

Tytuł

Patryk Wałach

January 2022

Contents

1	[Wstęp]	2
2	Ogólne wprowadzenie	2
2.1	Analiza leksykalna	2
2.2	Parser	3
2.3	System typów	3
2.4	System typów ML	4
2.5	Rekord z wariantami	4
3	Inferencja typów w teorii	5
3.1	Problem inferencji typów	5
3.2	Udowadnianie typów dla ML-the-calculus	5
3.3	Inferencja typów bazująca na ograniczeniach	6
3.4	System typów Hindley–Milner	6
3.5	Algorytm J	6
4	Rescript/ReasonML jako języki realizujące podobne zadania	7
5	Założenia i priorytety opracowanej aplikacji	7
5.1	Opis formalny składni języka	11
6	Narzędzia	11
6.1	Język python	11
6.2	Parsowanie i tokenizowanie przy użyciu biblioteki sly	11
6.2.1	Lexer	12
6.2.2	Parser	12
6.3	Środowisko nodejs do uruchomienia skompilowanego kodu	13
7	Implementacja	13
7.1	lexer	13
7.2	parser	14
7.3	Drzewo decyzyjne	15
7.4	inferencja typów	17

7.4.1	Context	17
7.4.2	Aplikowanie substytucji	17
7.4.3	Literały	19
7.4.4	Identyfikatory	19
7.4.5	Bloki wyrażeń	19
7.4.6	Operatory binarne	19
7.4.7	Wywołania funkcji	20
7.4.8	Wyrażenia warunkowe	20
7.4.9	Definicje funkcji	21
7.5	Pattern matching	21
7.6	Hoisting	24
7.7	Kompilacja	25
7.8	Literały	25
7.8.1	Wyrażenia warunkowe	26
7.8.2	Warianty	26
7.8.3	Pattern matching	26
8	Opis działania	27
8.1	Co działa	27
8.2	Uwagi do działania	27
8.2.1	Obsługa błędów	27
8.2.2	Pattern matching	27
9	[Podsumowanie]	28
10	[spisy – rysunków, tabel, listingów ipt.]	28

1 [Wstęp]

2 Ogólne wprowadzenie

2.1 Analiza leksykalna

Analiza leksykalna w informatyce jest to proces rozbijania program źródłowych na jednostki logiczne (zwane leksemami) złożone z jednego lub więcej znaków, które łącznie mają jakieś znaczenie[J.E. Hopcroft(2005)]. Przykładami leksemów mogą być słowa kluczowe (np. while), identyfikator lub liczba składająca się z cyfr. Rozdzielaniem programu źródłowego, na leksemy, zajmuje się lekser.

Token jest strukturą reprezentującą leksem i wprost go kategoryzującą[Aho et al.(1985)Aho, Sethi, and Ullman]. Tokeny ułatwiają pracę parserowi. Tokeny kategoryzuje się na komputerowy odpowiednik tego, co lingwiści określiliby mianem części mowy. Biorąc jako przykład poniższy kod w języku C:

```
1 x = a + b * 2;
```

analiza leksykalna, zwraca tokeny:

```
1 [(identifier , x), (operator , =), (identifier , a), (operator , +),  
  (identifier , b), (operator , *), (literal , 2), (separator , ;)]
```

Dwoma ważnymi przypadkami są znaki białe i komentarze. One również muszą być uwzględnione w gramatyce i przeanalizowane przez lexer, lecz mogą być odrzucone (nie produkować żadnych tokenów) i traktowane jako spełniające mało znaczące zadanie, rozdzielania dwóch tokenów (np. w `if x` zamiast `ifx`).

2.2 Parser

Analizator składniowy, parser – program dokonujący analizy składniowej danych wejściowych w celu określenia ich struktury gramatycznej w związku z określoną gramatyką formalną. Analizator składniowy umożliwia przetworzenie tekstu czytelny dla człowieka w strukturę danych przydatną dla oprogramowania komputera. Wynikiem analizy składni, dokonywanej przez parser, najczęściej jest drzewo składniowe nazywane czasami drzewem wyprowadzenia[Aho et al.(1985)Aho, Sethi, and Ullman].

Zadanie parsera sprowadza się do sprawdzenia czy i jak dane wejściowe mogą zostać wyprowadzone z symbolu startowego. To zadanie można zrealizować na dwa sposoby:

- Analiza zstępująca (ang. top-down parsing) to strategia znajdowania powiązań między danymi przez stawianie hipotez dotyczących drzewa rozbioru składniowego i sprawdzanie, czy zależności między danymi są zgodne z tymi hipotezami.
- Analiza wstępująca (ang. bottom-up parsing) – ogólna metoda analizy składniowej, w której zaczyna się od słowa wejściowego i próbuje się zredukować je do symbolu startowego. Drzewo wyprowadzenia jest konstruowane od liści do korzenia (stąd nazwa). W każdym momencie w trakcie tego procesu mamy formę zdaniową, która zawiera segment, powstały w ostatnim kroku wyprowadzenia. Segment ten nazywany uchwytem (ang. handle) jest prawą stroną produkcji i powinien zostać w tym kroku zredukowany do jej lewej strony, w wyniku czego powstanie poprzednia forma zdaniowa z wyprowadzenia. Główna trudność w analizie wstępującej polega właśnie na odpowiednim znajdowaniu uchwytów. Analiza wstępująca może przebiegać w określonym kierunku (np. od lewej do prawej), lub w sposób bezkierunkowy, wtedy analizowane jest całe słowo naraz. Jednym z bardziej znanych przedstawicieli metody bezkierunkowej jest algorytm CYK. Do metod kierunkowych zalicza się między innymi parsery shift-reduce czyli LR, LALR, SLR, BC, pierwszeństwa.

2.3 System typów

System typów jest to system klasyfikacji wyrażeń w zależności od rodzajów wartości, jakie one generują[Pierce(2002)]. Każdej obliczonej wartości przypisy-

wany jest pewien typ, który jednoznacznie definiuje, jakie operacje można na niej wykonać. Śledząc przepływ wartości, system typów stara się udowodnić, że w programie występuje poprawne typowanie, tzn. nie dochodzi do sytuacji, w której na wartości określonego typu próbujemy wykonać niedozwoloną operację.

2.4 System typów ML

System typów ML jest to silny system typów stosowany w językach rodziny ML (Ocaml, Standard ML) oparty na inferencji.

Podstawowy system typów jest następujący: istnieją typy proste, takie jak string, int, bool, unit (typ pusty) itd. Z dowolnych typów można też generować typy złożone – przez krotki ($\text{typ1} * \text{typ2}$, $\text{typ1} * \text{typ2} * \text{typ3}$ itd.), konstruktory typów (typ list, typ tree itd.) i funkcje ($\text{typ1} \rightarrow \text{typ2}$).

System próbuje nadać typy każdemu wyrażeniu języka, i nie licząc kilku rzadkich przypadków, udaje mu się to całkiem dobrze.

Generalnie system taki wyklucza polimorfizm (nie licząc typów polimorficznych), jednak w Standard ML stworzono specjalne reguły umożliwiające polimorfizm dla wyrażeń arytmetycznych.

System typów ML jest interesujący z teoretycznego punktu widzenia – wiele problemów ma bardzo wysoką złożoność, jednak w praktyce inferencja zachodzi bardzo szybko – typy, które są rzeczywiście używane, są zwykle bardzo proste – rzadko używa się funkcji rzędów wyższych niż trzeci-czwarty, oraz liczby argumentów większej niż kilkanaście.

W rzeczywistych implementacjach dochodzą do tego bardziej złożone problemy typizacji obiektów, modułów itd.

2.5 Rekord z wariantami

Rekord z wariantami jest to rodzaj rekordu, posiadającego tę właściwość, że zbiór rekordów posiada wspólny typ, lecz różną postać, określoną aktualną wartością specjalnego pola znacznikowego.

Przykład - Zakładając, że chcemy stworzyć drzewo binarne intów. W języku ML, zrobilibyśmy to tworząc nowy typ danych w ten sposób:

```
1 datatype tree = Leaf
2   | Node of (int * tree * tree)
```

Leaf i Node są konstruktorami, które pozwalają nam na stworzenie konkretnego drzewa, np.

```
1 Node(5, Node(1, Leaf, Leaf), Node(3, Leaf, Node(4, Leaf, Leaf)))
```

3 Inferencja typów w teorii

3.1 Problem inferencji typów

Język ML przyjmuje wiele form, najpopularniejszymi wariantami są Standard ML (SML), OCaml, i F#. Na potrzeby będziemy wzorować się na [Damas and Milner(1982)], i posługiwać się ML-the-calculus, drastycznie uproszczoną wersją języka co pozwoli na dojście do sedna w problemie rekonstrukcji typów.

Termy w języku ML-the-calculus są następujące:

$$\begin{array}{l}
 e ::= x \\
 \quad | \quad c \quad (\text{stałe}) \\
 \quad | \quad \lambda x.e \\
 \quad | \quad e \ e \\
 \quad | \quad \text{let } x = e \text{ in } e
 \end{array} \tag{1}$$

Typy które będziemy przypisywali do termów są następujące:

$$\begin{array}{l}
 \tau ::= \alpha \quad (\text{typ zmienny}) \\
 \quad | \quad B \quad (\text{typ podstawowy}) \\
 \quad | \quad \tau \rightarrow \tau
 \end{array} \tag{2}$$

3.2 Udowadnianie typów dla ML-the-calculus

$$\overline{\Gamma \vdash c : B}$$

$$\frac{\Gamma, x : \tau' \vdash e : \tau}{\Gamma \vdash \lambda x.e : \tau' \rightarrow \tau}$$

$$\frac{\Gamma \vdash e_1 : \tau_2 \rightarrow \tau \quad \Gamma \vdash e_2 : \tau_2}{\Gamma \vdash e_1 \ e_2 : \tau}$$

$$\frac{\Gamma(x) = \bigwedge \alpha_1, \dots, \alpha_n. \tau' \quad \tau = [\beta_i / \alpha_i] \tau' (\beta_1 \text{ fresh})}{\Gamma \vdash x : \tau}$$

$$\frac{\Gamma \vdash e_1 : \tau' \quad \Gamma, x : (\bigwedge \alpha_1, \dots, \alpha_n. \tau') \vdash e_2 : \tau}{\Gamma \vdash \text{let } x = e_1 \text{ in } e_2 : \tau} (\{\alpha_1, \dots, \alpha_n\} = \text{ftv}(\tau') \setminus \text{ftv}(\Gamma)) \tag{3}$$

Powyższe dowody nie mogą jednak być w prosty sposób przedstawione jako algorytm. Fakt, że typ τ' jest dowolnie wybierany w dowodzie dla λ prowadzi do nieskończonej ilości dowodów, nawet dla prostego wyrażenia $\lambda x.x$.

3.3 Inferencja typów bazująca na ograniczeniach

Algorytmy inferencji typów bazujące na ograniczeniach generują dużą liczbę zmiennych typów i zbiór ograniczeń dla tych zmiennych. W drugim kroku algorytmu dla każdej zmiennej, szuka typu, który spełnia wszystkie ograniczenia.

3.4 System typów Hindley–Milner

Algorytm inferencji typów przedstawiony w [Milner(1978)], oraz innych wcześniejszych publikacjach, opiera się na generowaniu ograniczeń i rozwiązywaniu ich w trakcie wykonywania prostego rekursywnego przejścia przez termy.

Warto zauważyć, że w ten sposób: dostajemy, prostą strukturalnie rekursywną definicję rekonstrukcji typów, ale tracimy modularność algorytmu. Rozszerzenie takiego algorytmu o dodatkowe funkcje jest o wiele trudniejsze, niż algorytmu, który oddziela kroki generowania ograniczeń i ich rozwiązywania.

3.5 Algorytm J

Algorytm W nie ma żadnych side-effectów, i świetnie zajmuje się aplikowaniem i komponowaniem substytucji w odpowiedniej kolejności. Częste aplikacje substytucji na wyrażeniach mogą znacznie zmniejszyć wydajność algorytmu rekonstruującego, dlatego też, Milner zaprezentował bardziej efektywną impereatywną wariację W nazwaną algorytmem J w [Milner(1978)].

Algorytm J jest funkcją, która dla dokonanych do tej pory substytucji S , środowiska Γ - zbioru przechowywującego pary identyfikator wraz z typem, i

wyrażnia e , zwraca kolejne substytucje S , oraz typ wyrażenia τ .

$$J : S \times \Gamma \times e \rightarrow S \times \tau$$

$$\begin{aligned} J(S, \Gamma, x) = & (S, [\beta_i/\alpha_i]\tau') \\ \text{where } \Gamma(x) = & \bigwedge \alpha_i, \dots, \alpha_n. \tau' \\ \text{and } \beta_i \text{ are fresh} \end{aligned}$$

$$\begin{aligned} J(S, \Gamma, e_1 \ e_2) = & (V, \beta) \\ \text{where } \Gamma(S_1, \tau_1) = & J(S, \Gamma, e_1) \\ \text{and } (S_2, \tau_2) = & J(S_1, \Gamma, e_2) \\ \text{and } V = \text{unify}'(\tau_1, \tau_2 \rightarrow \beta, S_2) \\ \text{and } \beta \text{ is fresh} \end{aligned}$$

$$\begin{aligned} J(S, \Gamma, \lambda x. e) = & (S_1, \beta \rightarrow \tau) \\ \text{where } (S_1, \tau) = & J(S, (\Gamma, x : \beta), e) \\ \text{and } \beta \text{ is fresh} \end{aligned}$$

$$\begin{aligned} J(S, \Gamma, \text{let } x = e_1 \text{ in } e_2) = & (S_2, \tau_w) \\ \text{where } (S_1, \tau_1) = & J(S, \Gamma, e_1) \\ \text{and } (S_2, \tau_2) = & J(S_1, (\Gamma, x : (\bigwedge \alpha_1, \dots, \alpha_n. \tau_1)), e_2) \\ \text{and } \{\alpha_1, \dots, \alpha_n\} = & \text{ftv}(S_2 \tau_1) \setminus \text{ftv}(S_2 \Gamma) \end{aligned} \tag{4}$$

Funkcja pomocnicza $\text{unify}'(\tau, \tau', S)$ rozszerza zbiór substytucji S o substytucje wynikające z unifikacji τ z τ' pod kontekstem S .

4 Rescript/ReasonML jako języki realizujące podobne zadania

5 Założenia i priorytety opracowanej aplikacji

Tworząc aplikację, chciałem, by język posiadał podstawowe typy danych (liczby, stringi, wartość logiczna), kilka typów generycznych (funkcje, tablice), typ ‘Option’, oraz możliwość tworzenia własnych typów. Dodatkowo nie powinno być potrzeby podawania typów zmiennych w większości przypadków, kompilator sam powinien wykrywać typy zmiennych na podstawie ich użycia.

Język poza zmiennymi, potrzebuje możliwości wykonywania operacji na danych, dlatego ważne było dla mnie, by zaimplementować operatory binarne, oraz

unarne. Operatory te miały też spełniać ważną rolę w trakcie inferencji typów. W języku javascript operator '+' może być wykorzystywany do dodawania liczb jak i konkatencji stringów, ważne więc było by stworzyć dwa oddzielne operatory.

Prymitywne typy danych:

- string

```
1 greeting = 'Hello world!'
```

examples/string.uwu

Do konkatencji stringów służy operator ++

- Wartość logiczna ma typ Bool i wartość True lub False. Powiązane operacje:

- <>, równość pomiędzy dwiema liczbami
- >, <
- != równość

- liczby Powiązane operacje: +, -, *, /, %, //

Chciałem również by funkcje wieloargumentowe kompilowane były jako funkcje jednoargumentowe zwracające kolejne funkcje, co pozwala na wywoływanie funkcji bez wszystkich argumentów w celu zwrócenia funkcji przyjmującej resztę argumentów tzw. currying.

```
1 def class_greeting(first_name, last_name) do
2   "The name's " ++ last_name ++ ', ' ++ first_name ++ ' ' ++
   last_name
3 end
4
5 def compose1(f, g, a) do
6   f(g(a))
7 end
8
9 def compose2(f, g, a, b) do
10  f(g(a, b))
11 end
12
13 yell_greetings = compose2(to_upper, class_greeting)
14 yell_greetings('James', 'Bond') # "THE NAME'S BOND, JAMES BOND"
15
16 compose1(compose1(abs, add(1)), multiply(2))(-4) # 7
```

examples/currying.uwu


```

1 const classy_greeting = (first_name) => (last_name) => {
2   return "The name's " + last_name + ", " + first_name + " " +
      last_name;
3 };
4 const compose1 = (f) => (g) => (a) => {
5   return f(g(a));
6 };
7 const compose2 = (f) => (g) => (a) => (b) => {
8   return f(g(a)(b));
9 };
10 const yell_greetings = compose2(to_upper)(classy_greeting);
11 yell_greetings("James")("Bond");
12 compose1(compose1(abs)(add(1.0)))(multiply(2.0))(-4.0);

```

examples/currying.uwu.js

Jednym z ważniejszych elementów każdego języka jest możliwość wykonywania różnego zbioru instrukcji, warunkowo. W tym celu planowałem zaimplementowanie instrukcji ‘if’, oraz ‘case’. Instrukcja ‘case’ wykonywać ma dopasowanie do wzorca (tzw. pattern-matching), wykonywać, odpowiedni zbiór instrukcji zależnie od wprowadzonych danych. Kompilator, powinien ostrzegać, jeżeli ścieżka dla jednego z typów danych nie została zaimplementowana.

- Przykład - funkcja łącząca dwie posortowane tablice

```

1 def merge<A>(a, b) do
2   merge2: Callable<Array<A>, Callable<Array<A>, Array<A>>> =
      'merge'
3
4   case Tuple(get_head(a), get_head(b)) of
5     Tuple(Empty(), _) do b end
6     Tuple(_, Empty()) do a end
7     Tuple(Head(head_a, rest_a), Head(head_b, rest_b)) do
8       if head_a < head_b then
9         [head_a] | merge2(rest_a, b)
10      else
11        [head_b] | merge2(a, rest_b)
12      end
13    end
14  end
15 end

```

examples/merge.uwu

```

1 const merge = (a) => (b) => {
2   const merge2 = merge;
3   return (() => {
4     const $ = { TAG: "Tuple", _0: get_head(a), _1: get_head(b)
5   };
6     if (typeof $ !== "string" && $.TAG === "Tuple") {
7       if ($. _0 === "Empty") {
8         const _ = $. _1;
9         return b;
10      }
11      if ($. _1 === "Empty") {
12        const _ = $. _0;
13        return a;
14      }
15      if (typeof $. _0 !== "string" && $. _0.TAG === "Head") {
16        if (typeof $. _1 !== "string" && $. _1.TAG === "Head") {
17          const head_b = $. _1. _0;
18          const rest_b = $. _1. _1;
19          const head_a = $. _0. _0;
20          const rest_a = $. _0. _1;
21          return (() => {
22            if (head_a < head_b) {
23              return [head_a].concat(merge2(rest_a)(b));
24            }
25            return [head_b].concat(merge2(a)(rest_b));
26          })();
27        }
28        throw new Error("Non-exhaustive pattern match");
29      }
30      throw new Error("Non-exhaustive pattern match");
31    }
32  })();
33 };

```

examples/merge.uwu.js

Kolejnym dość ważnym elementem języka jest brak wyrażenia ‘return’, które jest wykorzystywane do zwrócenia wartości z funkcji. Zamiast tego każdy blok instrukcji powinien zwracać ostatnie wyrażenie. Pozwoli to na łatwiejsze inicjowanie zmiennych, w przypadku gdy inicjalizacja wymaga więcej niż jednej linii kodu.

Przykład

```
1 message = if is_morning then
2   'Good morning!'
3 else
4   'Hello!'
5 end
6
7
8 result = do
9   arr1 = [1, 2, 3]
10  arr2 = map(arr1, add_one)
11
12  filter(arr2, def is_even(x) do
13    x % 2 == 0
14  end)
15 end
```

examples/return.uwu

```
1 const message = (() => {
2   if (is_morning) {
3     return "Good morning!";
4   }
5   return "Hello!";
6 })();
7 const result = (() => {
8   const arr1 = [1.0, 2.0, 3.0];
9   const arr2 = map(arr1)(add_one);
10  const is_even = (x) => {
11    return x % 2.0 == 0.0;
12  };
13  return filter(arr2)(is_even);
14 })();
```

examples/return.uwu.js

5.1 Opis formalny składni języka

6 Narzędzia

6.1 Język python

Do implementacji programu postanowiliśmy wykorzystać język python w wersji 3.10, ze względu na jego dynamiczność. W tej wersji języka pojawił się również pattern-matching, który znacząco ułatwia pracę z ast.

6.2 Parsowanie i tokenizowanie przy użyciu biblioteki sly

Biblioteka sly, jest pythonową implementacją narzędzi lex i yacc, wykorzystywanych do tworzenia parserów i kompilatorów. Tworzenie leksera i parsera jest bardzo proste.

6.2.1 Lexer

Tokeny są tworzone przy pomocy wyrażeń regularnych.

```
1 import sly
2 class Lex(sly.Lexer):
3     ID = r"\w+"
4     NUM = r"\d+"
```

Biblioteka udostępnia specjalny syntax do tworzenia tokenów, które są już opisane jako inny token.

```
1 ID["and"] = AND
```

Pojedyncze znaki mogą być ignorowane, poprzez ustawienie zmiennej 'ignore'. Dodatkowo ignorowane są też tokeny zaczynające się od 'ignore'.

```
1 ignore = " \t "
2 ignore_comm = r"\#.*"
```

Istnieje również możliwość tworzenia tokenów o typie równym ich wartości, tak długo jak składają się tylko z jednego znaku.

```
1 literals = {"(", ")"} 
```

6.2.2 Parser

W przypadku parsera każda reguła jest implementowana jako metoda, pod której argumentem mamy dostęp do tokenów i wartości zwróconych z innych metod.

```
1 import sly
2 class Parser(sly.Parser):
3     tokens = Lex.tokens
4
5     @_("NUM")
6     def expr(self, p):
7         return p[0] # Lub p.ID
8
9     @_("'(' expr AND expr ')'")
10    def expr(self, p):
11        return p.expr0 and p.expr1
```

Powyższy parser pozwala na opisanie wyrażeń logicznych typu: (1 and 0) and 4.

6.3 Środowisko nodejs do uruchomienia skompilowanego kodu

Javascript jest językiem programowania wykorzystywanym w przeglądarkach internetowych. W uruchomieniu skompilowanego kodu używaliśmy jednak środowiska nodejs. Pozwoli to na szybkie i wygodne testowanie wygenerowanego kodu. Po zainstalowaniu środowiska i skompilowaniu pliku 'index.uwu' otrzymamy plik 'index.uwu.js', który możemy uruchomić w terminalu komendą 'node index.uwu.js'.

7 Implementacja

7.1 lexer

Nasz lexer implementować będzie poniższy zbiór tokenów, warto zauważyć, że identyfikatory zaczynające się z dużej litery są identyfikatorami typów.

```
1 NOT_EQUAL = r"!="
2 EQUAL = r"=="
3 STRING = r"'[^']*'"
4 NUMBER = r"\d+"
5 CONCAT = r"\{2}"
6 INT_DIV = r"/{2}"
7 TYPE_IDENTIFIER = r"[A-Z\d][\w\d]*"
8 IDENTIFIER = r"[a-z_][\w\d]*"
9 IDENTIFIER["def"] = DEF
10 IDENTIFIER["do"] = DO
11 IDENTIFIER["end"] = END
12 IDENTIFIER["if"] = IF
13 IDENTIFIER["else"] = ELSE
14 IDENTIFIER["elif"] = ELIF
15 IDENTIFIER["case"] = CASE
16 IDENTIFIER["enum"] = ENUM
17 IDENTIFIER["then"] = THEN
18
19 IDENTIFIER["of"] = OF
20 EXTERNAL = r"'[^']*'"
```

parser.py

Nasz lexer zaimował się również będzie zwracaniem tokena dla znaku nowej linii

```
1 @_(r"\n([\s\t\n]|\\#.*)*")
2 def NEWLINE(self, t):
3     self.lineno += t.value.count("\n")
4     return t
```

parser.py

Lexer ignorował będzie tokeny komentarza, które zaczynają się od znaku '#'. Oraz znaki tabulacji i spacji.

```

1 ignore_comment = r"\#.*"
2 ignore = " \t"

```

parser.py

Pozatym Nasz lexer zwracał będzie poniższy zbiór tokenów

```

1 literals = {
2     "=",
3     ",",
4     "[",
5     "]",
6     ";",
7     "{",
8     "}",
9     "(",
10    ")",
11    ":",
12    "+",
13    "-",
14    ">",
15    "<",
16    "*",
17    "/",
18    "|",
19    "%",
20 }

```

parser.py

7.2 parser

Implementacja parsera jest bardzo prosta, każda z metod zajmuje jedynie liniijkę, gdzie zwracane jest odpowiednie wyrażenie drzewa ast.

```

1 def binary_expr(self, p):
2     return terms.EBinaryExpr(p[1], p[0], p[2])
3
4 @_(
5     "DO [ ':' type ] [ NEWLINE ] [ do_exprs ] END",
6 )
7 def do(self, p):
8     return terms.EDo(p.do_exprs or [], hint=terms.EHint.
9         from_option(p.type))

```

parser.py

W przypadku zbiorów wyrażeń, zwracane są listy.

```

1  @_("exprs ',' [ NEWLINE ] expr [ NEWLINE ]")
2  def exprs(self, p):
3      return p.exprs + [p.expr]
4
5  @_("expr [ NEWLINE ]")
6  def exprs(self, p):
7      return [p.expr]

```

parser.py

7.3 Drzewo decyzyjne

Wykorzystywany algorytm jest odrobinię inny od algorytmów opisanych w literaturze. Bierzemy pod uwagę następujące obserwacje. Część algorytmów w literaturze w celu uniknięcia wykładniczego wzrostu generowanego kodu, generuje zbędne testy [Augustsson(1985)]. Wykładniczy wzrost generowanego jednak praktycznie nie występuje w praktyce [Scott and Ramsey(2000)]. Najlepszą praktyką więc wydało nam się by nigdy nie generować niepotrzebnych sprawdzeń, i spróbować uniknąć duplikacji kodu używając heurystyki tak jak [Maranget(2008)]. Literatura udowadnia, że w przypadku praktycznie napisanego kodu, różne algorytmy generują prawie identyczny kod [Scott and Ramsey(2000), Maranget(2008)]. Naszą metodą więc jest, aby: (1) zawsze skupiać się na przypasowaniu pierwszego przypadku, by uniknąć niepotrzebnych testów i (2) zachłanie spróbować zminimalizować duplikację przy użyciu heurystyki.

Pattern matching zamieniany jest w drzewo decyzyjne przy pomocy rekurencyjnej funkcji. Rekursja ma dwa podstawowe przypadki:

- lista przypadków jest pusta, więc generujemy pustą gałąź końcową
- pierwszy przypadek, nie ma już więcej wzorów, więc zwracamy gałąź końcową z blokiem.

```

1  def gen_match2(clauses1: typing.Sequence[Clause]) -> CaseTree:
2
3      clauses = [subst_var_eqs(clause) for clause in clauses1]
4
5      match clauses:
6          case []:
7              return MissingLeaf()
8          case [(patterns, body), *_] if not patterns:
9              return Leaf(body)

```

case_tree.py

W innym przypadku wybieramy wzorec przy pomocy heurystyki

```

1      case [(patterns, body), * _]:
2          branch_var = branching_heuristic(patterns, clauses)
3          branch_pattern = patterns[branch_var]
4          yes = list[Clause]()
5          no = list[Clause]()
6
7          #
8          vars = [f"{branch_var}_{i}" for i in range(len(
          branch_pattern.patterns))]

```

case_tree.py

Następnie tworzymy dwa podproblemy yes i no iterując przez wszystkie przypadki. Dla każdego z nich robimy jedną z trzech rzeczy:

- Przypadek nie zawiera wzorca, więc dodajemy go do yes i no

```

1      for patterns, body in clauses:
2          clause = Clause(dict[str, terms.Pattern](
3              patterns), body)
4          match patterns.get(branch_var, None):
5              case None:
6                  yes.append(clause)
7                  no.append(clause)

```

case_tree.py

- Przypadek zawiera szukany wzorec, więc dodajemy go do yes

```

1      case terms.EMatchVariant(id, patterns) if
2      id == branch_pattern.id:
3          yes.append(
4              dataclasses.replace(
5                  clause,
6                  patterns={
7                      key: value
8                      for key, value in clause.
9                      patterns.items()
10                     if key != branch_var
11                     }
12                     | dict(zip(vars, patterns)),
13              )
14          )

```

case_tree.py

- Przypadek zawiera wzorec, więc dodajemy go do no

```

1      case terms.EMatchVariant():
2          no.append(clause)

```

case_tree.py

Rekursywnie generujemy kod dla yes i no i zwracamy gałąź decyzyjną


```

1         return Node(
2             branch_var, branch_pattern.id, vars, gen_match2(yes
3             ), gen_match2(no)
4         )

```

case_tree.py

7.4 inferencja typów

7.4.1 Context

Implementację inferencji typów zaczynamy przez stworzenie zmiennego typu, który będzie reprezentowany przez unikalny identyfikator.

```

1 counter = 0
2
3
4 def fresh_ty_var(kind=typed.KStar()) -> typed.TVar:
5     global counter
6     counter += 1
7     return typed.TVar(counter, kind)

```

algorithm_j.py

Substytucje to pary zmiennych i typów

```

1 Substitution: typing.TypeAlias = dict[int, typed.Type]

```

algorithm_j.py

Jednym z argumentów algorytmu J jest środowisko, w naszej implementacji reprezentowane jest one przez Context, będący mapą identyfikatorów i schem

```

1 Context: typing.TypeAlias = dict[str, Scheme]

```

algorithm_j.py

Gdzie Schema to zawierają informację o typie, oraz listę występujących w nim zmiennych typów

```

1 @dataclasses.dataclass
2 class Scheme:
3     vars: list[int]
4     ty: typed.Type

```

algorithm_j.py

7.4.2 Aplikowanie substytucji

Tworzymy funkcję, aplikującą substytucje na typie.

Przykład: dla $\text{subt}=\{a:\text{Num}\}$ i $\text{ty}=a \rightarrow b$ funkcja zwraca $\text{Num} \rightarrow b$

```

1 def apply_subst(subst: Substitution, ty: typed.Type) -> typed.Type:
2     match ty:
3         case typed.TVar(var):
4             return subst.get(var, ty)
5         case typed.TAp(arg, ret):
6             return typed.TAp(apply_subst(subst, arg), apply_subst(
7 subst, ret))
8         case typed.TCon():
9             return ty
10        case _:
11            raise TypeError(f"Cannot apply substitution to {ty=}")

```

algorithm_j.py

Tworzymy funkcję, unifikującą dwa typy

```

1 def unify(a: typed.Type, b: typed.Type) -> Substitution:
2     match (a, b):
3         case (typed.TCon(), typed.TCon()) if a == b:
4             return {}
5         case (
6             typed.TAp(arg0, ret0),
7             typed.TAp(arg1, ret1),
8         ):
9             subst = unify_subst(arg0, arg1, {})
10            subst = unify_subst(ret0, ret1, subst)
11
12            return subst
13        case (typed.TVar(u), t) | (t, typed.TVar(u)) if typed.kind(
14 a) != typed.kind(b):
15            raise UnifyException(f"Kind for {a=} and {b=} does not
16 match")
17        case (typed.TVar(u), t) | (t, typed.TVar(u)):
18            return var_bind(u, t)
19        case _:
20            raise UnifyException(f"Cannot unify {a=} and {b=}")

```

algorithm_j.py

wraz z funkcją generującą substytucje w przypadku unifikacji zmiennego typu

```

1 def var_bind(u: int, t: typed.Type) -> Substitution:
2     match t:
3         case typed.TVar(tvar2) if u == tvar2:
4             return {}
5         case typed.TVar(_):
6             return {u: t}
7         case t if u in free_type_vars(t):
8             raise TypeError(f"circular use: {u} occurs in {t}")
9         case t:
10            return {u: t}

```

algorithm_j.py

Teraz zaczynamy implementację algorytmu J.

7.4.3 Literały

W przypadku literałów typ, jest oczywisty.

```
1 def infer(  
2     subst: Substitution, ctx: Context, exp: terms.AstTree  
3 ) -> tuple[Substitution, typed.Type]:  
4     match exp:  
5         case terms.ELiteral(value=str()):  
6             return subst, typed.TStr()  
7         case terms.ELiteral(value=float()):  
8             return subst, typed.TNum()
```

algorithm_j.py

7.4.4 Identyfikatory

W przypadku identyfikatorów, musimy zastąpić występujące w nich zmienne typy nowymi zmiennymi typami.

```
1         case terms.EIdentifier(var):  
2             return subst, instantiate(ctx[var])
```

algorithm_j.py

7.4.5 Bloki wyrażeń

W przypadku bloku wyrażeń inferujemy typ każdego z nich i zwracamy ostatni typ.

```
1         case terms.EProgram(body) | terms.EBlock(body):  
2             ty = typed.TUnit()  
3             ctx = ctx.copy()  
4  
5             for exp in body:  
6                 subst, ty = infer(subst, ctx, exp)  
7  
8             return subst, ty
```

algorithm_j.py

7.4.6 Operatory binarne

Implementacja operatorów binarnych są bardzo podobne.

Na przykład operator `|` służący do kontkatenacji dwóch list, oczekuje by wyrażenie z lewej i prawej strony były listami tego samego typu i zwraca nową listę tego właśnie typu.

```

1      case terms.EBinaryExpr("|", left, right):
2          ty = typed.TArray(fresh_ty_var())
3          subst, ty_left = infer(subst, ctx, left)
4          subst = unify_subst(ty_left, ty, subst)
5          subst, ty_right = infer(subst, ctx, right)
6          subst = unify_subst(ty_right, ty, subst)
7          return subst, ty

```

algorithm_j.py

7.4.7 Wywołania funkcji

W przypadku wywołania funkcji musimy zaimplementować currying.

Przykład: $J(fn) = Num \rightarrow Str \rightarrow Unit \rightarrow Arr < Str >$

$J(args) = [Num, Str]$

- typy argumentów zbieramy w odwrotnej kolejności $ty_args=[Str,Num]$
- Tworzymy nowy zmienny typ $ty = \alpha$
- Następnie przy użyciu funkcji `reduce_args` aplikujemy ty_args na typie α co tworzy typ $tmp = Num \rightarrow Str \rightarrow \alpha$
- Unifikujemy typy tmp i $J(fn)$ co tworzy substytucję $\{\alpha : Unit \rightarrow Arr < Str >\}$
- I zwracamy α

```

1      case terms.ECall(fn, args):
2          subst, ty_fn = infer(subst, ctx, fn)
3
4          ty_args = list[typed.Type]()
5
6          for arg in reversed(args):
7              subst, ty_arg = infer(subst, ctx, arg)
8              ty_args.append(ty_arg)
9
10         ty = fresh_ty_var()
11         subst = unify_subst(ty_fn, reduce_args(ty_args, ty),
12         subst)
13
14         return subst, ty

```

algorithm_j.py

7.4.8 Wyrażenia warunkowe

W przypadku wyrażeń warunkowych musimy, sprawdzić, czy warunek jest typu Bool i czy, typy zwracane z każdej gałęzi są takie same, po czym zwracamy tą typ.

```

1      case terms.Elif(test, then, or_else, hint=hint):
2
3          subst, hint = infer(subst, ctx, hint)
4          subst, ty_condition = infer(subst, ctx, test)
5          subst = unify_subst(ty_condition, typed.TBool(), subst)
6
7          subst, ty_then = infer(subst, ctx, then)
8          subst = unify_subst(ty_then, hint, subst)
9
10         subst, ty_or_else = infer(subst, ctx, or_else)
11         subst = unify_subst(ty_or_else, hint, subst)
12
13     return subst, hint

```

algorithm_j.py

7.4.9 Definicje funkcji

W przypadku definicji funkcji musimy upenić się, że typy generyczne, są przypisywane od contextu. Musimy sprawdzić, czy typ zwracany z funkcji jest przypisywalny do zadeklarowanego typu, oraz zapisać funkcję do contextu.

```

1      case terms.EDef(id, params, body, hint, generics):
2          t_ctx = ctx.copy()
3
4          for generic in generics:
5              ty_generic = fresh_ty_var()
6              t_ctx[generic.name] = Scheme([], ty_generic)
7
8          subst, hint = infer(subst, t_ctx, hint)
9
10         ty_params = list[typed.Type]()
11
12         for param in reversed(params):
13             subst, ty_param = infer(subst, t_ctx, param)
14             ty_params.append(ty_param)
15
16         ty = reduce_args(ty_params, hint)
17
18         #
19         subst, ty_body = infer(subst, t_ctx, body)
20
21         subst = unify_subst(ty_body, hint, subst)
22
23         ctx[id] = Scheme.from_subst(subst, ctx, ty)
24
25     return subst, ty

```

algorithm_j.py

7.5 Pattern mathing

W trakcie pattern matching, musimy upewnić się, że konstruktory, posiadają odpowiednią ilość argumentów, oraz, przeprowadzą będzie exhaustivness check-

ing. W tym celu funkcja przyjmując argument `o_alts` będący bazą, zmiennych i ich konstruktorów.

W przypadku pustej gałęzi końcowej, sprawdzamy, czy każda z list jest pusta i wyrzucamy błąd, jeżeli jest.

```

1 def infer_case_tree(
2     subst: Substitution,
3     ctx: Context,
4     tree: case_tree.CaseTree,
5     o_alts: dict[str, list[str]] | None = None,
6 ) -> tuple[Substitution, typed.Type]:
7     alts = o_alts or {}
8
9     match tree:
10         case case_tree.MissingLeaf():
11             if any(alts.values()):
12                 raise NonExhaustiveMatchException()
13             return subst, fresh_ty_var()

```

algorithm_j.py

W przypadku gałęzi decyzyjnej inferujemy typ zmiennej i sprawdzanego konstruktora

```

1     case case_tree.Node(var, pattern_name, vars, yes, no):
2
3         subst, ty_var = infer(subst, ctx, terms.EIdentifier(var
4     )) # x | x.0
5         subst, ty_pattern_name = infer(
6             subst, ctx, terms.EIdentifier(pattern_name)
7         ) # Some() | None()
8         t_ctx = ctx.copy()

```

algorithm_j.py

Następne dwa kroki dopełniają nasz exhaustivness checking. Wpierw uzupełniamy naszą bazę konstruktorów przy użyciu funkcji `alternatives` zwracającej listę wszystkich konstruktorów dla danego typu. Następnie usuwamy z bazy konstruktorów konstruktor do którego odnosi się nasza gałąź decyzyjna.

```

1         if var not in alts:
2             alts[var] = alternatives(ty_pattern_name)
3
4         if pattern_name in alts[var]:
5             alts[var].remove(pattern_name)

```

algorithm_j.py

Kolejna część inferuje typy zmiennych dla każdego z argumentów konstruktora. Oraz, tworzy dla nich zmienne na kontekście.

```

1      ty_vars = list [typed.Type](fresh_ty_var() for _ in vars
2    )
3
4      pattern_name_con = typed.TCon(
5          pattern_name,
6          funtools.reduce(
7              flip(typed.KFun),
8              map(typed.kind, ty_vars),
9              typed.KStar(),
10         ),
11     )
12
13     subst = unify_subst(
14         ty_pattern_name,
15         typed.TDef(
16             funtools.reduce(
17                 typed.TAp,
18                 ty_vars,
19                 pattern_name_con,
20             ),
21             ty_var,
22         ),
23         subst,
24     )
25
26     for var2, ty_var in zip(vars, ty_vars):
27         t_ctx[var2] = Scheme.from_subst(subst, t_ctx,
28             ty_var)

```

algorithm_j.py

Reszta funkcji to sprawdzenie, czy typy zwracane z bloków są takie same.

```

1      ty = fresh_ty_var()
2
3      subst, ty_yes = infer_case_tree(
4          subst,
5          t_ctx,
6          yes,
7          {key: value for key, value in alts.items() if key
8             != var}
9      | {var2: alternatives(t_ctx[var2].ty) for var2 in
10         vars},
11  )
12      subst = unify_subst(ty_yes, ty, subst)
13
14      subst, ty_no = infer_case_tree(subst, ctx, no, alts)
15      subst = unify_subst(ty_no, ty, subst)
16
17      return subst, ty

```

algorithm_j.py

7.6 Hoisting

Ze względu, na to, że deklaracje zmiennych w javascript `const x = 1` w przeciwieństwie do naszego języka nie zwracają przypisywanej do identyfikatora wartości, wyrażenia typu `f(a = b = c)` muszą być przekonwertowane do

```
1  b=c
2  a=b
3  f(a)
```

W tym celu zaimplementowaliśmy funkcję, przyjmuje wyrażenie i zwraca, listę wyrażeń do zhoistowania, oraz zmodyfikowane wyrażenie.

W przypadku bloków wyrażeń, listy z wyrażeniami do zhoistowane są łączone z listą wyrażeń danego bloku.

```
1 def hoist (
2     node: terms.EProgram | terms.EBlock ,
3 ) -> terms.EProgram | terms.EBlock:
4     match node:
5         case terms.EBlock(body):
6             body2 = hoist_expr_list(body)
7             return terms.EBlock(filter_identifiers(body2[:-1]) +
8                                 body2[-1::])
9
10        case terms.EProgram(body):
11            body2 = hoist_expr_list(body)
12            return terms.EProgram(filter_identifiers(body2))
```

compile.py

```
1 def hoist_expr_list(body: list[terms.Expr]) -> list[terms.Expr]:
2     body2 = list[terms.Expr]()
3
4     for expr in body:
5         let , expr2 = hoist_expr(expr)
6         body2.extend(let)
7         body2.append(expr2)
8
9     return body2
```

compile.py

W przypadku deklaracji zmiennych i funkcji, dodajemy elementy do listy wyrażeń do zhoistowania.


```

1 def hoist_expr(node: terms.Expr) -> tuple[list[terms.Expr], terms.
  Expr]:
2     match node:
3         case terms.ELet(id, init):
4             let, init2 = hoist_expr(init)
5
6             return let + [dataclasses.replace(node, init=init2)],
7             terms.EIdentifier(id)
8         case terms.EDef(id, params, body, hint):
9
10            return [
11                terms.EDef(id, params, body=hoist_do(body), hint=
  hint)
12            ], terms.EIdentifier(id)

```

compile.py

W każdym innym przypadku lista wyrażeń do zhoistowania jest po prostu przepuszczana dalej.

```

1         case terms.EIf(test, then, terms.EIf() | terms.EIfNone() as
  or_else):
2             let_test, test2 = hoist_expr(test)
3             let_or_else, or_else2 = hoist_expr(or_else)
4             return let_test + let_or_else, dataclasses.replace(
5                 node,
6                 test=test2,
7                 then=hoist(then),
8                 or_else=or_else2,
9             )

```

compile.py

7.7 Kompilacja

Kompilacja opiera się na jednej funkcji zamieniającej, ast na kod javascript w postaci stringa.

7.8 Literały

Jak widać, dla literałów jest to dość trywialne.

```

1 def _compile(exp: terms.AstTree) -> str:
2     match exp:
3         case terms.EExternal(value=value):
4             return f"{value}"
5         case terms.ELiteral(value=str()):
6             return f'"{exp.value}"'
7         case terms.ELiteral(value=float()):
8             return f"{exp.value}"

```

compile.py

7.8.1 Wyrażenia warunkowe

Wyrażenia warunkowe, wymagają kompilacji, do postaci funkcji, gdyż w odróżnieniu do javascript w naszym języku, zwracają one wartość.

```
1     case terms.If(test, then, or_else):
2         return f"(()=>{{{if({_compile(test)}}){{{_compile(then)
    }}}{_compile(or_else)}}})"
```

compile.py

7.8.2 Warianty

Warianty, przyjmują postać obiektów z dyskryminatorem

```
1     case terms.EVariantCall(id, args) if args:
2         args = [f"_{i}:{_compile(arg)}" for i, arg in enumerate
    (args)]
3
4         return f"{{{TAG: ' {id} ', {', '.join(args)}}}"
```

compile.py

7.8.3 Pattern matching

W przypadku pustej gałęzi z blokiem kompilujemy blok. W przypadku pustej gałęzi końcowej generujemy wyjątek.

```
1 def _compile_case_tree(tree: case_tree.CaseTree):
2     match tree:
3         case case_tree.Leaf(body):
4             return _compile(terms.EBlock(body.body))
5         case case_tree.MissingLeaf():
6             return "throw new Error('Non-exhaustive pattern match')
    "
```

compile.py

W przypadku gałęzi decyzyjnej generowany jest blok if sprawdzający zawartość zmiennej. Można też uniknąć generowania bloku else, w przypadku, gdy wygenerowany byłby wyjątek.

```

1      case case_tree.Node(var, pattern_name, vars, yes, no):
2          conditions = []
3
4          if pattern_name == "True":
5              conditions.append(f"{var}")
6          if pattern_name == "False":
7              conditions.append(f"!{var}")
8          elif vars:
9              conditions.insert(0, f"{var}.TAG=='{pattern_name}'")
10         ")
11         conditions.insert(0, f"typeof {var} !== 'string'")
12     else:
13         conditions.insert(0, f"{var}=== '{pattern_name}'")
14
15     # if isinstance(no, case_tree.MissingLeaf):
16     #     return _compile_case_tree(yes)
17
18     return f"if(({ '&&'.join(conditions) }){{{
19         _compile_case_tree(yes)}}}{ _compile_case_tree(no)}}"
```

compile.py

8 Opis działania

Po uruchomieniu komendą `python3 main.py **/*.uwu` programu, program skompiluje wszystkie pliki kończące się rozszerzeniem `uwu` i wygeneruje dla każdego z nich korespondujący plik z rozszerzeniem `js`.

8.1 Co działa

W programie udało się zaimplementować dużą część funkcjonalności potrzebnych do pisania faktycznych programów. Zabrakło jednak kilku początkowo planowanych funkcji: między innymi: rekursji, krotek, rekordów.

8.2 Uwagi do działania

8.2.1 Obsługa błędów

Błędy zwracane przez program nie są optymalne, nie zaimplementowaliśmy żadnej obsługi błędów w parserze. Błędy wynikające z inferencji typów, zwracają jedynie informacje o błędnych typach, nie o miejscu wystąpienia błędu. Wynika to z braku propagacji numerów linii i kolumn z parsera do drzewa ast.

8.2.2 Pattern matching

Pattern matching mógłby wyświetlać warning w momencie, gdy jedna ze ścieżek jest nieosiągalna.

9 [Podsumowanie]

10 [spisy – rysunków, tabel, listingów ipt.]

References

- [Aho et al.(1985)Aho, Sethi, and Ullman] Alfred V Aho, Ravi Sethi, and Jeffrey D Ullman. *Compilers*. Addison Wesley, Boston, MA, January 1985. ISBN 9780201101942.
- [Augustsson(1985)] Lennart Augustsson. Compiling pattern matching. In *Proc. of a conference on Functional programming languages and computer architecture*, Berlin , Heidelberg, January 1985. Springer-Verlag.
- [Damas and Milner(1982)] Luis Damas and Robin Milner. Principal type-schemes for functional programs. In *Proceedings of the 9th ACM SIGPLAN-SIGACT symposium on Principles of Programming Languages*, pages 207–212, 1982.
- [J.E. Hopcroft(2005)] J.D. Ullman J.E. Hopcroft, R. Motwani. *Wprowadzenie do Teorii automatów, języków i obliczeń*. Wydawnictwo Naukowe PWN, 2005. ISBN 8301145021.
- [Maranget(2008)] Luc Maranget. Compiling pattern matching to good decision trees. In *Proceedings of the 2008 ACM SIGPLAN workshop on ML*, pages 35–46, USA, NY, September 2008. Association for Computing Machinery.
- [Milner(1978)] Robin Milner. A theory of type polymorphism in programming. *Journal of Computer and System Sciences*, 17:348–375, 1978.
- [Pierce(2002)] Benjamin C Pierce. *Types and Programming Languages*. The MIT Press. MIT Press, London, England, January 2002. ISBN 0262162091.
- [Scott and Ramsey(2000)] Kevin Scott and Norman Ramsey. When do match-compilation heuristics matter? Technical report, University of Virginia, USA, 2000.