



Università degli Studi di Camerino

SCUOLA DI SCIENZE E TECNOLOGIE

Corso di Laurea in Informatica (Classe L-31)

Compilazione di un linguaggio funzionale in Java

Laureando
Massimo Pavoni

Matricola 124377

Relatore
Prof. Luca Padovani

A.A. 2023/2024

Indice

1	Introduzione	1
1.1	Motivazione	1
1.2	Obiettivi	1
1.3	Struttura della Tesi	1
2	Funx	3
2.1	Linguaggi funzionali	3
2.1.1	ML, Haskell e Funx	4
2.2	Sintassi	5
2.2.1	Zucchero sintattico	6
3	Inferenza di tipo	9
3.1	Sistemi di tipo	9
3.1.1	λ -cubo	11
3.1.2	Sistema FC	12
3.2	Inferenza secondo Hindley–Milner	14
4	Java	17
4.1	Interfacce funzionali	17
4.2	Operatore ternario	19
4.3	Tipi generici	20
5	Compilatore	21
5.1	ANTLR	22
5.1.1	Analisi lessicale	22
5.1.2	Analisi sintattica	24
5.2	Albero sintattico astratto	26
5.2.1	Gerarchia delle classi	26
5.2.2	AST builder	28
5.3	Motore di inferenza	31
5.3.1	Sistema HM	31
5.3.2	Inferenza su espressioni	31
5.4	Traduzione in Java	31
6	Esempi di traduzione	33

7 Conclusioni	35
Bibliografia	37

Elenco dei codici

2.1	Esempio di programma	8
4.1	Semplice funzione in Funx	17
4.2	Corrispondente metodo in Java	17
4.3	Interfaccia funzionale per espressioni <code>let</code>	18
4.4	Espressione <code>let</code> in Funx	18
4.5	Corrispondente classe anonima in Java	18
4.6	If e operatori booleani in Funx	19
4.7	Corrispondenti operatori ternari in Java	19
4.8	Scrittura e utilizzo di funzioni polimorfe in Funx	20
4.9	Corrispondente traduzione in Java	20
5.1	Alcune <i>token</i> del <i>lexer</i>	23
5.2	Grammatica per il <i>parser</i>	25
5.3	Esempio di classe della gerarchia	27
5.4	Parte del codice di <code>PreludeFunction</code>	28
5.5	Metodi per astrazioni annidate e operatori simbolici binari	29
5.6	Alcuni metodi <i>visit</i> di <code>ASTBuilder</code>	29
5.7	Programma in Funx	30

Elenco delle figure

2.1	Grammatica del lambda calcolo	5
2.2	Grammatica di Funx	6
3.1	Alcuni linguaggi e loro sistemi di tipo	10
3.2	λ -cubo	11
3.3	Grammatica del sistema di tipo di Funx	13
3.4	Definizioni di contesto e variabili libere	14
3.5	Regole di inferenza del <i>sistema HM</i> in Funx	15
3.6	Algoritmo \mathcal{W}	16
3.7	Funzione \mathcal{U}	16
5.1	Possibili comandi e opzioni della <i>CLI</i>	21
5.2	Diagramma semplificato delle classi dell' AST	27
5.3	Oggetto AST in <i>debug</i> e visualizzazione dell'albero	30
5.4	Esempio di errore in Java dovuto a tipi generici	31

Elenco delle tabelle

2.1	Zucchero sintattico	7
3.1	Esempi di funzioni polimorfe	13

1. Introduzione

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrum exercitationem ullamco laboriosam, nisi ut aliquid ex ea commodo consequat. Duis aute irure reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint obcaecat cupiditat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Nel Capitolo 1 illustreremo prima le motivazioni che ci hanno spinto a perseguire l'obiettivo descritto e quindi la struttura della tesi.

1.1 Motivazione

1.2 Obiettivi

1.3 Struttura della Tesi

2. Funx

Questo capitolo descrive brevemente i linguaggi funzionali e le scelte effettuate durante l'ideazione del linguaggio usato per il progetto: **Funx**.

Il nome nasce dall'unione dei due termini anglosassoni *functional* e *expression*; viene quindi pronunciato [ˈfʌnɪk's] in inglese, [fan·èks] o [fan·iks] in italiano.

2.1 Linguaggi funzionali

Nonostante molti linguaggi non si possano confinare all'interno di un solo paradigma, parlando di linguaggi di programmazione si fa spesso riferimento a due grandi categorie: linguaggi imperativi e linguaggi dichiarativi.

I primi hanno caratteristiche direttamente legate al modello di calcolo di *John Von Neumann*, a sua volta non dissimile dalla macchina di *Alan Turing*. Questi linguaggi sono usati per scrivere codice che segue una precisa sequenza di istruzioni, la quale descrive più o meno esplicitamente i passi necessari per risolvere il problema affrontato. Appartengono alla famiglia dei linguaggi di programmazione imperativi sia linguaggi procedurali come Fortran, Cobol e Zig, sia i linguaggi orientati agli oggetti, tra cui Kotlin, C# e Ruby.

I linguaggi dichiarativi, invece, sono fondamentali per lo scopo del progetto: tali linguaggi sono generalmente di altissimo livello e permettono allo sviluppatore di concentrarsi sull'obiettivo da raggiungere piuttosto che sui dettagli implementativi.

Fanno parte di questa categoria linguaggi di interrogazione come SQL, linguaggi logici come Prolog e soprattutto i linguaggi funzionali: Lisp, Clojure, Elixir, OCaml e Haskell sono alcuni esempi.

Alla base di ogni linguaggio funzionale vi è il **lambda calcolo**¹: un sistema formale definito dal matematico *Alonzo Church* (supervisore di *Alan Turing* durante il dottorato), equivalente alla macchina di Turing, ma fondato sulle funzioni pure.

¹«A Set of Postulates for the Foundation of Logic» [Chu32] and «A Set of Postulates for the Foundation of Logic» (Second paper) [Chu33]

La grammatica del lambda calcolo verrà presentata poco più avanti (sezione 2.2), ma le regole che ne governano il funzionamento e il modo in cui queste vengano utilizzate per ridurre le espressioni ad una forma normale esulano dai fini di questo documento.

Rimane comunque rilevante elencare le principali qualità che un linguaggio funzionale usualmente matura grazie al lambda calcolo:

- **funzioni come entità di prima classe**: le funzioni possono essere passate come argomenti e restituite come risultato di altre funzioni;
- **immutabilità**: le variabili utilizzate sono immutabili;
- **purezza**: le funzioni sono libere da effetti collaterali (non modificano lo stato del programma) e restituiscono sempre lo stesso output per input identici;
- **ricorsione**: la ricorsione è il meccanismo più idiomático per esprimere l'iterazione su una struttura dati.

2.1.1 ML, Haskell e Funx

Nonostante le funzioni pure tipiche di un linguaggio funzionale siano un concetto molto attraente dal punto di vista della correttezza della computazione, i vincoli così imposti possono risultare stringenti a tal punto da rendere difficile, se non impossibile, la scrittura di programmi che interagiscano con il mondo reale.

Per questo motivo, molti linguaggi funzionali permettono invece di utilizzare particolari funzioni impure o di effettuare almeno operazioni di input/output. Inoltre, molti linguaggi prevalentemente imperativi adottano ormai da tempo alcune caratteristiche tipiche dei linguaggi funzionali (e.g. Rust, il linguaggio più amato² dagli sviluppatori secondo i sondaggi di *Stack Overflow*, eredita molto dal linguaggio con cui era scritto il suo primo compilatore, OCaml, ed è dotato quindi di funzioni di prima classe, immutabilità di default, strutture dati algebriche, ecc.).

ML è un linguaggio funzionale sviluppato negli anni '70 presso l'Università di Edimburgo, costituente la base per moltissimi dei linguaggi sviluppati in seguito. ML permette effettivamente l'uso di funzioni impure, ma fra i suoi discendenti vi è Haskell, uno dei pochi linguaggi invece completamente puri.

Haskell si avvale di un pattern di programmazione chiamato *monadi*³ per gestire le operazioni di input/output e altre operazioni impure, mantenendo le funzioni pure.

Nell'ideare **Funx** l'ispirazione viene proprio da Haskell, ma è presente la possibilità di dichiarare un'unica funzione impura (il cosiddetto *main*) per permettere di visualizzare a schermo un risultato. Il linguaggio non è quindi allo stesso livello di purezza di Haskell, e naturalmente non supporta molte delle funzionalità più avanzate di quest'ultimo (come le *classi di tipi* e il *pattern matching*), ma ne mutua altre comunque interessanti, tra cui l'uso di alcuni operatori infissi e il *polimorfismo parametrico*.

²Stack Overflow Developer Survey 2023 (<https://survey.stackoverflow.co/2023>), *Rust is the most admired language*

³«Notions of Computation and Monads» [Mog91]

2.2 Sintassi

La sintassi di **Funx** risulta molto simile a quella di `Haskell`, con poche differenze dovute a tre principali motivi:

- libera scelta di nomi e simboli per le parole chiave;
- necessità di successiva traduzione in `Java`;
- difficoltà e scarso valore all'interno del progetto dell'implementazione di un parser dipendente dall'indentazione.

A prescindere da ciò, il cuore del linguaggio è lo stesso di ogni altro linguaggio derivato dal lambda calcolo: la sua definizione si può agilmente comprendere visualizzando la grammatica del lambda calcolo e confrontandola con quella (leggermente semplificata) di **Funx**, facendo attenzione alle regole aggiuntive.

Espressione	E	$::=$	x	variabile
			$ \quad E_l \ E_r$	applicazione
			$ \quad \lambda x . E$	astrazione

Figura 2.1: Grammatica del lambda calcolo

Le tre regole in Figura 2.1 indicano le tre componenti indispensabili del lambda calcolo:

- **variabile**: simbolo rappresentante un parametro;
- **applicazione**: applicazione di funzione ad un argomento (entrambi espressioni);
- **astrazione**: definizione di una funzione anonima, con un solo input x (variabile vincolata) e un solo output E (espressione); per definire funzioni con più parametri si debbono usare molteplici astrazioni annidate (tecnica detta *currying*).

Modulo	M	$::=$	$nome \cdot L$	
Dichiarazione	D	$::=$	$?(schema\ di\ tipo) \cdot id = E$	funzione
Espressione	E	$::=$	c	costante
			$ \quad x$	variabile
			$ \quad E_l \ E_r$	applicazione
			$ \quad \lambda x \cdot E$	astrazione
			$ \quad L$	let
			$ \quad \mathbf{if} \ E_c \ \mathbf{then} \ E_t \ \mathbf{else} \ E_e$	if
Let	L	$::=$	$\mathbf{let} \cdot D (\cdot D)^* \cdot \mathbf{in} \ E$	

Figura 2.2: Grammatica di Funx

È facile constatare la presenza delle ulteriori produzioni per la definizione del modulo corrente (informazione inclusa a prescindere dal fatto che il linguaggio ad ora non supporti l'importazione di moduli esterni che non siano la libreria standard) e di funzioni con nome: lo *schema di tipo* è un'informazione opzionale relativa al tipo della funzione e di cui si parlerà più approfonditamente nella sezione 3.1.2.

Per quanto riguarda invece le espressioni, vengono introdotte tre nuove regole:

- **costante**: rappresenta un valore letterale, come un numero o una stringa;
- **let**: permette di avere dichiarazioni locali utilizzabili all'interno di un'espressione;
- **if**: la più classica istruzione condizionale controllata da un'espressione booleana.

2.2.1 Zucchero sintattico

Con lo scopo di rendere il codice più leggibile, conciso e semplice, **Funx** introduce dello zucchero sintattico (del tutto simile a quello di `Haskell`). In Tabella 2.1 sono riportati l'indispensabile per evitare il parsing dell'indentazione, le semplificazioni comuni utili all'arricchimento del lambda calcolo, e infine tutti gli operatori simbolici supportati al momento (assieme alla notazione per indicarne associatività e precedenza).

Zucchero	Sostituzione
$\backslash x \rightarrow e$	$\lambda x . e$
$\backslash x y \rightarrow e$	$\lambda x . \lambda y . e$
$f \ x \ y = e$	$f = \lambda x . \lambda y . e$
let f1 = e1 f2 = e2 in e3	let f1 = e1 · f2 = e2 in e3
f3 = e3 with f1 = e1 f2 = e2 out	f3 = let f1 = e1 · f2 = e2 in e3
main = e3 f1 = e1 f2 = e2	main = let f1 = e1 · f2 = e2 in e3
if b then e1 else e2 fi	if b then e1 else e2
e1 . e2 infixr 9 e1 / e2 infixl 7 e1 % e2 infixl 7 e1 * e2 infixl 7 e1 + e2 infixl 6 e1 - e2 infixl 6 e1 > e2 infix 4 e1 >= e2 infix 4 e1 < e2 infix 4 e1 <= e2 infix 4 e1 == e2 infix 4 e1 != e2 infix 4 !!e prefix 4 e1 && e2 infixr 3 e1 e2 infixr 2 e1 \$ e2 infixr 0	compose e1 e2 divide e1 e2 modulo e1 e2 multiply e1 e2 add e1 e2 subtract e1 e2 greaterThan e1 e2 greaterThanEquals e1 e2 lessThan e1 e2 lessThanEquals e1 e2 equalsEquals e1 e2 notEquals e1 e2 not e if e1 then e2 else False if e1 then True else e2 apply e1 e2

Tabella 2.1: Zucchero sintattico

Come già accennato, il Capitolo 5 illustrerà come l'albero sintattico astratto (**AST**) di un programma viene ottenuto, annotato e tradotto in Java; la sezione 4.2 esporrà invece il motivo della traduzione degli operatori booleani binari in `if`.

Alcuni esempi di funzioni sono presentati nel Codice 2.1; seppur superflua, l'indentazione è inclusa per maggiore chiarezza.

```
1 main = factorial 20
2
3 factorial : Int -> Int
4 factorial n = if n == 0 then 1 else n * factorial (n - 1) fi
5
6 even : Int -> Bool
7 even = let
8     even1 : Int -> Bool
9     even1 n = if n == 0 then True else odd (n - 1) fi
10
11     odd : Int -> Bool
12     odd n = if n == 0 then False else even1 (n - 1) fi
13     in even1
14
15 gcd : Int -> Int -> Int
16 gcd a b = if b == 0 then a else gcd b (a % b) fi
17
18 xor : Bool -> Bool -> Bool
19 xor a b = (a || b) && !(a && b)
```

Codice 2.1: Esempio di programma

3. Inferenza di tipo

Dopo aver discusso la sintassi di **Funx**, è importante far notare come i programmi non abbiano bisogno di annotazioni di tipo, nonostante siano stati adottati tipi statici.

In questo capitolo affronteremo l'argomento dei sistemi di tipo e dell'inferenza, meccanismo proprio di molti linguaggi, funzionali e non, che rende possibile la deduzione automatica del tipo di un termine basandosi sull'utilizzo delle variabili e delle funzioni.

3.1 Sistemi di tipo

Durante la genesi di ogni linguaggio di programmazione, una delle scelte più significative riguarda l'introduzione di un sistema per gestire i tipi di variabili ed espressioni.

Tali sistemi di tipo sono di fatto insiemi di regole logiche che permettono di assegnare una proprietà "*type*" a ciascuno dei termini del linguaggio che ne necessitano.

Sono principalmente suddivisi in due categorie:

- **tipizzazione statica:** i tipi sono definiti a tempo di compilazione e non possono cambiare mentre il programma è in esecuzione;
- **tipizzazione dinamica:** i tipi vengono stabiliti durante l'esecuzione e possono cambiare in qualsiasi momento.

Oltre a questa distinzione esistono varie sfumature e approcci differenti, informalmente classificati in base alla rigidità delle regole di tipizzazione. Si parla di *tipizzazione debole* quando ad esempio sono consentite conversioni implicite tra tipi diversi, *tipizzazione forte* se sono impedito, oppure qualora sia o meno disponibile l'aritmetica dei puntatori.

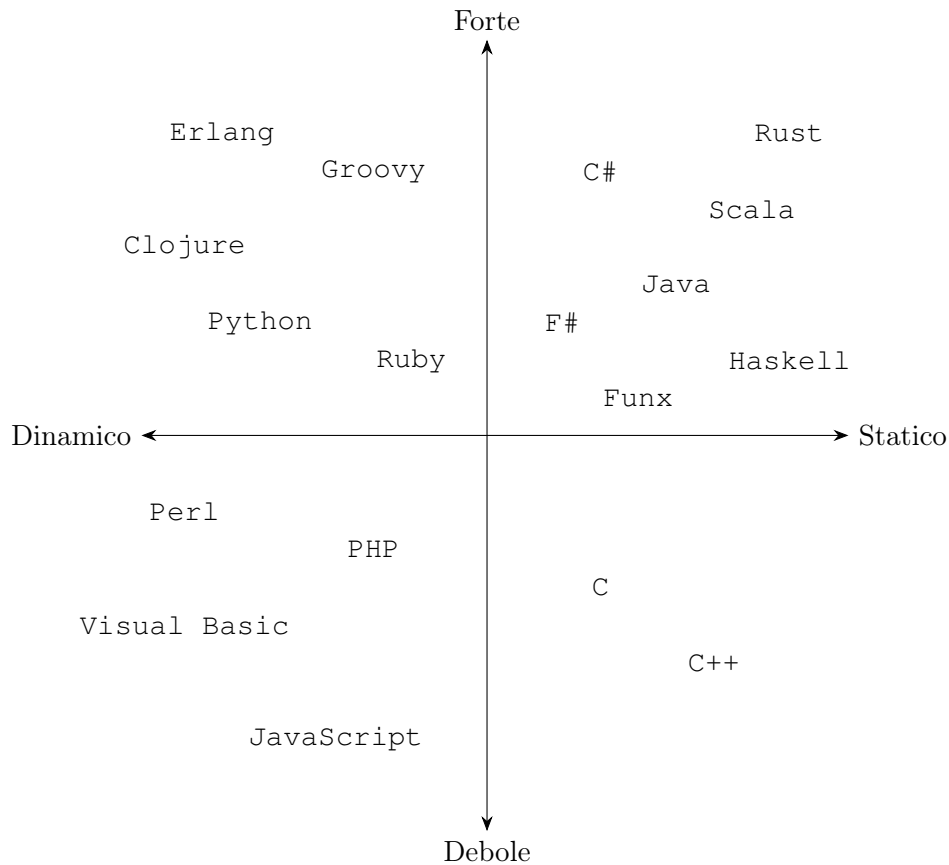


Figura 3.1: Alcuni linguaggi e loro sistemi di tipo

Grazie ai tipi dinamici, linguaggi quali Python e JavaScript permettono veloce prototipazione, flessibilità e codice più conciso, a discapito però di una più alta probabilità di incontrare errori importanti a runtime, piuttosto che in fase di compilazione.

Al contrario, i tipi statici spesso migliorano naturalmente la mantenibilità di un progetto: viene limitata la possibilità di scorciatoie nello sviluppo, ma si hanno maggiori garanzie di correttezza, in quanto il compilatore può implementare ulteriori controlli e segnalare errori semantici più precisi già prima dell'esecuzione del programma.

D'altro canto, l'obbligo di specificare i tipi di ogni variabile, oggetto, funzione e parametro può risultare tedioso e talvolta ridondante; molti linguaggi moderni, tra cui Haskell e Rust, ovviano magistralmente a quest'inconvenienza tramite l'uso dell'inferenza di tipo.

Gli algoritmi di inferenza introducono numerosi benefici, in particolare:

- la scrittura del codice è meno onerosa per lo sviluppatore a prescindere dal sistema di tipi utilizzato, e diviene quindi estremamente vantaggioso utilizzare tipi statici;
- le annotazioni ora opzionali possono essere aggiunte dal programmatore quando vi sono casi difficili da disambiguare automaticamente, oppure per migliorare la leggibilità del codice;
- gli strumenti di sviluppo per il linguaggio possono sfruttare informazioni fornite dal motore di inferenza per suggerire il tipo delle espressioni e arricchire i messaggi di errore e di warning.

3.1.1 λ -cubo

Al fine di comprendere quale sistema il linguaggio **Funx** implementi, prima di discutere l'inferenza si vuol descrivere brevemente il λ -cubo, lambda cubo¹, un modello introdotto per classificare i sistemi di tipo applicabili al lambda calcolo.

In Figura 3.2 è possibile osservare come la struttura del cubo abbia all'origine il *lambda calcolo semplicemente tipato* ($\lambda \rightarrow$) e come le tre dimensioni in cui si sviluppa rappresentino ciascuna un'estensione del sistema:

- **tipi dipendenti** (\rightarrow): la definizione dei tipi può dipendere dai valori delle variabili (implementati da linguaggi funzionali come Agda, Coq e Idris);
- **polimorfismo parametrico** (\uparrow): i tipi possono essere polimorfi, generalizzati tramite variabili di tipo (presenti nei sistemi adottati da ML, OCaml e Haskell);
- **costruttori di tipo** (\nearrow): capacità di costruire nuovi tipi a partire da tipi esistenti (Haskell ne fa grande uso poiché ogni nuovo tipo, dichiarato con la keyword `data`, è un nuovo costruttore di tipo).

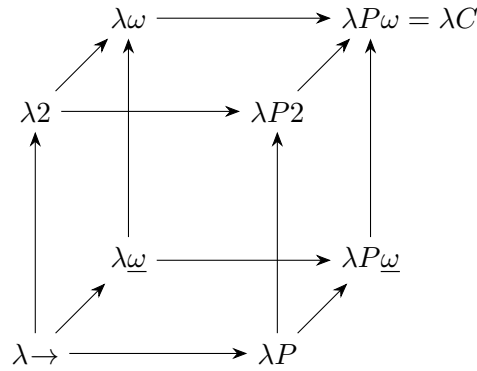


Figura 3.2: λ -cubo

Senza entrare troppo nei dettagli, in ordine crescente di potenza espressiva:

- $\lambda \rightarrow$ (*lambda calcolo semplicemente tipato*): tipi monomorfi;
- $\lambda \underline{\omega}$ (*lambda weak omega*): costruttori di tipo;
- $\lambda 2$ (*lambda due, lambda F, lambda calcolo polimorfico*): polimorfismo parametrico;
- λP (*lambda P*): tipi dipendenti;
- $\lambda P \underline{\omega}$ (*lambda P weak omega*): costruttori di tipo e tipi dipendenti;
- $\lambda \omega$ (*lambda omega*): costruttori di tipo e polimorfismo parametrico;
- $\lambda P 2$ (*lambda P due*): polimorfismo parametrico e tipi dipendenti;
- $\lambda P \omega = \lambda C$ (*lambda P omega, lambda C, calcolo delle costruzioni*): cstronglyombinazione di tutte le tre estensioni.

¹«Introduction to generalized type systems» [Bar91]

3.1.2 Sistema FC

Tra i vari sistemi di tipo per il lambda calcolo, uno dei più interessanti è *lambda F* (vertice $\lambda 2$ in Figura 3.2) poiché molto utile per la generalizzazione delle funzioni: un problema molto ricorrente nella programmazione con qualsiasi linguaggio è infatti la duplicazione di codice per funzioni che svolgono operazioni simili su tipi diversi.

Il *sistema F* risolve tale problema introducendo il **polimorfismo parametrico** e di conseguenza la distinzione tra tipi monomorfi (monotipi) e tipi polimorfi (politipi).

I tipi delle funzioni possono essere caratterizzati tramite quantificatori universali e variabili di tipo ove sia necessario un tipo generico (spesso vengono usate singole lettere dell'alfabeto greco o latino).

Tuttavia, $\lambda 2$ nella sua forma più pura, oltre a non essere un sistema Turing-completo (è possibile definire solamente la ricorsione primitiva), rende l'inferenza di tipo trattata nella sezione 3.2 un problema non decidibile².

Pertanto, il linguaggio `Haskell` non implementa semplicemente il *sistema F*, ma piuttosto una versione ristretta di $\lambda\omega$ chiamata *sistema FC*³.

Quest'ultima include anche i costruttori di tipo (*funzioni di tipo* in **Funx**), frenando però il polimorfismo ai cosiddetti *tipi polimorfici di rango 1* (*polimorfismo predicativo*): tale limitazione si manifesta nella scrittura di tutti i quantificatori universali all'inizio di un tipo polimorfo (che prende il nome di *schema di tipo*).

Le versioni invece più espressive e più vicine a $\lambda\omega$ sono:

- *polimorfismo di rango superiore*: supporta quantificatori universali in qualsiasi punto nelle definizioni delle funzioni (e.g. `Bool -> (forall b . b -> b)`); `Haskell` lo realizza con l'estensione `RankNTypes` del compilatore `GHC`, mentre offre anche l'estensione `Rank2Types`, per la quale l'inferenza rimane decidibile;
- *polimorfismo impredicativo*: permette di quantificare le variabili di tipo in modo arbitrario, anche e soprattutto all'interno dei costruttori di tipo (e.g. `Maybe (forall a . a -> a) -> Bool`, possibile in `Haskell` abilitando l'estensione `ImpredicativeTypes`).

Il linguaggio **Funx** ovviamente non è correntemente in grado di supportare queste estensioni del sistema di tipo, così come non è possibile definire nuovi tipi o fare uso di *classi di tipo* simili a quelle proprie di `Haskell`. Affermare che **Funx** adotti il *sistema FC* potrebbe lasciare intendere un linguaggio più espressivo di quanto non sia in realtà: è dunque più opportuno realizzare il *sistema HM*, di cui il *sistema FC* è un ampliamento, e che comunque ben si presta allo scopo principe di traduzione in `Java`.

In Tabella 3.1 si possono osservare i tipi di alcune funzioni polimorfe di `Haskell`: la sintassi di **Funx** è molto simile (identica in ognuno dei casi presentati), con l'eccezione che la parola chiave `forall` è completamente assente dal linguaggio, in quanto ogni identificatore che inizia con una lettera minuscola è considerato una variabile di tipo da quantificare universalmente (sezione 5.1.1).

²«Typability and type checking in System F are equivalent and undecidable» [Wel99]

³«System FC, as implemented in GHC» [Eis15]

Funzione	Schema
id	$\text{forall } a . a \rightarrow a$
const	$\text{forall } a \ b . a \rightarrow b \rightarrow a$
(.)	$\text{forall } b \ c \ a . (b \rightarrow c) \rightarrow (a \rightarrow b) \rightarrow a \rightarrow c$
flip	$\text{forall } a \ b \ c . (a \rightarrow b \rightarrow c) \rightarrow b \rightarrow a \rightarrow c$
(\$)	$\text{forall } a \ b . (a \rightarrow b) \rightarrow a \rightarrow b$
(&)	$\text{forall } a \ b . a \rightarrow (a \rightarrow b) \rightarrow b$

Tabella 3.1: Esempi di funzioni polimorfe

In Figura 3.3 è mostrata la grammatica per la definizione dei tipi implementati nel linguaggio **Funx**. Si noti come i tipi monomorfi siano solo variabili di tipo o applicazioni di funzioni ad altri tipi; al momento il linguaggio mette a disposizione le funzioni di tipo più elementari, la cui arietà è indicata in pedice.

Schema di tipo	σ	$::=$	τ	monotipo
			$\forall \alpha . \sigma$	politipo
Tipo	τ	$::=$	α	variabile di tipo
			$F \tau \dots \tau$	applicazione di funzione di tipo
Funzione di tipo	F	$::=$	\rightarrow_2	funzione
			$Bool_0$	booleano
			Int_0	intero

Figura 3.3: Grammatica del sistema di tipo di **Funx**

3.2 Inferenza secondo Hindley–Milner

Come già accennato, il sistema *Hindley–Milner* (HM)⁴, o *Damas–Hindley–Milner*⁵, è un sistema di tipo per il lambda calcolo con polimorfismo parametrico largamente utilizzato in molti moderni linguaggi di programmazione ad alto livello. Il maggiore punto di forza del sistema è il relativo metodo di inferenza, in grado di dedurre automaticamente il tipo di un termine senza annotazioni esplicite fornite dagli sviluppatori.

Prima di presentare l'algoritmo di inferenza è necessario complementare le nozioni di lambda calcolo esteso (Figure 2.1 e 2.2) e del sistema di **Funx** (Figura 3.3) con due concetti fondamentali (Figura 3.4):

- **contesto**: una insieme di associazioni tra variabili e schemi di tipo, che rappresenta lo stato corrente dell'ambiente in cui un termine viene tipato; l'unione tra la grammatica del linguaggio e il sistema di tipi è data da un giudizio di tipo effettuato nel contesto su un termine (espressione);
- **variabili libere**: l'insieme delle variabili libere di un tipo è semplicemente il complemento delle variabili vincolate, quantificate universalmente in un politipo.

Contesto	$\Gamma ::= \epsilon$	contesto vuoto
	$\mid \Gamma + x : \sigma$	aggiunta di associazione
Giudizio di tipo	$::= \Gamma \vdash E : \sigma$	
$free(\forall \alpha . \sigma)$	$= free(\sigma) - \{\alpha\}$	tipo polimorfo
$free(\alpha)$	$= \{\alpha\}$	variabile di tipo
$free(F \tau_1 \dots \tau_n)$	$= \bigcup_{i=1}^n free(\tau_i)$	applicazione di funzione di tipo
$free(\Gamma)$	$= \bigcup_{x : \sigma \in \Gamma} free(\sigma)$	contesto
$free(\Gamma \vdash E : \sigma)$	$= free(\sigma) - free(\Gamma)$	giudizio di tipo

Figura 3.4: Definizioni di contesto e variabili libere

Le regole di inferenza⁶ del sistema HM , riportate in Figura 3.5, informano il comportamento dell'algoritmo \mathcal{W} , implementazione dell'inferenza di tipo.

⁴«The Principal Type-Scheme of an Object in Combinatory Logic» [Hin69] and «A Theory of Type Polymorphism in Programming» [Mil78]

⁵«Principal type-schemes for functional programs» [DM82]

⁶«A Simple Applicative Language: Mini-ML» [Clé+86]

In aggiunta a principi già noti, le peculiarità non immediatamente chiare sono:

- *constantType*: funzione che restituisce il tipo di una costante;
- $\sigma \sqsubseteq \tau$: indica intuitivamente che σ è più generale di τ (τ è un'istanza di σ);
- $Clos_\Gamma$: ottiene la *chiusura*, ossia tipo più generale, di una variabile quantificando universalmente le variabili libere del tipo iniziale.

$$\begin{array}{c}
 \frac{\tau = \text{constantType}(c)}{\Gamma \vdash c : \tau} \quad \text{[costante]} \\
 \\
 \frac{x : \sigma \in \Gamma \quad \sigma \sqsubseteq \tau}{\Gamma \vdash x : \tau} \quad \text{[variabile]} \\
 \\
 \frac{\Gamma \vdash E_l : \tau \rightarrow \beta \quad \Gamma \vdash E_r : \tau}{\Gamma \vdash E_l E_r : \beta} \quad \text{[applicazione]} \\
 \\
 \frac{\Gamma + x : \beta \vdash E : \tau}{\Gamma \vdash \lambda x . E : \beta \rightarrow \tau} \quad \text{[astrazione]} \\
 \\
 \frac{\bar{\Gamma} = \Gamma + \vec{x} : \vec{\beta} \quad \bar{\Gamma} \vdash \vec{E} : \vec{\kappa} \quad \bar{\Gamma} \{Clos_\Gamma(\vec{\kappa})/\vec{\beta}\} \vdash E_i : \tau}{\bar{\Gamma} \vdash \text{let } \vec{x} = \vec{E} \text{ in } E_i : \tau} \quad \text{[let]} \\
 \\
 Clos_\Gamma(\tau) = \forall \hat{\alpha} . \tau \quad \hat{\alpha} = free(\tau) - free(\Gamma) \\
 \\
 \frac{\Gamma \vdash E_c : Bool \quad \Gamma \vdash E_t : \tau \quad \Gamma \vdash E_e : \tau}{\Gamma \vdash \text{if } E_c \text{ then } E_t \text{ else } E_e : \tau} \quad \text{[if]}
 \end{array}$$

Figura 3.5: Regole di inferenza del *sistema HM* in **Funx**

L'*algoritmo* \mathcal{W}^7 riportato in Figura 3.6 fornisce un'implementazione del metodo finora descritto, servendosi delle regole di inferenza e dettagliando gli step per ogni espressione disponibile in **Funx**.

L'algoritmo appare nel codice del compilatore in una forma più complessa, soprattutto nel caso dell'espressione `let` (sezione 5.3.2), volendo supportare la ricorsione senza far uso di alcun *fixed-point combinator*.

⁷«Proofs about a Folklore Let-Polymorphic Type Inference Algorithm» [LY98]

$$\begin{aligned}
 & \mathcal{W} : \text{Context} \times \text{Expression} \rightarrow \text{Substitution} \times \text{Type} \\
 & \mathcal{W}(\Gamma, c) = (id, \text{constantType}(c)) \\
 & \mathcal{W}(\Gamma, x) = \text{let } \forall \vec{\alpha} . \tau = \Gamma(x), \text{ new } \vec{\beta} \\
 & \quad \text{in } (id, \{\vec{\beta}/\vec{\alpha}\}\tau) \\
 & \mathcal{W}(\Gamma, E_l \ E_r) = \text{let } (S_l, \tau_l) = \mathcal{W}(\Gamma, E_l) \\
 & \quad (S_r, \tau_r) = \mathcal{W}(S_l \Gamma, E_r) \\
 & \quad S_a = \mathcal{U}(S_r \tau_l, \tau_r \rightarrow \beta), \text{ new } \beta \\
 & \quad \text{in } (S_a S_r S_l, S_a \beta) \\
 & \mathcal{W}(\Gamma, \lambda x . E) = \text{let } (S, \tau) = \mathcal{W}(\Gamma + x : \beta, E), \text{ new } \beta \\
 & \quad \text{in } (S, S\beta \rightarrow \tau) \\
 & \mathcal{W}(\Gamma, \text{let } \vec{x} = \vec{E} \text{ in } E_i) = \text{let } \bar{\Gamma} = \Gamma + \vec{x} : \vec{\beta}, \text{ new } \vec{\beta} \\
 & \quad (\vec{S}, \vec{\kappa}) = \mathcal{W}(\bar{\Gamma}, \vec{E}) \\
 & \quad (S_i, \tau) = \mathcal{W}(\vec{S} \bar{\Gamma} \{Clos_{\Gamma}(\vec{\kappa})/\vec{\beta}\}, E_i) \\
 & \quad \text{in } (S_i \vec{S}, \tau) \\
 & \mathcal{W}(\Gamma, \text{if } E_c \text{ then } E_t \text{ else } E_e) = \text{let } (S_c, \tau_c) = \mathcal{W}(\Gamma, E_c) \\
 & \quad (S_t, \tau_t) = \mathcal{W}(S_c \Gamma, E_t) \\
 & \quad (S_e, \tau_e) = \mathcal{W}(S_t S_c \Gamma, E_e) \\
 & \quad S_{cb} = \mathcal{U}(\tau_c, Bool) \\
 & \quad S_{te} = \mathcal{U}(S_e \tau_t, \tau_e) \\
 & \quad \text{in } (S_{te} S_{cb} S_e S_t S_c, S_{te} \tau_t)
 \end{aligned}$$

 Figura 3.6: Algoritmo \mathcal{W}

Dato un contesto in cui analizzare un'espressione, \mathcal{W} restituisce un tipo per l'espressione e una sostituzione. Quest'ultima può essere vuota (id), generata manualmente ($\{\tau_2/\tau_1\}$) o dall'unificazione di altri tipi, e rappresenta una mappatura tra variabili di tipo e altri tipi; può essere applicata a tipi e contesti.

La funzione *new* esprime la creazione di nuove variabili di tipo, con il vincolo che non siano state ancora usate dall'algoritmo.

Infine, l'unificazione (funzione \mathcal{U}) è un'operazione altrettanto importante per l'inferenza di tipo, con la quale si cerca di trovare una sostituzione che renda due tipi uguali.

$$\mathcal{U} : \text{Type} \times \text{Type} \rightarrow \text{Substitution}$$

$$\begin{aligned}
 \mathcal{U}(\alpha, \alpha) &= id \\
 \mathcal{U}(\alpha, \tau) &= \{\tau/\alpha\} \\
 \mathcal{U}(\tau, \alpha) &= \{\tau/\alpha\} \\
 \mathcal{U}(F \ \vec{\tau}, F \ \vec{\kappa}) &= \mathcal{U}(\vec{\tau}, \vec{\kappa})
 \end{aligned}$$

 Figura 3.7: Funzione \mathcal{U}

4. Java

Parallelamente alle prime fasi di sviluppo è stata svolta un'analisi di Java¹ per valutare quali fossero le caratteristiche del linguaggio utili alla traduzione di codice **Funx**.

Questo capitolo riporta pertanto una breve panoramica delle principali funzionalità impiegate, accompagnate da esempi di traduzione che illustrano alcuni dei risultati ottenibili con il compilatore (si ricorda che altri esempi sono esibiti nel Capitolo 6).

Le funzioni non dichiarate all'interno degli esempi stessi provengono dalla piccola libreria standard (suddivisa in `FunxPrelude` e `JavaPrelude`), parzialmente presentata nella Tabella 2.1 con gli operatori simbolici.

4.1 Interfacce funzionali

Nel tradurre un linguaggio funzionale viene naturale pensare immediatamente alle *interfacce funzionali* e *lambda espressioni* introdotte in Java 8² per rappresentare funzioni anonime: l'interfaccia generica `Function` è la più adatta a riprodurre il comportamento dell'astrazione, come mostrato nei Codici 4.1 e 4.2.

```
1 constant : a -> b -> a
2 constant x = \y -> x
```

Codice 4.1: Semplice funzione in **Funx**

```
1 public static <a, b> Function<a, Function<b, a>> constant() {
2     return (x -> (y -> x));
3 }
```

Codice 4.2: Corrispondente metodo in Java

Dato il naturale *currying* di oggetti `Function`, questo tipo di traduzione ha il vantaggio di permettere l'applicazione parziale di funzioni (tramite una sequenza di `apply()` in numero minore rispetto al totale degli input), ma il grande svantaggio della creazione di una nuova istanza della funzione per ogni chiamata.

Poiché `constant` ha tipo polimorfo (sezione 4.3), il metodo utilizza parametri di tipo e deve quindi necessariamente restituire un nuovo oggetto con ogni chiamata: nonostante la performance delle traduzioni non sia un obiettivo primario del progetto, la versione attuale del compilatore fa uso di alcune piccole ottimizzazioni nella trasposizione delle funzioni monomorfe, approfondite nella sezione ??.

¹OpenJDK (<https://openjdk.org>)

²OpenJDK 8 (<https://openjdk.org/projects/jdk8>)

In aggiunta alla classe `Function`, nella parte nativa (codice Java) della libreria standard di **Funx** è definita un'ulteriore interfaccia funzionale per creare espressioni `let`: in questo caso vi è un grande utilizzo di classi anonime, potenzialmente annidate, rappresentanti le dichiarazioni locali e l'espressione principale.

```
1 @FunctionalInterface
2 public interface Let<T> {
3     T _eval();
4 }
```

Codice 4.3: Interfaccia funzionale per espressioni `let`

```
1 hundredsSum : Int -> Int -> Int
2 hundredsSum = let
3     on : (a -> a -> b) -> (c -> a) -> c -> c -> b
4     on op f x y = op (f x) (f y)
5     in on add (multiply 100)
```

Codice 4.4: Espressione `let` in **Funx**

```
1 public static Function<Long, Function<Long, Long>> hundredsSum;
2
3 static {
4     hundredsSum = (new Let<>() {
5         private <a, b, c>
6             Function<
7                 Function<a, Function<a, b>>,
8                 Function<Function<c, a>, Function<c, Function<c, b>>>>
9                 on() {
10                     return (op -> (f -> (x -> (y -> op.apply(f.apply(x)).apply(f.apply(y))))));
11                 }
12             }
13         @Override
14         public Function<Long, Function<Long, Long>> _eval() {
15             return this.<Long, Long, Long>on().apply(add).apply(multiply.apply(100L));
16         }
17     })._eval();
18 }
```

Codice 4.5: Corrispondente classe anonima in Java

Nei Codici 4.4 e 4.5 si può vedere che la funzione `hundredsSum` è implementata attraverso la chiamata al metodo principale dell'interfaccia funzionale `Let`, realizzato internamente alla classe anonima con il supporto del metodo polimorfo `on`.

Inoltre, è immediatamente evidente come le traduzioni in Java siano progressivamente più complesse e meno leggibili con l'introduzione di nuove funzionalità: l'esempio più eclatante è dato proprio dalla *signature* del metodo locale `on`, divenuta di difficile comprensione rispetto alla sintassi molto concisa del linguaggio funzionale.

4.2 Operatore ternario

Una delle peculiarità della traduzione da **Funx** a Java è l'uso dell'operatore ternario (`condition ? thenBranch : elseBranch`) ogni qualvolta siano presenti espressioni condizionali `if-then-else`.

I linguaggi funzionali sfruttano spesso una caratteristica (non menzionata nella sezione 2.1) che prende il nome di *lazy evaluation* (valutazione pigra): fino a quando il risultato di un'espressione non è richiesto per un successivo calcolo, questa non verrà completamente valutata.

Oltre ad offrire molteplici possibilità di ottimizzazione dal punto di vista del tempo di esecuzione, tale comportamento è molto comodo nella scrittura di funzioni che per esempio potrebbero terminare prima del previsto o magari effettuare computazioni su strutture dati infinite. I linguaggi che sono *lazy evaluated* di default impegnano nella maggior parte dei casi un *garbage collector* per liberare la memoria occupata dalle espressioni non valutate e non più rilevanti.

Il linguaggio Java non adotta la *lazy evaluation* di default se non in casi particolari, tra cui gli operatori booleani binari, il costrutto `if-then-else` (e corrispondente operatore ternario) e altre funzionalità più avanzate tra cui gli *stream* e le *lambda espressioni* già viste. Utilizzando quest'ultime si potrebbero ottenere risultati simili, in termini di valutazione pigra, a quelli di un linguaggio funzionale; tuttavia, rendere **Funx** un linguaggio completamente pigro avrebbe comportato una traduzione indubbiamente ancora più complessa, molteplici rischi di peggiorare le prestazioni dei programmi e un'implementazione del compilatore che va oltre lo scopo di questo progetto.

Nonostante ciò, la scelta di ridurre gli operatori booleani binari (*and* e *or*) ad espressioni con operatore ternario è stata considerata quasi obbligatoria per conservarne la natura pigra: gli operatori ternari utilizzati a questo scopo derivano direttamente dalla costruzione dell'**AST** (sezione 5.2.2), motivo per cui non vengono riconvertiti in operatori nativi di Java in fase di traduzione.

```

1 power : Int -> Int -> Int
2 power b e = if e == 0 then 1 else b * power b (e - 1) fi
3
4 xor : Bool -> Bool -> Bool
5 xor a b = (a || b) && !(a && b)

```

Codice 4.6: If e operatori booleani in **Funx**

```

1 public static Function<Long, Function<Long, Long>> power;
2
3 public static Function<Boolean, Function<Boolean, Boolean>> xor;
4
5 static {
6     power = (b -> (e -> ((JavaPrelude.<Long>equalsEquals().apply(e).apply(0L))
7         ? (1L)
8         : (multiply.apply(b)
9             .apply(power.apply(b).apply(subtract.apply(e).apply(1L)))))));
10
11     xor = (a -> (b ->
12         (((a) ? (true) : (b))) ? (not.apply(((a) ? (b) : (false)))) : (false)));
13 }

```

Codice 4.7: Corrispondenti operatori ternari in Java

4.3 Tipi generici

Il sistema di tipo di **Funx** necessita la traduzione di funzioni polimorfe, e la soluzione più semplice e idiomatica in Java è l'utilizzo dei *generics*: tramite i parametri di tipo generici è possibile definire classi e metodi che agiscono su molteplici tipi di dati, implementando comportamenti che possono essere condivisi dai diversi elementi del dominio di tipi delle funzioni rappresentate.

Nel contesto del *sistema HM* di **Funx**, le variabili quantificate universalmente nei politipi hanno una diretta corrispondenza con i parametri di tipo che possono essere dichiarati tra i modificatori di visibilità e il tipo di ritorno di un metodo, il quale a sua volta combacia con la parte interna dello schema di tipo.

Nei Codici 4.8 e 4.9 si può notare come Java non sia in grado di inferire i tipi desiderati per le funzioni polimorfe: queste devono infatti essere istanziate esplicitamente usando la classe di appartenenza e le parentesi angolari (questa limitazione richiederà alcuni espedienti in casi limite illustrati nella sezione ??).

```

1 sumToN : Int -> Int
2 sumToN = let
3   ap : (a -> b -> c) -> (a -> b) -> a -> c
4   ap op f x = op x (f x)
5   in (flip divide 2) . ap multiply (add 1)

```

Codice 4.8: Scrittura e utilizzo di funzioni polimorfe in **Funx**

```

1 public static Function<Long, Long> sumToN;
2
3 static {
4   sumToN = (new Let<>() {
5     private <a, b, c>
6       Function<
7         Function<a, Function<b, c>>,
8         Function<Function<a, b>, Function<a, c>>>
9         ap() {
10          return (op -> (f -> (x -> op.apply(x).apply(f.apply(x))));
11        }
12
13    @Override
14    public Function<Long, Long> _eval() {
15      return FunxPrelude.<Long, Long, Long>compose()
16        .apply(FunxPrelude.<Long, Long, Long>flip().apply(divide).apply(2L))
17        .apply(this.<Long, Long, Long>ap().apply(multiply).apply(add.apply(1L)));
18    }
19  })._eval();
20 }

```

Codice 4.9: Corrispondente traduzione in Java

5. Compilatore

Il software sviluppato per il progetto è stato sviluppato mantenendo il codice sul repository GitHub **Funx-jt**¹, nel cui nome *"jt"* è l'acronimo per *"Java Transpiler"*.

La prima release stabile è disponibile sul repository (Funx-jt-0.1.0) con una *Command Line Interface* per traduzione ed esecuzione di programmi **Funx**; è stato utilizzato Gradle² per la compilazione e la gestione delle dipendenze, assieme alla più recente versione *Long Term Support (LTS)* di Java, OpenJDK 21³.

Nel corso di questo capitolo si discuteranno le fasi di compilazione e la struttura del software, analizzando nel dettaglio le parti più importanti.

```
Usage: Funx-jt [-hV] [COMMAND]
A Funx to Java source transpiler.
  -h, --help      Show this help message and exit.
  -V, --version   Print version information and exit.
Commands:
  t  Transpile a funx program
  r  Run a funx program
```

```
Usage: Funx-jt t [-adhIJp] [-o=<outputDir>] <input>
Transpile a funx program
  <input>          Input funx source file
  -a, --ast        Output abstract syntax tree visualization
  -d, --dot        Keep .dot file after AST visualization
  -h, --help       Show this help message and exit.
  -I, --no-inference Omit AST type inference annotations
  -J, --no-java    Skip Java transpilation
  -o, --output=<outputDir> Output directory
  -p, --parse-tree Output raw parse tree visualization
```

```
Usage: Funx-jt r [-h] <input>
Run a funx program
  <input>          Input java source file
  -h, --help       Show this help message and exit.
```

Figura 5.1: Possibili comandi e opzioni della *CLI*

¹massimopavoni/Funx-jt (<https://github.com/massimopavoni/Funx-jt>)

²Gradle Build Tool (<https://gradle.org>)

³OpenJDK 21 (<https://openjdk.org/projects/jdk/21>)

5.1 ANTLR

Al fine di semplificare lo sviluppo di *lexer* e *parser* per il linguaggio funzionale ideato è stato scelto il generatore di *parser* chiamato ANTLR⁴.

Grazie a tale strumento il processo iterativo di creazione della grammatica di **Funx** è stato notevolmente semplificato e accelerato, in quanto ANTLR mette a disposizione del programmatore un linguaggio per definire uno o più file di specifica per lessico e sintassi (directory `Funx-jt/src/main/antlr` nel repository): questi vengono poi processati per generare il codice sorgente del *lexer* e del *parser*.

5.1.1 Analisi lessicale

Data la probabile complessità delle regole della grammatica di **Funx**, fin dall'inizio la definizione dei *token* (lessemi) del linguaggio è stata separata dalla specifica del *parser*.

Il file `FunxLexer.g4` descrive i lessemi dividendoli nelle seguenti categorie:

1. *whitespace*: caratteri di spaziatura e tabulazione;
2. *comments*: commenti di linea e blocco;
3. *keywords*: parole chiave del linguaggio;
4. *Java keywords*: parole chiave del linguaggio Java, da evitare;
5. *types*: tipi di dato (funzioni di tipo con arità 0);
6. *literals*: costanti booleane e numeriche;
7. *variables*: identificatori per variabili di tipo o nomi di funzioni;
8. *module*: identificatori il modulo;
9. vari operatori simbolici per:
 - *bool*: valori booleani;
 - *comparison*: confronti tra numeri;
 - *arithmetic*: operazioni aritmetiche;
 - *other symbols*: simboli della sintassi (come `->`) e varie funzioni di libreria;
 - *delimiters*: parentesi tonde, quadre e graffe.

Le categorie 1 e 2 contengono token da scartare, tranne `NEWLINE`, mentre la categoria 4 è utile qualora eventualmente si permetta allo sviluppatore di utilizzare tali parole chiave riservate, effettuando una rinomina automatica; le categorie 7 e 8 devono necessariamente apparire dopo le categorie 3 e 4, poiché tra *keyword* e identificatori di ogni genere le prime devono avere la precedenza (la posizione della categoria 5 tiene conto di una possibile futura estensione per consentire la creazione di nuovi tipi).

Oltre alle categorie illustrate, in testa al file sono presenti dei cosiddetti *fragment* (frammenti) che semplificano le espressioni regolari dei *token* e complessivamente aumentano la leggibilità della specifica.

⁴ANother Tool for Language Recognition; «ANTLR: A predicated-LL(k) parser generator» [PQ95], *The Definitive ANTLR 4 Reference* [Par13] and ANTLRv4 (<https://www.antlr.org>)


```

1  lexer grammar FunxLexer;
2
3  // Fragments
4  fragment LALPHA: [a-z];
5  fragment UALPHA: [A-Z];
6  fragment ALPHA: LALPHA | UALPHA;
7  fragment ALPHA_: ALPHA | UnderScore;
8
9  fragment DIGIT: [0-9];
10 fragment DECIMAL: DIGIT+;
11
12 // Whitespace
13 NEWLINE: '\r'? '\n' | '\r';
14
15 TAB: [\t]+ -> skip;
16 WS: [\u0020\u00a0\u0168\u0200\u200a\u202f\u205f\u3000]+ -> skip;
17
18 // Comments
19 CloseMultiComment: '/*';
20 OpenMultiComment: '/*';
21 SingleComment: '//';
22
23 COMMENT: SingleComment ~[\r\n]* -> skip;
24 MULTICOMMENT: OpenMultiComment .*? CloseMultiComment -> skip;
25
26 // Keywords
27 ELSE: 'else';
28 FI: 'fi';
29 IF: 'if';
30 IN: 'in';
31 LET: 'let';
32
33 // Java keywords
34 RESERVED_JAVA_KEYWORD: 'abstract' | 'assert' | 'boolean' | 'break' | 'byte' | [...];
35
36 // Types
37 TYPE: BOOLTYPE | INTTYPE;
38 BOOLTYPE: 'Bool';
39
40 // Literals
41 INT: DECIMAL | OpenParen '-' DECIMAL CloseParen;
42
43 // Variables
44 VARID: LALPHA (ALPHA_ | DIGIT)*;
45
46 // Module
47 MODULEID: UALPHA (ALPHA_ | DIGIT)*;
48
49 // Bool
50 And: '&&';
51 Not: '!!';
52
53 // Comparison
54 EqualsEquals: '==';
55 NotEquals: '!=';
56
57 // Arithmetic
58 Add: '+';
59
60 // Other symbols
61 UnderScore: '_';
62 Arrow: '->';
63
64 // Delimiters
65 OpenParen: '(';
66 CloseParen: ')';

```

Codice 5.1: Alcune *token* del *lexer*

5.1.2 Analisi sintattica

Il file `FunxParser.g4` contiene le regole concrete della grammatica di **Funx**: nonostante la somiglianza con le grammatiche delle Figure 2.2 e 3.3, è evidente che queste non collimino esattamente a causa di zucchero sintattico e requisiti di ANTLR.

Lo strumento utilizzato, infatti, è un generatore di *parser* di tipo *top-down* per grammatiche *LL*, le quali in generale non supportano regole ricorsive a sinistra.

Essendo tali regole spesso comuni nella definizione di qualsiasi linguaggio di programmazione, **Funx** incluso, ANTLRv4 offre un diverso tipo di parsing, detto *Adaptive LL(*)*⁵: quest'ultimo è in grado di riscrivere automaticamente le grammatiche, eliminando la ricorsione a sinistra diretta (e.g. linee 36 e 38-43), così da non incorrere in regole ambigue che potrebbero causare *backtracking* e conseguente *overhead*.

Il Codice 5.2 riporta integralmente le regole concrete della sintassi di **Funx**, tra cui:

- *module*: nome del modulo, funzione *main* opzionale e dichiarazioni globali;
- *main*: funzione *main*, diversa dalle dichiarazioni classiche per l'assenza di schema di tipo e parametri lambda;
- *declaration*: funzione con nome, tipo e parametri (e opzionalmente *with* per funzioni locali);
- *typeElems*: tipo di una funzione, definito ricorsivamente secondo la grammatica del sistema di tipo di **Funx**;
- *statement*: per evitare ricorsione a sinistra indiretta, la separazione tra *statement* e *expression* forza l'uso di parentesi nei casi in cui lambda astrazioni, *let* e *if* siano usati all'interno di un'espressione;
- *expression*: racchiude l'applicazione funzionale, tutte le regole relative agli operatori simbolici, specificandone la priorità implicite (Tabella 2.1), e le espressioni primarie (costanti, variabili e parentesi per controllare la precedenza);
- *lambda*, *let*, *it*: corrispondenti alle produzioni per astrazione, *let* e *if* della grammatica formale.

⁵«LL(*)»: the foundation of the ANTLR parser generator» [PF11] and «Adaptive LL(*) parsing: the power of dynamic analysis» [PHF14]

```

1  parser grammar FunxParser;
2  options { tokenVocab = FunxLexer; }
3
4  // Module
5  module: (MODULE MODULEID (Dot MODULEID)* NEWLINE+)?
6         (main NEWLINE+)? declarations EOF;
7
8  declarations: declaration (NEWLINE declaration?)*;
9
10 main: id = MAIN Equals statement with?;
11
12 // Declaration
13 declaration: (declarationScheme NEWLINE)?
14             id = VARID lambdaParams? Equals statement with?;
15
16 declarationScheme: id = VARID Colon typeElems;
17
18 with: NEWLINE WITH localDeclarations OUT;
19
20 localDeclarations: NEWLINE declarations NEWLINE;
21
22 // Type
23 typeElems: OpenParen typeElems CloseParen # parenType
24           | VARID # typeVar
25           | TYPE # namedType
26           | <assoc = right> typeElems Arrow typeElems # arrowType;
27
28 // Statement
29 statement: expression # expressionStatement
30           | lambda # lambdaStatement
31           | let # letStatement
32           | ifs # ifsStatement;
33
34 // Expression
35 expression: primary # primExpression
36           | expression expression # appExpression
37           | <assoc = right> expression bop = Dot expression # composeExpression
38           | expression bop = (Divide | Modulo | Multiply) expression # divModMultExpression
39           | expression bop = (Add | Subtract) expression # addSubExpression
40           | expression
41             bop = (GreaterThan | GreaterThanEquals | LessThan | LessThanEquals)
42             expression # compExpression
43           | expression bop = (EqualsEquals | NotEquals) expression # eqExpression
44           | uop = Not expression # notExpression
45           | <assoc = right> expression bop = And expression # andExpression
46           | <assoc = right> expression bop = Or expression # orExpression
47           | <assoc = right> expression bop = Dollar expression # rightAppExpression;
48
49 primary: OpenParen statement CloseParen # parenPrimary
50         | constant # constPrimary | VARID # varPrimary;
51
52 // Lambda
53 lambda: Backslash lambdaParams? Arrow statement;
54
55 lambdaParams: VARID+;
56
57 // Let
58 let: LET localDeclarations IN statement;
59
60 // If
61 ifs: IF statement THEN statement ELSE statement FI;
62
63 // Constant
64 constant: BOOL | numConstant;
65
66 numConstant: INT;

```

Codice 5.2: Grammatica per il *parser*

5.2 Albero sintattico astratto

Il *parser* generato da ANTLR utilizza i token identificati dal *lexer* per costruire un albero di parsing concreto (*CST*, *concrete syntax tree*), molto complesso e poco comprensibile. È dunque opportuno astrarre i dettagli del parsing in una struttura più semplice e facile da interpretare, che rispecchia precisamente la grammatica del linguaggio (Figura 2.2). Tale operazione è effettuata dalla grande maggioranza dei compilatori moderni e il risultato è chiamato **AST**, *abstract syntax tree* (albero sintattico astratto).

5.2.1 Gerarchia delle classi

Il primo passo per la costruzione di un **AST** per la sintassi di **Funx** è la definizione di una gerarchia di classi Java che rappresentano i nodi dell'albero (`package com.github.massimopavoni.funx.jt.ast.node`).

La classe astratta `ASTNode` è la radice dell'ordinamento poiché sarà usata per gli oggetti creati a partire dal *CST*: contiene la proprietà `inputPosition` per facilitare la segnalazione di errori e vincola le classi figlie all'implementazione del metodo astratto `accept` per visitare i nodi.

Un'altra classe astratta, derivata dalla precedente, è `Expression`, la quale identifica un'espressione ed è dotata di alcuni campi e metodi utili all'inferenza di tipo.

Ogni altra sottoclasse (ad eccezione di `Declarations`, utilizzata per dichiarazioni globali e locali) è una trascrizione in Java delle produzioni della grammatica formale:

- `Module`: modulo del programma;
- `Declaration`: dichiarazione di funzione;
- `Constant`: termini costanti;
- `Variable`: simboli per variabili;
- `Application`: applicazione di funzione;
- `Lambda`: astrazione per le funzioni anonime;
- `Let`: contenitore di dichiarazioni locali;
- `If`: costrutto condizionale.

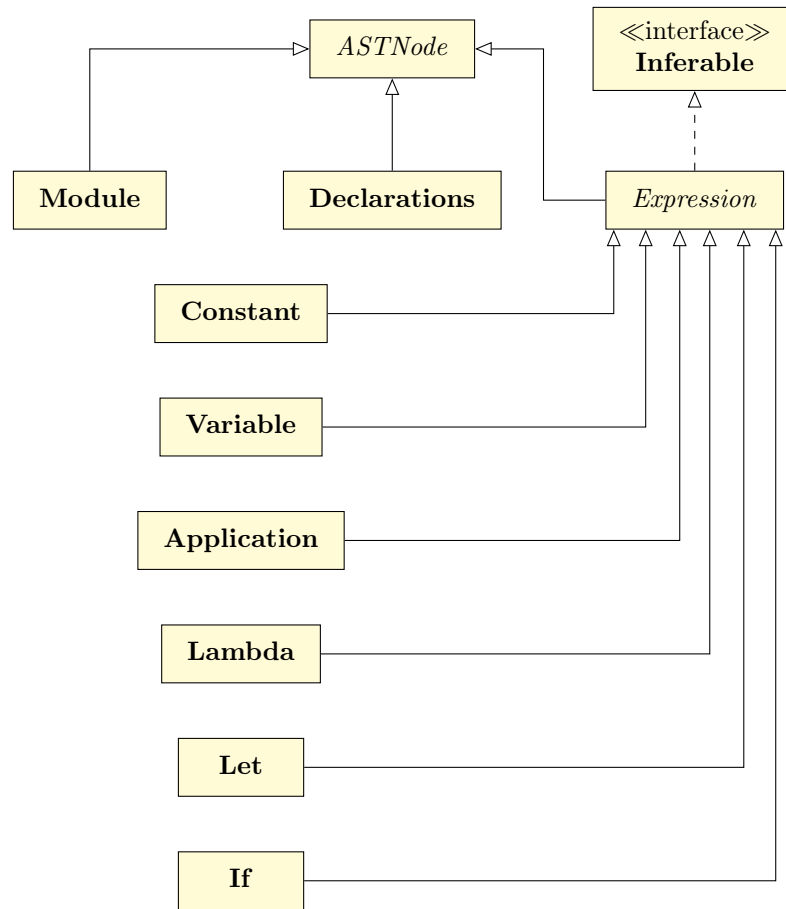


Figura 5.2: Diagramma semplificato delle classi dell'AST

```

1 public static final class Lambda extends Expression {
2     public final String paramId;
3
4     public final Expression expression;
5
6     public Lambda(InputPosition inputPosition, String paramId, ASTNode expression) {
7         super(inputPosition); // ASTNode constructor
8         this.paramId = paramId;
9         this.expression = (Expression) expression;
10    }
11
12    @Override // from Inferable interface
13    public Utils.Tuple<Substitution, Type> infer(Context ctx) { ... }
14
15    @Override // from Expression abstract class
16    protected void propagateSubstitution(Substitution substitution) { ... }
17
18    @Override // from ASTNode abstract class
19    public <T> T accept(ASTVisitor<? extends T> visitor) {
20        return visitor.visitLambda(this);
21    }
22 }

```

Codice 5.3: Esempio di classe della gerarchia

5.2.2 AST builder

ANTLR offre due diversi approcci per analizzare l'albero di parsing:

- *listener*: i metodi implementati per l'analisi dei nodi sono chiamati seguendo la ricerca in profondità (*DFS*, *depth-first search*);
- *visitor*: i nodi possono essere visitati arbitrariamente tramite il metodo `visit`, senza seguire un ordine specifico; non tutti i metodi devono essere implementati ed è possibile ignorare dei nodi o ri-visitarli.

La seconda opzione è indubbiamente più versatile, e può essere per esempio anche utilizzata qualora si desideri evitare completamente la costruzione di un **AST** e chiamare i metodi di visita per eseguire il codice (e.g. linguaggi interpretati).

Nel caso di **Funx-jt** la scelta ricade appunto sul *visitor* in virtù della probabilità di dover "saltare" alcuni nodi e ottenerne direttamente le informazioni interne.

Durante la compilazione del progetto Java, Gradle è configurato per generare le seguenti classi nel package `com.github.massimopavoni.funx.jt.parser`:

- `FunxLexer`: *lexer*;
- `FunxParser`: *parser*;
- `FunxParserVisitor`: interfaccia generica per differenti implementazioni del *visitor* (estende l'interfaccia `ParseTreeVisitor` di ANTLR);
- `FunxParserBaseVisitor`: classe predefinita che percorre l'intero albero di parsing (estende la classe astratta `AbstractParseTreeVisitor` per ereditare alcuni metodi standard come `visitChildren`).

La classe `ASTBuilder` estende a sua volta `FunxParserBaseVisitor` e sovrascrive quasi tutti i metodi ereditati; in particolare, `ASTBuilder` effettua la composizione degli schemi di tipo definiti dall'utente (sezione 5.3.1) e l'eliminazione dello zucchero sintattico mediante l'uso di alcune proprietà e metodi ausiliari.

Inoltre, all'interno del package `com.github.massimopavoni.funx.jt.ast` viene definita una classe enumeratore, `PreludeFunction` contenente le funzioni di libreria standard, con i relativi simboli e gli schemi di tipo corrispondenti.

```
1 public enum PreludeFunction {
2     COMPOSE(".", "compose",
3         new Scheme(Set.of(0L, 1L, 2L),
4             arrowOf(arrowOf(ZERO, ONE), arrowOf(TWO, ZERO), TWO, ONE)), false);
5
6     public final String symbol;
7     public final String id;
8     public final Scheme scheme;
9     public final boolean nativeJava;
10
11     PreludeFunction(
12         String symbol, String id, Scheme scheme, boolean nativeJava) { ... }
13
14     public static PreludeFunction fromSymbol(String symbol) {
15         return Utils.enumFromField(PreludeFunction.class,
16             f -> f.symbol.equals(symbol));
17     }
18 }
```

Codice 5.4: Parte del codice di `PreludeFunction`

```

1 private ASTNode createLambdaChain(
2     InputPosition position, Deque<String> params, ASTNode expression) {
3     if (params.size() == 1)
4         return new Expression.Lambda(
5             position,
6             params.getFirst(),
7             expression);
8     return new Expression.Lambda(
9         position,
10        params.pop(),
11        createLambdaChain(position, params, expression));
12 }
13
14 private ASTNode binarySymbolApplication(
15     InputPosition position, String symbol, ASTNode left, ASTNode right) {
16     return new Expression.Application(
17         position,
18         new Expression.Application(
19             position,
20             new Expression.Variable(
21                 position,
22                 PreludeFunction.fromSymbol(symbol).id),
23             left),
24         right);
25 }

```

Codice 5.5: Metodi per astrazioni annidate e operatori simbolici binari

```

1 public class ASTBuilder extends FunxParserBaseVisitor<ASTNode> {
2     @Override
3     public ASTNode visitAppExpression(FunxParser.AppExpressionContext ctx) {
4         return new Expression.Application(getInputPosition(ctx),
5             visit(ctx.expression(0)), visit(ctx.expression(1)));
6     }
7
8     @Override
9     public ASTNode visitComposeExpression(FunxParser.ComposeExpressionContext ctx) {
10        return binarySymbolApplication(getInputPosition(ctx),
11            Utils.fromLexerToken(ctx.bop.getType()),
12            visit(ctx.expression(0)), visit(ctx.expression(1)));
13    }
14
15    @Override
16    public ASTNode visitOrExpression(FunxParser.OrExpressionContext ctx) {
17        // transform logical disjunction into if statement for lazy behavior
18        return new Expression.If(getInputPosition(ctx),
19            visit(ctx.expression(0)),
20            new Expression.Constant(InputPosition.UNKNOWN, true),
21            visit(ctx.expression(1)));
22    }
23
24    @Override
25    public ASTNode visitLambda(FunxParser.LambdaContext ctx) {
26        return createLambdaChain(getInputPosition(ctx),
27            // lambda params are syntactic sugar for a lambda chain
28            ctx.lambdaParams().VARID().stream().map(ParseTree::getText)
29                .collect(Collectors.toCollection(ArrayDeque::new)),
30            visit(ctx.statement()));
31    }
32 }

```

Codice 5.6: Alcuni metodi *visit* di ASTBuilder

Il codice del compilatore include una classe astratta `ASTVisitor` atta a permettere l'analisi dell'**AST** in modo simile ai *visitor* di ANTLR. Il Codice 5.7 e la Figura 5.3 mostrano un **AST** come verrebbe creato da `ASTBuilder`; la *CLI* sviluppata fornisce delle opzioni per la visualizzazione dell'albero, generata utilizzando il software Graphviz⁶ con un *visitor* che costruisce progressivamente una stringa nel linguaggio DOT.

```
1 tautology : Bool -> Bool
2 tautology x = x || !!x
```

Codice 5.7: Programma in **Funx**

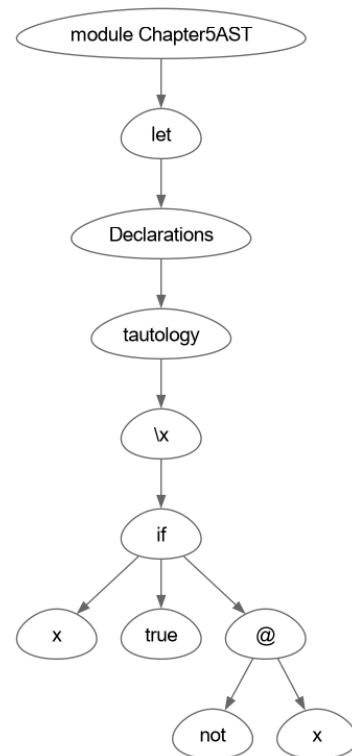
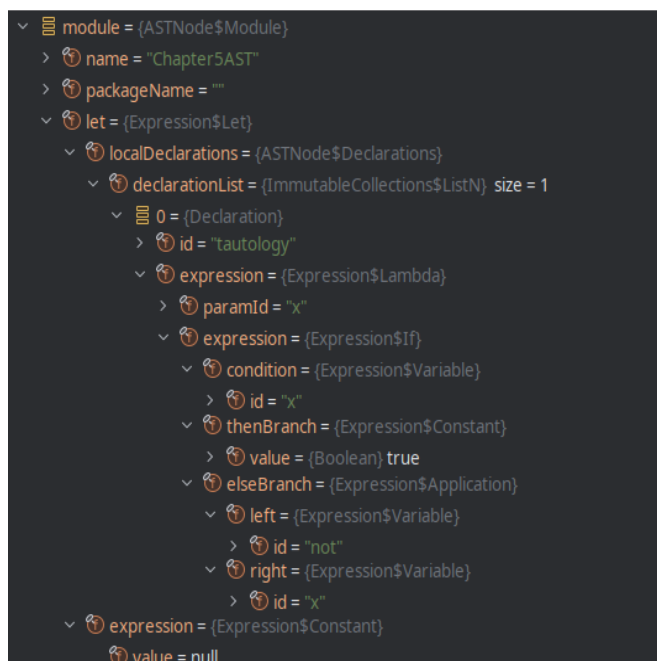


Figura 5.3: Oggetto **AST** in *debug* e visualizzazione dell'albero

⁶Graph Visualization Software (<https://graphviz.org>)

5.3 Motore di inferenza

I sistemi di tipo e l'inferenza descritti nel Capitolo 3 costituiscono una parte imprescindibile del software, data l'incapacità di simulare il polimorfismo parametrico del *sistema HM* lasciando al compilatore Java il compito di inferire e controllare i tipi dei metodi (o classi) generici. In tale caso, infatti, la compilazione fallirebbe a causa di tipi "troppo generici" e *type casting* non permesso in maniera implicita (e.g. non si è in grado di conciliare ad esempio un tipo generico con una lambda espressione).

```
Impossible.java:9: error: method apply in interface Function<T,R> cannot be applied to given types;
    id().apply(x -> y -> x || y).apply(true).apply(false);
           ^
    required: Object
    found:      (x)->(y)->x || y
    reason: argument mismatch; Object is not a functional interface
    where T,R are type-variables:
      T extends Object declared in interface Function
      R extends Object declared in interface Function
1 error
error: compilation failed
```

Figura 5.4: Esempio di errore in Java dovuto a tipi generici

La linea di codice citata in Figura 5.4 non è intrinsecamente errata, ma non può essere compilata in mancanza di un *cast* esplicito dall'oggetto lambda espressione al suo tipo concreto: il problema da risolvere è quindi conoscere il tipo di tale espressione.

Il motore di inferenza è la parte del compilatore che si occupa di stabilire i tipi delle espressioni seguendo le regole di inferenza e i passi dell'*algoritmo \mathcal{W}* ⁷ descritti nella sezione 3.2; la fase di inferenza avviene subito dopo il parsing e la costruzione dell'**AST** e precede la generazione del codice Java.

5.3.1 Sistema HM

L'implementazione del *sistema HM* è simile a quella della gerarchia di classi per l'albero sintattico astratto di **Funx**, seppur notevolmente più semplice: la complessità del

5.3.2 Inferenza su espressioni

5.4 Traduzione in Java

⁷ «Algorithm W Step by Step» [Gra06]

6. Esempi di traduzione

7. Conclusioni

Bibliografia

- [Bar91] Henk Barendregt. «Introduction to generalized type systems». In: *Journal of Functional Programming* 1.2 (apr. 1991), pp. 125–154. DOI: 10.1017/S0956796800020025.
- [Chu32] Alonzo Church. «A Set of Postulates for the Foundation of Logic». In: *Annals of Mathematics* 33.2 (apr. 1932), pp. 346–366. DOI: 10.2307/1968337.
- [Chu33] Alonzo Church. «A Set of Postulates for the Foundation of Logic». In: *Annals of Mathematics* 34.4 (ott. 1933). Second paper, pp. 839–864. DOI: 10.2307/1968702.
- [Clé+86] Dominique Clément et al. «A Simple Applicative Language: Mini-ML». In: *Proceedings of the 1986 ACM conference on LISP and functional programming - LFP '86* (8 ago. 1986), pp. 13–27. DOI: 10.1145/319838.319847.
- [DM82] Luis Damas e Robin Milner. «Principal type-schemes for functional programs». In: *Proceedings of the 9th ACM SIGPLAN-SIGACT symposium on Principles of programming languages - POPL '82* (25 gen. 1982), pp. 207–212. DOI: 10.1145/582153.582176.
- [Eis15] Richard A. Eisenberg. «System FC, as implemented in GHC». In: (2015). Last updated in 2020. URL: <https://github.com/ghc/ghc/blob/master/docs/core-spec/core-spec.pdf?raw=true>.
- [Gra06] Martin Grabmüller. «Algorithm W Step by Step». In: (26 set. 2006). URL: <https://github.com/mgrabmueller/AlgorithmW/blob/master/pdf/AlgorithmW.pdf?raw=true>.
- [Hin69] J. Roger Hindley. «The Principal Type-Scheme of an Object in Combinatory Logic». In: *Transactions of the American Mathematical Society* 146 (dic. 1969), pp. 29–60. DOI: 10.2307/1995158.
- [LY98] Oukseh Lee e Kwangkeun Yi. «Proofs about a Folklore Let-Polymorphic Type Inference Algorithm». In: *ACM Transactions on Programming Languages and Systems (TOPLAS)* 20.4 (1 lug. 1998), pp. 707–723. DOI: 10.1145/291891.291892.
- [Mil78] Robin Milner. «A Theory of Type Polymorphism in Programming». In: *Journal of Computer and System Sciences* 17.3 (dic. 1978), pp. 348–375. DOI: 10.1016/0022-0000(78)90014-4.
- [Mog91] Eugenio Moggi. «Notions of Computation and Monads». In: *Information and Computation* 93.1 (lug. 1991), pp. 55–92. DOI: 10.1016/0890-5401(91)90052-4.
- [Par13] Terence J. Parr. *The Definitive ANTLR 4 Reference*. Pragmatic Bookshelf, 22 gen. 2013. ISBN: 978-1-934356-99-9.

- [PF11] Terence J. Parr e Kathleen Fisher. «LL(*): the foundation of the ANTLR parser generator». In: *Proceedings of the 32nd ACM SIGPLAN Conference on Programming Language Design and Implementation - PLDI '11* (4 giu. 2011), pp. 425–436. DOI: 10.1145/1993498.1993548.
- [PHF14] Terence J. Parr, Sam Harwell e Kathleen Fisher. «Adaptive LL(*) parsing: the power of dynamic analysis». In: *Proceedings of the 2014 ACM International Conference on Object Oriented Programming Systems Languages and Applications - OOPSLA '14* (15 ott. 2014), pp. 579–598. DOI: 10.1145/2660193.2660202.
- [PQ95] Terence J. Parr e Russell W. Quong. «ANTLR: A predicated-LL(k) parser generator». In: *Software: Practice and Experience* 25.7 (lug. 1995), pp. 789–810. DOI: 10.1002/spe.4380250705.
- [Wel99] Joe B. Wells. «Typability and type checking in System F are equivalent and undecidable». In: *Annals of Pure and Applied Logic* 98.1-3 (30 giu. 1999), pp. 111–156. DOI: 10.1016/s0168-0072(98)00047-5.

Ringraziamenti

Ringrazio...