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

Elenco dei codici

Elenco delle figure

Elenco delle tabelle

1. Introduzione

Lorem ipsum dolor sit amet, consectetur adipisci elit, sed do eiusmod tempor incidunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrum exercitationem ullamco laboriosam, nisi ut aliquid ex ea commodi consequatur. 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 ['fʌnɪk's] in inglese, o comunque [fàn·èx] 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**: un sistema formale definito dal matematico *Alonzo Church* (supervisore di *Alan Turing* durante il dottorato), equivalente alla macchina di Turing, ma fondato sul concetto di funzione pura.

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 idiomatico 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 amato dagli sviluppatori secondo gli ultimi 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*[1] 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*.

[1] *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
		$\mid$	$E_l\ E_r$	applicazione
		$\mid$	$\lambda x\ .\ E$	astrazione

Figura 2.1: Grammatica del lambda calcolo

Le tre regole presenti 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, potenzialmente un'altra astrazione); per definire funzioni con più parametri si debbono usare molteplici astrazioni annidate (tecnica detta *currying*).

Modulo	M	::=	$nome \cdot L$	dichiarazione del modulo
Dichiarazione	D	::=	$?(schema\ di\ tipo) \cdot id = E$	dichiarazione di funzione
Espressione	E	::=	c	costante
		|	x	variabile
		|	$E_l\ E_r$	applicazione
		|	$\lambda x\ .\ E$	astrazione
		|	L	let
		|	$if\ E\ then\ E\ else\ E$	if
Let	L	::=	$let \cdot D\ (\cdot D)^* \cdot in\ E$	let

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'arrichimento del lambda calcolo, e infine tutti gli operatori simbolici supportati al momento (assieme alla notazione per indicarne associatività e precedenza).

Zucchero	**Sostituzione**
`if b then e1 else e2 fi`	`if b then e1 else e2`

`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`

`\x -> e`	`λx . e`
`\x y -> e`	`λx . λy . e`
`f x y = e`	`f = λx . λy . e`
`e1 . e2` `infixr 9`	`compose e1 e2`
`e1 / e2` `infixl 7`	`divide e1 e2`
`e1 % e2` `infixl 7`	`modulo e1 e2`
`e1 * e2` `infixl 7`	`multiply e1 e2`
`e1 + e2` `infixl 6`	`add e1 e2`
`e1 - e2` `infixl 6`	`subtract e1 e2`
`e1 > e2` `infix 4`	`greaterThan e1 e2`
`e1 >= e2` `infix 4`	`greaterThanEquals e1 e2`
`e1 < e2` `infix 4`	`lessThan e1 e2`
`e1 <= e2` `infix 4`	`lessThanEquals e1 e2`
`e1 == e2` `infix 4`	`equalsEquals e1 e2`
`e1 != e2` `infix 4`	`notEquals e1 e2`
`!!e` `prefix 4`	`not e`
`e1 && e2` `infixr 3`	`if e1 then e2 else False`
`e1 || e2` `infixr 2`	`if e1 then True else e2`
`e1 $ e2` `infixr 0`	`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 **??** in particolare esporrà 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  xor a b = (a || b) && !!(a && b)
4
5  factorial n = if n == 0 then 1 else n * factorial (n - 1) fi
6
7  gcd a b = if b == 0 then a else gcd b (a % b) fi
8
9  even = let
10    even1 n = if n == 0 then True else odd (n - 1) fi
11
12    odd n = if n == 0 then False else even1 (n - 1) fi
13  in even1
```

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 *"tipo"* 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 impedite, oppure qualora sia o meno disponibile l'aritmetica dei puntatori.

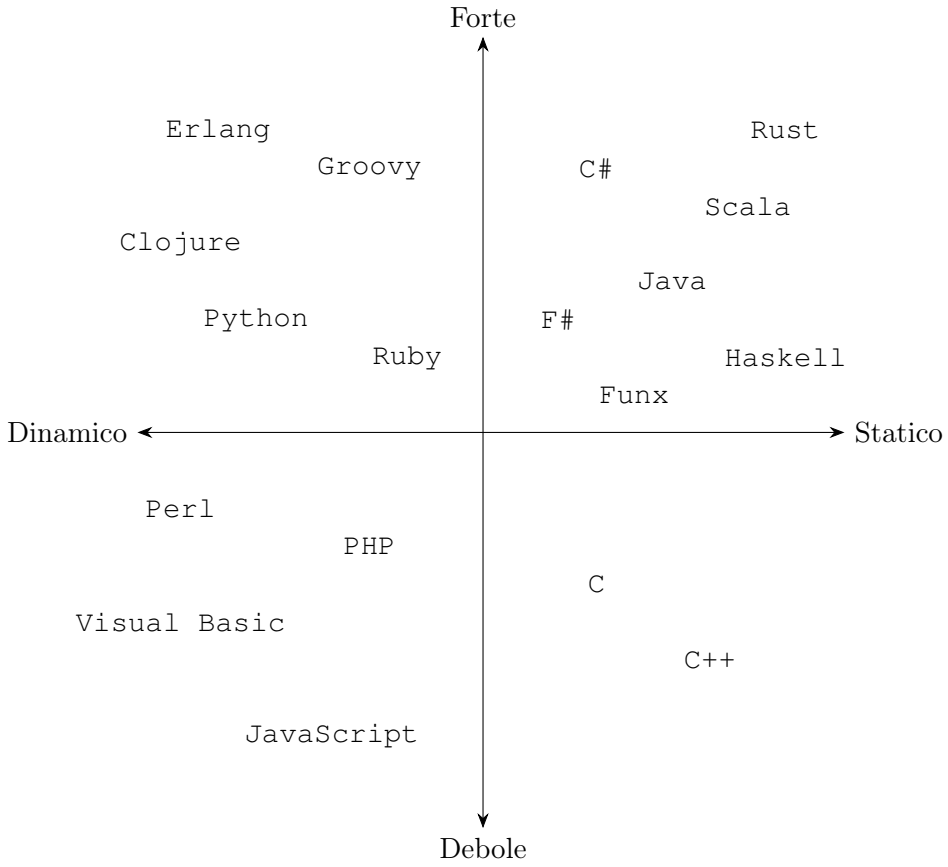

Figura 3.1: Alcuni linguaggi e loro sistemi di tipo

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 cubo[1], 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).

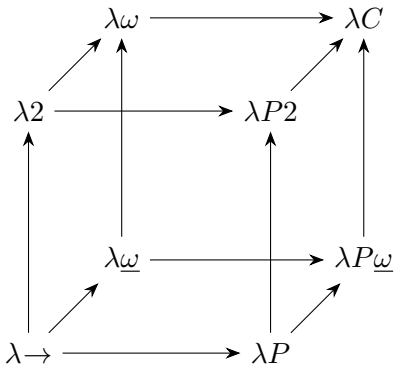

Figura 3.2: λ-cubo

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 pi weak omega*): costruttori di tipo e tipi dipendenti;
- $\lambda\omega$ (*lambda omega*): costruttori di tipo e polimorfismo parametrico;
- $\lambda P2$ (*lambda P due*): polimorfismo parametrico e tipi dipendenti;
- λC (*lambda C, calcolo delle costruzioni*): combinazione 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 di quelli più interessanti è il *sistema F* (vertice $\lambda 2$ in Figura 3.2) poiché molto utile per la generalizzazione delle funzioni: uno dei problemi più ricorrenti 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 e tipi polimorfi.

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 decidibile[2].

Pertanto, il linguaggio `Haskell`, così come **Funx**, 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. `(forall a. a -> a) -> (forall b. b -> b)`); `Haskell` lo realizza grazie all'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 dunque lasciare intendere un linguaggio più espressivo di quanto non sia in realtà: in ogni caso, il fine di questi paragrafi è anche difendere la scelta di utilizzare l'algoritmo di inferenza descritto successivamente, 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 (vedi sezione **??**).

[2] *Typability and type checking in System F are equivalent and undecidable*, [Wel99]
[3] *System FC, as implemented in GHC*, [Eis15]

Funzione	**Schema**
`id`	`a -> a`
`const`	`a -> b -> a`
`(.)`	`(b -> c) -> (a -> b) -> a -> c`
`flip`	`(a -> b -> c) -> b -> a -> c`
`($)`	`(a -> b) -> a -> b`
`(&)`	`a -> (a -> b) -> b`
`on`	`(b -> b -> c) -> (a -> b) -> a -> a -> c`

Tabella 3.1: Esempi di funzioni polimorfe

In Figura 3.3 è mostrata la grammatica per la definizione dei tipi nel *sistema FC* così come implementato nel linguaggio **Funx**. Si noti come i tipi monomorfi possano solo essere variabili di tipo o applicazioni di funzioni ad altri tipi; al momento il linguaggio mette a disposizione le funzioni di tipo più elementari: funzione (l'unica con notazione infissa), booleano e intero.

Schema di tipo	S	$::=$	$\forall\ \alpha.\ S$	tipo polimorfo
		$\mid$	T	tipo monomorfo
Tipo	T	$::=$	α	variabile di tipo
		$\mid$	$F\ T^*$	applicazione di funzione di tipo
Funzione di tipo	F	$::=$	$\rightarrow_2$	funzione
		$\mid$	$Bool_0$	booleano
		$\mid$	Int_0	intero

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

3.2 Inferenza secondo Hindley–Milner

4. Java

5. Compilatore

6. Esempi di traduzione

7. Conclusioni

Bibliografia

[Bar91] Henk Barendregt. «Introduction to generalized type systems». In: *Journal of Functional Programming* 1 (2 1991), pp. 125–154.

[Eis15] Richard A. Eisenberg. *System FC, as implemented in GHC*. Last revised in 2020. 2015. URL: https://github.com/ghc/ghc/blob/master/docs/core-spec/core-spec.pdf?raw=true.

[Mog91] Eugenio Moggi. «Notions of computation and monads». In: *Information And Computation* 93 (1 1991), pp. 55–92.

[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 1999), pp. 111–156.

Indice analitico

Ringraziamenti

Ringrazio...