language Recognition* [1], u daljem tekstu *ANTLR*. Da bi se dobijena stabla koristila, neophodno je poznavati *obrasce za projektovanje* opisane u odeljku 2.3 koji pružaju gotova rešenja za česte probleme koji se u ovom radu koriste za pružanje interfejsa obilaska stabala i izračunavanja vrednosti nad istim. O raznovrsnosti programskih jezika i opisu koncepata koji programski jezici u okviru iste paradigme pružaju biće reči u odeljku 2.4.

2.1 Apstraktna sintaksička stabla — AST

Kako bi se kôd pisan u nekom programskom jeziku (*izvorni fajl*) preveo u kôd koji će se izvršavati na nekoj mašini (*izvršivi fajl*), prevodilac izvršava određene korake. Prvi korak je prepoznavanje gradivnih elemenata programskog jezika koji se nazivaju *leksema*, nalik na prepoznavanje reči u rečenicama govornog jezika, i

GLAVA 2. GENERISANJE I KORIŠĆENJE APSTRAKTHNIH SINTAKSIČKIH STABALA

raspodela leksema u grupe po funkciji (subjekti, predikati, objekti itd.). Nakon *leksike*, procesa prepozavanja leksema, proverava se da li niz prepoznatih leksema ispunjava pravila odnosno *sintaksu* programskog jezika, nalik na proveru da li se složene govorne rečenice sastoje od subjekta i predikta u određenom redosledu. Na kraju se proverava značenje odnosno *semantika* koda, nalik na proveru da li govorne rečenice, iako sintaksno pravilne, imaju smisla. Deo prevodioca koji određuje da li je izvorni kôd ispravno formiran u terminima leksike, sintakse i semantike se naziva *prednji deo* (engl. *front end*). Ukoliko je izvorni kôd ispravan, prednji deo kreira *međureprezentaciju* koda (engl. *intermediate representation*, u daljem tekstu *IR*). Ukoliko to nije slučaj, prevođenje ne uspeva i programeru se daje poruka sa obrazloženjem zašto prevođenje nije uspelo [5].

Za potrebe ovog rada, što se procesa prevođenja tiče, dovoljno je poznavanje prednjeg dela, stoga neće biti reči o daljim koracima u fazi prevođenja (semantička provera, optimizacije, generisanje *IR*). Zainteresovani čitalac može više detalja pronaći u [5] i [25].

Pretpostavimo da želimo da prevedemo izvorni kôd pisan u programskom jeziku C sa slike 2.1. Primetimo da postoji greška u datom kodu — simbol *c* koji se koristi u dodeli u liniji 7 će biti prepoznat kao identifikator koji ne odgovara nijednoj deklarisanom promenljivoj — stoga ne možemo prevesti ovaj kôd. Ovo, doduše, nije greška u sintaksi — izraz *a+c* je sasvim validan u programskom jeziku C bez analize konteksta u kom se javlja. Problem će postati vidljiv tek nakon parsiranja izvornog koda i provere ispunjenosti sintaksnih pravila, tačnije u fazi semantičke provere. Stoga se ovakve greške nazivaju *semantičke greške*, dok se greške u sintaksi nazivaju *sintaksne greške*.

```
1  #include<stdio.h>
2  #define T int
3
4  int main()
5  {
6      T a, b;
7      a = a + c;          // c nije deklarisano
8      printf("%d", a);
9      return 0;
10 }
```

Slika 2.1: Primer izvornog koda sa semantičkom greškom (C).

GLAVA 2. GENERISANJE I KORIŠĆENJE APSTRAKTHNIH SINTAKSIČKIH STABALA

Pre nego što prednji kraj prevodioca uopšte dobije kôd koji treba da prevede u međureprezentaciju, vrši se *pretprocesiranje* od strane programa koji se naziva *pretprocesor*. U fazi pretprocesiranja se izvode samo tekstualne operacije kao što su brisanje komentara ili zamena makroa u jezicima kao što je C. Rezultat rada pretprocesora za izvorni kôd sa slike 2.1 se može videti na slici 2.2 ¹.

```
1  # 1 "<stdin>"
2  # 1 "<built-in>"
3  # 1 "<command-line>"
4  # 31 "<command-line>"
5  # 1 "/usr/include/stdc-predef.h" 1 3 4
6
7  ...
8
9  extern char *ctermid (char *__s) __attribute__ ((__nothrow__ ,
    __leaf__));
10 # 840 "/usr/include/stdio.h" 3 4
11 extern void flockfile (FILE *__stream) __attribute__ ((__nothrow__
    , __leaf__));
12 extern int ftrylockfile (FILE *__stream) __attribute__
    ((__nothrow__ , __leaf__));
13 extern void funlockfile (FILE *__stream) __attribute__
    ((__nothrow__ , __leaf__));
14 # 868 "/usr/include/stdio.h" 3 4
15 # 2 "<stdin>" 2
16 # 2 "<stdin>"
17
18 int main()
19 {
20     int a, b;
21     a = a + c;
22     printf("%d", a);
23     return 0;
24 }
```

Slika 2.2: Prikaz rezultata rada pretprocesora za izvorni kôd sa slike 2.1. Pritom, prikazano je samo par linija sa početka i kraja izlaza pretprocesora — kôd iznad main funkcije je uključen iz `stdio.h` zaglavlja.

U toku faze leksičke analize, kako prevodilac ne bi radio nad sirovim karakterima izvornog koda, potrebno je izvršiti transformaciju karaktera izvornog koda.

¹U nekim implementacijama C standardne biblioteke, moguće je da se poziv funkcije `printf` zameni pozivom funkcije `fprintf` sa ispisom na `stdout`. U standardu se propisuje da funkcije kao što je `printf` mogu biti implementirane kao makroi. Izlaz na slici 2.2 je generisan od strane GCC 7.4.0 po C11 standardu i ovo nije slučaj u datom okruženju.

GLAVA 2. GENERISANJE I KORIŠĆENJE APSTRAKTNIH SINTAKSIČKIH STABALA

Prevodilac ima u vidu elemente programskog jezika i šablone za njihovo prepoznavanje. Lekseme su one sekvence karaktera izvornog koda koje zadovoljavaju ove šablone. Lekseme se dalje razvrstavaju u kategorije, tzv. *tokene* — ključne reči, operatore, promenljive itd. Proces dobijanja tokena od izvornog koda se naziva *tokenizacija*. Komponenta prednjeg dela koja vrši tokenizaciju se naziva *skener* ili *lekser*. Pojednostavljen primer šablona koje lekser pokušava da prepozna se mogu videti na slici 2.3. Primer izlaza leksera za izlaz pretprocesora sa slike 2.2 se može videti na slici 2.4.

```
1 Identifier
2     : IdentifierNondigit (IdentifierNondigit | Digit)*
3     ;
4 IdentifierNondigit
5     : Nondigit
6     | UniversalCharacterName
7     ;
8 Nondigit
9     : [a-zA-Z_]
10    ;
11 Digit
12    : [0-9]
13    ;
```

Slika 2.3: Primer delimične definicije tokena za ime promenljive po C11 standardu.

Nakon faze skeniranja potrebno je proveriti da li niz pročitanih tokena zadovoljava sintaksu programskog jezika. Pevodilac stoga mora da uporedi strukturu koda sa unapred definisanom strukturom za određeni programski jezik što zahteva formalnu definiciju sintakse jezika. Programski jezik možemo posmatrati kao skup *pravila* pod imenom *gramatika* [4]. Na slici 2.5 se može videti deo gramatike programskog jezika C. Proces provere zadovoljenosti sintaksnih pravila se naziva *parsiranje* a komponenta prednjeg dela koja vrši parsiranje se naziva *parser*².

Parser, imajući u vidu gramatiku jezika, kreira *stablo parsiranja* (eng. *parse tree* ili *derivation tree*). Takvo stablo i dalje sadrži sve relevantne informacije o izvornom kodu. Vizuelni prikaz rada parsera za gramatiku sa slike C11 i izvornog koda sa slike 2.2 je dat na slici 2.6. Stablo parsiranja se koristi u narednim fazama prevođenja.

²Moderni kompajleri često nemaju odvojene faze skeniranja i parsiranja, već se skeniranje odvija paralelno sa fazom parsiranja. Međutim, to nas ne sprečava da ispišemo tokene onda kada se oni prepoznaju, što to je demonstrirano na slici 2.4.

GLAVA 2. GENERISANJE I KORIŠĆENJE APSTRAKTNIH SINTAKSIČKIH STABALA

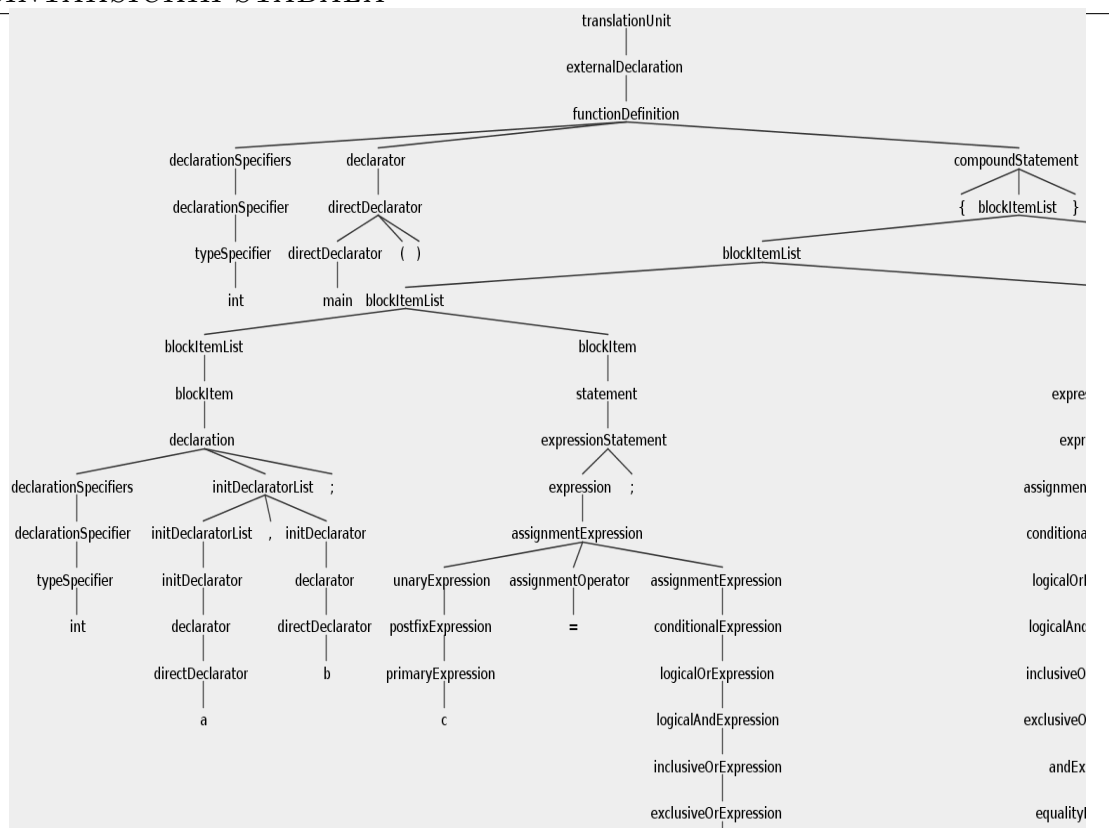
```
1 identifier 'main'      [LeadingSpace] Loc=<sample.c:3:5>
2 l_paren '('          Loc=<sample.c:3:9>
3 r_paren ')'          Loc=<sample.c:3:10>
4 l_brace '{'          [StartOfLine] Loc=<sample.c:4:1>
5 int 'int'            [StartOfLine] [LeadingSpace] Loc=<sample.c:5:5>
6 identifier 'a'        [LeadingSpace] Loc=<sample.c:5:9>
7 comma ','            Loc=<sample.c:5:10>
8 identifier 'b'        [LeadingSpace] Loc=<sample.c:5:12>
9 semi ';'             Loc=<sample.c:5:13>
10 identifier 'a'        [StartOfLine] [LeadingSpace] Loc=<sample.c:6:5>
11 equal '='            [LeadingSpace] Loc=<sample.c:6:7>
12 identifier 'a'        [LeadingSpace] Loc=<sample.c:6:9>
13 plus '+'             [LeadingSpace] Loc=<sample.c:6:11>
14 identifier 'c'        [LeadingSpace] Loc=<sample.c:6:13>
15 semi ';'             Loc=<sample.c:6:14>
16 identifier 'printf'   [StartOfLine] [LeadingSpace]
    Loc=<sample.c:7:5>
17 l_paren '('          Loc=<sample.c:7:11>
18 string_literal '%"d"' Loc=<sample.c:7:12>
19 comma ','            Loc=<sample.c:7:16>
20 identifier 'a'        [LeadingSpace] Loc=<sample.c:7:18>
21 r_paren ')'          Loc=<sample.c:7:19>
22 semi ';'             Loc=<sample.c:7:20>
23 return 'return'      [StartOfLine] [LeadingSpace] Loc=<sample.c:8:5>
24 numeric_constant '0' [LeadingSpace] Loc=<sample.c:8:12>
25 semi ';'             Loc=<sample.c:8:13>
26 r_brace '}'          [StartOfLine] Loc=<sample.c:9:1>
27 eof ''               Loc=<sample.c:9:2>
```

Slika 2.4: Proces tokenizacije koda sa slike 2.2. Generisano uz pomoć clang [3] kompajlera.

```
1 declarationList
2     : declaration
3     | declarationList declaration
4     ;
5 declaration
6     : declarationSpecifiers initDeclaratorList ';'
7     | declarationSpecifiers ';'
8     | staticAssertDeclaration
9     ;
```

Slika 2.5: Prikaz para pravila iz gramatike programskog jezika C po standardu C11.

GLAVA 2. GENERISANJE I KORIŠĆENJE APSTRAKTHNIH SINTAKSIČKIH STABALA

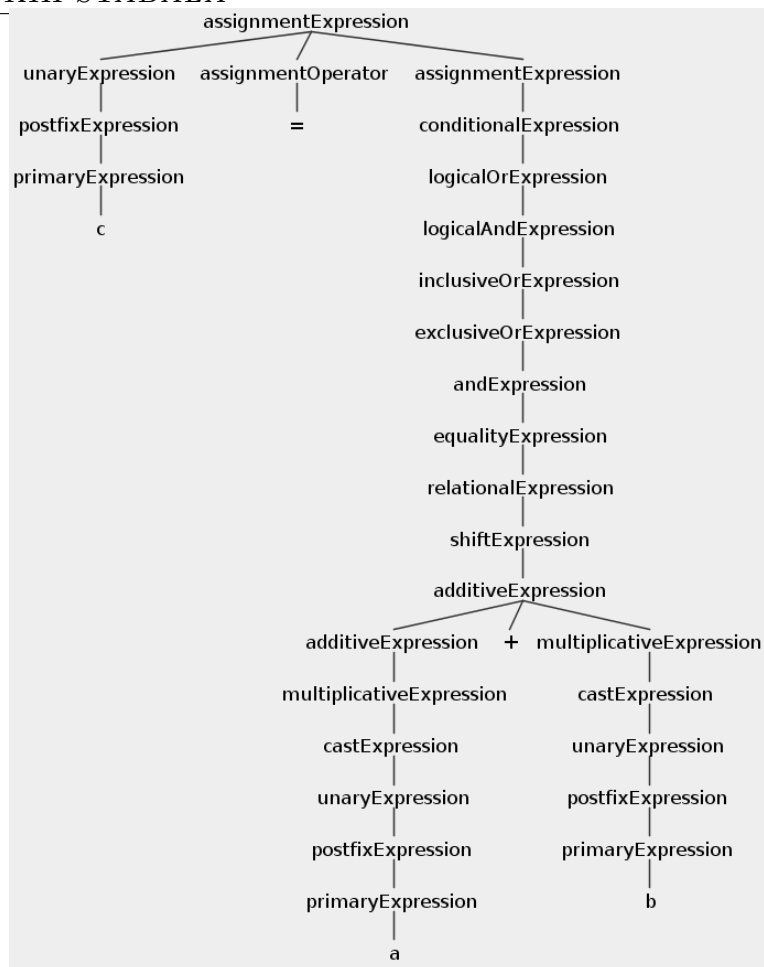


Slika 2.6: Prikaz dela stabla parsiranja koje generiše parser kreiran od strane alata ANTLR4 [1] za kôd sa slike 2.2.

Stablo parsiranja sadrži sve informacije potrebne u fazi parsiranja uključujući detalje korisne samo za parser prilikom provere ispunjenosti gramatičkih pravila. Sa druge strane, *apstraktno sintaksičko stablo* sadrži samo sintaksičku strukturu u jednostavnijoj formi. Na slici 2.7 se može videti koliko stablo parsiranja može biti kompleksno čak i za jednostavne aritmetičke izraze. Razlog kompleksnosti u ovom slučaju dolazi iz rekurzivnih pravila iz C11 gramatike. Parseru su sve ove informacije neophodne ali na apstraktnijem nivou nisu potrebne. Na primer, jedina važna semantička odlika izraza $a+c$ je da je to zbir vrednosti nekih promenljivih. Na slici 2.8 se mogu videti različita apstraktna sintaksička stabla za pomenuti izraz, ali takođe i za malo složenije izraze. Podrazumeva se, naravno, da je ulaz već tokenizovan.

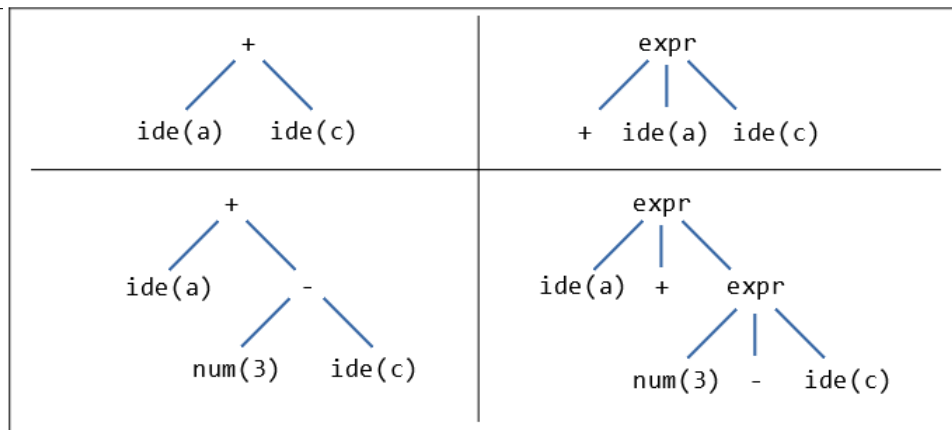
Uloga apstraktnog sintaksičkog stabla [21] je da pokaže semantiku strukture koda preko stabla. Kao što se vidi na slici 2.8, postoji određeni nivo slobode prilikom njihovog kreiranja. Generalno, *terminalni simboli* — simboli koji pred-

GLAVA 2. GENERISANJE I KORIŠĆENJE APSTRAKTHNIH SINTAKSIČKIH STABALA



Slika 2.7: Prikaz kompleksnosti stabla parsiranja za izraz $a+c$ u C11 gramatici.

stavljaju listove stabla parsera — koji odgovaraju operatorima i naredbama se podižu naviše i postaju koreni podstabala, dok se njihovi operandi ostavljaju kao njihovi potomci u stablu. Desna stabla sa slike ne prate u potpunosti ovaj princip, ali se takođe koriste zbog regularnosti izraza — ukoliko binarni izraz posmatramo kao apstrakciju, za implementaciju je lakše koristiti ovakav tip apstraktnih sintaksičkih stabala i stoga će biti korišćen kasnije u implementaciji opšte apstrakcije. Primetimo takođe da se u stablima za izraz $a+(3-c)$ (dole) implicitno sačuvala informacija o prioritetu operacije oduzimanja u izrazu. Jasno je, dakle, da se računanje vrednosti aritmetičkih izraza onda vrši kretanjem od listova stabla ka korenu. Takođe, pošto je apstraktno sintaksičko stablo apstrakcija stabla parsiranja, više semantički ekvivalentnih izraza može imati isto apstraktno sintaksičko stablo ali različito stablo parsiranja; na primer, ako razmatramo izraz $(a+5)-x/2$



Slika 2.8: AST varijante bez regularnosti (levo) i sa regularnošću (desno) za izraze $a+c$ (gore) i $a+(3-c)$ (dole).

i izraz $a+5-(x/2)$.

2.2 Parsiranje gramatika programskih jezika

Ukoliko imamo gramatiku proizvoljnog programskog jezika, postavlja se pitanje:

Da li je moguće definisati postupak i zatim napraviti program koji će generisati kodove leksera i parsera napisane u nekom specifičnom programskom jeziku za proizvoljnu gramatiku datu na ulazu?

Odgovor je potvrđan i postoji veliki broj alata koji se mogu koristiti u ove svrhe, od kojih je navedeno par njih u odeljcima ispod.

Flex

Flex [8] je kreiran kao alternativa *lex*-u [15]. *Flex* generiše samo leksera pa se stoga najčešće koristi u kombinaciji sa drugim alatima koji mogu da generišu parsera, kao što su npr. *GNU Bison* ili *BYACC*.

GNU Bison

GNU Bison [11] je generator parsera i deo GNU projekta [12], često referisan samo kao *Bison*. *Bison* generiše parser na osnovu korisnički definisane kontekstno

GLAVA 2. GENERISANJE I KORIŠĆENJE APSTRAKTNIH SINTAKSIČKIH STABALA

slobodne gramatike [4], upozoravajući pritom na dvosmislenosti prilikom parsiranja ili nemogućnost primene gramatičkih pravila. Generisani parser je najčešće C a ređe C++ program, mada se u vreme pisanja ovog rada eksperimentiše sa Java podrškom. Generisani kodovi su u potpunosti prenosivi i ne zahtevaju specifične kompajlere. Bison može da, osim podrazumevanih *LALR(1)* [14] parsera, generiše i kanoničke *LR* [18], *IELR(1)* [6] i *GLR* [10] parsere.

BYACC

Berkeley YACC, skraćeno *BYACC* [2], je generator parsera pisan po ANSI C standardu i otvorenog je koda. Posmatra se od strane mnogih kao *najbolja varijanta YACC-a* [15]. *BYACC* dozvoljava tzv. *reentrant* kôd — omogućava bezbedno konkurentno izvršavanje koda na način kompatibilan sa Bison-om i to je delom razlog njegove popularnosti.

ANTLR

Another Tool for Language Recognition, ili kraće *ANTLR* [1], je generator *LL(*)* [16] leksera i parsera pisan u programskom jeziku Java sa intuitivnim interfejsom za obilazak stabla parsiranja. Verzija 3 podržava generisanje parsera u jezicima Ada95, ActionScript, C, C#, Java, JavaScript, Objective-C, Perl, Python, Ruby, i Standard ML, dok verzija 4 u vreme pisanja ovog rada samo generiše parsere u narednim jezicima: Java, C#, C++, JavaScript, Python, Swift i Go.

ANTLR verzije 4 je izabran u ovom radu zbog svoje popularnosti, jednostavnosti, intuitivnosti i podrške za mnoge moderne programske jezike. Verzija 4 je izabrana umesto verzije 3 po preporuci autora, na osnovu eksperimentalne analize brzine i pouzdanosti te verzije u odnosu na prethodnu. Lekseri i parseri za ulazne gramatike će u implementaciji biti generisani u programskom jeziku C#.

Parseri generisani koristeći *ANTLR* koriste novu tehnologiju koja se naziva *Prilagodljiv LL(*)* (engl. *Adaptive LL(*)*) ili *ALL(*)* [19], dizajniranu od strane Terensa Para, autora *ANTLR-a*, i Sema Harvela. *ALL(*)* vrši *dinamičku analizu* gramatike u fazi izvršavanja, dok su starije verzije radile analizu pre pokretanja parsera. Ovaj pristup je takođe efikasniji zbog značajno manjeg prostora ulaznih sekvenci u parser.

Najbolji aspekt *ANTLR-a* je lakoća definisanja gramatičkih pravila koji opisuju sintaksne konstrukte. Primer jednostavnog pravila za definisanje aritmetičkog

izraza je dat na slici 2.9. Pravilo `exp` je levo rekurzivno jer barem jedna od njegovih alternativnih definicija referiše na baš pravilo `exp`. ANTLR4 automatski zamenjuje levo rekurzivna pravila u nerekurzivne ekvivalente. Jedini zahtev koji mora biti ispunjen je da levo rekurzivna pravila moraju biti *direktna* — da pravila odmah referišu sama sebe. Pravila ne smeju referisati drugo pravilo sa leve strane definicije takvo da se eventualno kroz rekurziju stigne nazad do pravila od kog se krenulo bez poklapanja sa nekim tokenom.

```

1  exp : (exp)
2      | exp '*' exp
3      | exp '+' exp
4      | INT
5      ;

```

Slika 2.9: Definicija uprošćenog aritmetičkog izraza po ANTLR4 gramatici.

Preduslovi za pokretanje ANTLR4

Kako bi se ANTLR koristio, potrebno je instalirati ANTLR i imati *Java Runtime Environment* (skr. *JRE*) instaliran na sistemu i dostupan globalno pokretanjem putem komande `java`. Instalacija se sastoji od preuzimanja najnovijeg `.jar` fajla ³, sa zvanične stranice [1] ili recimo korišćenjem `curl` alata:

```

1  $ curl -O http://www.antlr.org/download/antlr-4-complete.jar

```

Na UNIX sistemima moguće je kreirati alias `antlr4` ili *shell* skript unutar direktorijuma `/usr/local/bin` sa imenom `antlr4` koji će pokrenuti `.jar` fajl na sledeći način (pretpostavljajući da se `.jar` fajl nalazi u direktorijumu `/usr/local/lib`):

```

1  #!/bin/sh
2  java -cp "/usr/local/lib/antlr4-complete.jar:$CLASSPATH"
    org.antlr.v4.Tool $*

```

Na Windows sistemima moguće je kreirati *batch* skript sa imenom `antlr4.bat` koji će pokrenuti ANTLR4, na sledeći način (pretpostavljajući da se `.jar` fajl nalazi u direktorijumu `C:\lib`):

³Takođe je moguće prevesti izvorni kôd dostupan na servisu GitHub <https://github.com/antlr/antlr4>

GLAVA 2. GENERISANJE I KORIŠĆENJE APSTRAKTNIH SINTAKSIČKIH STABALA

```
1 java -cp C:\lib\antlr-4-complete.jar;%CLASSPATH% org.antlr.v4.Tool  
   %*
```

Ukoliko su aliasi ili skript fajlovi imenovani kao iznad, moguće je iz komandne linije pojednostavljeno pokretati ANTLR4:

```
1 $ antlr4  
2 ANTLR Parser Generator Version 4.0  
3 -o ___ specify output directory where all output is generated  
4 -lib ___ specify location of .tokens files  
5 ...
```

Dodatno, za Unix sisteme ⁴, moguće je kreirati dodatni alias `grun` (ili alternativno, kreirati `shell script`) za biblioteku `TestRig`. Biblioteka `TestRig` se može koristiti za brzo testiranje parsera - moguće je pokrenuti parser od bilo kog pravila i dobiti izlaz parsera u raznim formatima. `TestRig` dolazi uz ANTLR `.jar` fajl i moguće je napraviti prečicu za brzo pokretanje (nalik na ANTLR alias):

```
1 $ alias grun='java -cp  
   "/usr/local/lib/antlr-4-complete.jar:$CLASSPATH"  
   org.antlr.v4.gui.TestRig '
```

Generisanje parsera koristeći ANTLR4

U ovom odeljku će biti opisan proces kreiranja interfejsa za parsiranje programa pisanih u pseudo-programskom jeziku (u nastavku *pseudo-jezik*), nalik na pseudokod. Ovako dobijeni interfejs će moći da se koristi u opšte svrhe, a za potrebe ovog rada će se koristiti za generisanje apstraktnog sintaksičkog stabla za izvorni kôd pisan u pseudo-jeziku.

Definišimo gramatiku pseudo-jezika prateći ANTLR pravila za definisanje gramatika. Kao i za svaki drugi programski jezik, treba podržati neke osnovne koncepte: *identifikatore*, *izraze*, *naredbe*, *funkcije* i slično. Za sada se fokusirajmo na naredbe, kao samostalne izvršive jedinice kôda. Stoga program možemo posmatrati kao niz naredbi. U nekim slučajevima će biti potrebno definisanje kompleksnih naredbi koje se sastoje od više drugih naredbi, i ovakve složene naredbe ćemo zvat *blok* ili *blok naredbi*. Stoga, radi konzistentnosti, program će biti blok naredbi.

⁴Za Windows operativni sistem je moguće kreirati *batch* skript po opisu na <https://github.com/antlr/antlr4/blob/master/doc/getting-started.md>.

GLAVA 2. GENERISANJE I KORIŠĆENJE APSTRAKTHNIH SINTAKSIČKIH STABALA

Kako bismo označili da su naredbe deo bloka naredbi, korišćemo reči `begin` i `end`, osim ukoliko je reč o samo jednoj naredbi. Ovakve situacije rešavamo definisanjem *alternativa* u definiciji pravila — više definicija razdvojenih simbolom `|`. Specijalne reči kao što su `begin` i `end` će biti rezervisane reči našeg pseudo-jezika, tzv. *ključne reči*. Na slici 2.10 se može videti definicija programa ⁵ i bloka naredbi pseudo-jezika, pri čemu se ključne reči u pravilima navode između apostrofa. ANTLR dozvoljava jednostavne definicije pravila u kojima figuriše promenljiv broj drugih pravila, pri čemu se koriste simboli kao u regularnim izrazima ⁶, što je iskorišćeno za definiciju pravila bloka naredbi. `NAME` je identifikator koji predstavlja ime programa. Identifikatore ćemo definisati kasnije, za sada možemo posmatrati identifikator kao nisku karaktera s tim što će postojati restrikcije vezane za to koji karakteri se mogu naći unutar identifikatora.

```
1 unit
2     : 'algorithm' NAME block EOF
3     ;
4 block
5     : 'begin' statement+ 'end'
6     | statement
7     ;
```

Slika 2.10: Definicija jedinice prevođenja i bloka naredbi za pseudo-jezik.

Sledeći korak je definisanje naredbi pseudo-jezika. Slično kao i u drugim programskim jezicima, potrebno je podržati koncept deklaracije promenljive, dodele vrednosti izraza promenljivoj, naredbe kontrole toka — grananje i petlje. Na slici 2.11 je definisano šta se sve smatra jednom naredbom. Naredbe mogu biti i prazne, što je označeno ključnom rečju `pass`. Iz definicije sa slike se jasno vidi šta sve može biti naredba (prateći redosled alternativa pravila): deklaracija, dodela, poziv funkcije (označen kao `cexp`, skraćeno od *function call expression*) ⁷,

⁵Drugim rečima, jedan program u pseudo-jeziku je jedinica prevođenja, pa je zato pravilo nazvano *unit*.

⁶U regularnim izrazima, simbol `a?` označava opciono pojavljivanje simbola `a`, simbol `a+` označava jedno ili više pojavljivanja simbola `a`, a simbol `a*` označava proizvoljan broj pojavljivanja simbola `a` — kombinacija simbola `?` i `+`.

⁷Funkcije mogu vratiti vrednosti pa se stoga njihovi pozivi mogu naći u izrazima — dakle poziv funkcije je validan izraz (stoga `expression` u imenu `function call expression`). Naravno, ta vrednost se može ignorisati ili pak sama funkcija može biti takva da nema povratnu vrednost već je samo neophodno izvršiti je zbog sporednih efekata.

GLAVA 2. GENERISANJE I KORIŠĆENJE APSTRAKTNIH SINTAKSIČKIH STABALA

vraćanje vrednosti izraza (ključna vrednost `return`) iz funkcije, prekidanje izvršavanja davanjem poruke o grešci, naredba grananja, *while* petlja, *repeat-until* petlja i inkrementiranje/dekrementiranje vrednosti promenljive.

```
1 statement
2     : 'pass'
3     | declaration
4     | assignment
5     | cexp
6     | 'return' exp
7     | 'error' STRING
8     | 'if' exp 'then' block ('else' block)?
9     | 'while' exp 'do' block
10    | 'repeat' block 'until' exp
11    | ('increment' | 'decrement') var
12    ;
```

Slika 2.11: Definicija naredbe za pseudo-jezik.

Deklaracija, prikazana na slici 2.12, uvodi pojavljivanje simbola sa identifikatorom `NAME` kao oznaku za promenljivu, funkciju ili proceduru — funkciju bez povratne vrednosti. Svaka promenljiva mora biti određenog tipa, što se postiže pravilom `type`. Promenljivoj se, opciono, može pridružiti početna vrednost, drugim rečima promenljiva se može *inicijalizovati* tako da joj se pridruži vrednost nekog izraza. Procedure i funkcije imaju opcione parametre, vrednosti izraza koje im se prosleđuju kasnije u pozivu kao argumenti. Lista parametara, takođe prikazana na slici 2.12, se navodi kao lista proizvoljno mnogo parova `NAME : type`, što se vidi iz definicije pravila `parlist`.

```
1 declaration
2     : 'declare' type NAME ('=' exp)?
3     | 'procedure' NAME '(' parlist? ')' block
4     | 'function' NAME '(' parlist? ')' 'returning' type block
5     ;
6 parlist
7     : NAME ':' type (',' NAME ':' type)*
8     ;
```

Slika 2.12: Definicija deklaracije za pseudo-jezik.

GLAVA 2. GENERISANJE I KORIŠĆENJE APSTRAKTHNIH SINTAKSIČKIH STABALA

Identifikatori su niske karaktera koje predstavljaju oznaku koja odgovara određenoj memorijskoj adresi. Identifikatori se koriste umesto sirovih vrednosti adresa kako bi kôd bio čitljiviji i lakši za pisanje — na nivou assemblera se većinom koriste adrese ili automatski generisane oznake. Na slici 2.13 se može videti definicija identifikatora. Identifikator se sastoji od slova, cifara i simbola `_`, s tim što ne sme početi cifrom. Ovo je konvencija koju prati dosta jezika, uključujući programski jezik C. Primetimo da je identifikator nešto što bi lekser trebalo da prepozna tokom tokenizacije. Međutim, kada definišemo gramatiku od koje će ANTLR praviti lekser i parser, možemo i tokene definisati na isti način kao i gramatička pravila dajući regularni izraz za njihovo poklapanje. Listovi stabla parsiranja su uvek tokeni, drugim rečima se nazivaju i *terminalni simboli*. Tokeni se, osim u listovima, mogu naći bilo gde u stablu parsiranja.

```
1  NAME
2      : [a-zA-Z_][a-zA-Z_0-9]*
3      ;
```

Slika 2.13: Definicija identifikatora za pseudo-jezik.

Pošto želimo da pseudo-jezik bude strogo tipiziran, potreban je koncept tipa (videti definiciju deklaracije), čija je definicija data na slici 2.14. Tip može biti *primitivan* (drugim rečima *prost*) ili *složen*. Primitivni tipovi su podržani u samoj sintaksi jezika — u našem slučaju brojevni tipovi i niske. Brojevi mogu biti celi ili realni. U složene tipove spadaju korisnički definisani tipovi (sa imenom `NAME`, u četvrtoj alternativni pravila `typename` sa slike 2.14) i kolekcije. Od kolekcija su podržani nizovi, liste i skupovi. Prilikom definicije kolekcije mora se navesti tip elemenata kolekcije i taj tip mora biti uniforman — isti za sve elemente kolekcije.

Izrazi, iako definisani rekursivno, se mogu posmatrati kao kombinacija promenljivih, operatora i poziva funkcija sa odlikom da se mogu *evaluirati*, tj. moguće je izračunati njihovu vrednost. Iz definicije pravila `exp` sa slike 2.15, mogu se uočiti tipovi izraza, pri čemu nije vođeno računa o matematičkom prioritetu operatora, radi jednostavnosti. Izraz može biti *literal*, koji predstavlja konstantu, bilo brojevu, logičku ili nisku karaktera. Promenljive, definisane pravilom `var` su takođe izrazi, jer se trenutna vrednost promenljive posmatra kao vrednost izraza. Primetimo da promenljiva može biti kolekcijskog tipa, u kom slučaju se navodi redni broj elementa nakon identifikatora promenljive — taj redni broj može biti rezultat

```
1  type
2      : typename 'array'?
3      | typename 'list'?
4      | typename 'set'?
5      ;
6  typename
7      : 'integer'
8      | 'real'
9      | 'string'
10     | NAME
11     ;
```

Slika 2.14: Definicija tipa podataka za pseudo-jezik.

evaluacije drugog izraza, ali ne bilo kakvog, stoga se u pravilu `iexp` definiše šta sve može biti korišćeno da se indeksira element kolekcije. Izrazima se može dati prioritet pomoću zagrada, što se vidi u trećoj alternativni pravila `exp`. U naredne tri alternative su opisani tipovi izraza: aritmetički, relacioni i logički. Aritmetički izrazi su vezani aritmetičkim operatorima definisanim preko pravila `aop`, slično važi i za ostala dva tipa. Svi tipovi izraza navedeni iznad su binarni, što znači da operatori zahtevaju dva argumenta. Postoje i unarni izrazi, od kojih su podržane promena znaka i logička negacija, što se vidi iz pravila `uop`.

Definicija literala je prikazana na slici 2.16. Literali mogu bili istinitosne konstante `True` i `False`, brojeve konstante ili niske karaktera. Brojeve konstante mogu bili celobrojni ili realni dekadni brojevi. Realne konstante je moguće definisati u fiksnom ili pokretnom zarezu. Niske se mogu definisati između navodnika ili apostrofa. Pritom, kao i u modernim programskim jezicima, moguće je navesti sekvence koje predstavljaju specijalne karaktere kao što su novi red, tabulator itd. Oznaka `fragment` označava optimizaciju, naime nije potrebno da postoji na primer pravilo `Digit`, već samo dajemo simbol za regularni izraz koji će se koristiti u više drugih pravila i poklapati jednu dekadnu cifru.

Poslednje što treba definisati je sve ono što lekser treba da preskoči tokom prolaska kroz izvorni kôd programa. To su beline (nevidljivi karakteri kao što su razmaci, tabulatori i novi redovi) i komentari. Definicije ovih pravila se mogu videti na slici 2.17. Vidimo da se u njima koristi posebna oznaka `-> skip`, koja predstavlja instrukcije lekseru da preskoči sve ono što ovo pravilo poklopi. Komentari su u stilu kao u programskom jeziku C (ali naravno, isti stil se koristi

GLAVA 2. GENERISANJE I KORIŠĆENJE APSTRAKTNIH SINTAKSIČKIH STABALA

```
1  exp
2      : literal
3      | var
4      | '(' exp ')'
5      | exp aop exp
6      | exp rop exp
7      | exp lop exp
8      | uop exp
9      | cexp
10     ;
11  var
12     : NAME ('[' iexp ']')?
13     ;
14  iexp
15     : literal
16     | var
17     | aexp
18     ;
19  cexp
20     : 'call' NAME '(' explist? ')'
21     ;
22  explist
23     : exp (',' exp)*
24     ;
25  aop : '+' | '-' | '*' | '/' | 'div' | 'mod' ;
26  rop : '>' | '>=' | '<' | '<=' | '==' | '!=' ;
27  lop : 'and' | 'or' ;
28  uop : '-' | 'not' ;
```

Slika 2.15: Definicija izraza za pseudo-jezik.

i u mnogim jezicima) i mogu biti jednolinijski ili višelinijski. Beline koje treba preskočiti su definisane u pravilu WS, skraćeno od *whitespace*, što u prevodu sa engleskog znači *beli prostor*, *belina*.

Ovako definisanu gramatiku možemo sačuvati u fajl sa imenom `Pseudo.g4`, potrebno je samo navesti ime gramatike na početku fajla, kao na slici 2.18. Naredni korak je kreiranje leksera i parsera koristeći ANTLR4, predpostavljajući da je instaliran na način opisan u ???. Pokretanjem ANTLR-a generišemo lekser i parser za gramatiku pseudo-jezika:

```
1 $ antlr4 Pseudo.g4
```

ANTLR4 će generisati lekser i parser podrazumevano napisane u programskom jeziku Java. Ukoliko želimo to da promenimo, možemo koristiti opciju

GLAVA 2. GENERISANJE I KORIŠĆENJE APSTRAKTNIH SINTAKSIČKIH STABALA

```
1 literal : 'True' | 'False' | INT | FLOAT | STRING ;
2 STRING : '"' ( EscapeSequence | ~('\\"'|"'') )* '"' ;
3 INT : Digit+ ;
4 FLOAT
5     : Digit+ '.' Digit* ExponentPart?
6     | '.' Digit+ ExponentPart?
7     | Digit+ ExponentPart
8     ;
9
10 fragment
11 ExponentPart : [eE] [+-]? Digit+ ;
12 fragment
13 Digit : [0-9] ;
14 fragment
15 EscapeSequence : '\\\' [abfnrtvz"'\\" | '\\\' '\\r'? '\\n' ;
```

Slika 2.16: Definicija konstanti za pseudo-jezik.

```
1 BlockComment
2     :   '/*' .*? '*/'   -> skip
3     ;
4 LineComment
5     :   '//' ~[\r\n]*   -> skip
6     ;
7 WS
8     :   [ \t\u000C\r\n]+ -> skip
9     ;
```

Slika 2.17: Definicija komentara i belina za pseudo-jezik.

```
1 grammar Pseudo;
```

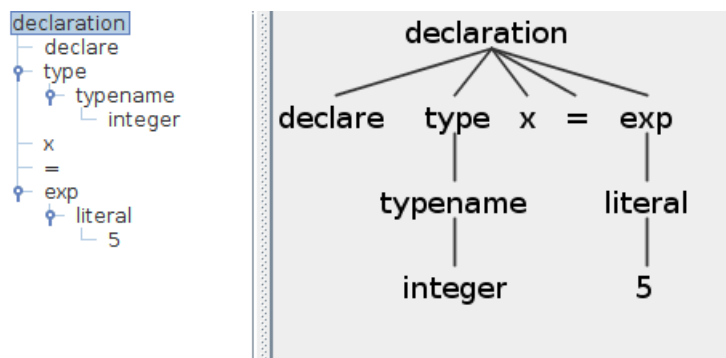
Slika 2.18: Definicija imena gramatike za pseudo-jezik.

-Dlanguage=.... Kako bismo testirali generisani lekser i parser, možemo koristiti ANTLR TestRig da vizualno prikažemo stablo parsiranja, s tim što moramo prvo kompajlirati generisane Java kôdove. TestRig pozivamo navođenjem ime gramatike (koje se poklapa sa imenom leksera i parsera) i imenom pravila od koga će parser krenuti. Opcija -gui pokreće vizualni prikaz stabla parsiranja pokazan na slici 2.19 (vizualni prikaz je moguće preskočiti i samo ispisati stablo u LISP

GLAVA 2. GENERISANJE I KORIŠĆENJE APSTRAKTHNIH SINTAKSIČKIH STABALA

formi koristeći opciju `-tree`), mada je moguće i ispisati samo tokene koristeći opciju `-tokens`. Ulaz se prosleđuje programu dok se ne nađe na simbol EOF, ili alternativno se može preneti ulaz putem UNIX pipeline-a (na slici 2.19 se može videti izlaz koji se dobija korišćenjem opcije `-gui`):

```
1 $ javac *.java
2 $ echo "declare integer x = 5" | grun Pseudo declaration -tokens
3 [@0,0:6='declare',<'declare'>,1:0]
4 [@1,8:14='integer',<'integer'>,1:8]
5 [@2,16:16='x',<NAME>,1:16]
6 [@3,18:18='=',<'='>,1:18]
7 [@4,20:20='5',<INT>,1:20]
8 [@5,22:21='<EOF>',<EOF>,2:0]
9 $ echo "declare integer x = 5" | grun Pseudo declaration -tree
10 (declaration declare (type (typename integer)) x = (exp (literal
    5)))
11 $ echo "declare integer x = 5" | grun Pseudo declaration -gui
```



Slika 2.19: Grafički prikaz stabla parsiranja koje generiše parser kreiran od strane alata ANTLR4 TestRig za kôd pisan u pseudo-jeziku.

Obilazak stabla parsiranja

ANTLR, osim leksera i parsera za datu gramatiku, može da kreira interfejs i bazne klase koji prate obrasce za projektovanje *posetilac* (engl. *visitor*) i osluškivač (engl. *listener*)⁸ opisane u 2.3. Tako kreirani interfejsi i klase imaju metode za obilazak stabla parsiranja. ANTLR podrazumevano generiše interfejs osluškivača (slika 2.20) kao i baznu klasu koja implementira generisani interfejs tako što su

⁸Osluškivač je varijanta obrasca *posmatrač* (engl. *observer*)

GLAVA 2. GENERISANJE I KORIŠĆENJE APSTRAKTHNIH SINTAKSIČKIH STABALA

sve implementirane metode prazne. Stoga, ukoliko korisnik želi da definiše operaciju samo u slučaju da se prilikom obilaska stabla parsiranja naiđe na određeni tip čvora, nije potrebno implementirati ceo interfejs osluškivača već je moguće naslediti baznu klasu i predefinisati samo jedan metod. ANTLR može da generiše i posetilac (slika 2.21), ukoliko se navede odgovarajuća opcija `-visitor` prilikom pokretanja. Slično, ukoliko nije potrebno generisati osluškivač, može se koristiti opcija `-no-listener`.

```
1 public interface IPseudoListener : IParseTreeListener
2 {
3     void EnterUnit([NotNull] PseudoParser.UnitContext context);
4     void ExitUnit([NotNull] PseudoParser.UnitContext context);
5     void EnterBlock([NotNull] PseudoParser.BlockContext context);
6     void ExitBlock([NotNull] PseudoParser.BlockContext context);
7     void EnterStatement([NotNull] PseudoParser.StatementContext
8         context);
9     void ExitStatement([NotNull] PseudoParser.StatementContext
10        context);
11
12     ...
13 }
```

Slika 2.20: Delimični prikaz interfejsa osluškivača generisanog od strane ANTLR4 za pseudo-jezik definisan u prethodnom odeljku (C#).

Sa slike 2.20 se vidi da je moguće definisati metode koje će se pozivati prilikom ulaska ali i prilikom izlaska iz čvora određenog tipa prilikom obilaska stabla parsiranja. Pritom je važno kako se stablo obilazi. U slučaju ANTLR, to je pretraga u dubinu (engl. *depth-first search*, *DFS*)⁹, stoga će se metod `Exit` za proizvoljni čvor pozvati tek kad se obiđu sva deca tog čvora — dakle nakon poziva njihovih `Enter` i `Exit` metoda. Pošto se DFS obično implementira putem LIFO strukture¹⁰, može se reći da se `Enter` metod poziva onog trenutka kad se čvor ubaci u strukturu, a `Exit` metod onda kada se čvor ukloni iz strukture.

⁹DFS je obilazak stabla takav da se obilazak duž grane stabla nastavlja sve dok je moguće ići dublje, a ako to nije moguće vratiti se unazad i obići druge grane.

¹⁰*Last In, First Out* struktura podataka je apstraktna struktura podataka sa operacijama ubacivanja i izbacivanja elemenata, pri čemu je element koji se izbacuje onaj koji je poslednji ubačen. Primer LIFO strukture je držač za CD-ove — ne mogu se ukloniti CD-ovi ispod CD-a na vrhu (poslednji ubačen) a da se ne ukloni isti. U slučaju opisanom iznad, implementacija LIFO strukture se naziva stek *stack*.

GLAVA 2. GENERISANJE I KORIŠĆENJE APSTRAKTHNIH SINTAKSIČKIH STABALA

```
1 public interface IPseudoVisitor<T> : IParseTreeVisitor<T>
2 {
3     T VisitUnit([NotNull] PseudoParser.UnitContext context);
4     T VisitBlock([NotNull] PseudoParser.BlockContext context);
5     T VisitStatement([NotNull] PseudoParser.StatementContext
6         context);
7     T VisitDeclaration([NotNull] PseudoParser.DeclarationContext
8         context);
9     ...
10 }
```

Slika 2.21: Delimični prikaz interfejsa posetioca generisanog od strane ANTLR4 za pseudo-jezik definisan u prethodnom odeljku (C#).

Za razliku od osluškivača, posetilac je prirodnije koristiti ukoliko je potrebno izvršiti neko izračunavanje nad strukturom koja se obilazi. Interfejs posetioca (slika 2.21) je šablonski, i metodi imaju povratnu vrednost šablonskog tipa za razliku od metoda osluškivača i, u odnosu na osluškivač, nema para metoda za svaki čvor već samo jedan metod. Dodatna razlika, ali i najveća, je ta što se metodi posetioca ne pozivaju automatski. Stoga je na programeru da nastavi obilazak i da odluči u koje čvorove želi da se spusti. Jasno je da i osluškivač i posetilac imaju svoje primene — ukoliko je potrebno obići stablo parsiranja i dovući neke informacije može se iskoristiti osluškivač jer onda ne moramo brinuti o obilasku. S druge strane, ukoliko je potrebno izračunati neku vrednost prirodno je iskoristiti rekursiju i iskoristiti posetilac — rekursivni pozivi prilikom obilaska nam idu u prilog jer koristimo povratne vrednosti tih metoda da gradimo rezultat od listova ka korenu stabla parsiranja. U nastavku će se koristiti posetilac zbog kontrole obilaska ali i činjenice da se stablo parsiranja obilazi sa ciljem da se izgradi AST, koji je takođe rekursivna struktura i gradi se inkrementalno kroz rekursiju.

Bilo da se koristi osluškivač ili posetilac, potrebno je nekako proslediti informacije o samom čvoru na koji se naišlo tokom obilaska stabla parsiranja. Te informacije se metodima osluškivača i posetioca prosleđuju putem potklasa apstrakne klase konteksta pravila `ParserRuleContext` — u primeru iznad `UnitContext`, `BlockContext` itd. Svaki kontekst pravila po imenu odgovara pravilima definisanim u gramatici i sadrži informacije bitne za trenutni čvor u stablu parsiranja koji odgovara tipu konteksta. Takođe, u svakom kontekstu su prisutne i metode čija imena odgovaraju pravilima koja se javljaju u definiciji samog pravila koje odgo-

GLAVA 2. GENERISANJE I KORIŠĆENJE APSTRAKTHNIH SINTAKSIČKIH STABALA

vara kontekstu. Tako da, za `BlockContext`, imajući u vidu definiciju sa slike 2.10, pošto se u definiciji osim tokena koristi i pravilo `statement`, u okviru `BlockContext` klase biće implementiran i metod `statement()` koji vraća kontekst pravila u ovom slučaju tipa `StatementContext[]` jer u prvoj alternativu stoji `statement+` — dakle možemo imati više `statement` poklapanja. Sa ovim u vidu, moguće je odrediti kako će se obilazak nastaviti (u slučaju posetioca) ili dovući informacije o delovima definicije pravila. Ukoliko pravilo ima više alternativa, metodi koje vraćaju kontekst pravila koje figuriše u alternativu koja nije korišćena za poklapanje pravila će vratiti `null`. Pošto se `statement()` pravilo javlja u obe alternative pravila `block` (i nije opciono), možemo biti sigurni da povratna vrednost `statement()` metoda nikada neće biti `null`.

U poglavlju 4 će se koristiti posetilac za obilazak stabla parsiranja i kreiranja AST apstrakcije od istog. Pritom, koristiće se implementacija posetioca u programskom jeziku C#.

2.3 Korišćenje generisanih stabala

Kako bi se stabla parsiranja i apstraktna sintaksička stabla mogla koristiti, potrebno je pružiti i uniformni interfejs za njihov obilazak. Postoje situacije kada se stablo obilazi sa ciljem izvršavanja operacija prilikom ulaska ili izlaska iz čvorova određenog tipa, ili pak sa ciljem izračunavanja neke konkretne vrednosti. Prilikom razvoja softvera se često nailazi na ovakve probleme i stoga su kreirana ponovno upotrebljiva rešenja za te probleme.

Obrasci za projektovanje (engl. *design patterns* [7], drugačije nazvani i *projektni šabloni*, *uzorci*) predstavljaju opšte i ponovno upotrebljivo rešenje čestog problema, obično implementirani kroz koncepte objektno-orijentisanog programiranja. Svaki obrazac za projektovanje ima četiri osnovna elementa:

- ime — ukratko opisuje problem, rešenje i posledice
- problem — opisuje slučaj u kome se obrazac koristi
- rešenje — opisuje elemente dizajna i odnos tih elemenata
- posledice — obuhvataju rezultate i ocene primena obrasca

GLAVA 2. GENERISANJE I KORIŠĆENJE APSTRAKTHNIH SINTAKSIČKIH STABALA

Obrasce za projektovanje je moguće grupisati po situaciji u kojoj se mogu iskoristiti ili načinu na koji rešavaju zadati problem. Stoga je opšte prihvaćena podela na tri grupe:

- *gradivni obrasci* (engl. *creational patterns*)
- *strukturni obrasci* (engl. *structural patterns*)
- *obraci ponašanja* (engl. *behavioral patterns*)

Gradivni obrasci apstrahuju proces pravljenja objekata i važni su kada sistemi više zavise od sastavljanja objekata nego od nasleđivanja. Neki od najvažnijih gradivnih obrazaca su *apstraktna fabrika* (engl. *abstract factory*), *graditelj* (engl. *builder*), *proizvodni metod* (engl. *factory method*), *prototip* (engl. *prototype*) i *unikat* (engl. *singleton*). Strukturni obrasci se bave načinom na koji se klase i objekti sastavljaju u veće strukture. Neki od najvažnijih strukturnih obrazaca su *adapter* (engl. *adapter*), *most* (engl. *bridge*), *sastav* (engl. *composite*), *dekorater* (engl. *decorator*), *fasada* (engl. *facade*), *muva* (engl. *flyweight*) i *proksi* (engl. *proxy*). Obrasci ponašanja se bave načinom na koji se klase i objekti sastavljaju u veće strukture. Neki od najvažnijih strukturnih obrazaca su *lanac odgovornosti* (engl. *chain of responsibility*), *komanda* (engl. *command*), *interpretator* (engl. *interpreter*), *iterator* (engl. *iterator*), *posmatrač* (engl. *observer*), *strategija* (engl. *strategy*) i *posetilac* (engl. *visitor*).

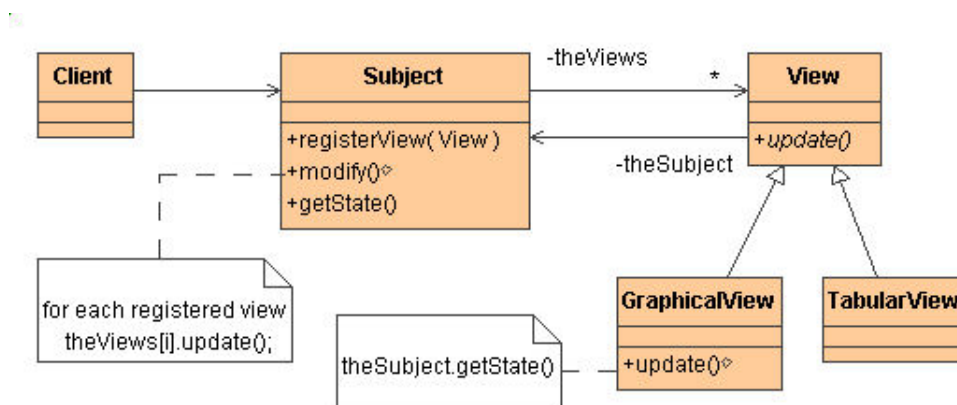
Za potrebe ovog rada, obrasci za projektovanje će se koristiti kao opšte prihvaćeno i programerski intuitivno rešenje određenih problema. Takođe, u kontekstu stabala parsiranja i apstraktnih sintaksičkih stabala opisanih u poglavlju 2.1, obrasci *Posmatrač* i *Posetilac* su od velikog značaja jer pružaju interfejs za obilazak takvih stabala. Ovi obrasci, opisani u narednim odeljcima, se koriste od strane alata korišćenih u ovom radu kao što je ANTLR koji je opisan u odeljku 2.2. Takođe, s obzirom da su ovi obrasci opšte-prihvaćeno rešenje za pružanje interfejsa obilaska stabala, biće korišćeni i u implementaciji opšte apstrakcije. U nastavku će zbog opisanih razloga biti opisani samo obrasci posmatrač i posetilac, dok zainteresovani čitalac može pročitati više u [9].

Obrazac „Posmatrač”

Obrazac za projektovanje *Posmatrač* je strukturni obrazac za projektovanje koji se koristi kada je potrebno definisati jedan-ka-više vezu između objekata tako

GLAVA 2. GENERISANJE I KORIŠĆENJE APSTRAKTHNIH SINTAKSIČKIH STABALA

da ukoliko jedan objekat promeni stanje (subjekat) svi zavisni objekti su obavješteni o izmeni i shodno ažurirani. Posmatrač predstavlja *pogled* (engl. *View*) u MVC (engl. *Model-View-Controller*) arhitekturi. Na slici 2.22 se može videti UML dijagram [24] ovog obrasca.



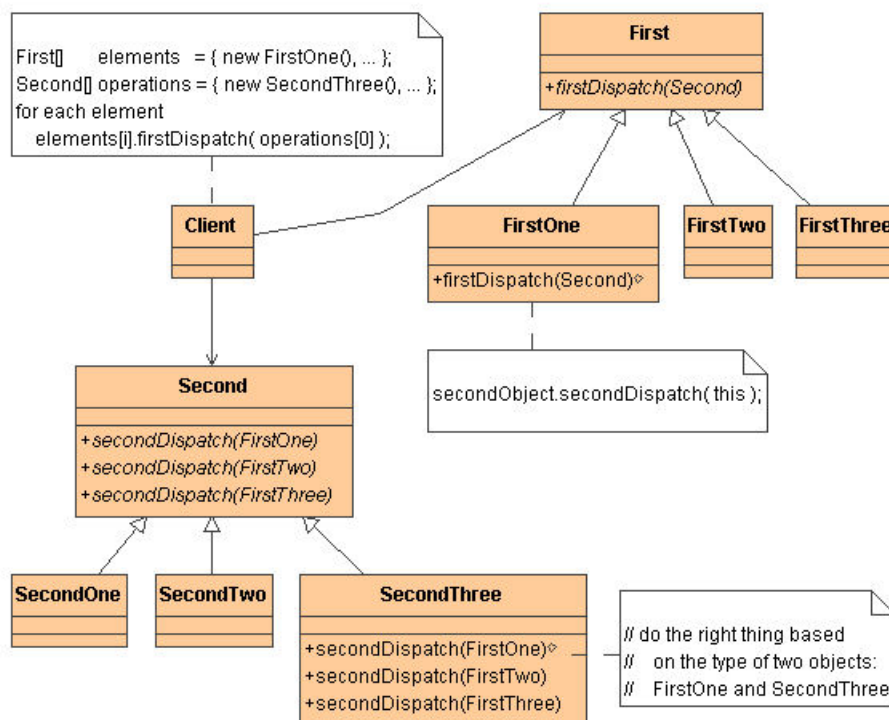
Slika 2.22: UML dijagram obrasca za projektovanje „Posmatrač”.

Primer upotrebe ovog obrasca može biti aukcija gde je aukcionar subjekat i započinje aukciju, dok učesnici aukcije (objekti) posmatraju aukcionera i reaguju na podizanje cene. Prihvatanje promene cene menja trenutnu cenu i aukcioner oglašava promenu iste, a svi učesnici aukcije dobijaju informaciju da se izmena izvršila. Za potrebe ovog rada, primer upotrebe može biti obilazak stablolike kolekcije (recimo stabla parsiranja) i obaveštavanje o nailasku na čvorove određenih tipova. Te informacije se dalje mogu iskoristiti za izračunavanja nad pomenutom strukturom ili generisanje novih struktura (recimo AST).

Obrazac „Posetilac”

Obrazac za projektovanje *Posetilac* je strukturni obrazac za projektovanje koji predstavlja operaciju koju je potrebno izvesti nad elementima objektne strukture. Posetilac omogućava definisanje nove operacije bez izmena klasa elemenata nad kojima operiše. Operacija koja će se izvesti zavisi od imena zahteva, tipa posetioca i tipa elementa kog posećuje. Na slici 2.23 se može videti UML dijagram [24] ovog obrasca.

Primer upotrebe ovog obrasca može biti operisanje taksi kompanija. Kada osoba pozove taksi kompaniju (prihvatanje posetioca), kompanija šalje vozilo osobi



Slika 2.23: UML dijagram obrasca za projektovanje „Posetilac”.

koja je pozvala kompaniju. Nakon ulaska u vozilo (posetilac), mušterija ne kontroliše svoj transport već je to u rukama taksiste (posetioca). Za potrebe ovog rada, primer upotrebe može biti prikupljanje informacija o kolekciji stablolike strukture (recimo stablo parsiranja) i korišćenje istih za neko izračunavanje ili generisanje novih struktura (recimo AST).

2.4 Programske paradigme i gramatičke razlike programskih jezika

Iako se u suštini svode na mašinski jezik ili assembler, viši programski jezici mogu imati velike razlike međusobno — kako u načinu pisanja koda, tako i u efikasnosti izvršavanja. Način, ili stil programiranja se naziva *programska paradigma* [20]. Može se pokazati da sve što je rešivo putem jedne, može i da se reši i putem ostalih; međutim neki problemi se prirodnije rešavaju koristeći specifične paradigme. Neke poznatije programske paradigme su navedene u nastavku zajedno sa njihovim odlikama i primerima upotrebe.

Imperativna paradigma

Imperativna paradigma pretpostavlja da se promene u trenutnom stanju izvršavanja mogu sačuvati kroz promenljive. Izračunavanja se vrše putem niza koraka, u svakom koraku se te promenljive referišu ili se menjaju njihove trenutne vrednosti. Raspored koraka je bitan, jer svaki korak može imati različite posledice s obzirom na trenutne vrednosti promenljivih na početku tog koraka. Primer koda pisanog po imperativnoj paradigmi se može videti na slici 2.24.

```
1      result = []
2      i = 0
3  start:
4      numPeople = length(people)
5      if i >= numPeople goto finished
6      p = people[i]
7      nameLength = length(p.name)
8      if nameLength <= 5 goto nextOne
9      upperName = toUpper(p.name)
10     addToList(result, upperName)
11  nextOne:
12     i = i + 1
13     goto start
14  finished:
15     return sort(result)
```

Slika 2.24: Primer koda pisanog po imperativnoj paradigmi.

Stariji programski jezici najčešće prate ovu paradigmu više nego bilo koju drugu iz par razloga. Prvi je taj što imperativna paradigma najbliže oslikava samu mašinu na kojoj se program izvršava, pa je programer mnogo „bliži” istoj. Posledica ovog pristupa, a to je i drugi razlog za popularnost imperativne paradigme, je omogućila da je ova paradigma bila najefikasnija zbog ograničenja u hardveru. Danas, zbog mnogo bržeg razvoja i mnogo jačih računara, efikasnost se sve manje i manje uzima u obzir.

Naravno, imperativna paradigma ima i svoje nedostatke. Naime, najveći problem je razumevanje i verifikovanje semantike programa zbog postojanja sporednih efekata ¹¹. Stoga je i pronalaženje grešaka u kodovima koji prate imperativnu

¹¹Sporedni efekti (promena stanja mašine) ne poštuju *referencijalnu transparentnost* koja se definiše na sledeći način: *Ako važi $P(x)$ i $x = y$ u nekom trenutku, onda $P(x) = P(y)$ važi tokom čitavog vremena izvršavanja programa.*

GLAVA 2. GENERISANJE I KORIŠĆENJE APSTRAKTHNIH SINTAKSIČKIH STABALA

paradigmu znatno komplikovanije. Apstrakcija je takođe više ograničena u imperativnoj nego u ostalim paradigmama. Na kraju, redosled izvršavanja je vrlo bitan, što neke probleme čini težim ukoliko se pokušaju rešiti pomoću imperativne paradigme.

Strukturna paradigma

Strukturna paradigma je vrsta imperativne paradigme gde se kontrola toka vrši putem ugnježenih petlji, uslovnih grananja i podrutina. Promenljive su obično lokalne za blok u kome su definisane, što određuje i njihov životni vek i vidljivost. Primer koda pisanog po imperativnoj paradigmi se može videti na slici 2.25. Najpopularniji derivat strukturne paradigme je *proceduralna paradigma*, bazirana na konceptu poziva *procedure* — podrutine ili funkcije koja sadrži seriju koraka koje je potrebno izvršiti redom.

```
1 result = [];  
2 for (i = 0; i < length(people); i++) {  
3     p = people[i];  
4     if (length(p.name)) > 5 {  
5         addToList(result, toUpper(p.name));  
6     }  
7 }  
8 return sort(result);
```

Slika 2.25: Primer koda pisanog po strukturnoj paradigmi.

Logička paradigma

Logička paradigma koristi deklarativni pristup rešavanju problema. Umesto zadavanja instrukcija koje treba da dovedu do rezultata, opisuje se sam rezultat kroz činjenice — skup logičkih pretpostavki. Taj opis se zatim prevodi u upit koji se dalje koristi. Uloga računara je održavanje i logička dedukcija. Logička paradigma se deli u tri sekcije:

- niz deklaracija i definicija koje opisuju problem iz nekog domena,
- relevantne činjenice i
- relevantni ciljevi u formi upita.

GLAVA 2. GENERISANJE I KORIŠĆENJE APSTRAKTNIH SINTAKSIČKIH STABALA

Bilo koji rezultat dedukcije rešenja upita predstavlja rezultat izvršavanja. Deklaracije i definicije se konstruišu iz relacija, npr. $X \in Y$ ili $X \in [a, b]$. Prednost ovog pristupa je mala količina programiranja, pošto dedukcioni sistem traži rešenje problema. Takođe, verifikovanje validnosti je stoga trivijalno. Primer koda pisanog po logičkoj paradigmi se može videti na slici 2.26.

```
1 domains
2     being = symbol
3 predicates
4     animal(being)      % sve životinje su ziva bica
5     dog(being)         % svi psi su ziva bica
6     die(being)         % svi ziva bica umiru
7 clauses
8     animal(X) :- dog(X) % svi psi su zivotinje
9     dog(fido).         % fido je pas
10    die(X) :- animal(X) % sve životinje umiru
```

Slika 2.26: Primer koda pisanog po logičkoj paradigmi.

Funkcionalna paradigma

Funkcionalna paradigma posmatra sve potprograme kao funkcije u matematičkom smislu — uzimaju argumente i vraćaju jedinstven rezultat. Povratna vrednost zavisi isključivo od argumenata, što znači da je nebitan trenutak u kom je funkcija pozvana. Izračunvanja se vrše primenom aplikacije funkcija, kompozicijom funkcija i redukcijom. Primer koda pisanog po funkcionalnoj paradigmi se može videti na slici 2.27.

```
1 people
2     |> map      (extract_name . to_upper)
3     |> filter  (\name -> length name > 5)
4     |> sort
5     |> take 5
6     |> join  ", "
```

Slika 2.27: Primer koda pisanog po funkcionalnoj paradigmi.

Funkcionalni programski jezici se baziraju na funkcionalnoj paradigmi. Takvi jezici dozvoljavaju tretiranje funkcija kao *građana prvog reda* — mogu biti tretirane

kao podaci pa se mogu proslediti drugim funkcijama ili vratiti kao rezultat izračunavanja drugih funkcija. Prednosti funkcionalnih jezika su visok nivo apstrakcije, što prevazilazi mnogo detalja programiranja i stoga eliminiše pojavu velikog broja grešaka, nezavisnost od redosleda izračunavanja, što omogućava paralelizam, i formalnu matematičku verifikaciju. Mane su potencijalna manja efikasnost, što danas predstavlja manji problem, kao i teškoća implementacije specifične sekvencijalne aktivnosti ili potreba za stanjem, što bi se lako implementiralo imperativno ili preko OO paradigme.

Objektno-orijentisana paradigma

Objektno-orijentisana paradigma (kraće *OO*) je paradigma u kojoj se objekti stvarnog sveta posmatraju kao zasebni entiteti koji imaju sopstveno stanje koje se modifikuje samo pomoću procedura ugrađenih u same objekte — tzv. *metode*. Posledica zasebnog operisanja objekata omogućava njihovu enkapsulaciju u module koji sadrže lokalnu sredinu i metode. Komunikacija sa objektom se vrši prosleđivanjem poruka. Objekti su organizovani u klase, od kojih nasleđuju metode i ekvivalentne promenljive. *OO* omogućava ponovnu iskorišćenost koda i ekstenzibilnost koda. Primer koda po strukturnoj paradigmi je dat na slici 2.28.

Nova klasa (*A*) može *naslediti* ili *konkretizovati* drugu klasu (*B*). *B* se zove *bazna klasa* ili *natklasa*, dok se *A* naziva *izvedena klasa* ili *potklasa*. Izvedena klasa nasleđuje sve odlike bazne klase — strukturu i ponašanje — postaje specijalizacija bazne klase dok bazna klasa postaje generalizacija svoje potklase. Osim nasleđenih osobina, potklasa može imati dodatna stanja (instancne promenljive) ili dodatna ponašanja (metode). Dozvoljeno je i predefinisanje ponašanja bazne klase. Mehanizam nasleđivanja je dozvoljen i ukoliko nije dozvoljen pristup izvornom kodu bazne klase.

Idealno, stanju objekta može da pristupi i modifikuje samo pomoću metoda tog objekta. Većina OO jezika dozvoljava direktnu manipulaciju stanja ali taj pristup nije preporučen. Kako bi se enkapsulacija i skrivanje informacija kao najveće prednosti OO paradigme mogle iskoristiti, interfejs klase (kako se pristupa objektima) bi trebalo da bude odvojen od implementacije klase (izvornog koda metoda klase).

```
1  abstract class Employee
2  {
3      private String name;
4
5      Employee(String name) {
6          this.name = name;
7      }
8
9      abstract void work();
10 };
11
12 class WageEmployee extends Employee
13 {
14     private double wage;
15     private double hours;
16
17     WageEmployee(String name) {
18         super(name);
19         this.name = name;
20     }
21
22     void work() {
23         wage += 200;
24         hours += 8;
25     }
26 };
27
28 var bill = new WageEmployee("Bill Gates");
29 bill.work();
```

Slika 2.28: Primer koda pisanog po OO paradigmi u programskom jeziku Java.

Skript paradigma i njen odnos sa proceduralnom paradigmom

Čak i unutar jedne paradigme kao što je proceduralna, mogu se naći veoma velike varijacije u izgledu koda pisanog u različitim programskim jezicima koji prate proceduralnu paradigmu. Kako hardver postaje moćniji, više se ceni vreme koje programer provede u procesu pisanja koda nego koliko je taj kod efikasan. Štaviše, u nekim slučajevima je dobitak u efikasnosti veoma mali u poređenju sa vremenom koje je potrebno utrošiti da bi se ta efikasnost postigla. Ukoliko se program pokreće veoma retko, možda nije ni bitno da li se on izvršava sekundu sporije od efikasnog programa, ako je za njegovo pisanje utrošeno znatno manje vremena. Ovo je pristup koji prate *skript* jezici kao što su Python, Perl, bash

GLAVA 2. GENERISANJE I KORIŠĆENJE APSTRAKTNIH SINTAKSIČKIH STABALA

itd. Iako proceduralni, oni se razlikuju od klasičnih predstavnika proceduralne paradigme i njihove razlike su vremenom postale tolike da nije neuobičajeno da se skript jezici svrstaju u zasebnu, *skript paradigmu*. Stoga će se u nastavku, pod terminom *proceduralni jezik* smatrati tradicionalni proceduralni jezik, ukoliko nije naznačeno drugačije. Na slici 2.29 se mogu uočiti navedene razlike.

```
1  int main() {
2      int k = 0;
3      for (int i = 0; i < 1000000; i++)
4          k++;
5      return 0;
6  }
```

```
1  $ time: 0.03s user 0.00s system 70% cpu 0.044 total
```

```
1  k = 0
2  for i in range(1000000):
3      k += 1
```

```
1  $ time: 0.16s user 0.03s system 93% cpu 0.200 total
```

Slika 2.29: Primer koda pisanog po tradicionalnoj proceduralnoj paradigmi (gore, C) i po modernoj skript paradigmi (gore, Python 3) kao i odgovarajuća vremena izvršavanja dobijena komandom `time`.

Promenljive predstavljaju jedan od osnovnih koncepata na kojem se zasnivaju i proceduralni i skript jezici. Promenljivu odlikuje, između ostalog, i njen *tip* koji određuje količinu memorije potrebnu za njeno skladištenje. Proceduralni programski jezici zahtevaju definisanje tipa promenljive i obično su i *statički*, što znači da promenljive ne mogu menjati svoj tip tokom izvršavanja programa. Proces uvođenja imena promenljive se u naziva *deklaracija promenljive*. Slično kao i za promenljive, potrebno je deklarirati i funkcije pre trenutka njihovog korišćenja kako bi prevodilac znao broj i tipove parametara funkcije kao i njihove povratne vrednosti. Skript jezici žrtvuju strogu tipiziranost kako bi proces pisanja koda bio brži. Stoga su oni obično *dinamički* — promenljive mogu menjati tip tokom izvršavanja programa. Pošto promenljive mogu menjati svoj tip, definisanje ti-

GLAVA 2. GENERISANJE I KORIŠĆENJE APSTRAKTNIH SINTAKSIČKIH STABALA

pa prilikom uvođenja imena promenljive postaje redundantno jer prevodilac može to sam da zaključi. Stoga i sam proces uvođenja imena promenljive postaje redundantan. Slično, parametri funkcija takođe nisu fiksnog tipa. Slično važi i za povratnu vrednost funkcije.

Kod proceduralnih jezika, pošto su obično strogo tipizirani, mogu se iskoristiti strukture podataka koje omogućavaju brz pristup svojim elementirama. To su obično nizovi koji predstavljaju kontinualni blok memorije u kom su elementi niza smešteni jedan do drugog. Pristup se vrši na osnovu indeksa *i*, pošto su svi elementi istog tipa (zauzimaju jednaku količinu memorije), može se u konstantnom vremenu izračunati memorijska lokacija na kojoj se nalazi element niza sa datim indeksom. Kompleksnije strukture podataka obično nisu podržane u samom jeziku. Neki proceduralni jezici dozvoljavaju veoma niski pristup kroz *pokazivače* ili *reference* na memorijske adrese (C i C++). Većina modernih proceduralnih jezika ne dozvoljava rad sa pokazivačima, ne brinući puno o efikasnosti, dok neki dozvoljavaju korišćenje pokazivača u specijalnim situacijama sa eksplicitnom naznakom (C#).

Pored dinamičnosti kad je u pitanju tip promenljivih, skript jezici često imaju neke specifične strukture podataka ugrađene u sam jezik kao olakšice prilikom programiranja. Primarna struktura podataka je *jednostruko ulančana lista* ¹², za razliku od niza kod proceduralnih jezika. Razlog zašto se koriste liste je delimično zbog toga što, kao i ostale promenljive, liste nisu strogo tipizirane. Moguće je u listu ubacivati elemente različitih tipova — što onemogućava skladištenje u kontinualnom bloku memorije (osim ukoliko je lista imutabilna, što nije obično slučaj). Skript jezici uglavnom omogućavaju indeksni pristup elementima liste pa programeru izgleda kao da radi nad običnim nizom. Neki skript jezici omogućavaju kreiranje *asocijativnih nizova*, gde indeks niza ne mora biti ceo broj već može uzimati vrednost iz domena bilo kog tipa. Osim listi, obično su podržane i torke, i za njih važe iste slobode kao i za liste. Kompleksnije strukture podataka uključuju skupove i rečnike (drugačije nazivane i *heš mape*, engl. *hash map*) koji su kolekcija ključ-vrednost parova gde je dozvoljen indeksni pristup po vrednosti ključa. Skript programski jezici su skoro uvek interpretirani, iako se neki jezici mogu kompajlirati po potrebi za efikasnije ponovno izvršavanje. S obzirom da efikasnost nije u glavnom planu, u skript jezicima nije dozvoljen direktan pristup

¹²Lista je rekurzivna kolekcija podataka koja se sastoji od glave koja sadrži vrednost određenog tipa, i pokazivača na rep — drugu listu. Specijalno, praznim pokazivačima se označava kraj liste (prazna lista).

*GLAVA 2. GENERISANJE I KORIŠĆENJE APSTRAKTNIH
SINTAKSIČKIH STABALA*

memoriji putem pokazivača ili referenci.

Glava 3

Simboličko izračunavanje

Današnji softver je veoma komplikovan i često funkcioniše na različitim nivoima arhitekture velikih projekata. Stoga je proces verifikacije softvera danas veoma značajan i delikatan. U procesu verifikacije se najčešće koriste ručno pisani testovi i pregledi koda od strane drugih programera. Uprkos svim ovim merama, greške su i dalje nezaobilazne — jedan test može proveriti ponašanje koda za samo jedan ulaz. S obzirom da je nemoguće testirati sve ulaze zbog njihovog ogromnog broja (ukoliko posmatramo samo funkciju jedne promenljive koja prima 32-bitni ceo broj, broj mogućih ulaza je 2^{32}) potrebno je da testovi dobro *generalizuju* — da pokrivaju opšte ali i neke specijalne ulaze. To se postiže uočavanjem da se vrednosti ulaza mogu razvrstati u klase po tome kakav izlaz uzrokuju. Ukoliko imamo funkciju koja deli dva broja, te klase mogu biti celi brojevi, realni brojevi, neke specijalne vrednosti specifične za operaciju deljenja (recimo 0), kao i granice za tip podataka iz čijeg domena argumenti funkcije mogu uzeti vrednost. Čak i ovakav pristup, iako drastično smanjuje broj testova i eliminiše redundantne testove, i dalje zahteva relativno veliki broj testova u slučaju većih projekata i stoga je teško pronaći sve greške, pogotovo u slučajevima koji se retko dešavaju i ako ispoljavanje istih zavisi od stanja drugih komponenti ili pak nekih nedeterminističkih ponašanja samog sistema. Poželjno je čitav izvorni kôd pokriti testovima (engl. *code coverage*) — iako dostignuta pokrivenost koda od 100% i dalje ne znači da taj kôd ispravno radi.

Statička analiza koda predstavlja analizu izvornog koda bez pokretanja istog sa ciljem ispitivanja stanja u kojima se može naći program i proveru rada jedinice koja se testira za mnogobrojne ulaze. Iako teorijski valjana, u praksi nailazi na puno problema — osnovni je razlika u apstrakcijama koju prave statički analizator

i programer, što otežava korišćenje ovakve analize u praksi.

Simboličko izvršavanje [23] predstavlja sredinu između klasične verifikacije putem pisanja testova i statičke analize koda. Prilikom simboličkog izvršavanja, umesto stvarnih vrednosti ulaza koriste se *simboličke promenljive*. Simbolička promenljiva nije vezana za specifičnu vrednost i analiza se dalje vrši samo nad njom — samim tim se istovremeno mogu testirati višestruke klase sličnih ulaza.

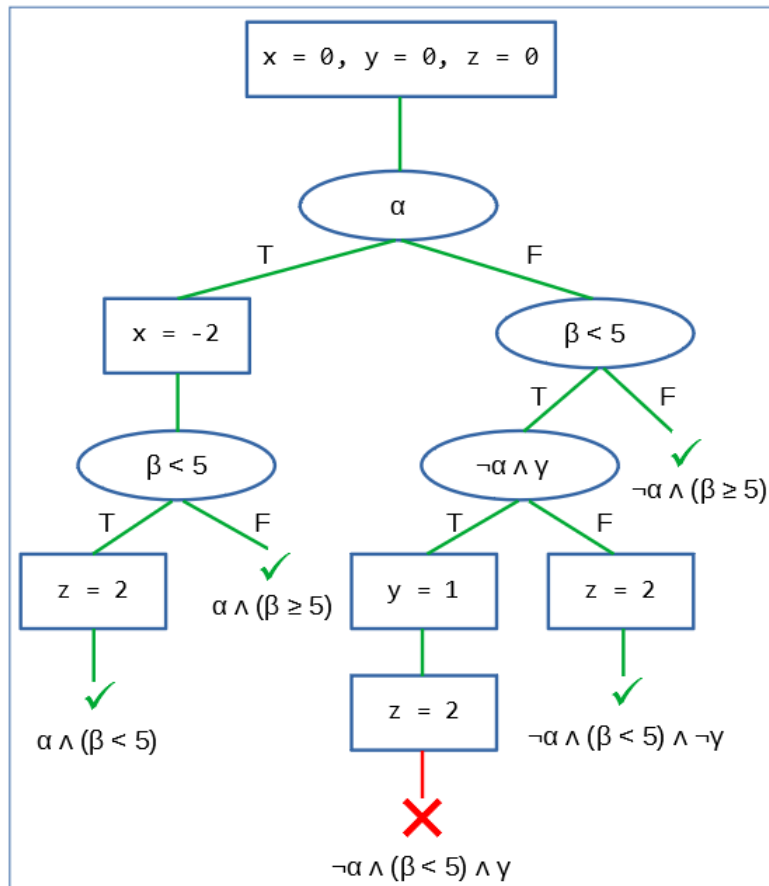
Primer simboličkog izvršavanja će biti opisan na isečku C koda sa slike 3.1. Pretpostavimo da imamo deklarisanje promenljive a , b i c i da se neke operacije izvršavaju nad njima, reprezentovano komentarom u liniji 3. U nekom trenutku se vrednosti tih promenljivih koriste kao uslovi od kojih zavisi prolaznost testa u poslednjoj liniji. Dodelimo svakoj promenljivoj simboličku vrednost — $a = \alpha$, $b = \beta$, $c = \gamma$. Možemo izgraditi stablo izvršavanja i uslove koji moraju da važe nad simboličkim vrednostima α , β i γ kako bi test u poslednjoj liniji prošao.

```
1  int a, b, c;
2
3  // ...
4
5  int x = 0, y = 0, z = 0;
6  if (a)
7      x = -2;
8  if (b < 5) {
9      if (!a && c)
10         y = 1;
11         z = 2;
12 }
13
14 assert(x + y + z != 3);
```

Slika 3.1: Isečak C koda dat kao primer nad kojim će se prikazati simboličko izvršavanje.

Ukoliko put izvršavanja programa zavisi od simboličke promenljive, kao što je to slučaj za izvorni kôd sa slike 3.1, simbolička promenljiva se konceptualno „grana” i analiza se nastavlja za oba slučaja posebno. Tako se dobija drvo izvršavanja, gde svaki put odgovara mnogim individualnim testovima koji bi uzrokovali prolazak izvršavanja tim putem. Vrednosti promenljivih u tim testovima moraju zadovoljiti uslove na kraju svakog puta — tzv. *uslove puta* (engl. *path conditions*). Odgovarajuće stablo izvršavanja za izvorni kôd sa slike 3.1 sa definisanim simboličkim vrednostima α , β i γ se može videti na slici 3.2. Svaka naredba dodele je

uokvirena pravougaonikom dok je uslov uokviren elipsom. Boje grana odgovaraju istinitosnoj vrednosti uslova iz poslednje linije u tom trenutku. Na kraju svake grane se nalazi uslov puta za tu granu koje u nekim slučajevima može jedinstveno odrediti vrednost simboličke promenljive koja dovodi do prolaska tim putem ili u opštem slučaju generisati test primer koji dovodi do prolaska tim putem. Dakle, ukoliko je stablo simboličkog izvršavanja poznato, moguće je trivijalno generisati kontra-primere koji dokazuju da program ne radi kao što je očekivano.



Slika 3.2: Drvo simboličkog izvršavanja na kom su prikazane sve putanje koje se razmatraju. Na kraju svake grane je napisan uslov koji mora da važi da bi se došlo do lista te grane.

Simboličko izvršavanje, iako konceptualno moćno, ima par problema koji su uzrokovani pre svega kompleksnošću problema koji se rešava:

- *Eksplorzija putanja* — Broj puteva izvršavanja eksponencijalno zavisi od broja uslovnih grananja u kodu. Ukoliko imamo 3 naredbe grananja, broj puteva

izvršavanja je $2^3 = 8$. Štaviše, petlje su još komplikovanije, jer ukoliko u petlji postoji uslov koji zavisi od simboličke vrednosti koja uzima vrednosti iz opsega 32-bitnog celog broja, broj puteva kroz petlju je u tom slučaju 2^{31} , a u komplikovanijim slučajevima i beskonačan. Slično važi i za rekurziju.

- *Ograničenja rešavača* — Kako broj puteva raste, povećava se i broj uslova koji se moraju zadovoljiti prilikom nalaženja kontraprimera. Trivijalni algoritmi kao što su pretrage u dubinu i širinu nisu dovoljno efikasni. U nekim slučajevima je moguće osloniti se na *SMT rešavače* [22] za nalaženje kontra-primera kako bi se ublažio ovaj problem.
- *Modelovanje podataka* (engl. *heap modelling*) — Kreiranje simboličkih struktura podataka i pokazivača nije jednostavan proces.
- *Modelovanje okruženja* (engl. *environment modelling*) — Nije uvek jednostavno adaptirati mehanizam za česte potrebe prilikom dizajna softvera kao što je korišćenje eksternih biblioteka i sistemskih poziva ali i specifičnosti sistema.

Postoji dosta alata i biblioteka koje pružaju interfejs za simboličko izračunavanje u raznim programskim jezicima — jedan od najpoznatijih alata za simboličko izračunavanje je *KLEE* [13], izgrađen nad *LLVM* infrastruktorom [17] i dizajniran za analizu koda pisanog u programskom jeziku C. U ovom radu će simboličko izvršavanje biti korišćeno za detekciju razlika u vrednostima promenljivih uz pomoć biblioteka za programski jezik C#.

Glava 4

Opis opštih AST apstrakcija za imperativne jezike

Kao što je opisano u odeljku 2.4, dosta različitih „pod-paradigmi” potiče iz imperativne paradigme. Strukturna, proceduralna i skript paradigma, iako naizgled različite, poseduju veliki broj sličnih osobina i koncepata. Moderni programski jezici uzimaju korisne koncepte iz svakakvih paradigmi pa je teško vezati jezik za jednu konkretnu paradigmu. Ovo je motivacija za apstrahovanje koncepata različitih paradigmi, ali pre svega imperativne i njenih „derivata” — proceduralne, skript i objektno orijentisane. U ovom poglavlju će biti opisana opšta apstrakcija za imperativnu paradigmu i njene derivate. To uključuje i skript jezike koji, kako će biti pokazano u ovom radu, mogu da se posmatraju na istom nivou kao i svoji proceduralni „rođaci”.

Svaki programski jezik ima svoju gramatiku i na osnovu toga ima svoja gramatička pravila koja se oslikavaju u apstraktnim sintaksnim stablima tih jezika. Na slikama 4.1 i 4.2 se mogu videti razlike jezika Lua i Go, kao primere skript odnosno proceduralne paradigme, kad se posmatra njihov AST.

Kako bi se kreirala smisljena apstrakcija stabla parsiranja, potrebno je identifikovati bitne informacije u stablu parsiranja ali i koncepte same gramatike koji su ponovno upotrebljivi. Najjednostavnije rešenje je mimikovati čvorove stabla parsiranja, ukoliko su gramatička pravila kreirana tako da oslikaju koncepte jezika koji gramatika definiše. Na primer, ukoliko u gramatici imamo pravilo deklaracija sa alternativama deklaracijaPromenljive i deklaracijaFunkcije, možemo kreirati apstraktni koncept Deklaracija sa konkretizacijama DeklaracijaPromenljive i DeklaracijaFunkcije. Kako se definišu deklaracije promenljivih i funkcija za-

```

5
6 function allwords ()
7   local line = io.read()
8   local pos = 1
9   return function ()
10    while line do
11      local s, e = string.find(line, "%w+", pos)
12      if s then
13        pos = e + 1
14        return string.sub(line, s, e)
15      else
16        line = io.read()
17        pos = 1
18      end
19    end
20    return nil
21  end
22 end

```

```

Chunk {
  type: "Chunk"
  - body: [
    - FunctionDeclaration {
      type: "FunctionDeclaration"
      + identifier: Identifier {type, name, range}
      isLocal: false
      parameters: [ ]
    - body: [
      + LocalStatement {type, variables, init, range}
      + LocalStatement {type, variables, init, range}
      - ReturnStatement {
        type: "ReturnStatement"
        - arguments: [
          - FunctionDeclaration {
            type: "FunctionDeclaration"
            identifier: null
            isLocal: false
            parameters: [ ]
            + body: [2 elements]
            + range: [2 elements]
          }
        ]
        + range: [2 elements]
      }
    ]
    + range: [2 elements]
  ]
  + range: [2 elements]
  + comments: [1 element]
}

```

Slika 4.1: AST isečka koda pisanog u programskom jeziku Lua. Prikazano putem <https://astexplorer.net/>

visi dalje od definicija pravila deklaracijaPromenljive i deklaracijaFunkcije. Naravno, nije uvek moguće primeniti ovakav postupak. Takođe, nekada u gramatici definišemo pomoćna pravila kako bismo se izborili sa rekurzijom ili izbegli neke tipove rekurzije — ta pravila ne bi trebalo da imaju odgovarajuće tipove u opštoj apstrakciji.

Pošto su u pitanju gramatike programskih jezika, onda je jasno da dosta različitih gramatika dele slične koncepte i da je moguće definisati tipove čvorova koji odgovaraju tim konceptima. Neki od njih mogu biti: naredba, izraz, deklaracija, poziv funkcije, dodela itd. Može se uočiti i hijerarhija između navedenih koncepta, međutim poziv funkcije se može smatrati kao samostalna naredba ali može biti i deo izraza. Dakle, prilikom definisanja hijerarhije ne treba dozvoliti nešto

```

package main

import "fmt"

const TIPS = `abc`;

func PrintTips() {
    fmt.Println(TIPS)
}

```

```

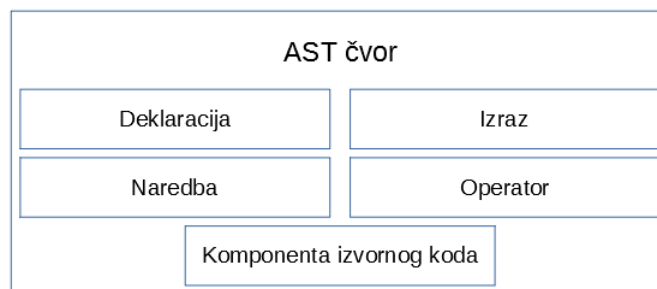
- File {
  Comments: [ ]
  - Decls: [
    + GenDecl {Loc, Lparen, Rparen, Specs, Tok}
    - GenDecl {
      + Loc: {End, Start}
      Lparen: 0
      Rparen: 0
      - Specs: [
        - ValueSpec {
          + Loc: {End, Start}
          + Names: [1 element]
          - Values: [
            - BasicLit {
              Kind: "STRING"
              + Loc: {End, Start}
              Value: "`abc`"
            }
          ]
        }
      ]
      Tok: "const"
    }
    + FuncDecl {Body, Loc, Name, Type}
  ]
  + Imports: [1 element]
  + Loc: {End, Start}
  + Name: Ident {Loc, Name}
  Package: 6
  + Unresolved: [1 element]
}

```

Slika 4.2: AST isečka koda pisanog u programskom jeziku Go. Prikazano putem <https://astexplorer.net/>

što nema smisla (npr. ako je dozvoljeno višestruko nasleđivanje i poziv funkcije je i naredba ali i izraz, onda se izrazi u kojima figurišu pozivi funkcija sastoje od više naredbi.).

Osim naredbi i izraza (koje vezuju operatori), kao osnovnih koncepata imperativnih jezika, deklaracije se ne pojavljuju u skript jezicima zbog slabe tipiziranosti. Moguće je, međutim, posmatrati i promenljive u kodovima skript jezika kao promenljive deklarisanе neposredno pre trenutka njihove upotrebe — detaljnije opisano u 4.1. Što se tiče njihovog tipa, može biti dozvoljena promena istog, ili, kako je izabrano u ovom radu, biće iskorišćen specijalni tip od kog potiču svi ostali tipovi.



Slika 4.3: Prikaz osnovnih vrsta AST čvorova.

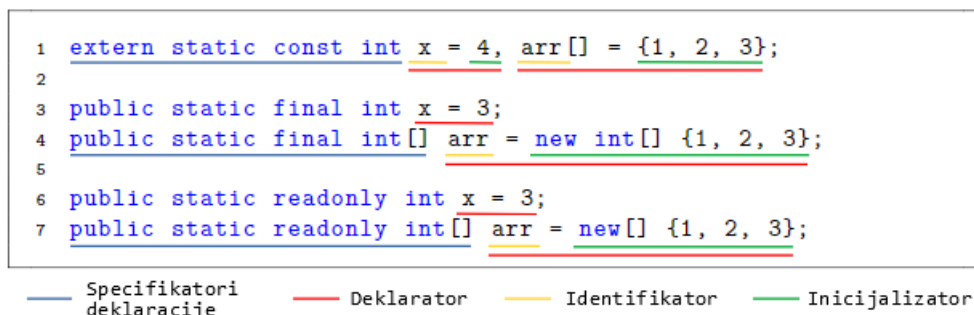
Na slici 4.3 se mogu videti osnovni tipovi AST čvorova zasnovani na konceptima opisanim iznad. U nastavku će po odeljcima biti detaljnije opisan svaki od prikazanih tipova.

4.1 Čvorovi deklaracija

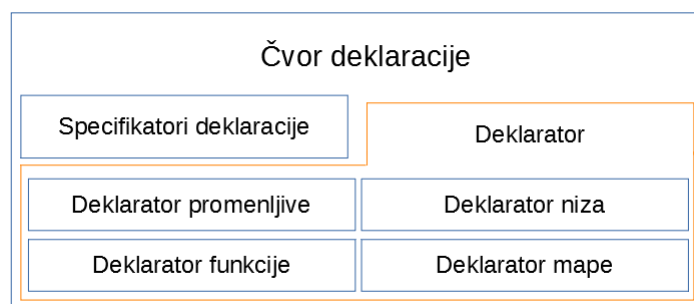
Kao što je opisano u odeljku 2.4, u striktno tipiziranim proceduralnim jezicima promenljive i funkcije koje se koriste se moraju deklarirati pre trenutka njihovog korišćenja. Prateći kvalifikatori (statičnost, konstantnost itd.) i modifikatori pristupa (javni, privatni itd.) će se u nastavku nazivati *specifikatori deklaracije* (engl. *declaration specifiers*). Nakon specifikatora deklaracije dolazi konkretan *deklarator*, koji ima specifičan oblik u zavisnosti od toga šta se deklarira. Oba imena su uzeta po uzoru na imena pravila gramatike programskog jezika C.

Veliki broj proceduralnih jezika dozvoljava deklarisanje više promenljivih odjednom koje dele iste specifikatore deklaracije. Stoga specifikatore neće pratiti jedan deklarator, nego *lista deklaratora*. Takođe, deklaratori u listi ne moraju biti samo deklaratori promenljivih — moguće je deklarirati i nizovnu promenljivu zajedno sa deklaracijama običnih promenljivih. Na slici 4.4 se može videti dekompozicija deklaracije promenljive i niza u različitim proceduralnim programskim jezicima a na slici 4.5 uočena hijerarhija sa podvrstama deklaratora.

Kao što je prikazano na slici 4.4, specifikatori deklaracije pokrivaju kvalifikatore, specifikatore pristupa i ime tipa. Pošto se pravi zajednička apstrakcija, potrebno je uočiti ekvivalentne ključne reči u različitim programskim jezicima — u primeru sa slike to su `const`, `final` i `readonly`. Imena tipova u programskim jezicima Java i C# uzeta su po uzoru na programski jezik C, tako da tu ne vidimo razlike. U opštem slučaju, moguće je definisati mapiranje imena tipa u apstraktni



Slika 4.4: Delovi deklaracije promenljive i niza prikazani na isečcima koda pisanog u programskim jezicima C, Java i C#.



Slika 4.5: Prikaz vrsti AST čvorova deklaracije.

tip. Ukoliko, na primer, posmatramo tipove koji predstavljaju realne brojeve, osim tipova `float` i `double`, postoji i tip `decimal` prisutan u programskom jeziku C#¹. Sva tri ova tipa mogu da se posmatraju na istom nivou apstrakcije kao tip realnih brojeva. Za korisnički definisane tipove isto ne može da se primeni.

Deklaratori za proceduralne jezike mogu biti deklaratori promenljive, niza ili funkcije i od toga zavisi njihov sastav. Svi deklaratori moraju sadržati informaciju o identifikatoru. Ukoliko je reč o deklaratoru niza, dodatno se očekuje i oznaka za niz (obično par srednjih zagrada — `[]`) i opcioni izraz koji predstavlja dimenziju niza, obično unutar oznake niza. Ukoliko je reč o deklaratoru funkcije, pored identifikatora se očekuje i lista parametara funkcije obično navedena unutar para običnih zagrada. Lista parametara funkcije se može posmatrati rekurzivno — svaki parametar se može posmatrati kao varijanta deklaracije — sadrži specifikatore deklaracije (koji uključuju i tip) i deklarator, s tim što u ovom slučaju nije

¹Tip `decimal` predstavlja 128-bitni realan broj sa povećanom veličinom mantise a smanjenom veličinom eksponenta u odnosu na tip `double`. Koristi se u numeričkim izračunavanjima gde preciznost primitivnih tipova realnih brojeva nije dovoljna.

dozvoljeno da taj deklarator bude deklarator funkcije (pošto funkcije nisu građani prvog reda u imperativnoj paradigmi).

Deklaratori promenljive i niza mogu dodatno sadržati i *inicijalizator*. Inicijalizator možemo posmatrati kao opcioni izraz u slučaju deklaratora promenljive. U slučaju deklaratora niza, inicijalizator može biti lista izraza. Deklaratori funkcije ne mogu imati inicijalizatore.

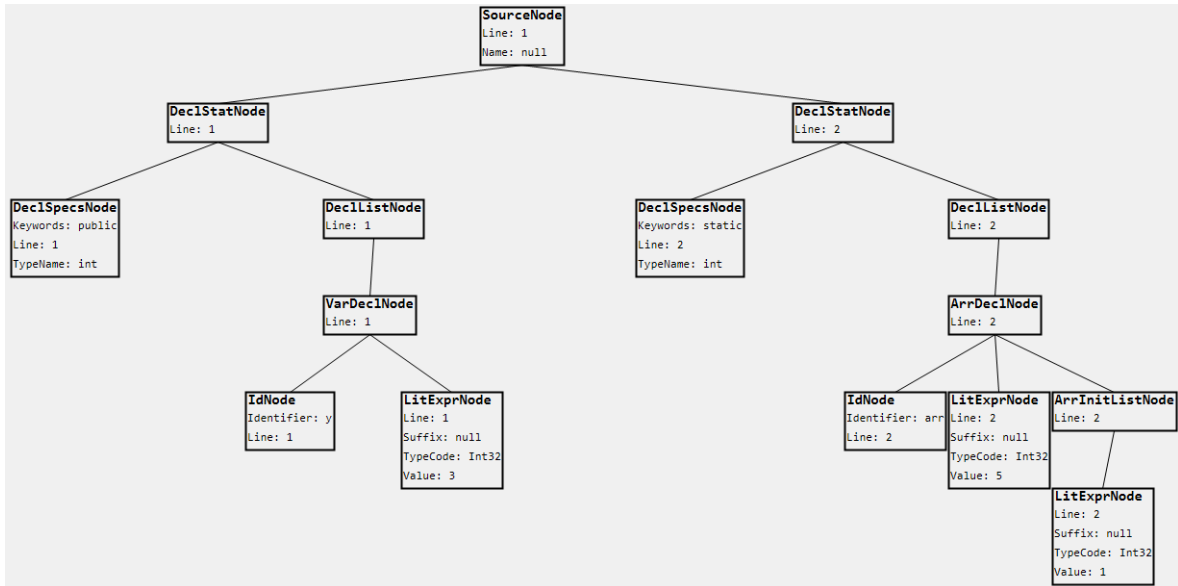
U skript jezicima su uobičajeno podržane strukture podataka kao što su skupovi i mape. Stoga, kako bi se i mape mogle predstaviti apstraktno, dodat je tip deklaratora koji predstavlja deklarator mape. Mapa se sastoji od skupa ključeva pri čemu je svakom ključu dodeljena vrednost ne nužno istog tipa kao što je tip ključa. Mape postoje i u proceduralnim jezicima, ali ključna razlika je ta što tipovi u skript jezicima nisu striktni — ključevi međusobno, ali i vrednosti mogu biti različitog tipa. Vredi naglasiti da se mape mogu porediti sa objektima određenih klasa — svaki objekat se može serijalizovati u mapu gde su ključevi imena javnih atributa klase a vrednosti su vrednosti javnih atributa objekta koji se serijalizuje. Neki jezici (kao što je Python), imaju funkcije koje od objekta vraćaju baš ovakvu mapu. Ova ideja se dalje može proširiti kako bi se serijalizovale i metode klase, označila statička i privatna polja kao i sačuvala informacije o definisanim konstruktorima. Zatim je moguće porediti mape definisane u skript jezicima sa objektima iz proceduralnih jezika.

Na slici 4.6 se mogu videti kreirani AST za nekoliko deklaracija pisanih u programskom jeziku C a na slici 4.7 se može isto videti demonstracija *automatske deklaracije* promenljivih za skript programski jezik Lua. Naime, pre prvog pojavljivanja identifikatora veštački će biti deklarisan taj identifikator, kako bi razlika između apstrakcija dobijenih iz proceduralnih i skript jezika bila što manja. U ovom slučaju će se deklaracija i dodela spojiti u deklaraciju sa inicijalizatorom.

4.2 Čvorovi operatora

Svrha operatora je da vezuju izraze i da tako grade nove izraze. Operator se karakteriše simbolom i *arnošću*, tj. brojem argumenata koje taj operator prima. Na osnovu arnosti, svaki operator se može apstraktno posmatrati kao članica grupe operatora sa istom arnošću. Na slici 4.8 se može videti hijerarhija operatora korišćena dalje u apstrakciji. Binarni operatori zahtevaju dva operanda i pišu se infiksno, dok unarni zahtevaju jedan operand i pišu se prefiksno. Ternarni ope-

```
1 extern int y = 3;
2 static int arr[5] = { 1 };
```



Slika 4.6: Primer deklaracije promenljive i niza u programskom jeziku C i odgovarajući AST.

ratori koji postoje u nekim programskim jezicima nisu razmatrani jer se mogu posmatrati kao druge strukture ².

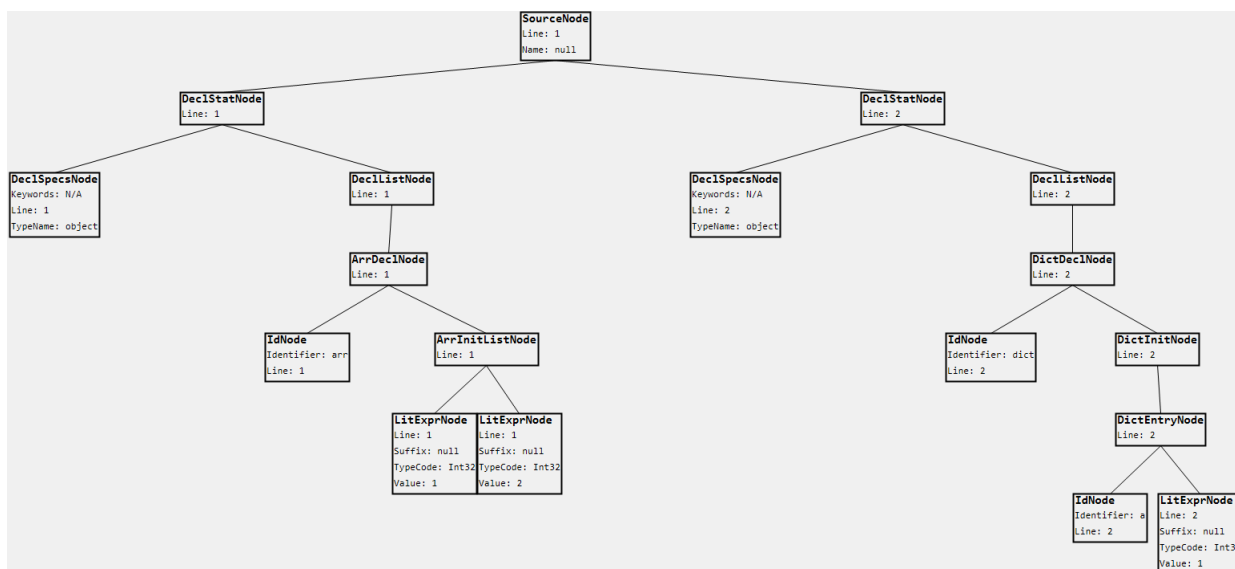
Unarni aritmetički operatori su unarni operatori koji figurišu u aritmetičkim izrazima, npr. operator promene znaka, operator bitovske negacije *Bitovski izrazi se mogu posmatrati kao vrsta aritmetičkih izraza*, operatori kastovanja ili inkrementiranja odnosno dekrementiranja. Unarni logički operatori su unarni operatori koji figurišu u logičkim izrazima, npr. operator negacije. Možemo sve ove unarne operatore posmatrati apstraktno ukoliko definišemo unarni operator kao strukturu koja definiše unarnu funkciju koja transformiše svoj argument na osnovu logike konkretnog unarnog operatora. Tip argumenta i povratne vrednosti pomenute funkcije zavisi od tipa unarnog operatora — aritmetički unarni operatori mogu primiti vrednost bilo kog tipa ³ i vraćaju vrednost proizvoljnog, ne nužno istog

²Na primer, ternarni operator `?:` prisutan u jezicima zasnovanim na sintaksi programskog jezika C se može zameniti naredbom uslovnog grananja.

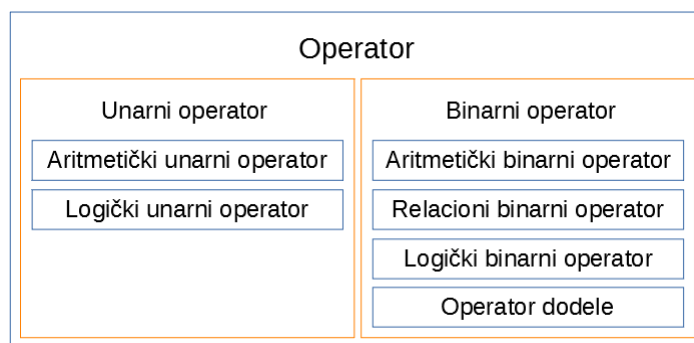
³Ne postoji ograničenje na brojeve tipove jer se u nekim jezicima operatori mogu predefinisati tako da rade i za korisnički definisane tipove (engl. *operator overloading*).

GLAVA 4. OPIS OPŠTIH AST APSTRAKCIJA ZA IMPERATIVNE JEZIKE

```
1 arr = { 1, 2 }
2 dict = { a = 1 }
```



Slika 4.7: Primer deklaracije promenljive i niza u programskom jeziku Lua i odgovarajući AST.



Slika 4.8: Podela operatora na osnovu njihove arnosti.

tipa; dok unarni logički operatori primaju i vraćaju bulovsku vrednost ⁴. Koristeći ovaj pristup, nije potrebno praviti novi AST čvor za svaki mogući operator, već je dovoljno da postoji samo jedan čvor koji predstavlja unarni operator. Ovakav pristup odgovara varijanti AST sa regularnošću (videti sliku 2.8), omogućava opi-

⁴U nekim programskim jezicima postoji implicitna konverzija brojevnih tipova u bulovski tip, što se jednostavno može posmatrati kao poređenje vrednosti po jednakosti sa nulom.

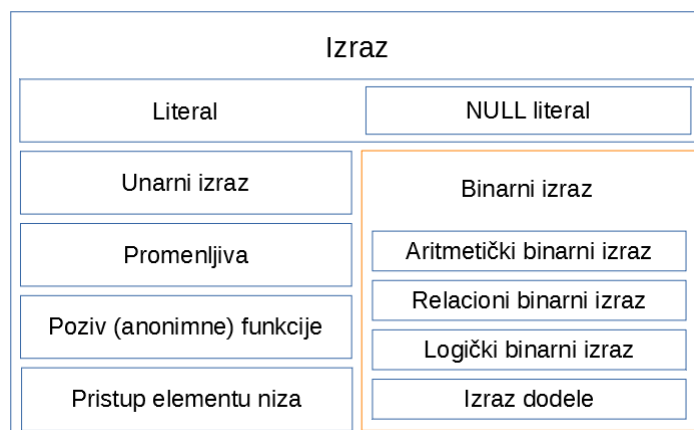
sivanje proizvoljnih operatora i nije vezan za konkretnu programsku paradigmu.

Binarni aritmetički operatori su binarni operatori koji figurišu u aritmetičkim izrazima, npr. operatori koji odgovaraju matematičkim operacijama ali i bitovski binarni operatori. Binarni relacioni operatori su binarni operatori koji figurišu u relacionim izrazima, npr. operatori poretka ($<$, $>$, \leq , \geq) i poređenja po jednakosti ili različitosti ($=$, \neq). Binarni logički operatori su binarni operatori koji figurišu u logičkim izrazima, npr. bulovske operacije (\wedge , \vee). Slično kao i za unarne operatore, moguće je apstraktno posmatrati sve binarne operatore tako što ih definišemo kao strukturu koja definiše binarnu funkciju koja transformiše argumente na osnovu logike konkretnog binarnog operatora. Tip argumenata i povratne vrednosti te funkcije zavisi od tipa binarnog operatora, kao i u slučaju unarnih operatora — aritmetički binarni operatori primaju dva argumenta proizvoljnog tipa i vraćaju rezultat proizvoljnog, ne nužno istog tipa; relacioni binarni operatori primaju iste tipove argumenata kao i aritmetički binarni operatori, međutim povratna vrednost mora biti bulovskog tipa; dok logički binarni operatori zahtevaju da argumenti i povratna vrednost budu bulovskog tipa. Pritom, na prvi pogled nije jasno kako se operator dodele može uklopiti u ovaj šablon ali, na osnovu toga da je dodela zapravo sporedni efekat i da se posmatra kao izraz čija je vrednost jednaka vrednosti izraza sa desne strane operatora, može se primeniti isti princip kao i za aritmetičke binarne izraze. Neki programski jezici dozvoljavaju i složene operatore dodele, koji se mogu dekomponovati na više jednostavnijih izraza.

4.3 Čvorovi izraza

Izraz, kao što se može videti na primeru gramatike sa slike 2.9, se definiše rekursivno i izraze mogu proširiti razni operatori. Na slici 4.9 se mogu videti tipovi apstraktnih konstrukcija koje će se koristiti da bi se predstavili izrazi. Dodatno, za vezivanje izraza će se koristiti apstrakcije operatora definisane u odeljku 4.2.

Najjednostavniji izraz predstavlja konstante ili *literali*. Literali mogu biti brojeve konstante, karakterske konstante ili konstantne niske. Literali često mogu imati i sufiks (najčešće za brojeve literale), koji određuje tip literala u slučajevima gde postoji dvosmislenost. Na primer, literal 5 možemo posmatrati kao 32-bitni ceo broj ili kao 64-bitni ceo broj (ali i kao realan broj, ako ne zahtevamo da realne brojeve moramo pisati u nepokretnom ili pokretnom zarezu). Da bi se ova dvosmislenost uklonila, možemo eksplicitno naznačiti da se govori o 64-bitnom



Slika 4.9: Vrste čvorova izraza.

celom broju dodavanjem sufiksa L, ako je u pitanju programski jezik C ili njemu slični jezici. Takođe, pošto većina programskih jezika dozvoljava rad sa pokazivačima ili neposredno koristi alokaciju memorije za kreiranje objekata, uobičajeno je korišćenje prazne adrese kao specijalne vrednosti (null ili nil). Za ovu specijalnu vrednost je moguće kreirati poseban tip literala, jer ovaj literal može biti bilo kog tipa koji nije primitivni tip.

Osim literala, samostalne promenljive mogu predstavljati validan izraz, u kom slučaju je vrednost izraza trenutna vrednost te promenljive. Slično važi i za indeksni pristup nizu ⁵. U slučaju indeksnog pristupa, potrebno je navesti izraz čija vrednost označava indeks (to ne mora biti jednostavni literal). Postoje smisljena ograničenja šta sve sme da se nađe unutar izraza koji predstavlja indeks elementa niza tako da to semantički ima smisla, ali se na ovom nivou ne bavimo semantičkom analizom.

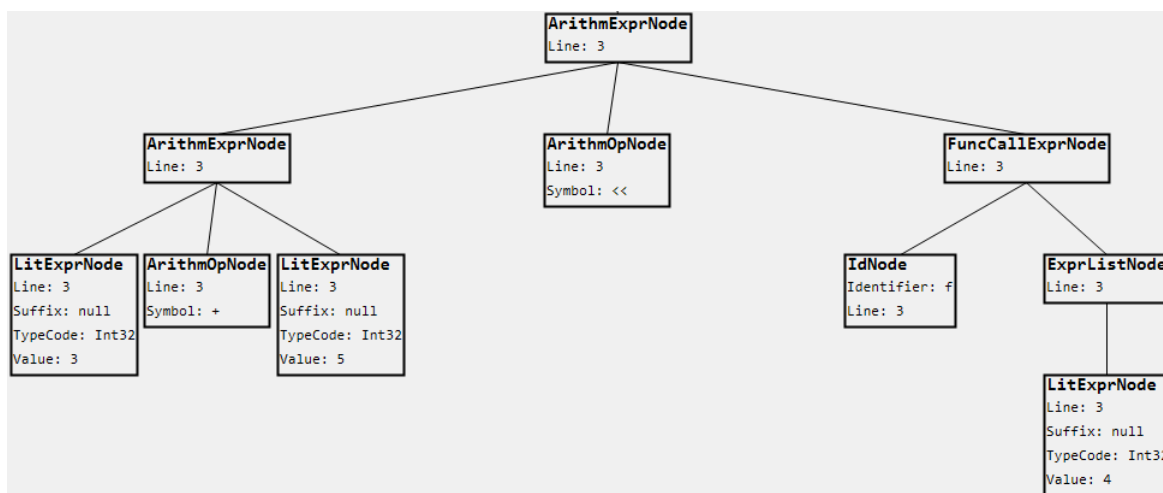
Unutar izraza se mogu naći i pozivi funkcija. Naravno, pretpostavljamo da funkcija ima povratnu vrednost, koja će se iskoristiti nakon poziva funkcije u kontekstu iz kojeg je ona pozvana. Pod uticajem funkcionalne paradigme, veliki broj programskih jezika dozvoljava definisanje anonimnih (lambda) funkcija, čija se definicija može smatrati izrazom koji se može dodeliti nekom simbolu. Stoga se i anonimne funkcije mogu smatrati validnim izrazima.

Operatori opisani u 4.2 mogu vezati sve tipove iznad i formirati složenije izraze. U zavisnosti od broja izraza koje operator vezuje, izraze možemo podeliti na unarne i binarne. Unarne izraze nadograđuju unarni operatori dok su binar-

⁵Isto važi i za bilo koju drugu kolekciju, ukoliko je nad njom definisan operator indeksnog pristupa, predefinisane ovog operatora nije razmatrano u ovom radu.

ni izrazi dobijeni primenom binarnog operatora na dva izraza. U zavisnosti od tipa binarnog operatora (videti sliku 4.8), binarne izraze delimo na sličan način. Naravno, svaki od tipova binarnog izraza zahteva odgovarajući tip binarnog operatora. Slično se može uraditi i za unarne izraze, ali takođe i napraviti podela na prefiksne i postfiksne, ali to nije rađeno u ovom radu zarad pojednostavljenja — u nekim jezicima su određeni operatori prefiksni dok su u drugim jezicima ti isti operatori postfiksni pa stoga nije pravljen ova podela.

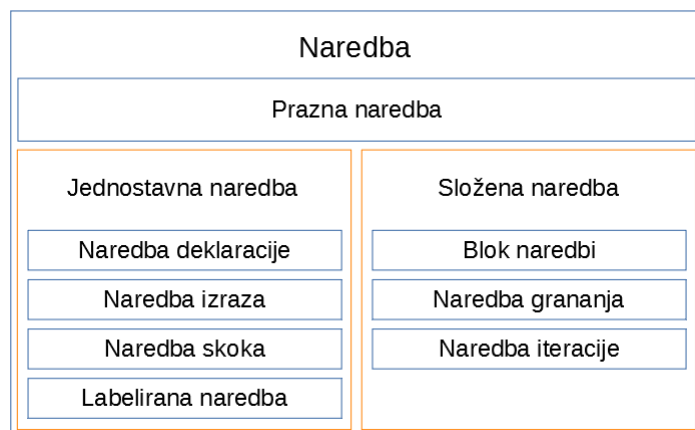
Na slici 4.10 se mogu videti kreirani AST za izraz $(3 + 5) << f(4)$. Ovaj izraz poprima isti oblik bez obzira na to koji je programski jezik u pitanju, ali iako se sintaksa bude razlikovala ili operatori budu imali drugi simbol, logika operatora opisana putem funkcije će ostati ista.



Slika 4.10: AST generisan od izraza $(3 + 5) << f(4)$.

4.4 Čvorovi naredbi

Naredbe su najkomplikovanije za apstrahovanje zbog njihove raznovrsnosti. Programski jezici često uvode nove sintaksne strukture i naredbe koje nisu do tada viđene u ostalim jezicima. Uprkos svemu tome, ipak je moguće uočiti neke sličnosti sa već postojećim konceptima i svesti ih na isti nivo. Na slici 4.11 se mogu videti tipovi apstraktnih konstrukcija koje će se koristiti da bi se predstavile naredbe.



Slika 4.11: Vrste čvorova naredbi.

Veliki broj programskih jezika podržava praznu naredbu, sa semantikom ne izvršavanja nikakvih operacija. U programskim jezicima koji su zasnovani na sintaksi jezika C, praznu naredbu navodimo samo korišćenjem simbola za kraj naredbe (;), dok u programskom jeziku Python koristimo ključnu reč `pass`.

Naredbe su podeljene na *jednostavne* i *složene*, koje se sastoje od više drugih naredbi. Primer jednostavne naredbe može biti deklaracija promenljive, dok primer složene naredbe može biti definicija funkcije koja se sastoji od više jednostavnih deklaracija ali možda i drugih složenih naredbi kao što su grananja i petlje.

Jednostavne naredbe uključuju naredbe deklaracije i izraza. Razlog zašto se deklaracije i izrazi opet pojavljuju je taj što izrazi sami po sebi mogu biti deo drugih naredbi. Ukoliko se naredba sastoji samo od izraza, onda nju zovemo naredbom izraza. Primer može biti izraz dodele — vrednost izraza dodele se može koristiti u drugim izrazima ali, ukoliko samo želimo da izvršimo dodelu i ništa više u okviru iste naredbe, onda izraz dodele „umotavamo” u naredbu izraza. Slično važi i za deklaracije, ukoliko razmotrimo idiomsku `for` petlju (od standarda C99) — moguće je deklarirati promenljive koje se koriste unutar ciklusa ali to nije naredba deklaracije već deklaracija koja se koristi unutar druge naredbe.

Naredbe se mogu označiti, po uzoru na koncept *label* u imperativnim jezicima — identifikatora koji označava lokaciju u izvornom kodu. Labele se u imperativnim jezicima najviše koriste da bi se izvršili skokovi na određene lokacije u kodu ali su takođe prisutne i u proceduralnim jezicima (npr. kroz naredbu višestrukog grananja — `switch` ili u nekim jezicima `case`). Labelirana naredba se sastoji od naredbe i identifikatora koji predstavlja labelu.

Naredbe skoka se koriste obično u paru sa labeliranim naredbama, ali to ne mora uvek biti slučaj. Iako ove čvorove koristimo da bismo predstavili naredbe skoka prisutne u imperativnim jezicima, one predstavljaju i naredbe prekida (`break` ili `continue`) ili povratka vrednosti funkcije (`return`). U slučaju da je u pitanju skok na određenu labelu, onda se sastoji i od identifikatora koji predstavlja labelu na koju se skače. Ukoliko je u pitanju naredba prekida, nisu potrebne nikakve dodatne informacije (mada se i u tom slučaju može iskoristiti činjenica da su u pitanju skokovi pa se može labelirati petlja na koju se odnosi naredba prekida). U slučaju povratka vrednosti funkcije, sadrži opcioni izraz čija vrednost predstavlja povratnu vrednost funkcije.

Složene naredbe se sastoje od više drugih naredbi (ne nužno samo od jednostavnih). Često je potrebno izvršiti više naredbi u okviru jednog konteksta i za to se koristi blok naredba. Blok naredba grupiše više drugih naredbi u jednu. Blok naredba se u proceduralnim jezicima obično navodi eksplicitno, za programski jezik C pomoću velikih zagrada (`{}`), a za skript jezike često je implicitna ili se navodi korišćenjem različitih nivoa indentacije (Python).

Naredbe uslovnog grananja se sastoje od *uslova*, koji može biti relacioni ili logički izraz, naredbe koja se vrši ukoliko je uslov ispunjen (*then* grama), i opcionalne naredbe koja se izvršava ako uslov nije ispunjen (*else* grana). Rezultat uslovnog izraza, iako mora biti istinitosna vrednost, je dozvoljeno da bude bilo kog tipa (dakle nema ograničenja samo na relacione i logičke izraze) iz razloga što određeni programski jezici dozvoljavaju automatsku konverziju brojevnih tipova u logički (C). Štaviše, nekada je moguća i implicitna konverzija određenih tipova u logički tip definisanjem implicitnih operatora konverzije (C#). Zato će u apstrakciji uslov biti bilo koji izraz. Što se *then* i *else* grana tiče, one mogu biti bilo koje naredbe, ali zarad konzistentnosti će obe biti blokovi naredbi. Na slici 4.13 se može videti AST za naredbu grananja.

Naredbe iteracije imaju raznovrsni oblik u programskim jezicima. Najčešće podržane naredbe iteracije su *for*, *while* i *foreach* petlje. U opštem slučaju, dovoljno je koristiti samo jedan tip petlji, ali zarad jednostavnosti i prisutnosti ovih tipova u velikoj većini programskih jezika oba će biti podržana. Ostali tipovi petlji, kao što su *do-while* ili *repeat-until* petlje, će se svoditi na njih. *do-while* petlja se može svesti na *while* petlju jednostavnim ponavljanjem tela petlje pre same petlje i kreiranjem obične *while* petlje sa istim uslovom i telom. Slično se može uraditi i za *repeat-until* petlju, s tim što je potrebno samo negirati uslov u dobijenoj *while*

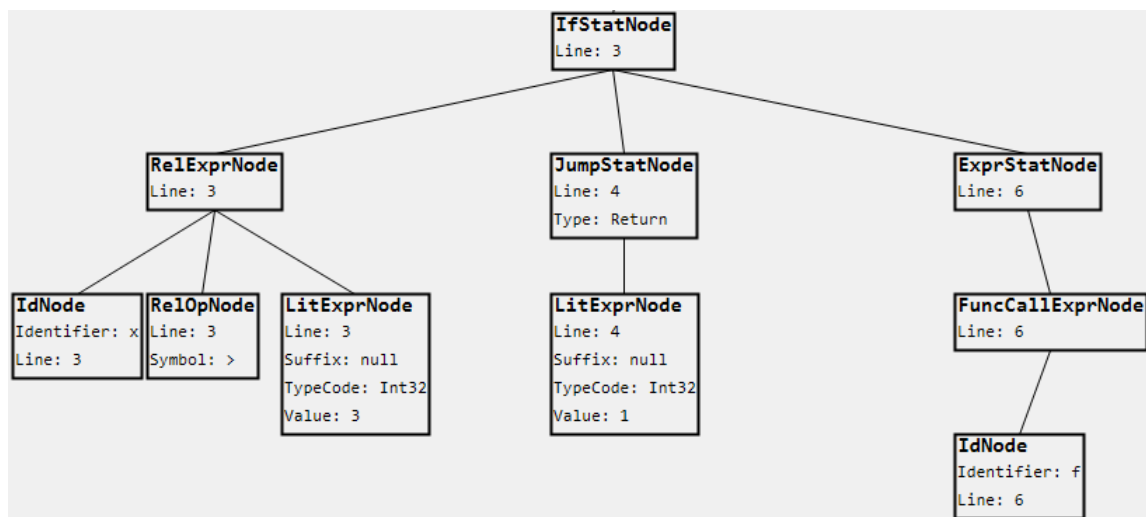
GLAVA 4. OPIS OPŠTIH AST APSTRAKCIJA ZA IMPERATIVNE JEZIKE

```
1 do something()
2   something() while (condition) do
3 while (condition) something()
```

```
1 repeat something()
2   something() while (not condition) do
3 until (condition) something()
```

Slika 4.12: Procedura svođenja redih tipova petlji (levo) na *while* petlju (desno) prikazana u pseudo-jeziku.

petlji. Ovaj proces je ilustrovan na slici 4.12.



Slika 4.13: AST naredbe grananja.

Glava 5

Semantičko poređenje opštih AST

Jedna od motivacija svođenja imperativnih jezika na isti nivo apstrakcije (opisane u poglavlju 4) može biti poređenje kodova pisanih u različitim programskim jezicima. Pritom, s obzirom da su u pitanju stabla, moguće je koristiti razne algoritme za poređenje stabala (ali i grafova uopšte) nad ovakvim apstrakcijama. Pritom, potrebno je i definisati kriterijum poređenja — moguće je porediti kodove *strukturno*, *semantički* itd. U ovom radu je od značaja semantička ekvivalentnost, koja je u opštem slučaju neodlučiv problem. Međutim, ukoliko se ograničimo samo na strukturno slične kodove, moguće je dobiti smislene rezultate u praksi, nalik na one dobijene u ovom radu. Precizna definicija strukturne sličnosti se obično iskazuje u terminima sličnosti strukture njihovih apstrakcija. Naravno, postoje kodovi koji nisu strukturno ekvivalentni ali su semantički ekvivalentni — takvi slučajevi se onda neće razmatrati zbog neispunjenosti pretpostavke o strukturnoj sličnosti.

Pretpostavka strukturne sličnosti je velika i znatno smanjuje broj slučajeva upotrebe takvog upoređivača. Međutim, danas je od velikog značaja verifikacija programa dobijenih sitnim refaktorisanjem već postojećih i verifikovanih programa, često ne menjajući strukturu uopšte (a ako se struktura koda menja onda se to obično radi u manjim koracima između kojih i dalje važi pretpostavka strukturne sličnosti kodova u susednim koracima). Slično, prepisivanja programa sa jednog programskog jezika na drugi su jako uobičajena, pa je i u takvim situacijama implicitno prisutna strukturna slučajnost, što zavisi od konkretnih programskih jezika ali se u praksi često smanjuje napor tako što se održava struktura koda, barem u inicijalnim verzijama.

Definicija strukturne sličnosti za potrebe ovog rada će se odnositi na sličnosti

u rasporedu blokova naredbi dok će raspored naredbi u blokovima biti nebitan. Na primer, ukoliko se prvi program sastoji od bloka naredbi u kome se nalaze dva druga bloka, očekuje se da i drugi program ima istu organizaciju blokova, pri čemu se pod terminom blok podrazumevaju i složene naredbe koje se sastoje od bloka naredbi u sebi, kao što su definicije funkcija, uslovna grananja, petlje itd. Pod ovim uslovima, moguće je porediti blokove prvog programa sa odgovarajućim blokovima drugog programa.

U ovom radu je poređenje vršeno pomoću algoritma pisanog specifično za rad sa opštim apstrakcijama opisanim u ovom radu. Grubi opis algoritma za poređenje, u daljem tekstu *upoređivač*, je opisan na slici 5.1. Upoređivač se sastoji od više upoređivača koje porede specifične tipove čvorova. Za početak, potreban je jedan adapter koji će dobiti pokazivače na korene stabala koje je potrebno uporediti. S obzirom da tipovi čvorova mogu biti različiti, potrebno je proveriti da li su tipovi isti. Ukoliko to nije slučaj, prijavljuje se greška i rad se prekida. U protivnom, potrebno je odrediti tip čvorova i pozvati konkretni algoritam za poređenje.

```

1: procedure UPOREDI( $n_1, n_2$ )
2:   if  $n_1$  i  $n_2$  su istog tipa then
3:      $t \leftarrow$  tip čvora  $n_1$ 
4:     if postoji definisan upoređivač za čvorove tipa  $t$  then
5:        $U \leftarrow$  upoređivač čvorova tipa  $t$ 
6:       return  $U(n_1, n_2)$ 
7:     else
8:       if BrojDece( $n_1$ )  $\neq$  BrojDece( $n_2$ ) thend
9:         return False
10:      else
11:        if Atributi( $n_1$ )  $\neq$  Atributi( $n_2$ ) then
12:          return False
13:        for  $i \leftarrow 0$  to BrojDece( $n_1$ ) do
14:           $d_1 \leftarrow$  dete  $i$  čvora  $n_1$ 
15:           $d_2 \leftarrow$  dete  $i$  čvora  $n_2$ 
16:          if not Uporedi( $d_1, d_2$ ) then
17:            return False
18:          return True
19:      else
20:        return False

```

Slika 5.1: Osnovni AST upoređivač.

Podrazumevana implementacija poredenja može biti takva da se uporede atri-

buti svih čvorova a zatim se svako dete prvog čvora rekursivno uporedi sa odgovarajućim detetom drugog čvora (ukoliko imaju isti broj dece). Ako neki par dece nije ekvivalentan, onda to ne važi ni za njihove roditelje. Za većinu tipova čvorova ovakvo poređenje je dovoljno. Međutim, poređenje blokova naredbi je fundamentalno drugačije i za njega će biti definisane posebna procedura poređenja opisana u nastavku.

5.1 Upoređivač blokova naredbi

Podrazumevani način poređenja dece svakog čvora nije dobar u opštem slučaju za blokove naredbi jer je osetljiv na izmene redosleda naredbi — na primer promena redosleda deklaracija. Stoga je upoređivač blokova potrebno napisati tako da može da uoči semantičku ekvivalentnost iako naredbe nisu nužno jednake, a možda ih čak ima i različit broj.

Upoređivač se zasniva na poređenju vrednosti promenljivih na kraju svakog bloka naredbi. Apstrakcije dva programa će se porediti paralelno — *blok-po-blok*. Naredbe svakog bloka će se izvršavati i pratiće se izmene vrednosti promenljivih deklariranih do sada (bilo u trenutnom bloku, ili u roditeljskim blokovima). Na kraju svakog bloka će se izvršiti provera vrednosti promenljivih — svaka razlika će se prijaviti kao potencijalna greška a finalnu presudu o jednakosti će dati analiza jednakosti promenljivih. Ceo algoritam je prikazan na slici 5.2.

```

1: procedure UPOREDIBLOKOVE( $b_1, b_2$ )
2:    $gds_1 \leftarrow$  simboli iz svih predaka bloka  $b_1$ 
3:    $gds_2 \leftarrow$  simboli iz svih predaka bloka  $b_2$ 
4:    $lds_1 \leftarrow$  lokalni simboli za blok  $b_1$ 
5:    $lds_2 \leftarrow$  lokalni simboli za blok  $b_2$ 
6:   UporediSimbole( $lds_1, lds_2$ )
7:   IzvrsiNaredbe( $b_1, b_2, lds_1, lds_2, gds_1, gds_2$ )
8:   return UporediSimbole( $lds_1, lds_2$ )  $\wedge$  UporediSimbole( $gds_1, gds_2$ )

```

Slika 5.2: Upoređivač blokova naredbi.

U opisu algoritma se koristi termin *simbol* koji se sastoji od identifikatora i simboličke vrednosti promenljive. Lokalni simboli su deklarirani unutar bloka a globalni su svi simboli koji su deklarirani van trenutnog bloka a koji se mogu referisati iz njega. Pronalaženje deklariranih simbola u bloku podrazumeva prolaz kroz naredbe bloka i registrovanje svih naredbi deklaracije, izvlačenje deklaratora

iz njih i, uzimajući u obzir opcione inicijalizatore, kreiranje simboličke vrednosti za upravo deklarirani identifikator. Identifikator i opcioni simbolički inicijalizator čine *simbol*. Isto se ponavlja za sve naredbe deklaracije u bloku i rezultat je skup deklariranih simbola.

Nakon registrovanja svih lokalnih simbola proverava se njihova ekvivalentnost u funkciji `UporediSimbole`. Ova funkcija proverava da li se svi simboli iz prvog bloka nalaze u drugom i prijavljuje ukoliko neki simboli fale ili ukoliko postoje simboli koji su višak. Zatim, za simbole koji se nalaze u oba skupa, proverava njihove simboličke vrednosti. Ukoliko su te vrednosti različite, prijavljuje se potencijalna greška i na osnovu toga da li je bilo konflikata vraća se istinitosna vrednost. Razlog zašto se ta vrednost ne koristi dalje nakon prvog poziva ove funkcije je ta što različiti inicijalizatori ne znače nužno da postoji problem. Problem postoji ukoliko se nakon izvršavanja svih naredbi i dalje dešavaju konflikti u simboličkim vrednostima za neke promenljive.

Procedura `IzvršiNaredbe` izvršava paralelno naredbe iz oba bloka i na osnovu toga koje su naredbe u pitanju može i da ažurira simboličke vrednosti unutar skupa deklariranih simbola. Pseudokod ove procedure je dat na slici 5.3. Naredbe se za svaki blok izvršavaju dok se ne nađe do naredbe iz koje se može izvući novi blok — to mogu biti naredbe grananja, iteracije, definicije funkcija i slično. Sve naredbe do pronađene naredbe se izvršavaju. Procedura `IzvršiNaredbu` će proveriti tip naredbe i, u zavisnosti od toga da li je to naredba dodele, eventualno promeniti vrednosti u skupovima prosleđenih simbola. Nakon izvršavanja svih naredbi do pronađene naredbe koja sadrži blok, izvlači se blok iz nje (to isto se radi i za drugi program). Kad se blokovi izvuku, rekurzivno se poziva upoređivač blokova za pronađene parnjake. Po povratku iz rekurzivnog poziva nastavlja se isti postupak sve dok se ne izvrše sve naredbe. Pritom, algoritam se oslanja na strukturnu sličnost — ukoliko jedan AST ima više blokova na istoj dubini u odnosu na drugi, poređenje možda neće uočiti neke razlike.

```

1: procedure IzvrsiNaredbe( $b_1, b_2, lds_1, lds_2, gds_1, gds_2$ )
2:    $n_1 \leftarrow$  niz naredbi bloka  $b_1$ 
3:    $n_2 \leftarrow$  niz naredbi bloka  $b_2$ 
4:    $i \leftarrow j \leftarrow 0$ 
5:    $ni \leftarrow nj \leftarrow 0$ 
6:    $eq \leftarrow \text{True}$ 
7:   while  $\text{True}$  do
8:      $ni \leftarrow$  indeks prve naredbe koja sadrži blok u  $n_1$  počev od indeksa  $ni$ 
9:      $nj \leftarrow$  indeks prve naredbe koja sadrži blok u  $n_2$  počev od indeksa  $nj$ 
10:    for  $naredba \in \{n_1[x] \mid x \in [i..ni]\}$  do
11:      IzvrsiNaredbu( $naredba, lds_1, gds_1$ )
12:     $i \leftarrow i + ni$ 
13:    for  $naredba \in \{n_2[x] \mid x \in [j..nj]\}$  do
14:      IzvrsiNaredbu( $naredba, lds_2, gds_2$ )
15:     $j \leftarrow j + nj$ 
16:    if  $i > \text{Duzina}(n_1) \vee j > \text{Duzina}(n_2)$  then
17:      prekini petlju
18:     $nb_1 \leftarrow$  izvuci blok iz naredbe  $n_1[i]$ 
19:     $nb_2 \leftarrow$  izvuci blok iz naredbe  $n_2[j]$ 
20:     $eq \leftarrow eq \wedge \text{UporediBlokove}(nb_1, nb_2)$ 
21:     $i \leftarrow i + 1$ 
22:     $j \leftarrow j + 1$ 
23:  return  $eq$ 

```

Slika 5.3: Upoređivač blokova naredbi.

Glava 6

Implementacija

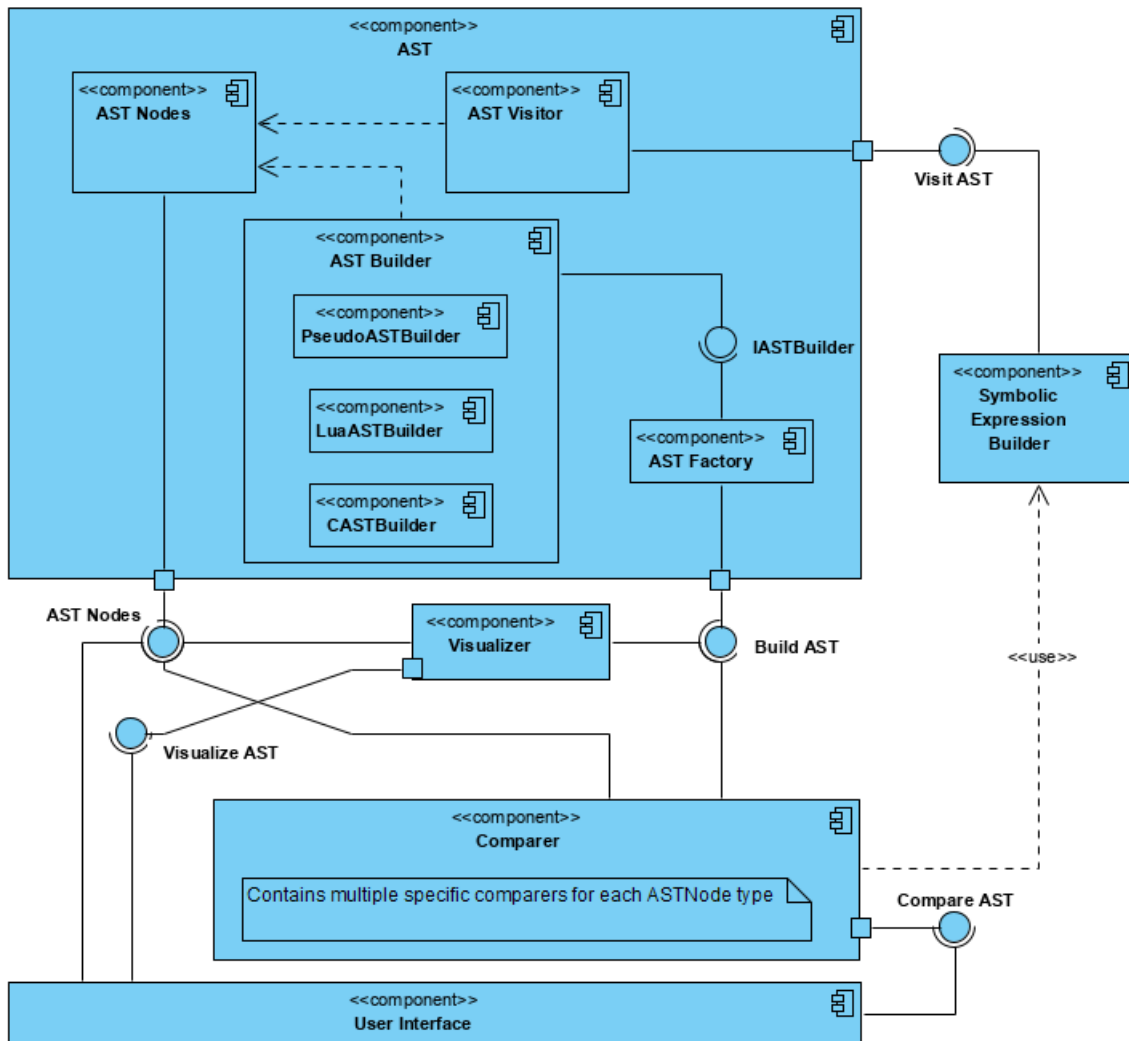
U ovom poglavlju će biti opisana implementacija pratećeg projekta nazvanog *Language Invariant Code Comparer* (skr. *LICC*), pisanog u programskom jeziku C# 8.0, koristeći *.NET Core 3.1* radni okvir. C# je izabran zbog lakoće implementacije velikih projekata i velike podrške paketa koji se mogu preuzeti, od kojih su korišćeni *ANTLR Runtime* paket koji daje potrebne biblioteke za rad sa ANTLR generisanim parserima i *Math.NET Symbolics* paket za rad sa simboličkim vrednostima. Rezultat je konzolna aplikacija koja može da generiše, serijalizuje ili prikaže opšti AST za dati izvorni kôd, ali i da poredi takav AST sa drugim. Čitav projekat je dostupan u potpunosti na servisu GitHub na adresi <https://github.com/ivan-ristovic/LICC>.

Jedan od glavnih ciljeva aplikacije je modularnost i jednostavna proširivost. U tom duhu se, pored implementacije klasa potrebnih za predstavljanje opšte AST apstrakcije, pruža i interfejs za kreiranje adaptera koji će od proizvoljnog stabla parsiranja kreirati opšti AST. Kao primer, adapteri su kreirani za programske jezike C i Lua, a za primer potpune slobode u izboru gramatike je kreirana gramatika za pseudo-jezik i adapter za istu, što dozvoljava poređenje kodova sa specifikacijom datom u obliku pseudo-koda. Čitav projekat se sastoji od više komponenti, organizovanih po prostorima imena, od kojih su značajnije:

- *LICC* — Glavni program (korisnički interfejs) koji omogućava generisanje, prikaz, serijalizaciju i poređenje AST.
- *LICC.AST* — Biblioteka klasa za rad sa opštom AST apstrakcijom.
- *LICC.Core* — Upoređivač opštih AST — konzolni izlaz.

- `LICC.Visualizer` — Komponenta za vizualizaciju — grafički prikaz AST.
- `LICC.Tests` — Prateći testovi jedinica koda i integracioni testovi.

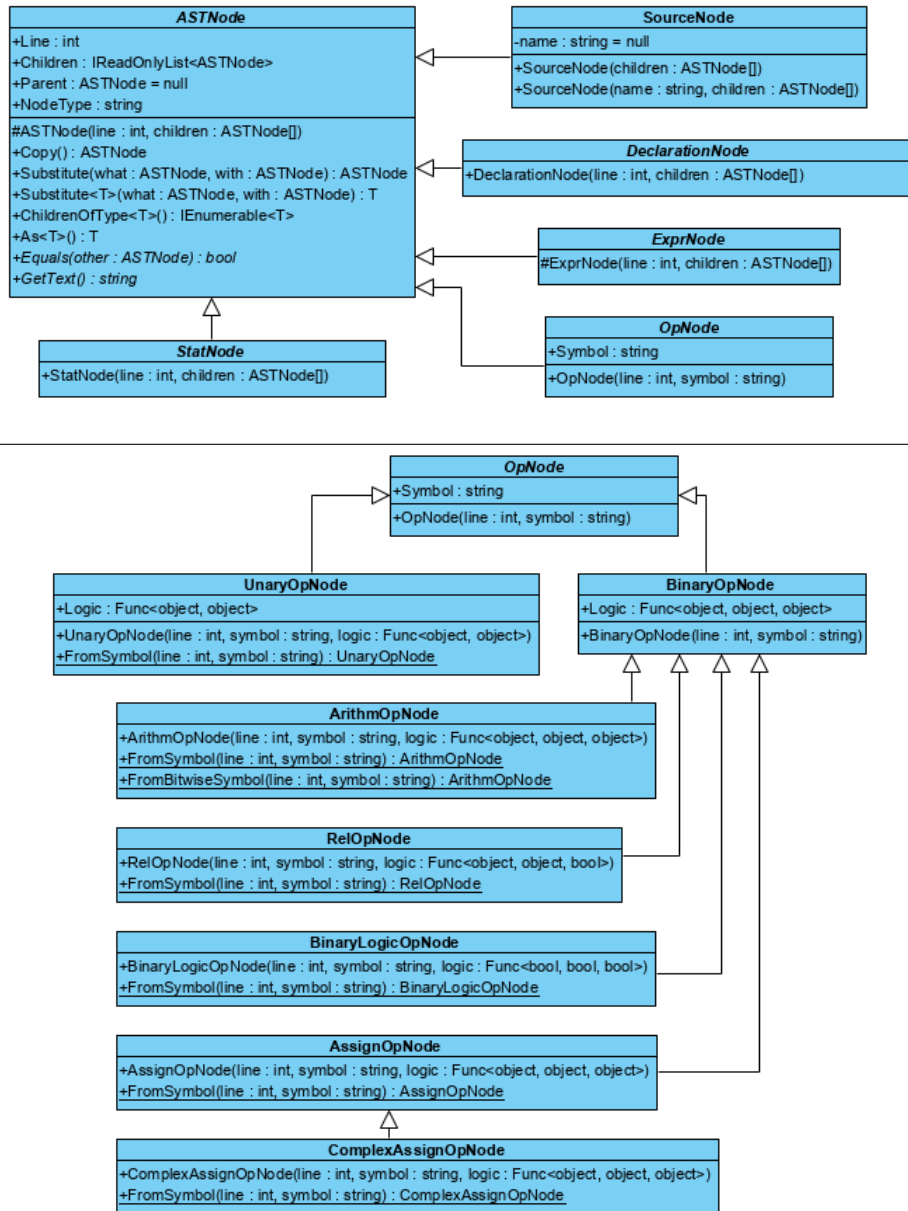
Čitava arhitektura data putem UML dijagrama komponenti se može videti na slici 6.1. Osim implementacije same aplikacije, svaki funkcionalni deo projekta prate i testovi jedinica koda, koji su povezani sa *GitHub Actions* porškom za neprekidnu integraciju (engl. *continuous integration*, srk. *CI*). CI omogućava prevođenje izvornog koda nakon svake izmene kao i izvršavanje akcija nakon prevođenja kao što su testiranje ili generisanje predmeta za upotrebu (engl. *artifacts*) koji predstavljaju rezultat procesa prevođenja i mogu se direktno isporučiti.



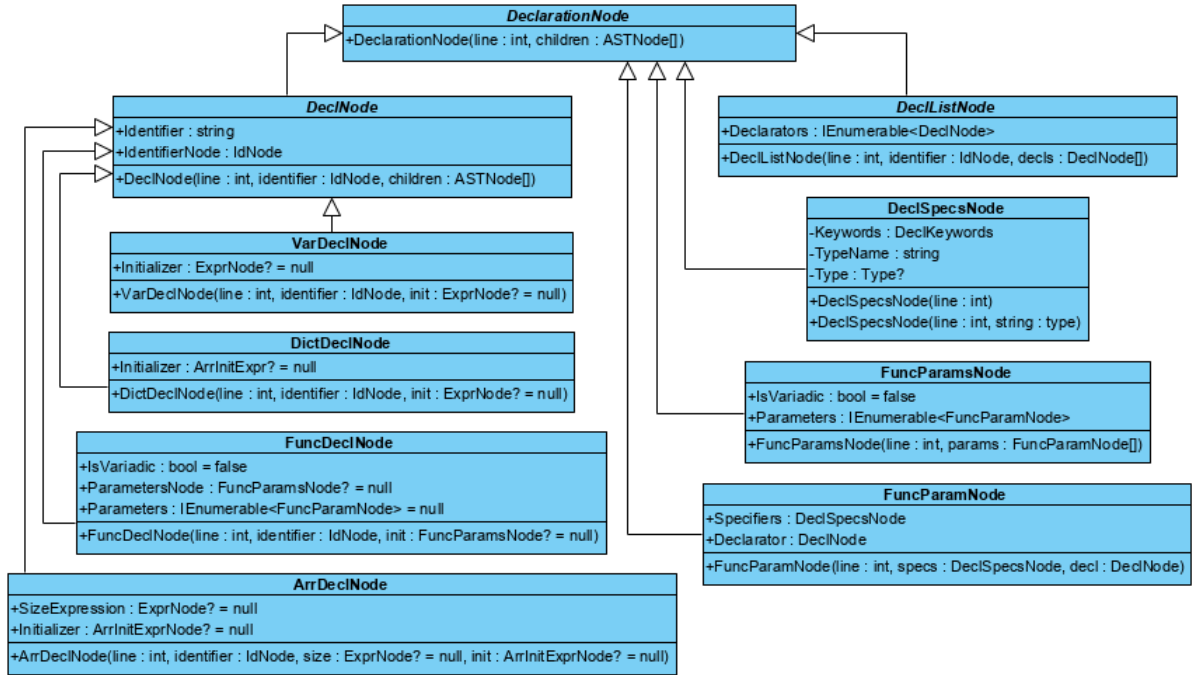
Slika 6.1: UML dijagram komponenti implementacije.

6.1 Implementacija apstrakcije

Implementacija prati hijerarhije opisane u poglavlju 4 kroz mehanizam nasleđivanja. Svaki tip čvora je implementiran kao zasebna klasa koja direktno ili tranzitivno nasleđuje apstraktnu klasu `ASTNode`. Klase koje čine apstrakciju zajedno sa njihovom hijerarhijom se može videti na slikama 6.2, 6.3 i 6.4.



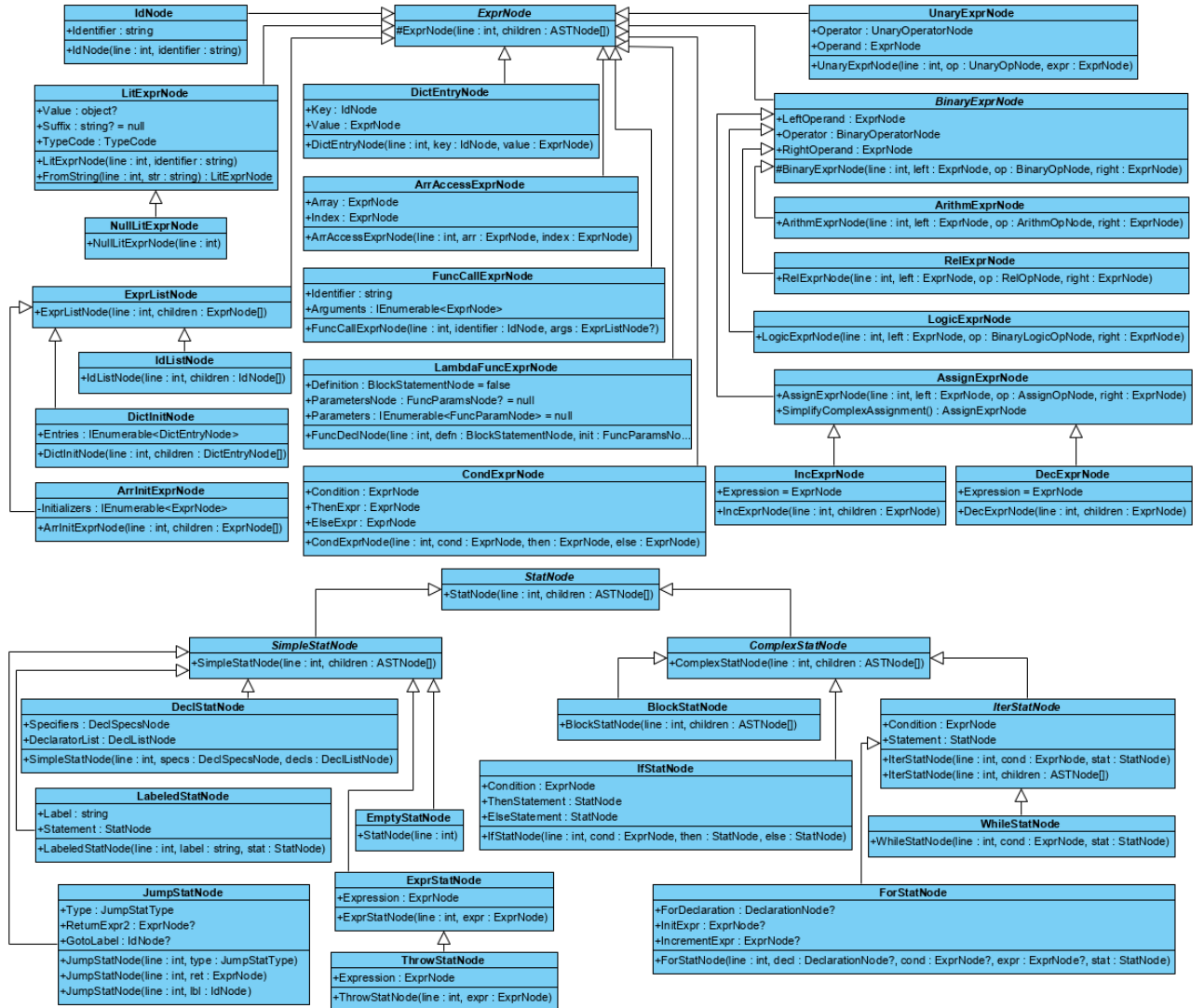
Slika 6.2: UML klasni dijagram opšte apstrakcije (deo 1).



Slika 6.3: UML klasni dijagram opšte apstrakcije (deo 2).

Jednom kreirani AST čvor je *imutabilan*, što znači da je i ceo AST imutabilan — ne mogu se dodavati ili uklanjati deca čvorovima u stablu. Međutim, moguće je klonirati AST čvorove ili vršiti zamenu određenog podstabla drugim podstablom ne menjajući original. Svaki AST čvor se može porediti po jednakosti sa drugim AST čvorom po intuitivnoj logici poređenja pruženom kroz predefinisane operatore poređenja. Dostupne su i ekstenzije koje omogućavaju serijalizaciju stabla u JSON format.

Pored implementacije klasa koje predstavljaju AST čvorove, kreiran je i javno dostupni način za obilazak AST putem obrazca posetilac kroz apstraktnu klasu `BaseASTVisitor<TResult>`. Ova klasa ima implementirane javne virtualno predefinisani metod `TResult Visit(ASTNode node)` jednog argumenta za svaki mogući tip koji nasleđuje tip `ASTNode`. Svi ti metodi su definisani tako da se pozove metod `VisitChildren(ASTNode)` koji će posetiti svu decu trenutnog čvora. Svaki naredni rezultat poziva metoda `Visit()` za svako dete se agregira pozivanjem metoda `TResult AggregateResult(TResult curr, TResult next)` koji podrazumevano samo vraća `next`. Dodatno, moguće je predefinisati metod `bool ShouldVisitNextChild(ASTNode node, TResult curr)` koji određuje da li



Slika 6.4: UML klasni dijagram opšte apstrakcije (deo 3).

je potrebno prekinuti posećivanje pre nego što se posete sva deca. Ovaj metod se poziva svaki put pre nego što se poseti dete i podrazumevano vraća `true`. `BaseASTVisitor` klasa je u implementaciji korišćen za kreiranja graditelja simboličkog izraza od AST koji predstavlja izraz kao i evaluatora takvog AST.

Adapteri ili graditelji se koriste kao spona između stabla parsiranja za konkretni programski jezik i opšte AST apstrakcije. Uloga adaptera je da od stabla parsiranja kreiraju opšti AST. Kako bi se pružio ujednačen način za kreiranje adaptera, pružena su dva interfejsa prikazana na slici 6.5.

Postupak kreiranja AST za dati izvorni fajl se sastoji iz više koraka. Pr-

```
1 public interface IAbstractASTBuilder
2 {
3     ASTNode BuildFromSource(string code);
4 }
5
6 public interface IASTBuilder<TParser> : IAbstractASTBuilder where
    TParser : Parser
7 {
8     TParser CreateParser(string code);
9     ASTNode BuildFromSource(string code, Func<TParser,
        ParserRuleContext> entryProvider);
10 }
```

Slika 6.5: Interfejs graditelja opštog AST od stabla parsiranja.

vo se izvlači ekstenzija izvornog fajla. Zatim se putem refleksije pronalazi klasa koja implementira interfejs `IAbstractASTBuilder` i ima u sebi prisutan atribut `ASTBuilderAttribute(string)`¹ čija se niska poklapa sa ekstenzijom izvornog fajla. Naposljetku se kreira instanca pronađene klase i, na osnovu toga što implementira interfejs `IAbstractASTBuilder`, poziva se metod `BuildFromSource()` za sadržaj izvornog fajla.

Dedukovani adapter implementira uvek `IASTBuilder` interfejs na osnovu tipa parsera koji je generisao ANTLR alat za gramatiku konkretnog programskog jezika. Metod `CreateParser()` instancira parser odgovarajućeg tipa. Razlog zašto ovaj proces nije uniforman je taj što se pre kreiranja parsera uvek kreira i lekser koji prvi dobija tok podataka pa se tek onda kreira parser. Dodatno, prilikom kreiranja parsera se definiše i pravilo gramatike od kojeg parser kreće (tzv. *ulazno pravilo*, engl. *entry rule*) koje je specifično za programski jezik. Stoga je nemoguće kreirati jedinstveni niz operacija koji će kreirati svaki parser pa se to radi prilikom implementacije adaptera.

Predefinisani metod `BuildFromSource()` interfejsa `IASTBuilder` treba da obide stablo parsiranja počev od pravila koje se dobija pozivom funkcije prosledene kao drugi argument za dati tip parsera. Ovo je generalizacija kreiranja parsera i na prvi pogled izgleda suvišno jer, pošto se parsira validan izvorni kôd, mora se uvek krenuti od podrazumevanog ulaznog pravila za konkretni programski jezik.

¹Atributi u C#-u su deklarativni tagovi (nalik na anotacije u Java programskom jeziku) koji se koriste da se raznim elementima pridruže informacije dostupne u toku prevođenja. Ti elementi mogu biti klase, metode, osobine itd. Atributi se definišu kao klase i mogu imati polja i konstruktore s tim što vrednosti atributa moraju biti poznate za vreme prevođenja — konstante.

Međutim, ovaj način kreiranja AST je pogodan prilikom testiranja jedinica koda, jer ukoliko se npr. testira generisanje AST od izraza, nije potrebno pisati čitave funkcije sa naredbama samo da bi se testiralo generisanje apstrakcije izraza, već je moguće samo navesti izraz i instrukovati adapter da kreira parser tako da on počne od pravila koje definiše izraz.

6.2 Implementacija upoređivača

Implementacija algoritma upoređivača opisana u poglavlju 5 se svodi na implementaciju funkcija za poređenje za svaki tip AST čvora. Te funkcije su enkapsulirane u klase koje implementiraju interfejs za upoređivač čvorova. Te klase nisu javne, tako da se poređenje vrši kroz upoređivač koji poredi instance tipa `ASTNode`, a koji putem refleksije određuje konkretni tip čvorova i, ukoliko su tipovi isti, pronalazi konkretni upoređivač i poziva operaciju interfejsa upoređivača. Upoređivači međusobno pozivaju jedni druge, kako bi se logika poređenja uprostila — pošto se naredbe deklaracije sastoje od specifikatora deklaracije i liste deklaratora, upoređivač naredbi deklaracije može pozivati upoređivač za specifikatore deklaracije i upoređivač za listu deklaratora.

Upoređivač kao rezultat svog rada vraća kolekciju potencijalnih problema (upozorenja ili grešaka) koje je detektovao prilikom analize. Ovakav pristup je odabran zbog lakoće testiranja upoređivača, s obzirom da se može očekivati određena kolekcija problema za određeni izvorni kôd. Problem se modeluje kao apstraktna klasa `BaseIssue` dok se upozorenja ili greške modeluju kroz njene konkretizacije `BaseWarning` i `BaseError`. Konkretno greške, kao što su npr. nedostatak deklaracija se mogu onda modelovati kao konkretizacije ovih klasa u zavisnosti od ozbiljnosti problema. Primer izlaza upoređivača za algoritam *swap* se može videti na slici 6.6.

Slika 6.6: Rezultat rada upoređivača za *swap* algoritam.

6.3 Implementacija vizualnog prikaza AST

Osim serijalizacije AST u JSON format, kreiran je potprogram za vizualni prikaz AST-a (u daljem tekstu *vizualizator*). Primeri izlaza vizualizatora se mogu

videti na slikama iz poglavlja 4 — 4.6, 4.7, 4.10 i 4.13. Prikaz je izvršen koristeći nativni `Graphics` paket i, s obzirom da je u pitanju *.NET Core 3.1* radni okvir, moguće je dobiti vizualni prikaz nezavisno od sistema.

Vizualizacija počiva na rekurzivnom algoritmu prikaza u dubinu — za svaki čvor se prikažu potomci, rasporede jednako po širini, a onda se roditelj centrira u odnosu na ukupnu širinu koju zauzimaju deca. Ovaj pristup nije prostorno optimalan, zbog varijacija u broju dece za čvorove različitih tipova. Što se informacija za svaki čvor tiče, prikazuju se vrednosti svih atributa čvora zajedno sa njegovim tipom u zaglavlju kao i grane do njegovih potomaka.

6.4 Implementacija korisničkog interfejsa

S obzirom da je aplikacija konzolna (osim dela komponente za vizualizaciju), korisnički interfejs se sastoji od argumenata komandne linije. Pokretanje programa bez argumenata pruža prikaz za pomoć u kome su nabrojane sve moguće opcije. Pomoć koja se pruža korisniku se može videti na slici 6.7.

Program zahteva da se kao prvi argument prosledi *glagol* (engl. *verb*) koji će odrediti operaciju koju program treba da izvrši. Od glagola zavisi broj i tip ostalih argumenata u nastavku. Mogući glagoli, sa svojim dodatnim opcijama, su:

- `ast [-v -c -t] source-path [-o output-path]`

Generiše opšti AST za izvorni kôd na putanji `source-path` u JSON formatu i ispisuje isti na standardni izlaz, ili u fajl na putanji `output-path` ako je prisutna opcija `-o`. Moguće je generisati kompaktan JSON (bez poravnanja) navođenjem opcije `-c`. Vizualizacija u obliku stabla se prikazuje u novom prozoru ukoliko je prisutna opcija `-t`.

- `cmp [-v] specification-path test-path`

Generiše opšti AST za izvorne kodove na putanjama `specification-path` i `test-path` i poredi generisana stabla.

Dodatno, svi glagoli podržavaju opciono detaljno logovanje operacija koje program izvršava navođenjem opcije `-v`. Takođe, navođenjem opcije `-help` nakon glagola se ispisuje uputstvo specifično za taj glagol. Verzija programa se može proveriti opcijom `--version`. Opcije `-o`, `-c`, `-t` i `-v` imaju svoje duže sinonime — `--output`, `--compact`, `--tree` i `--verbose`, redom.

```
1 $ ./licc
2 ERROR(S):
3 No verb selected.
4
5 ast          AST generation commands
6 cmp          Compare source against the specification source
7 help        Display more information on a specific command.
8 version      Display version information.
9
10 $ ./licc ast
11 ERROR(S):
12 A required value not bound to option name is missing.
13
14 -v, --verbose (Default: false) Verbose output
15 -t, --tree   (Default: false) Visualize AST tree
16 -o, --output Output path
17 -c, --compact (Default: false) Compact AST output
18 --help      Display this help screen.
19 --version    Display version information.
20
21 value pos. 0 Required. Source path
22
23 $ ./licc cmp
24 ERROR(S):
25 A required value not bound to option name is missing.
26
27 -v, --verbose Set output to verbose messages.
28 --help      Display this help screen.
29 --version    Display version information.
30
31 value pos. 0 Required. Specification path.
32 value pos. 1 Required. Test source path.
```

Slika 6.7: Korisniči interfejs programa pružen kroz argumente komandne linije.

6.5 Implementacija testova

Komponentu za kreiranje AST i komponentu za poređenje AST prate testovi jedinica koda. Testovi su organizovani u zasebnom projektu na sledeći način:

- `LICC.Tests.AST` — Testovi za adaptere i posetioce, kao i testovi funkcionalnosti metoda klase `ASTNode`.
- `LICC.Tests.Core` — Testovi upoređivača.

Radni okvir koji se koristi za testiranje je NUnit ² koji pruža tzv. *model ograničenja* (engl. *constraint model*) i time omogućava pisanje čitljivog koda. Pisanje testova po modelu ograničenja se sastoji od korišćenja jednog metoda za pisanje svih testova koji kao argumente prima objekat koji se testira i složeni objekat koji predstavlja ograničenje koje objekat koji se testira treba da zadovoljava. Primer testa pisanog u ovom radnom okviru uz model ograničenja u kontekstu implementacije ovog rada se može videti na slici 6.8.

Osim testova jedinica koda, prisutni su i testovi integracije svih komponenti. Kao što je opisano u prethodnim odeljcima, rezultat rada adaptera je AST, dok je rezultat data upoređivača za data dva stabla kolekcija problema. Ta dva odvojena procesa se onda mogu spojiti kako bi se testirala integracija te dve komponente — dakle, od dva programa očekivati određenu kolekciju problema. Primer za *swap* algoritam se može videti na slici 6.9.

²<https://nunit.org/>


```
1 [Test]
2 public void ComplexDefinitionTest()
3 {
4     FuncDefNode f = this.AssertFunctionSignature(@"
5         float f(const unsigned int x, ...) {
6             int z = 4;
7             return 3.0;
8         }",
9         2, "f", "float", isVariadic: true,
10        @params: ("unsigned int", "x")
11    );
12    Assert.That(f.IsVariadic);
13    Assert.That(f.Definition.Children, Has.Exactly(2).Items);
14 }
15
16 protected FuncDefNode AssertFunctionSignature(
17     string src, int line, string fname,
18     string returnType = "void", bool isVariadic = false,
19     AccessModifiers access = AccessModifiers.Unspecified,
20     QualifierFlags qualifiers = QualifierFlags.None,
21     params (string Type, string Identifier)[] @params)
22 {
23     FuncDefNode f = this.GenerateAST(src).As<FuncDefNode>();
24     this.AssertChildrenParentProperties(f);
25     this.AssertChildrenParentProperties(f.Definition);
26     Assert.That(f, Is.Not.Null);
27     Assert.That(f.Line, Is.EqualTo(line));
28     Assert.That(f.Declarator, Is.Not.Null);
29     Assert.That(f.Declarator.Parent, Is.EqualTo(f));
30     Assert.That(f.Keywords.AccessModifiers, Is.EqualTo(access));
31     Assert.That(f.Keywords.QualifierFlags, Is.EqualTo(qualifiers));
32     Assert.That(f.Identifier, Is.EqualTo(fname));
33     Assert.That(f.ReturnTypeName, Is.EqualTo(returnType));
34     Assert.That(f.IsVariadic, Is.EqualTo(isVariadic));
35     if (@params?.Any() ?? false) {
36         Assert.That(f.Parameters, Is.Not.Null);
37         Assert.That(f.Parameters,
38             Has.Exactly(@params.Length).Items);
39         Assert.That(f.ParametersNode, Is.Not.Null);
40         Assert.That(
41             f.Parameters.Select(
42                 p => (p.Specifiers.TypeName,
43                     p.Declarator.Identifier)
44             ),
45             Is.EqualTo(@params)
46         );
47     }
48     return f;
49 }
```

Slika 6.8: Primer jediničnog testa za proveru generisanog AST čvora za datu funkciju.

```
1  [Test]
2  public override void DifferenceTests()
3  {
4      this.Compare(
5          this.FromPseudoSource(@"
6              algorithm Swap
7              begin
8                  declare integer x = vx
9                  declare integer y = vy
10                 procedure swap()
11                 begin
12                     declare integer tmp = x
13                     x = y
14                     y = tmp
15                 end
16             end
17         "),
18          this.FromCSource(@"
19             int x = vx, y = vy;
20             void swap() {
21                 int tmp = x; y = tmp; x = y;
22             }
23         "),
24          new MatchIssues()
25              .AddError(
26                  new BlockEndValueMismatchError("x", 1, "vy", "vx")
27              )
28              .AddError(
29                  new BlockEndValueMismatchError("x", 3, "vy", "vx")
30              )
31      );
32  }
33
34  protected void Compare(ASTNode src, ASTNode dst,
35      MatchIssues? expectedIssues = null)
36  {
37      expectedIssues ??= new MatchIssues();
38      MatchIssues issues = new ASTNodeComparer(src, dst)
39          .AttemptMatch();
40      Assert.That(issues, Is.EquivalentTo(expectedIssues));
41  }
```

Slika 6.9: Primer kompletnog testa za algoritam *swap*.

Glava 7

Zaključak

Kako softver postaje složeniji i hardver se sve više razvija, potreba za pouzdanim verifikacionim alatima postaje sve veća i uzima sve veći značaj u procesu razvoja softvera. Održavanje koda je najduža faza u razvoju softvera i prilikom održavanja koda se vrše refaktorisanja koja ne menjaju puno izvorni kôd. Ponekad je potrebno i prepisati izvorni kôd sa jednog programskog jezika na drugi. U svim ovim situacijama je moguće izvršiti semantičko poređenje s obzirom da struktura koda ostaje pretežno ista.

U tezi je opisan način posmatranja apstrakne sintakse programa kroz AST, opisan je proces kreiranja AST od proizvoljne gramatike programskog jezika i opšte-prihvaćen interfejs za obilazak istog. Opisan je model opšte AST apstrakcije sa ciljem dovođenja imperativnih i skript jezika na isti nivo apstrakcije. Ova apstrakcija je korišćena za određivanje semantičke ekvivalentnosti strukturno sličnih segmenata koda kroz naivni algoritam poređenja vrednosti na krajevima blokova.

Kao glavni doprinos teze, implementiran je računarski program za kreiranje opšteg AST od izvornog fajla, serijalizaciju i prikaz istog. Mehanizam dobijanja opšteg AST omogućava jednostavno proširenje za već postojeće ali i za proizvoljne gramatike, kroz implementaciju adaptera za tu gramatiku koji služi kao posrednik između stabla parsiranja i opšteg AST. Dodatno, kao jedna primena opšte AST apstrakcije, implementiran je opisani algoritam za semantičko poređenje kroz proširiv model uporedivača tipova opštih AST čvorova.

Naredni koraci u dizajniranju modela opšte apstrakcije bi bili usmereni na podršku za korisnički definisane tipove kroz čvorove za opis klasa, struktura i enumeracija. Takođe, klase se u skript jezicima često izbegavaju tako što se podaci smeste u mapu gde ključevi imitiraju attribute klase. Stoga bi bilo poželjno imati

i interfejs za kreiranje mape objekta od datog klasnog čvora ali i obrnuto. Osim korisnički definisanih tipova, oseća se i potreba za apstrahovanjem čestih struktura podataka kao što su skupovi, torke i redovi sa prioritetom. Na taj način, ako se u jednom programu koristi niz a u drugom jednostruko ulančana lista sa definisanim indeksnim pristupom, moguće je posmatrati te programe kao jednake uz potencijalno upozorenje o gubitku na efikasnosti. U planu je i proširivanje modela tako da se uključe i druge programske paradigme osim imperativne.

Literatura

- [1] ANTLR4. <https://www.antlr.org/>.
- [2] BYACC. <http://byaccj.sourceforge.net/>.
- [3] Clang. <https://clang.llvm.org/>.
- [4] Context-Free Grammars. https://www.cs.rochester.edu/~nelson/courses/csc_173/grammars/cfg.html.
- [5] Keith Cooper and Linda Torczon. *Engineering a Compiler, 2nd Edition*. 2013.
- [6] Joel Denny and Brian Malloy. IELR(1): practical LR(1) parser tables for non-LR(1) grammars with conflict resolution. pages 240–245. Association for Computing Machinery, New York, NY, United States.
- [7] Design Patterns. <http://www.vincehuston.org/dp/>.
- [8] Flex. <https://www.gnu.org/software/flex/>.
- [9] Erich Gamma, Richard Helm, Ralph Johnson, and John Vlissides. *Design Patterns: Elements of Reusable Object-Oriented Software*. 1995.
- [10] GLR. https://www.gnu.org/software/bison/manual/html_node/GLR-Parsers.html.
- [11] GNU Bison. <https://www.gnu.org/software/bison/>.
- [12] GNU Project. <https://www.gnu.org/gnu/thegnuproject.en.html>.
- [13] KLEE. <https://klee.github.io/>.
- [14] LALR1. <https://web.cs.dal.ca/~sjackson/lalr1.html>.
- [15] LexYacc. <http://dinosaur.compilertools.net/>.

- [16] LLStar. <https://wwwantlr.org/papers/LL-star-PLDI11.pdf>.
- [17] LLVM. <https://llvm.org/>.
- [18] LR. <http://pages.cpsc.ucalgary.ca/~robin/class/411/LR.1.html>.
- [19] Terrence Parr. *The Definitive ANTLR 4 Reference*. 2012.
- [20] ProgrammingParadigms. <https://cs.lmu.edu/~ray/notes/paradigms/>.
- [21] Slonneger and Kurtz. *Formal syntax and semantics of programming languages*. Addison-Wesley Pub. Co, Reading, Mass, 1995.
- [22] SMT. https://resources.mpi-inf.mpg.de/departments/rg1/conferences/vtsa08/slides/barret2_smt.pdf.
- [23] Symbolic Execution. <https://www.cs.umd.edu/~mwh/se-tutorial/symbolic-exec.pdf>.
- [24] Design Patterns. <https://www.uml.org/>.
- [25] William M. Waite and Gerhard Goos. *Compiler Construction*. 1995.

Biografija autora

Ivan Ž. Ristović rođen je 17.01.1995. godine u Užicu. Osnovnu školu, kao i prirodno-matematički smer Užičke gimnazije, završio je kao nosilac Vukove diplome. Tokom navedenog perioda školovanja isticao se u oblastima matematike, informatike, fizike, hemije i engleskog jezika, što potvrđuje veći broj nagrada na Državnim takmičenjima.

Smer Informatika na Matematičkom fakultetu Univerziteta u Beogradu upisuje 2014. godine. Na navedenom smeru je diplomirao 2018. godine, posle tri godine studija sa prosečnom ocenom 9,17. Master studije upisuje na istom fakultetu odmah nakon diplomiranja.

U avgustu 2018. biva izabran u zvanje „Saradnik u nastavi“ na Matematičkom fakultetu paralelno sa master studijama. Drži vežbe iz kurseva „Računarske mreže“, „Funkcionalno programiranje“, „Programske paradigme“ i „Objektno orijentisano programiranje“ na kasnijim godinama osnovnih studija.

Oblasti interesovanja uključuju pre svega razvoj i verifikaciju softvera, mikroservise i računarske mreže.