Παρουσίαση του εργαλείου

flex

γεννήτρια λεκτικών αναλυτών

για το μάθημα:

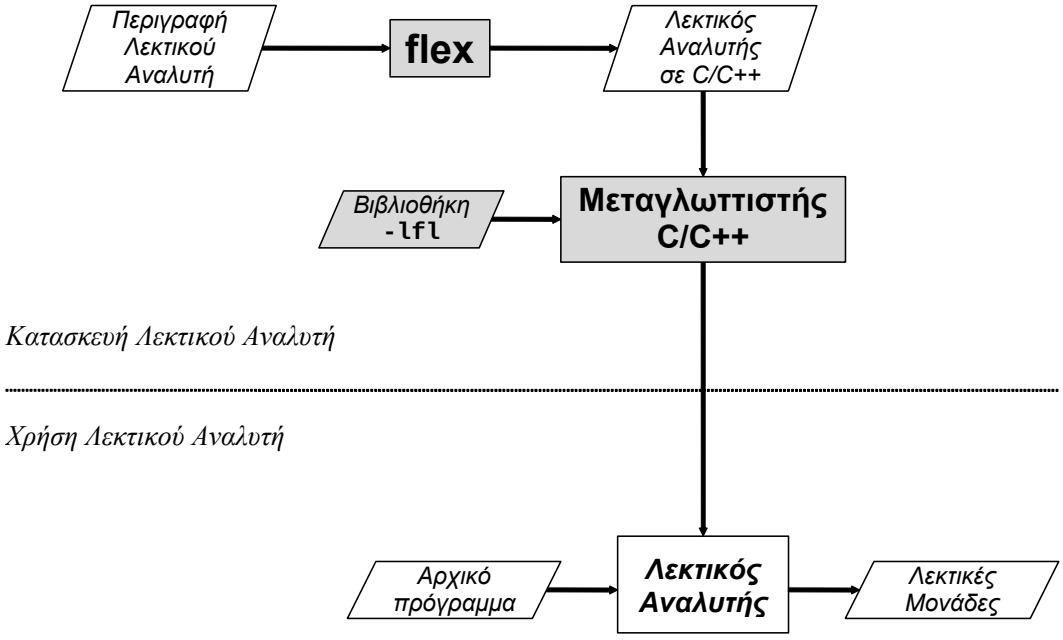
Μεταγλωττιστές

Χανιά, 2005

Χαρακτηριστικά του flex

- Γεννήτρια λεκτικών αναλυτών σε C/C++ (fast lexical analyzer generator).
- Βασισμένο στο εργαλείο του Unix lex.
- Σχετικά εύκολο στη χρήση.
- Υψηλές επιδόσεις.
- Μεγάλη ευελιξία και εκφραστικότητα.
- Συνεργασία με γεννήτριες συντακτικών αναλυτών.

Λειτουργία του flex



Μορφή αρχείου περιγραφής

Μέρος Α

%%

Μέρος Β

%%

Μέρος Γ

Περιγραφή μέρους Α

- Σχόλια ανάμεσα σε **/*** και ***/**
- Κώδικας C/C++ που θα ενσωματωθεί όπως είναι, ανάμεσα σε **%{** και **%}**, π.χ.

```
%{
    #define TK_EOF      0
    #define TK_ID       1

    int lineCount = 0;
%}
```

- Ορισμοί μνημονικών ονομάτων για οικογένειες χαρακτήρων ή κανονικές εκφράσεις, π.χ.

```
letter          [A-Za-z]
digit           [0-9]
identifier      {letter}({letter}|{digit})*
```

- Ορισμοί αρχικών καταστάσεων.
- Οδηγίες προς το ***flex*** , π.χ.

  %option noyywrap

Περιγραφή μέρους Β

- Περιέχει κανόνες της μορφής:

  κανονική-έκφραση ενέργεια

- Δεν μπορούν να παρεμβάλλονται σχόλια.
- Η σύνταξη είναι αυστηρή και μικρές απροσεξίες καταλήγουν σε διαφορετικά αποτελέσματα.
- Κώδικας C/C++ ανάμεσα σε **%{** και **%}** στην αρχή του μέρους Β χρησιμεύει για τη δήλωση τοπικών μεταβλητών στη συνάρτηση του λεκτικού αναλυτή yylex().
- Default κανόνας ισχύει ακόμη κι αν δεν υπάρχει κανείς αλλος.

Περιγραφή μέρους Γ

- Προαιρετικό (όπως και ο διαχωριστής **%%**).
- Περιέχει κώδικα C/C++ που ενσωματώνεται όπως είναι. Χρησιμεύει συνήθως για τον ορισμό βοηθητικών συναρτήσεων που πρέπει να καλούνται από το λεκτικό αναλυτή (στις ενέργειες του μέρους Β).

Κανονικές εκφράσεις – 1

Έκφραση	**Ταιριάζει με** *(matches)*
x	το χαρακτήρα **x**
\x	αν **x** είναι ένας από τους χαρακτήρες **a**, **b**, **f**, **n**, **r**, **t**, **v**, την ANSI ερμηνεία του, αλλιώς τον ίδιο το χαρακτήρα **x** (χρησιμοποιείται για τους ειδικούς χαρακτήρες)
\123	το χαρακτήρα με οκταδική τιμή **123**
\x2a	το χαρακτήρα με δεκαεξαδική τιμή **2a**
.	Οποιονδήποτε χαρακτήρα εκτός του (new line).
[xyz]	έναν από τους χαρακτήρες **x**, **y**, **z.** (κλάση χαρακτήρων)
[abm-sx]	έναν από τους χαρακτήρες **a**, **b**, **m** ως **s**, **x .**(εμβέλεια m –s)
[^A-Z]	Οποιονδήποτε χαρακτήρα εκτός από τους **A** ως **Z.** Συμπεριλαμβανομένου και του **\n.**
r*	καμιά ή περισσότερες εμφανίσεις της κανονικής έκφρασης **r**
r+	μια ή περισσότερες εμφανίσεις της κανονικής έκφρασης **r**
r?	καμιά ή μια εμφάνιση της κανονικής έκφρασης **r** (προαιρετικά **r**)
r{2,5}	2 μέχρι 5 εμφανίσεις της κανονικής έκφρασης **r**.
r{2,}	2 ή περισσότερες εμφανίσεις της κανονικής έκφρασης **r**.
r{2}	2 ακριβώς εμφανίσεις της κανονικής έκφρασης **r**.
{name}	την επέκταση του μνημονικού ονόματος **name** όπως αυτό ορίζεται στο μέρος Α.
"[ab]\"c"	τη συμβολοσειρά **[ab]"c**
(r)	την κανονική έκφραση **r** (παράκαμψη προτεραιότητας).

Κανονικές εκφράσεις – 2

Έκφραση	Ταιριάζει με *(matches)*
r|s	την κανονική έκφραση **r** ή την **s**
rs	Την κανονική έκφραση **r** ακολουθούμενη από την κανοκή έκφραση **s (**παράθεση).
r/s	την κανονική έκφραση **r** αλλά μόνο όταν ακολουθείται από την κανονική έκφραση **s** (η οποία δε διαβάζεται στη μεταβλητή **yytext**).
^r	την κανονική έκφραση **r** αλλά μόνο στην αρχή μιας γραμμής.
r$	την κανονική έκφραση **r** αλλά μόνο στο τέλος μιας γραμμής.
<s>r	την κανονική έκφραση **r** αλλά μόνο με αρχική κατάσταση **s.**
<s1,s2>r	την κανονική έκφραση **r** αλλά μόνο με αρχική κατάσταση **s1** ή **s2.**
<*>r	την κανονική έκφραση **r** με οποιαδήποτε αρχική κατάσταση.
<<EOF>>	το τέλος του αρχείου.
<s1,s2><<EOF>>	το τέλος του αρχείου αλλά μόνο με αρχική κατάσταση **s1** ή **s2.**

Ειδικές ενέργειες

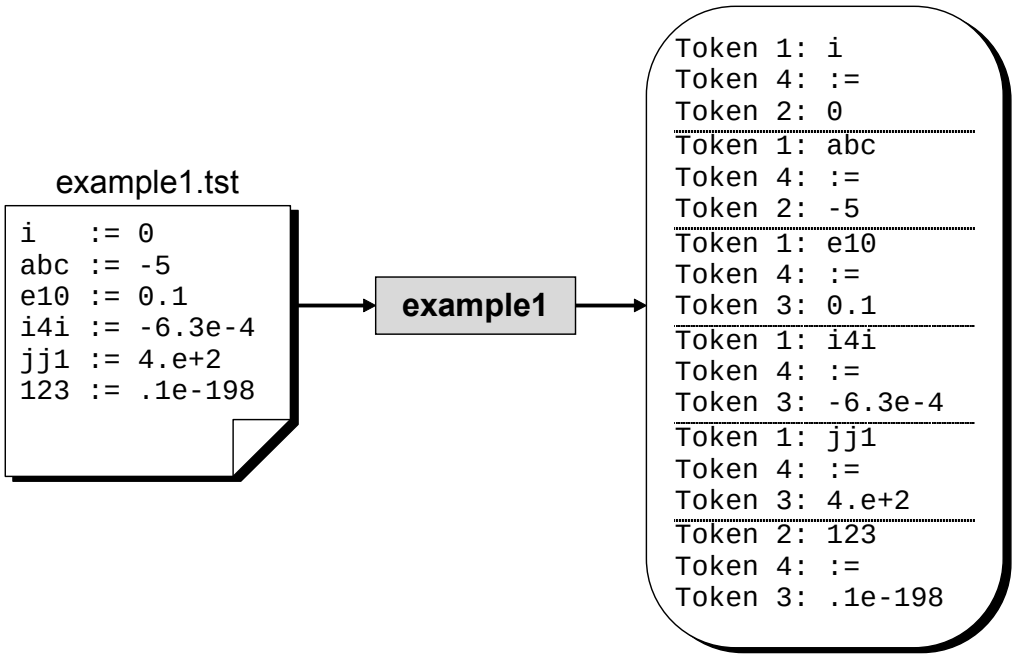
Ενέργεια	Σημασιολογία
ECHO	Εκτυπώνει το τμήμα του αρχείου εισόδου που αναγνωρίστηκε.
REJECT	Απορρίπτει την αναγνώριση που συνέβη και προχωράει στην αμέσως επόμενη δυνατή αναγνώριση.
|	Ίδια με την ενέργεια του επόμενου κανόνα (για αποφυγή επανάληψης).

Ένα απλό παράδειγμα

```
%{
  #define TK_EOF          0
  #define TK_ID           1
  #define TK_INT          2
  #define TK_FLOAT        3
  #define TK_ASSGN        4
%}

white               [ \n\t]
letter              [A-Za-z]
digit               [0-9]
dot                 [\.]
sign                [\+\-]
exp                 [Ee]{sign}?{digit}{1,3}

%%

{white}*                { /* nothing */    }

{letter}({letter}|{digit})*  { return TK_ID;    }

":="                    { return TK_ASSGN; }

{sign}?{digit}+             { return TK_INT;   }

{sign}?{digit}+{dot}{digit}*{exp}?   |

{sign}?{digit}*{dot}{digit}+{exp}?   {
  return TK_FLOAT;
}

%%

int main ()
{
  int token;

  while ((token = yylex()) != TK_EOF)
    printf("Token %d: %s\n", token, yytext);
}
```

Επίδειξη του παραδείγματος

```
> flex example1.l
> gcc -o example1 lex.yy.c -lfl
> example1 < example1.tst
```



Περισσότερες πληροφορίες και τεκμηρίωση του flex

Για περισσότερες πληροφορίες σχετικά με το flex χρήστης παραπέμπεται στο manual που βρίσκεται στην ιστοσελίδα του μαθήματος: www.ced.tuc.gr/~manolis/courses/compilers/top.html

Μεταβλητές και συναρτήσεις του παραγόμενου λεκτικού αναλυτή

Σύμβολο	Περιγραφή
int yylex()	Η κύρια συνάρτηση του λεκτικού αναλυτή που παράγεται βάσει του αρχείου περιγραφής. Επιστρέφει έναν ακέραιο αριθμό, που συνήθως αντιστοιχεί σε μια λεκτική μονάδα.
char * yytext	Περιέχει το τμήμα του αρχείου εισόδου που αναγνωρίστηκε κατά την τελευταία κλήση της **yylex**.
int yyleng	Περιέχει το πλήθος των χαρακτήρων που περιέχονται στο **yytext**.
void yymore()	Διατηρεί το **yytext** μεταξύ διαδοχικών κλήσεων της **yylex**.
void yyless(n) int n;	Διατηρεί στο **yytext** τους αρχικούς **n** χαρακτήρες και επιστρέφει τους υπόλοιπους στο αρχείο εισόδου.
FILE * yyin	Το τρέχον αρχείο εισόδου.
FILE * yyout	Το τρέχον αρχείο εξόδου.
void yyrestart(f) FILE * f;	Χρησιμοποιεί ως τρέχον αρχείο εισόδου το **f**.
int input()	Διαβάζει ένα χαρακτήρα από το τρέχον αρχείο εισόδου.
void unput(c) char c;	Επιστρέφει το χαρακτήρα **c** στο τρέχον αρχείο εισόδου.
Macro **yyterminate()**	Τερματίζει τη λειτουργία της συνάρ-τησης **yylex** επιστρέφοντας **0**.
int yywrap()	Καλείται κάθε φορά που συναντάται τέλος αρχείου. Αν επιστραφεί **1**, η λεκτική ανάλυση τερματίζεται, ενώ αν επιστραφεί **0** συνεχίζει.

Αρχικές καταστάσεις

- Επιτρέπουν τον ορισμό ειδικής συμπεριφοράς, ανάλογα με το τί έχει προηγηθεί κατά την αναγνώριση του αρχείου εισόδου.
- Διακρίνονται σε ***εγκλειστικές*** (inclusive) και ***αποκλειστικές*** (exclusive).
- Δηλώνονται στο μέρος Α, με τις οδηγίες:

```
%s INCLUSIVE
%x EXCLUSIVE
```

- Στην πραγματικότητα δεν πρόκειται για ονόματα αλλά για macros που αντιστοιχούν σε ακέραιες τιμές.
- Η αρχική κατάσταση ονομάζεται **INITIAL** και αντιστοιχεί στην τιμή **0**.
- Η τρέχουσα κατάσταση είναι γνωστή μέσω του macro **YY_START**.
- Η τρέχουσα κατάσταση μπορεί να αλλάξει με την εκτέλεση του macro **BEGIN(new)**.
- Κανόνες που εξαρτώνται από την τρέχουσα κατάσταση:

```
<SPECIAL>[A-Za-z]+ { ECHO; }
```

- Διαφορά εγκλειστικών–αποκλειστικών καταστάσεων.