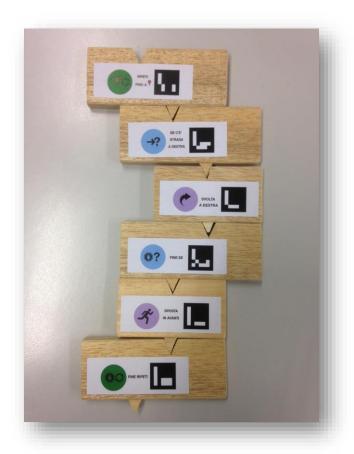


Informatica e Laboratorio di Programmazione linguaggi Alberto Ferrari



o linguaggi formali

- o linguaggi di *programmazione*
- o linguaggi di *marcatura* (es. HTML, Latex)
- o *interazione* uomo macchina (es. ricerca <u>Google</u>)
- o linguaggi nel software di sistema
 - $\circ \ compilatori$
 - o interpreti ...



http://www.ce.unipr.it/~aferrari/codowood

descrizione di un linguaggio

o grammatica

- o quali frasi sono corrette?
- o lessico
 - o analisi lessicale (sequenze corrette di simboli dell'alfabeto (token))
- o sintassi
 - o analisi sintattica (sequenze corrette di token)

o semantica

- o cosa significa una frase corretta?
- o attribuzione del significato

o pragmatica

o come usare una frase corretta?

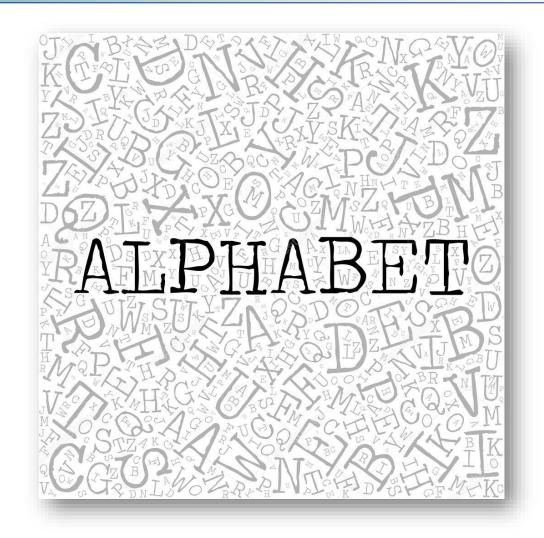
o implementazione

o come eseguire una frase corretta rispettando la semantica?



linguaggi formali

GRAMMATICA







- o alfabeto Σ : insieme di simboli
- \circ stringa **s**: sequenza di simboli di Σ
 - o $\mathbf{s} \in \Sigma^*$, insieme di tutte le stringhe
 - ο ε: stringa vuota
 - o |s|: lunghezza della stringa s
- o linguaggio $L \subseteq \Sigma^*$
 - o **sottoinsieme** di tutte le stringhe possibili
 - o grammatica: regole formali per definire le "stringhe ben formate" di L
- o esempio: numeri romani da 1 a 1000
 - \circ alfabeto $\{I, V, X, L, C, D, M\}$ + regole ...





- o operazione di concatenazione
 - o proprietà associativa: $(x \cdot y) \cdot z = x \cdot (y \cdot z)$
 - o non commutativa: $\mathbf{x} \cdot \mathbf{y} \neq \mathbf{y} \cdot \mathbf{x}$
 - o Σ^* chiuso rispetto alla concatenazione: $\Sigma^* \bullet \Sigma^* \to \Sigma^*$
- o potenza

$$\circ$$
 $x^n = x \cdot x \cdot x \cdot x \dots (n \text{ volte})$

- o elemento neutro ε
 - o stringa vuota, $\forall x \in \Sigma^*, \varepsilon \cdot x = x \cdot \varepsilon = x$





- \circ L₁ ed L₂ linguaggi su Σ^* (due insiemi di stringhe)
- o unione: $L_1 \cup L_2 = \{x \in \Sigma^* : x \in L_1 \lor x \in L_2\}$
- o *intersezione*: $L_1 \cap L_2 = \{x \in \Sigma^* : x \in L_1 \land x \in L_2\}$
- o complementazione: $\overline{\mathbf{L_1}} = \{\mathbf{x} \in \Sigma^* : \mathbf{x} \notin \mathbf{L_1}\}$
- concatenazione o *prodotto*: $L_1 \cdot L_2 = \{x \in \Sigma^* : x = x1 \cdot x2, x1 \in L_1, x2 \in L_2\}$
- o **potenza**: $L^n = L \cdot L^{n-1}$, $n \ge 1$; $L^0 = \{\epsilon\}$ per convenzione
 - o concatenazione di n stringhe qualsiasi di L
- $stella\ di\ Kleene$: L* = \cup Lⁿ, n = $0..\infty$
 - o concatenazione arbitraria di stringhe di L

 $m{L}$ *: chiusura riflessiva e transitiva di $m{L}$ rispetto a •





- o approccio algebrico
 - o linguaggio costruito a partire da linguaggi più elementari, con *operazioni* su linguaggi
- o approccio generativo
 - o *grammatica*, regole per la *generazione* di stringhe appartenenti al linguaggio
- o approccio riconoscitivo
 - o *macchina* astratta o *algoritmo* di *riconoscimento*, per decidere se una stringa appartiene o no al linguaggio



un esempio di grammatica

espressioni regolari - regex



- o dato un alfabeto Σ , chiamiamo $espressione \ regolare$ una stringa \mathbf{r} sull'alfabeto $\Sigma \cup \{+, *, (,), \bullet, \emptyset\}$ tale che:
 - o r = Ø: linguaggio vuoto; oppure
 - \circ r \in Σ : linguaggio con un solo simbolo; oppure
 - \circ r = s + t: unione dei linguaggi L(s), L(t); oppure
 - o r = s t: concatenazione dei linguaggi L(s), L(t); oppure
 - \circ r = s*: chiusura del linguaggio L(s)
 - o (con s e t espressioni regolari; simbolo spesso implicito)

esempio
$$\Sigma = \{a, b\}$$
: $a \cdot (a + b) \cdot b$



- o una *espressione regolare* (*regular expression*) è una sequenza di simboli che identifica un insieme di stringhe
- o esistono *regole condivise* fra i vari linguaggi relative alle regex 😊
- o i vari linguaggi utilizzano notazioni diverse per rappresentare le stesse espressioni regolari 🖰
- o le espressioni regolari possono definire tutti e soli i *linguaggi regolari*
 - o nella gerarchia di Chomsky la classe dei linguaggi generati da grammatiche di tipo 3
 - o riconosciuti da automi a stati finiti
- o espressioni regolari nei linguaggi di programmazione
 - o prima implementazione *PERL* '80





- o [...] per includere uno qualsiasi dei caratteri in parentesi
- o singoli caratteri o intervalli di caratteri adiacenti
 - o [A-Z] = qualsiasi lettera maiuscola
 - o [a-zABC] = qualsiasi lettera minuscola oppure A, B, o C
- o [^...] per escludere uno qualsiasi dei caratteri in parentesi
 - o [^0-9] = qualsiasi carattere non numerico
- o simboli speciali per identificare classi di caratteri
 - $\circ \ \mathbf{d} = \text{numerico}, \text{ equivale a } [0-9]$

 - $\circ \ \mathbf{w} = [0.9a zA Z_{-}]$
 - \circ **D** = non numerico, equivale a [^0-9]



- o . per un carattere qualsiasi
 - o A.B riconosce la stringa AoB, AwB, AOB ecc.
- \ escape, per sequenze speciali o caratteri speciali
 - \? cerca il ?
- o ^ corrisponde all'inizio del testo
- \$ corrisponde alla fine del testo
- | per alternativa tra due espressioni (unione)
 - A | B = carattere A o carattere B
- o (...) per raggruppare sotto-espressioni
 - o ga(zz|tt)a trova sia gazza che gatta



- o {...} per specificare il numero di ripetizioni
 - o \d{3,5} sequenze di almeno tre cifre, al massimo cinque
- o * zero o più occorrenze di un'espressione
 - o (ab)* riconosce ab, abab, la stringa vuota, ma non riconosce abba
- o + una o più occorrenze
 - o (ab)+ non riconosce la stringa vuota
- o ? zero o al più una occorrenza (parte opzionale)
 - o (ab)? riconosce ab ma non abab





- o codice fiscale:
 - \circ ^[A-Z]{6}[0-9]{2}[A-Z][0-9]{2}[A-Z][0-9]{3}[A-Z]\$
 - o dominio:
 - o ^[\w\-]+\.(it|com|org|net|eu|mobi)\$
- o e-mail:
 - o ^[\w\-\.]+@[\w\-\.]+\.[a-z]+\$
- o file: ^.+\.zip\$
- o data: ^\d{2}/\d{2}/\d{4}\$

html & JavaScript - regex

```
<script>
function validate(val) {
    var expr = /^{d{2}}//d{2}//d{4}$/;
    if (! expr.test(val)) {
        window.alert("Wrong date format");
        return false;
    return true;
</script>
<form onsubmit="return validate(usrdate.value)">
    Date (dd/mm/yyyy): <input name="usrdate">
    <input type="submit">
</form>
```



- o modulo standard re
- o le espressioni regolari vengono *compilate* in oggetti RegexObject
 - o gli oggetti RegexObject contengono metodi per varie operazioni
 - o *match()*
 - o determina se la RE corrisponde all'inizio della stringa (None se non trovata)
 - o search()
 - o ricerca tutte le posizioni corrispondenti alla RE all'interno di una stringa (None se non trovata)
 - o findall()
 - o trova tutte le sottostringhe corrispondenti alla RE, e le restituisce in una lista
 - 0 ...

```
>>> import re
>>> p = re.compile('ab*')
```

https://docs.python.it/html/lib/re-syntax.html



```
import re
ro = re.compile('[a-z]+') # uno o più caratteri alfabetici minuscolo
#f = ro.match('sistemi09informativi')
f = ro.search('09sistemi09informativi')
if f:
    #group() restituisce la stringa corrispondente alla RE
   print('f.group()',f.group())
    #start() restituisce la posizione iniziale della corrispondenza
   print('f.start()',f.start())
    #end() restituisce la posizione finale della corrispondenza
   print('f.end()',f.end())
    #span() restituisce una tupla contenente la (start, end) posizione della corrispondenza
   print('f.span()',f.span())
else:
   print('no match')
1 = ro.findall('09sistemi09informativi')
print(1)
```



```
import re
# Remove anything other than digits
phone = 'Phone: +39 0521 905708'
num = re.sub(r')D', "", phone)
print("Phone Num : ", num)
# Hide password
string = 'Password: mypwd'
hidden = re.sub(r'^(Password:\s^*).+\$', r'\1(******)', string)
print(hidden)
# part of a match
address = 'Please mail it to alberto.ferrari@unipr.it.com'
match = re.search(r'([\w.-]+)@([\w-]+).([\w]+)',address)
print(match.group(0))
print(match.group(1))
print (match.group (2))
                                                   https://docs.python.org/3/library/re.html
print(match.group(3))
```



```
import re
print('match only upper and lowercase letters, numbers, and underscores')
patterns = r'^[a-zA-Z0-9]*$'
text1 = 'sistemi informativi'
if re.search(patterns,text1):
   print('ok')
print('minimum eight characters, at least one uppercase letter, one lowercase letter and
one number')
patterns = r'^(?=.*[a-z])(?=.*[A-Z])(?=.*d)[a-zA-Zd]{8,}$'
pwd = input('password ')
while not re.search(patterns,pwd):
   pwd = input('password ')
```



```
def is alphanumeric(string):
    characherRegex = re.compile(r'[^a-zA-z0-9.]')
    string = characherRegex.search(string)
    return bool(string)
def is mailAddress(string):
    characherRegex = re.compile(r'^[ a-z0-9-]+(\.[ a-z0-9-]+)*@[a-z0-9-]+(\.[a-z0-9-]+)*(\.[a-z]{2,4})$')
    string = characherRegex.search(string)
    return bool(string)
def is CF(string):
    \# characherRegex = re.compile(r'^[a-z]{6}[0-9]{2}[a-z][0-9]{2}[a-z][0-9]{3}[a-z]$')
    characherRegex = re.compile(r'^[A-Z] \{6\})
                                r'[0-9LMNPQRSTUV]{2}[ABCDEHLMPRST]{1}[0-9LMNPQRSTUV]{2}'
                                r'[A-Z]{1}[0-9LMNPQRSTUV]{3}[A-Z]{1}$')
    string = characherRegex.search(string)
    return bool(string)
```



recursively enumerable

context-sensitive

context-free

regular

approccio generativo

grammatiche di Chomsky





- o grammatica $G = \langle V_T, V_N, P, S \rangle$
- \circ \mathbf{V}_{π} : alfabeto finito di simboli terminali
- \circ $\mathbf{V_N}$: alfabeto finito di simboli *non terminali*
 - o (variabili, categorie sintattiche)
- o P: insieme di produzioni
 - \circ < α , β > \in P si indica con $\alpha \rightarrow \beta$
 - \circ α \in V
 - \circ $\beta \in V$
- \circ S \in V_N : assioma
- L(G): insieme delle stringhe di simboli terminali ottenibili con finite operazioni di riscrittura





- · la radice dell'albero è l'*assioma*
- ogni nodo ha tanti *figli*, quante sono le *produzioni* applicabili
- ogni figlio è una forma di frase
- · le foglie sono stringhe del linguaggio
- primo esempio:

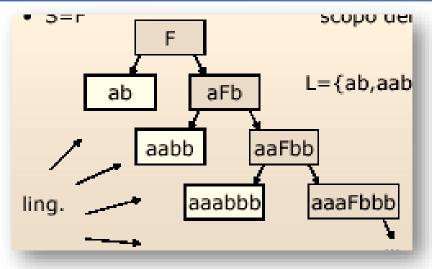
-
$$G_1 = \{a,b\}, \{F\}, P_1, F>$$

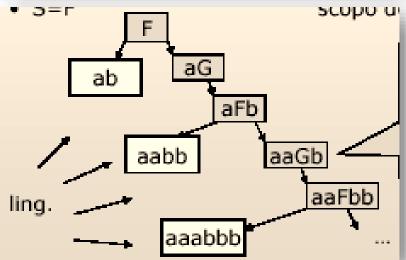
- $P_1 = \{F \rightarrow ab, F \rightarrow aFb\}$

secondo esempio:

$$- G_2 = \{a,b\}, \{F,G\}, P_2, F>$$

$$- P_2 = \{F \rightarrow ab, F \rightarrow aG, G \rightarrow Fb\}$$





esempio grammatica context-free

$$\circ$$
 G = $\{a, b, +, *, -, (,)\}, \{E, I\}, P, E>$

- $(1) \quad \mathbf{E} \quad \rightarrow \quad \mathbf{I}$
- $(2) E \rightarrow E + E$
- $(3) E \rightarrow E \star E$
- $(4) \quad \mathbf{E} \quad \rightarrow \quad \mathbf{E} \quad \quad \mathbf{E}$
- $(5) \quad \mathbf{E} \quad \rightarrow \quad -\mathbf{E}$
- $(6) \quad \mathbf{E} \quad \rightarrow \quad (\mathbf{E})$
- (7) $I \rightarrow a$
- (8) $I \rightarrow b$
- (9) $I \rightarrow Ia$
- (10) I \rightarrow Ia
- o ... genera espressioni aritmetiche con operatori +,*,- unario e binario e simboli a b

esempio grammatica contextsensitive

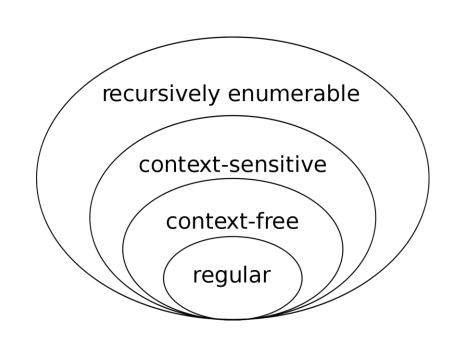
- \circ G = $\{a, b, c\}, \{S, B, C\}, P, S>$
 - (1) $S \rightarrow aSBC$
 - (2) $S \rightarrow aBC$
 - (3) $CB \rightarrow BC$
 - (4) $aB \rightarrow ab$
 - (5) $bB \rightarrow bb$
 - (6) $bC \rightarrow bc$
 - (7) $cC \rightarrow cc$
- o ... genera il linguaggio {aⁿbⁿcⁿ : n≥1}
- o per generare aaabbbccc:
 - o applicare 1-1-2-3-3-4-5-5-6-7-7



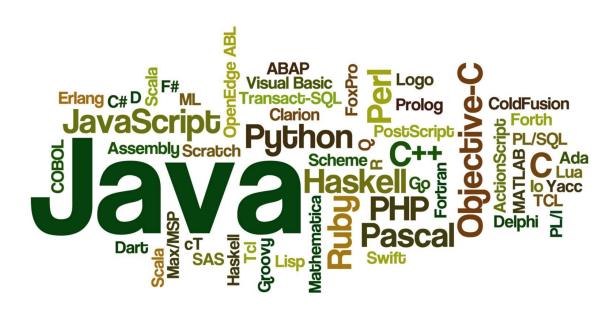


- o tipo 0: grammatiche ricorsivamente enumerabili (RE)
 - \circ $\alpha A\beta \rightarrow \gamma$ (non limitate)
- o tipo 1: grammatiche contestuali (CS)
 - $\bigcirc \quad \alpha A\beta \quad \rightarrow \quad \alpha \gamma \beta$
- o tipo 2: grammatiche non contestuali (CF)
 - \circ **A** \rightarrow γ
- o *tipo 3*: grammatiche *regolari (REG)*
 - \circ **A** \rightarrow **aB**, oppure **A** \rightarrow **b**, oppure **A** \rightarrow ϵ
 - o coincide con classe dei linguaggi definiti da regex

A, $B \in V_N$; a, $b \in V_T$; α , θ , $\gamma \in V^*$







linguaggi di programmazione

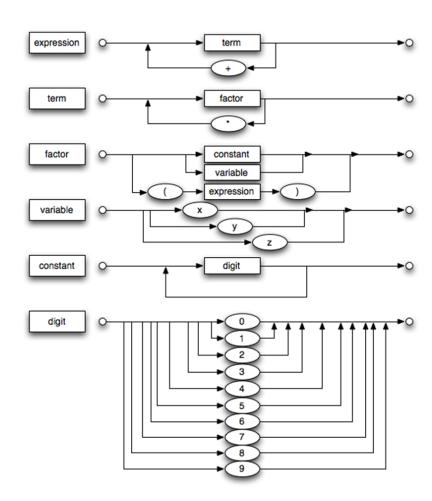


linguaggio di programmazione

- o notazione *formale* per definire algoritmi
 - o algoritmo: sequenza di istruzioni per risolvere un dato problema in un tempo finito
- o ogni linguaggio è caratterizzato da:
 - o sintassi
 - o semantica



- insieme di *regole formali* per scrivere *frasi ben formate* (*programmi*)
- · lessico
 - parole riservate, operatori, variabili,
 costanti ecc. (token)
- grammatiche non contestuali (...) espresse con notazioni formali:
 - Backus-Naur Form
 - Extended BNF
 - Diagrammi sintattici





- o attribuisce un **significato** alle frasi (sintatticamente corrette) costruite nel linguaggio
- o una frase può essere sintatticamente corretta ma non aver alcun significato
 - o soggetto predicato complemento
 - o "La mela mangia il bambino"
 - o "Il bambino mangia la mela"
- o avere un significato diverso da quello previsto...
 - \circ GREEK_PI = 345



semantica nei linguaggi di programmazione

o correttezza sui *tipi*

- o quali tipi di dato possono essere elaborati?
- o quali operatori applicabili ad ogni dato?
- o quali regole per definire nuovi tipi e operatori?

o semantica operazionale

- o qual è l'effetto di ogni azione elementare?
- o qual è l'effetto dell'aggregazione delle azioni?
 - o qual è l'effetto dell'esecuzione di un certo programma?





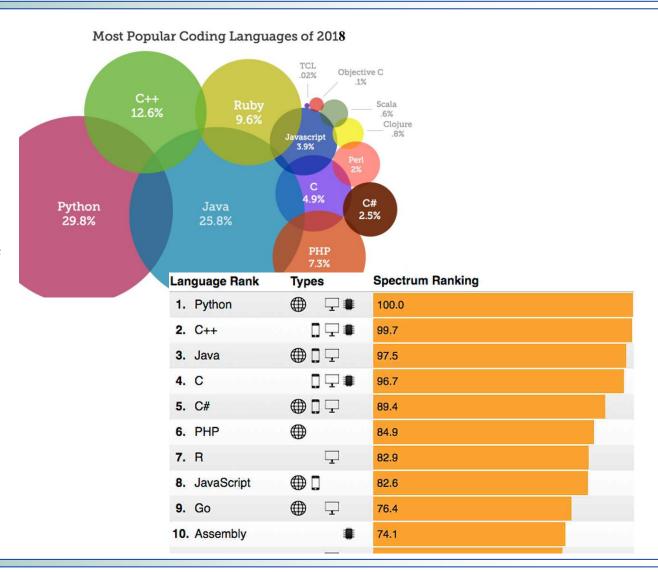
- più *orientati alla macchina* che ai problemi da trattare
- · linguaggi macchina
 - solo operazioni eseguibili direttamente dall'elaboratore
 - Op. molto elementari, diverse per ogni processore, in formato binario
- linguaggi assembly
 - prima evoluzione
 - codici binari → mnemonici

```
; Example of IBM PC assembly language
; Accepts a number in register AX;
  subtracts 32 if it is in the range 97-122;
; otherwise leaves it unchanged.
SUB32
      PROC
                    ; procedure begins here
                    ; compare AX to 97
       CMP
           AX,97
            DONE
                    ; if less, jump to DONE
           AX,122 ; compare AX to 122
            DONE
                    ; if greater, jump to DONE
       SUB
           AX,32
                    : subtract 32 from AX
DONE:
                    ; return to main program
      RET
SUB32
      ENDP
                    ; procedure ends here
             FIGURE 17. Assembly language
```



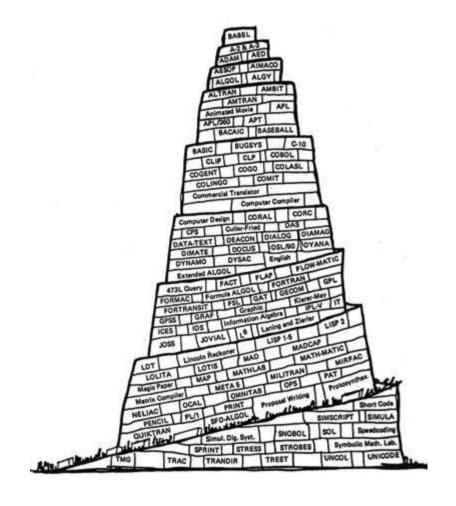
linguaggi di alto livello

- introdotti per *facilitare* la scrittura dei programmi
- definizione della soluzione in modo intuitivo
- astrazione rispetto al calcolatore su cui verranno eseguiti
- necessaria *traduzione* in linguaggio macchina



storia ed evoluzione dei linguaggi di programmazione

https://www.cs.toronto.edu/~gpenn/csc324/PLhistory.pdf
http://www.levenez.com/lang/history.html
http://www.cs.brown.edu/~adf/programming_languages.html





paradigmi di programmazione

- o definiscono la filosofia e la metodologia con cui si scrivono i programmi
- o definiscono il concetto (astratto) di computazione
- o ogni linguaggio consente (o spinge verso) l'adozione di un particolare paradigma
 - o imperativo/procedurale
 - o orientato agli oggetti
 - o **scripting** (tipizzazione dinamica, principio DRY Don't Repeat Yourself)
 - o funzionale (funzioni come "cittadini di prima classe")
 - o **logico** (base di conoscenza + regole di inferenza)



o imperativi/procedurali

o Cobol, Fortran, Algol, C, Pascal

o orientati agli oggetti

o Simula, Smalltalk, Eiffel, C++, Delphi, Java, C#, VB.NET

o scripting

o Basic, Perl, PHP, Javascript, Python

o funzionali

o Lisp, Scheme, ML, Haskell, Erlang

o logici

o Prolog...



- o il *paradigma logico* utilizza la *logica del primo ordine* (clausole di Horn) sia per rappresentare sia per elaborare l'informazione
 - o assiomi e regole di inferenza
- o si descrive la struttura logica del problema piuttosto che il modo di risolverlo
 - o focus sugli aspetti logici del problema e sul modo migliore per rappresentarli
- o **prolog** (**pro**grammation en **log**ique)
 - o ideato da Alain Colmerauer negli anni '70
 - o linguaggio di programmazione di tipo dichiarativo
 - \circ il programmatore dice cosa fare e $non\ come$ farlo



o un programma prolog è composto da

?- fratello sorella(X,giovanni)

o *fatti* (affermazioni sempre vere) padre(aldo,giovanni). padre (aldo, lucia). madre(giulia, giovanni). madre (giulia, lucia). o regole genitore(X,Y) := padre(X,Y). genitore(X,Y) :- madre(X,Y). fratello sorella(X,Y) :padre (Z,X), padre (Z,Y), madre (W,X), madre (W,Y), X = Y. o domande (goal) ?- padre(luca,pippo) -- YES

-- X=lucia



- o per soddisfare un **goal** (rispondere a una domanda)
 - o cerca un *fatto* che unifica il goal
 - o identificare una sostituzione delle variabili che rende identici il fatto e il goal
 - o se un fatto viene trovato, il goal è **soddisfatto**
 - o cerca una *regola* la cui testa (parte sinistra) unifica il goal
 - o se la trova, cerca di soddisfare tutti i sotto-goal di cui è composta la regola (parte destra)
 - o usa una stessa sostituzione delle variabili per tutti i sotto-goal

https://swish.swi-prolog.org/



example: mortal philosophers

```
consider the following argument:
```

all men are mortal.

Socrates is a man.

hence, Socrates is mortal.

it has two premises and a conclusion

the two premises can be expressed as a little prolog program:

```
mortal(X) :- man(X).
```

man(socrates).

the conclusion can then be formulated as a query:

```
?- mortal(socrates).
```

Yes



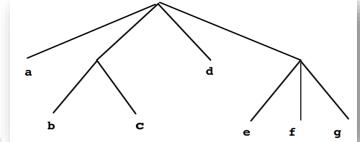
paradigma funzionale

- o paradigma di programmazione in cui l'operazione fondamentale è l'applicazione di funzioni ai loro argomenti
- o un programma è una funzione, definita in termini di *altre funzioni*, che elaborano l'input e restituiscono il risultato finale
- o le funzioni sono prive di effetti collaterali (no side effects)
 - $\circ \Rightarrow funzioni pure$
- o le funzioni sono trattate come un qualsiasi altro tipo di dato
 - $\circ \Rightarrow first\ class\ objects$
- o la *ricorsione* è la principale struttura di controllo
 - o non esistono costrutti iterativi
- o high order functions: funzioni che operano su funzioni, che operano su funzioni, ...



paradigma funzionale - lisp

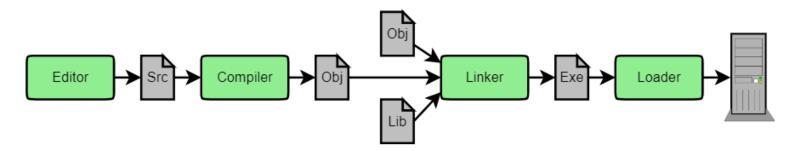
- o *lisp (list processing)* è un linguaggio di programmazione (*John McCarthy 1958*) per il trattamento di *liste* e, in generale, di informazione non numerica
 - o utilizzato in progetti di intelligenza artificiale
- o le liste in lisp sono in realtà alberi con l'informazione (i simboli atomici) memorizzata nelle foglie
 - o es. la lista (a (b c) d (e f g)) corrisponde all'albero in figura
- o un *programma* lisp è una *lista*
 - o è possibile scrivere programmi che ricevono in input altri programmi



funzione mylength che, data una lista lst, ne calcola il numero di elementi



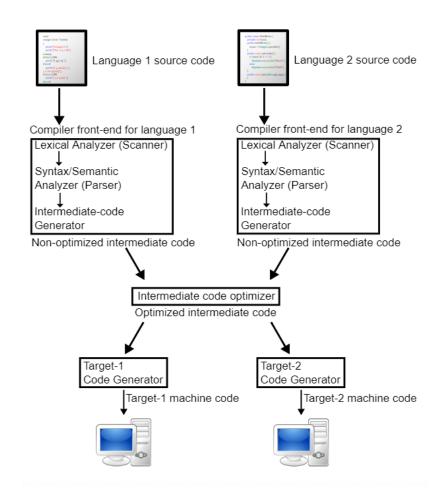
- o linguaggio ad alto livello → passi necessari:
 - o compilazione, traduzione in linguaggio macchina
 - o *collegamento* (*linking*) con librerie di supporto
 - o caricamento (loading) in memoria
- o programmi *compilati*: applicati i 3 passi...
 - o a tutto il codice; prima dell'esecuzione
- o programmi *interpretati*: applicati i 3 passi...
 - o in sequenza, *su ogni istruzione*; a tempo di esecuzione







- traduzione da linguaggio alto livello a linguaggio macchina
- analisi: lessicale, grammaticale, contestuale
- rappresentazione *intermedia*: albero sintattico annotato (AST)
- generazione codice oggetto
 - non ancora eseguibile
 - linker, loader

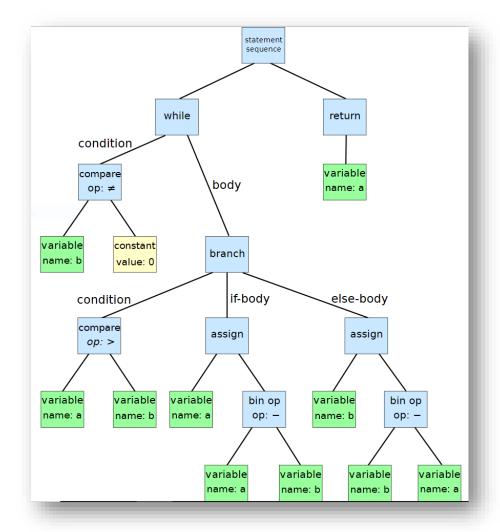






```
while b != 0:
    if a > b:
        a = a - b
    else:
        b = b - a
return a
```

Algoritmo di Euclide per MCD





- o il *linker* collega diversi moduli oggetto
 - o simboli irrisolti → riferimenti esterni
- o il collegamento può essere *statico* o *dinamico*
 - o collegamento **statico**
 - o libreria inclusa nel file oggetto, eseguibile stand-alone
 - o dimensioni maggiori, ma possibile includere solo funzionalità utilizzate
 - \circ collegamento dinamico
 - o librerie condivise da diverse applicazioni
 - o installazione ed aggiornamento unici
 - o caricate in memoria una sola volta





- o il *loader* carica in memoria un programma *rilocabile*
 - o risolti tutti gli *indirizzi* relativi (variabili, salti ecc.)
 - o caricati eventuali programmi di supporto
- o rilocazione *statica*
 - o indirizzi logici trasformati in *indirizzi assoluti*
- o rilocazione dinamica
 - o *indirizzi logici* mantenuti nel programma in esecuzione
 - o programma compilato: indirizzamento relativo
 - o tramite registro base: locazione in memoria del codice, dei dati e dello stack
 - Memory Management Unit



- o compilazione in *codice intermedio*
 - o Bytecode (Java), Common Intermediate Lang. (.NET), ...
 - o Python: compilato per una macchina virtuale (file .pyc), ma in modo trasparente
- o esecuzione su una *macchina virtuale*, che gestisce la memoria (garbage collection)
 - o Java Virtual Machine, Common Language Runtime, ...
 - o spesso compilazione "al volo" (Just In Time) in codice nativo