

Prezentacja – kompilatory

22.03.2018



Program, który czyta kod napisany w jednym języku (*źródłowym*) i tłumaczy go na równoważny kod w drugim języku (*wynikowym*)

- Analiza leksykalna
- Analiza składniowa
- Analiza semantyczna
- *Generacja kodu pośredniego*
- *Optymalizacja*
- Generacja kodu

Kod źródłowy

```
// Ustaw wartosc a  
var a = 4 + 0.2
```

⇒

Tokeny

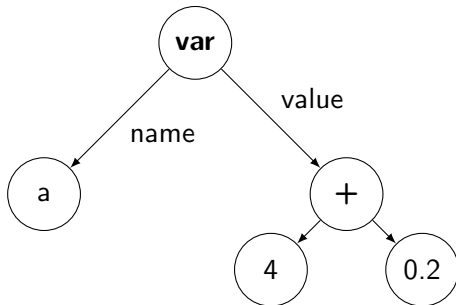
```
keyword(var),  
id(a),  
operator(=),  
integer(4),  
operator(+),  
real(0.2)
```

Tokeny

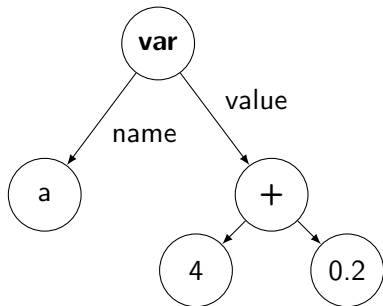
keyword(var),
id(a),
operator(=),
integer(4),
operator(+),
real(0.2)

⇒

AST

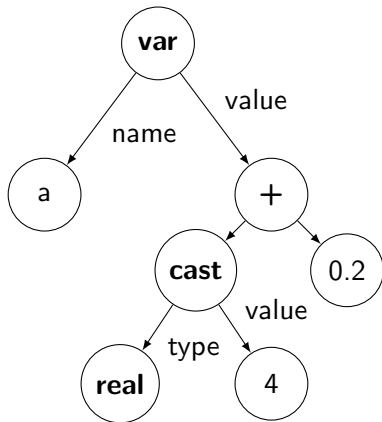


AST

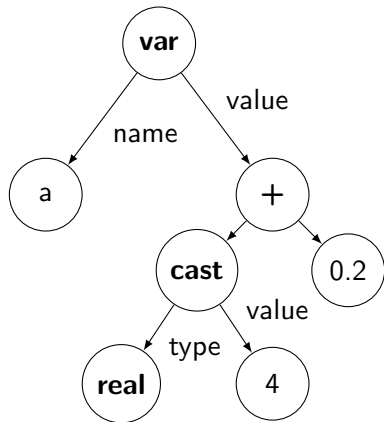


\Rightarrow

AST'



AST'



⇒ ... ⇒

Kod wynikowy

```
.LC0:  
    .long 1082549862  
_start:  
    movss xmm0, .LC0[rip]  
    movss -4[rbp], xmm0  
    mov eax, 0  
    ret
```

Notacja infiksowa \rightarrow Notacja postfiksowa

- Notacja infiksowa: $e_1 + e_2$
- Notacja postfiksowa (odwrotna notacja polska): $e_1 e_2 +$
 - Przykłady:
 - $1 + 2 * 3 \Rightarrow 1\ 2\ 3\ *\ +$
 - $(1 + 2) * 3 \Rightarrow 1\ 2\ +\ 3\ *$
 - Nie wymaga nawiasów
 - Łatwo obliczać wartość wyrażenia (implementacja na stosie)
 - Niestety nieczytelna dla ludzi :(

Notacja infiksowa \rightarrow Notacja postfiksowa

- Notacja infiksowa: $e_1 + e_2$
- Notacja postfiksowa (odwrotna notacja polska): $e_1 e_2 +$
 - Przykłady:
 - $1 + 2 * 3 \Rightarrow 1\ 2\ 3\ *\ +$
 - $(1 + 2) * 3 \Rightarrow 1\ 2\ +\ 3\ *$
 - Nie wymaga nawiasów
 - Łatwo obliczać wartość wyrażenia (implementacja na stosie)
 - Niestety nieczytelna dla ludzi :(

Napiszmy kompilator!

Definicja składni v1 (Składnia abstrakcyjna)

wyrażenie \rightarrow wyrażenie + wyrażenie |
wyrażenie * wyrażenie |
(wyrażenie) | liczba

Definicja składni v1 (Składnia abstrakcyjna)

wyrażenie \rightarrow wyrażenie + wyrażenie |
wyrażenie * wyrażenie |
(wyrażenie) | liczba

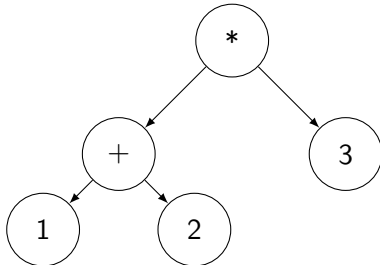
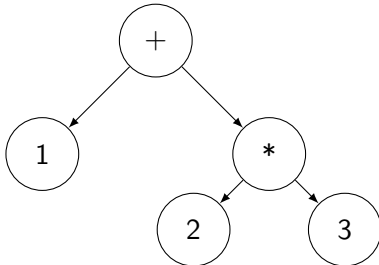
Jaki jest problem?

Definicja składni v1 (Składnia abstrakcyjna)

wyrażenie \rightarrow wyrażenie + wyrażenie |
wyrażenie * wyrażenie |
(wyrażenie) | liczba

Jaki jest problem?

1 + 2 * 3



Definicja składni v2 (Składnia konkretna)

wyrażenie \rightarrow składnik | składnik + wyrażenie

składnik \rightarrow czynnik | czynnik * składnik

czynnik \rightarrow (wyrażenie) | liczba

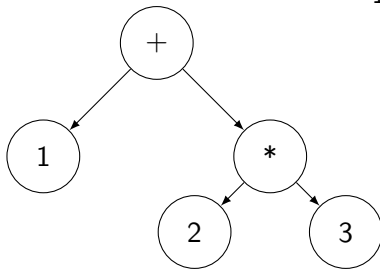
Definicja składni v2 (Składnia konkretna)

wyrażenie \rightarrow składnik | składnik + wyrażenie

składnik \rightarrow czynnik | czynnik * składnik

czynnik \rightarrow (wyrażenie) | liczba

1 + 2 * 3



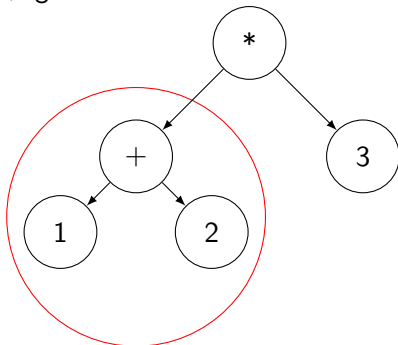
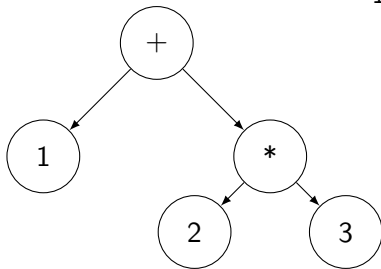
Definicja składni v2 (Składnia konkretna)

wyrażenie \rightarrow składnik | składnik + wyrażenie

składnik \rightarrow czynnik | czynnik * składnik

czynnik \rightarrow (wyrażenie) | liczba

1 + 2 * 3



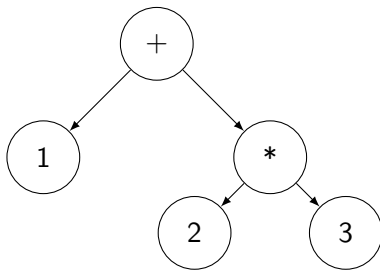
Definicja składni v2 (Składnia konkretna)

wyrażenie \rightarrow składnik | składnik + wyrażenie

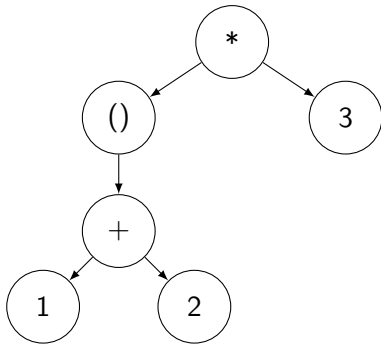
składnik \rightarrow czynnik | czynnik * składnik

czynnik \rightarrow (wyrażenie) | liczba

1 + 2 * 3



(1 + 2) * 3



Dygresja: składnia abstrakcyjna vs składnia konkretna

Składnia abstrakcyjna

- Prostsza
- Każde drzewo wyprowadzenia wyznacza jednoznacznie program
- Potencjalnie wiele drzew wyprowadzeń dla danego programu

Składnia konkretna

- Definiowanie wymaga więcej wysiłku
- Drzewo wyprowadzenia może zawierać nadmiarowe węzły
- Zwykle spełnia dodatkowe własności, umożliwiającą wydajną analizę składniową (np. jednoznaczność, LL, LR)

- `obecny()` – zwraca obecny znak
- `nastepny()` – wczytuje kolejny znak z wejścia
- `wypisz()`, `wypisz_liczbe()` – emituje ciąg znaków/liczbę
- `error()` – sygnalizuje błąd i kończy program

Lekser (analiza leksykalna)

```
int liczba() {  
    int wynik = 0;  
    if (!isdigit(obecny())) error();  
    while (isdigit(obecny())) {  
        wynik = wynik*10 + (obecny() - '0');  
        nastepny();  
    }  
    return wynik;  
}
```

Lekser (analiza leksykalna)

```
int liczba() {  
    int wynik = 0;  
    if (!isdigit(obecny())) error();  
    while (isdigit(obecny())) {  
        wynik = wynik*10 + (obecny() - '0');  
        nastepny();  
    }  
    return wynik;  
}  
  
bool probuj_znak(char c) {  
    if (obecny() != c) return false;  
    nastepny();  
    return true;  
}
```

Lekser (analiza leksykalna)

```
int liczba() {
    int wynik = 0;
    if (!isdigit(obecny())) error();
    while (isdigit(obecny())) {
        wynik = wynik*10 + (obecny() - '0');
        nastepny();
    }
    return wynik;
}

bool probuj_znak(char c) {
    if (obecny() != c) return false;
    nastepny();
    return true;
}

void znak(char c) {
    if (!probuj_znak(c)) error();
}
```

Parser (analiza składowa) + generator kodu

wyrażenie \rightarrow składnik | składnik + wyrażenie

```
void wyrażenie() {  
    składnik();  
    if (próbuj_znak('+')) {  
        wyrażenie();  
        wypisz("+ ");  
    }  
}
```

Parser (analiza składowa) + generator kodu

składnik \rightarrow czynnik | czynnik * składnik

```
void skladnik() {  
    czynnik();  
    if (probuj_znak('*')) {  
        skladnik();  
        wypisz("* ");  
    }  
}
```

Parser (analiza składniowa) + generator kodu

czynnik \rightarrow (wyrażenie) | liczba

```
void czynnik() {  
    if (próbuj_znak('(')) {  
        wyrażenie();  
        znak(')');  
    } else {  
        int n = liczba();  
        wypisz_liczbe(n);  
    }  
}
```

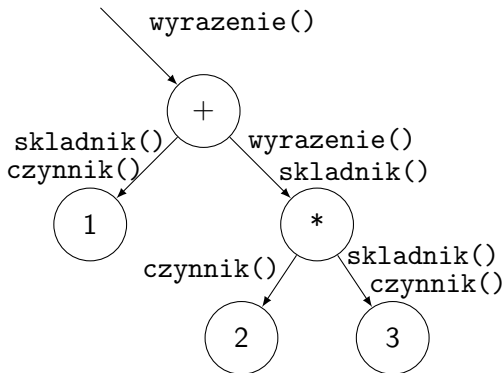
Brakujące AST

Ok, tylko gdzie tu jest drzewo AST?

Brakujące AST

Ok, tylko gdzie tu jest drzewo AST?
Jest konstruowane na bieżąco!

1 + 2 * 3



Demo – kompilator do ONP

Bardziej skompilowane gramatyki

- Powyższa gramatyka była bardzo prosta – nigdy nie potrzebowaliśmy więcej niż następnego znaku
- Prawdziwe gramatyki są bardziej skomplikowane
- Ogólne metody analizy języków bezkontekstowych – $O(n^3)$
- Języki $LL(k)$ – mogą być sparsowane zstępująco używając co najwyżej najbliższych k znaków
- Języki $LR(k)$ – mogą być sparsowane wstępująco używając co najwyżej najbliższych k znaków

```
def parse_block() {  
  eat('{');  
  var ret = [];  
  while (!try_eat('}')) {  
    try var tmp = parse_cmd();  
    array::push(ref ret, tmp);  
  }  
  return :block(ret);  
}
```

Parser NL

```
def parse_cmd() {  
  if (try_eat('if')) {  
    var tmp = {};  
    try tmp->cond = parse_cond(ref state);  
    try tmp->if_cmd = parse_block(ref state);  
    if (try_eat('else')) {  
      try tmp->else_cmd = parse_block(ref state);  
    } else {  
      tmp->else_cmd = :nop;  
    }  
    return :if(tmp);  
  } elsif (try_eat('for')) {  
    ...  
  }  
}
```

Automatyczne generowanie parserów

- lex – generator lekserów
- yacc (yet another compiler compiler) – generator parserów
- bison – nowsza wersja programu yacc

Automatyczne generowanie parserów

- lex – generator lekserów
- yacc (yet another compiler compiler) – generator parserów
- bison – nowsza wersja programu yacc



Notacja infksowa → Notacja postfiksowa raz jeszcze

lex

```
[0-9]+ {yyval=atoi(yytext); return LICZBA;}  
\n      return 0;  
      return *yytext;
```

Notacja infksowa → Notacja postfiksowa raz jeszcze

lex

```
[0-9]+ {yylval=atoi(yytext); return LICZBA;}  
\n      return 0;  
      return *yytext;
```

yacc

```
%token LICZBA  
%left '+' '*'  
E:    E '+' E {printf("+ ");}  
      | E '*' E {printf("* ");}  
      | '(' E ')'  
      | LICZBA {printf("%d ", yyval);}
```

- Bardzo zależna od języka
- Wykrywanie konstrukcji poprawnych składniowo, ale nie mających sensu w semantyce języka
- Obliczanie różnych cech poszczególnych węzłów i zapisywanie ich w drzewie
- Sprawdzanie typów
- Modyfikacja drzewa AST dla operacji implicite (np. rzutowanie w językach słabo typowanych)

```
def check_cmd(cmd) {  
  match (cmd) case :if(var as_if) {  
    var cond_type = check_value(as_if->cond);  
    if (!type::is_accepted(cond_type, type::bool())) {  
      error();  
    }  
    check_cmd(as_if->if_cmd);  
    check_cmd(as_if->else_cmd);  
  } case :for(var as_for){  
    ...  
  }  
}
```

Generowanie kodu pośredniego

- Kod dla pewnej abstrakcyjnej maszyny, bliższej maszynie docelowej, ale ukrywającej jej szczegóły
- Ułatwia zmianę języka docelowego (np. na inną architekturę procesora)
- Wnioskowanie i optymalizacje często są łatwiejsze niż na drzewie AST
- Umożliwia wydajniejszą interpretację (uproszczone parsowanie i kompilacja)

Rodzaje kodu pośredniego

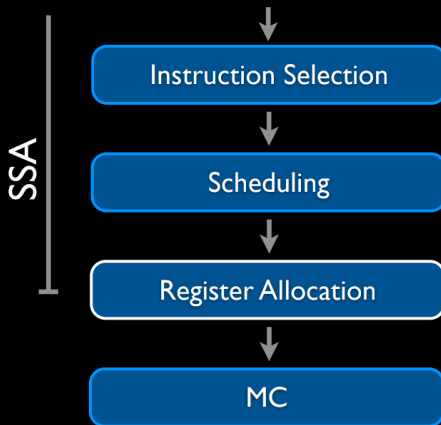
- Kod dla maszyny rejestrowej
 - LLVM
 - nasm
- Kod dla maszyny stosowej
 - bajtkod .NET – CIL
 - bajtkod JVM
 - bajtkod CPythona
 - bajtkod WebAssembly

- Silnie typowany kod dla maszyny rejestrowej z funkcjami
- Standardowy backend dla różnych języków (C, C++, Haskell, Fortran)
- Kod trójadresowy: większość instrukcji jest postaci
wynik = operator typ arg1 arg2
- SSA – static single assignment, na dowolny rejestr można przypisać wartość tylko raz

LLVM – przykład (ze smurfa [4])

```
define i32 @func(i32 %n) {  
    %c0 = icmp eq i32 %n, 0  
    br i1 %c0, label %L0, label %L1  
L0:  
    ret i32 1  
L1:  
    %i1 = sub i32 %n, 1  
    %i2 = call i32 @func(i32 %i1)  
    %i3 = mul i32 %n, %i2  
    ret i32 %i3  
}
```


Register Allocation in LLVM



.NET Common Intermediate Language

- Silnie typowany kod dla maszyny stosowej
- Punkt integracji języków C#, F#, VB.NET
- Wspiera ideę programowania obiektowego, na której oparta jest cała platforma
- Wszystkie operacje dzieją się przy użyciu stosu (dodawanie, porównywanie, wołanie metod)

```
.class public abstract auto ansi beforefieldinit Sample
extends [mscorlib]System.Object
{
    .field public static string SomeText
    .method public hidebysig static void Main() cil
        managed
    {
        .entrypoint
        .maxstack 1
        .locals init (int32 V_0)
        IL_0000: ldc.i4.0
        IL_0001: stloc.0
        IL_0002: ldloc.0
        IL_0003: call      void
                    [mscorlib]System.Console::WriteLine(int32)
        IL_0008: ret
    } // end of method Sample::Main
} // end of class Sample
```

- Upraszczanie kodu wynikowego (np. zamiana mnożenia przez 2 przez left-shift)
- Zmniejszanie wielkości pliku wynikowego (XOR na rejestrze, zamiast przypisania 0)
- Zmiana ogonowych funkcji rekurencyjnych na iteracyjne
- Cache'owanie wyniku funkcji bez skutków ubocznych przy wielokrotnym jej wołaniu
- Zmiana kolejności zagnieżdżonych pętli, aby czytać wartości z pamięci blisko siebie (najlepiej jak się zmieszczą w cache'u procesora)

Bibliografia

- ❶ A.V. Aho, R. Sethi, J.D. Ullman, Kompilatory. Reguły, metody i narzędzia
- ❷ http://wazniak.mimuw.edu.pl/index.php?title=SW_wyk%C5%82ad_2_-_Slajd2
- ❸ http://wazniak.mimuw.edu.pl/index.php?title=SW_wyk%C5%82ad_2_-_Slajd3
- ❹ <http://smurf.mimuw.edu.pl/node/797>
- ❺ https://llvm.org/devmtg/2011-11/0lesen_RegisterAllocation.pdf
- ❻ <https://llvm.org/devmtg/2016-09/slides/Absar-SchedulingInOrder.pdf>