Project Report: Flex/Bison 2025

Καρκάνη Μαρία, ΑΜ1108398, 2° έτος, <u>st1108398@ceid.upatras.gr</u>
Παπανικολάου Παναγιώτης Αντώνης, ΑΜ1108379, 2° έτος, st1108379@ceid.upatras.gr
Μπακλαβάς Λουκάς, ΑΜ1108340, 2° έτος, <u>st1108340@ceid.upatras.gr</u>

Η εργασία είχε ως στόχο την δημιουργία ενός lexer/parser για την γλώσσα MyHTML.

Δομή των αρχείων:

```
project_directory

CMakeLists.txt

examples (directory)

grammar.bnf

project_report.pdf

src (directory)

lexer.l

parser.y

main.c

ckecks.c

checks.h
```

Επεξήγηση αρχείων:

grammar.bnf: περιέχει την γραμματική σε BNF μορφή.

Examples (directory): Παραδείγματα για τον έλεγχο των ερωτημάτων.

CMakeLists.txt: Για το build system χρησιμοποιήθηκε το cmake, το οποίο ενώνει τα αρχεία flex/bison, το main.c, καθώς και τα checks.c, checks.h ως static library. Με αυτόν τον τρόπο το build του προγράμματος γίνεται κατευθείαν από το make, χωρίς περεταίρω εντολές όπως bison ..., flex ..., gcc ...

Οδηγίες για το build του project υπάρχουν παρακάτω.

lexer.l: περιέχει τον λεκτικό αναλυτή. Η δομή του είναι σχετικά απλή, καθώς το μόνο που κάνει είναι να μετατρέπει μία λέξη σε ένα token και να τα στέλνει στον parser.

parser.y: περιέχει τον συντακτικό αναλυτή. Ο σκοπός του είναι να δέχεται τα tokens από τον λεκτικό αναλυτή και να ελέγχει την ορθότητα της σύνταξης. Οι κανόνες που ορίζονται σε αυτόν περιέχουν κώδικα C για περεταίρω ελέγχους.

Περισσότερα για τον συντακτικό αναλυτή αναφέρονται παρακάτω.

main.c: Περιέχει την main() συνάρτηση, η οποία δέχεται ως είσοδο το όνομα του MyHTML αρχείου. Στη συνέχεια ανοίγει το αρχείο, το εκτυπώνει στην οθόνη, επιστρέφει το input στην αρχή του αρχείου και καλεί τον parser. Αφού αναλυθεί το αρχείο, καλείται η συνάρτηση show errors() για την εμφάνιση των προβλημάτων που μπορεί να υπήρξαν.

rewind(input);
yyrestart(input);

Επιστροφή του input στην αρχή του αρχείου

checks.c: αποτελεί αρχείο πηγαίου κώδικα για την βιβλιοθήκη που υλοποιεί τους ελέγχους για τα ερωτήματα 2 και 3.

Περισσότερες λεπτομέρειες υπάρχουν παρακάτω.

checks.h: αποτελεί αρχείο header για την βιβλιοθήκη checks.

Οδηγίες building:

Στο directory του project:

1. mkdir build (δημιουργία build directory)

2. cmake .. (εκτέλεση cmake)

3. make (compile το πρόγραμμα)

Χρήση του προγράμματος:

Στο directory του project:

1. ./myhtml example.myhtml (εκτέλεση του προγράμματος με το όνομα του αρχείου που θα αναλυθεί)

Σημείωση: Το αρχείο εισόδου θα πρέπει να βρίσκεται στο examples directory

Σημείωση: Τα αρχεία C που παράγονται από τα flex και bison θα βρίσκονται στο build directory

Έξοδος του προγράμματος:

Το πρόγραμμα έχει ως έξοδο το αρχείο που παίρνει ως είσοδο, καθώς και τυχόν προβλήματα.

Τα προβλήματα μπορεί να είναι συντακτικού τύπου, δηλαδή να προέρχονται από τον parser, ή προβλήματα που περιγράφονται στα ερωτήματα 2 και 3 και υλοποιούνται σε C.

Screenshot παραδειγμάτων εφαρμογής για κάθε ερώτημα υπάρχουν παρακάτω.

Επιτυχής ανάλυση του προγράμματος

```
====== INPUT ======
<MYHTML>
<head>
    <title>Test Title</title>
    <!--this is a comment-->
    <meta charset="test">
    <meta name="something" content="something">
</head>
<body>
    <form id="f1" style="color:black" input_count=2>
       <!--this is a comment-->
       <input id="i1" type="text" value="something">
       <input id="i2" type="checkbox" value="something">
       <input id="i5" type="checkbox" value="something">
       <input id="i3" type="submit" value="something">
       <label for="i1"></label>
       <label for="i2"></label>
       <!--this is a comment-->
    </form>
    <!--this is a comment-->
    <!--this is a comment-->
   Hello
    <div style="color:red; font_size:12px;">
       <a id="a1" href="./something.com">
           An image
           <img src="http://image.png" alt="example">
       </a>
       </div>
</body>
</MYHTML>
_____
Program has been parsed successfully%
```

```
<MYHTML>
<head>
    <title>Test Title</title>
    <!--this is a comment-->
    <meta charset="test">
    <meta name="something" content="something">
</head>
<body>
     <form id="f1" style="color:black" input_count=1>
         <!--this is a comment-->
         <input id="i2" type="text" value="something">
         <input id="12" type="cekt" value="something">
<input id="15" type="checkbox" value="something">
<input id="i5" type="checkbox" value="something">
<input id="i3" type="submit" value="something">
         <label for="i1"></label>
         <label for="i2"></label>
         <!--this is a comment-->
    </form>
    <!--this is a comment-->
    <!--this is a comment-->
    Hello
    An image
            <img src="http://image.png" alt="example">
         </div>
</body>
</MYHTML>
ERROR: Input checkbox count does not match the input_count attribute at line 20
ERROR: For label is not valid at line 16
ERROR: ID is not unique at line 13
```

Αποτυχημένη ανάλυση του προγράμματος με ένδειξη του προβλήματος και της γραμμής στην οποία βρίσκεται

Περιγραφή του lexer:

Ο σκοπός του lexer είναι να δημιουργήσει tokens από τις λέξεις του αρχείου εισόδου.

Αυτό γίνεται εφικτό μέσω των κανόνων που ορίζονται στο .l αρχείο.

Η παραπάνω εντολή επιστρέφει ένα "START_HEAD" token κάθε φορά που η λέξη "<head>" εμφανίζεται στο αρχείο.

Σημείωση: Το macro debug_token(tok) χρησιμοποιήθηκε για λόγους debugging (δήλωνε μία συνάρτηση printf()).

Ιδιαίτερη προσοχή ίσως θέλει το comment token, το οποίο λέει στον lexer να μπει σε κατάσταση comment κάθε φορά που το συναντάει:

Υλοποιήθηκε με αυτόν τον τρόπο ώστε ο κανόνας για το comment_body να μην επεκτείνεται στους υπόλοιπους κανόνες.

Μερικά regular expressions χρησιμοποιήθηκαν για τον ορισμό κειμένου, κειμένου που περικλείεται από "", ''.

Για ταίριασμα κειμένου:

```
TEXTDATA [a-zA-Z][^"<>\n=]+
```

Το κείμενο θα πρέπει να ξεκινάει με ένα γράμμα (κεφαλαίο ή πεζό) και συνεχίζεται με έναν ή περισσότερους χαρακτήρες, εκτός από ", <>, \n (νέα σειρά), =.

Για ταίριασμα κειμένου μέσα σε σχόλιο:

```
COMMENT_BODY ([^-]|(-[^-]))*
```

Ταιριάζει κείμενο χωρίς τον χαρακτήρα "-" ή κείμενο με "-" χωρίς να συνεχίζεται ξανά από χαρακτήρα "-".

Περιγραφή του parser:

O parser ξεκινάει με τον ορισμό των tokens τα οποία θα δεχτεί. Μερικά από αυτά έχουν και τύπο δεδομένου, ο οποίος ορίζει τον τύπο που θα επιστρέφει το token.

π.χ.

```
%token <str> QUOTED_TEXT TEXT
%token <intval> POSITIVE_INT
```

Ακολουθεί ο ορισμός των κανόνων με την γενική μορφή myhtml, η οποία αποτελείται από τα στοιχεία START_MYHTML, head_opt, body και END_MYHTML. Αυτά επεκτείνονται στη συνέχεια, ορίζοντας την γραμματική της γλώσσας.

Σε αρκετά σημεία των κανόνων υπάρχει κώδικας C με συναρτήσεις από το αρχείο checks, όπως:

check_id(): έλεγχος αν το id έχει χρησιμοποιηθεί

is url(): έλεγχος αν το string είναι url

is_valid_href(): έλεγχος του format του χαρακτηριστικού href

is_valid_style() : έλεγχος αν η τιμή του χαρακτηριστικού style ακολουθεί τους κανόνες του ερωτήματος 2.g

type_is_valid(): έλεγχος του format του χαρακτηριστικού type

Αν κάποιος έλεγχος από τους παραπάνω δεν είναι επιτυχής, ένα error προστίθεται στην στοίβα των error μέσω της συνάρτησης

void add_error(int line_number, err_type_t err);

Οι παράμετροι της συνάρτησης αποτελούνται από την γραμμή στην οποία συναντήθηκε το error, καθώς και τον τύπο του προβλήματος που υπήρξε.

Η γραμμή στην οποία συναντήθηκε το πρόβλημα μετριέται από τον lexer, ο οποίος αυξάνει την μεταβλητή line_number κάθε φορά που συναντά τον χαρακτήρα \n (new line).

Για τον ορισμό ενός η παραπάνω στοιχείου χρησιμοποιήθηκε left recursion:

```
div_content_list:
    /*nothing*/
    | div_content_list body_element
;
```

Περιγραφή του checks.c:

Το αρχείο checks.c περιέχει τις συναρτήσεις για την υλοποίηση των ελέγχων των ερωτημάτων 2, 3, καθώς και για την διαχείρηση των errors.

Για την αποθήκευση των errors χρησιμοποιήθηκε ένα array από error_t, το οποίο το πρόγραμμα το διαχειρίζεται σαν στοίβα.

To error_t είναι τύπος που περιέχει έναν ακέραιο αριθμό για την γραμμή που εμφανίστηκε το error, καθώς και ένα err_type_t το οποίο είναι ένα enum με τα πιθανά προβλήματα.

```
typedef enum
{
    title_err,
    id_err,
    href_err,
    src_err,
    type_err,
    for_err,
    style_err,
    input_count_err,
    input_count_used_err
} err_type_t;

typedef struct
{
    int line;
    err_type_t type;
} error_t;
```

Ορισμός των τύπων.

Για την αποθήκευση των id χρησιμοποιήθηκαν δύο 2D πίνακες από χαρακτήρες. Ο ένας περιέχει όλα τα id του προγράμματος και ο άλλος τα id που έχουν χρησιμοποιηθεί. Ο διαχωρισμός αυτός έγινε ώστε να εξασφαλιστεί η σύνδεση ενός στοιχείου label με ένα μοναδικό στοιχείο input μέσω του id.

Οι δομές για την αποθήκευση των παραπάνω στοιχείων, καθώς και οι μεταβλητές για την καταμέτρηση των id, των error κλπ. ορίζονται ως static ώστε να είναι προσβάσιμες μόνο μέσα στο αρχείο checks.c. Με αυτόν τον τρόπο, το check.c αρχείο παρέχει στον parser μόνο ένα σύνολο συναρτήσεων για την υλοποίηση των ελέγχων.

Screenshots παραδειγμάτων εφαρμογής:

```
in build — -zsh — 117×48
 loukas@MacBookPro build % ./myhtml example_1.myhtml
  ====== INPUT ==
1: <MYHTML>
2: head
3: <titl
4: <!--t
5: <meta
           <title>Test Title</title>
             <!--this is a comment-->
           <meta charset="test">
<meta name="something" content="something">
6: <met
7: </head>
8: <body>
9: <fc
10: 11:
12: 13: 14: 15: 16: 17: 18: 19: 28: </td>

10: 4: 4: 4: 22: <br/>22: <br/>23: 24: <br/>25: Hee 26: 

26: 4: 4: 25: Hee 26: <br/>27: <br/>28: 29: 30: 

             <form id="f1" style="color:black" input_count=2>
                    <!--this is a comment-->
<input id="i1" type="text" value="something">
<input id="i2" type="checkbox" value="something">
<input id="i5" type="checkbox" value="something">
<input id="i5" type="submit" value="something">
<label for="i1"></label>
<label for="i2"></label>
<!--this is a comment-->
              </form>
              <!--this is a comment-->

              <!--this is a comment-->
              Hello
              An image
                            <img src="http://image.png" alt="example">
31:
32:
                     </a>

</div>
 33:
 34:
 35: </body>
 36: </MYHTML>
 37:
 ERROR: Syntax is not correct at line 2 loukas@MacBookPro build %
```

Ερώτημα 1:

Η σύνταξη δεν είναι σωστή.

```
build — -zsh — 117×48
         loukas@MacBookPro build % ./myhtml example_2a.myhtml
                     ===== INPUT =:
       1: <MYHTML>
2: <head>
3: <tit!
4: <!--t
                                      <!--this is a comment-->
                                         <meta charset="test">
6: <met
7: </head>
8: <body>
9: <for
10:
11:
12:
13:
14:
15:
16:
17:
18:
19:
20: </h>
21: <!
22: <p
23:
24: <!
25: He
26: <,
27: <6
28:
29:
30:
31:
32:
                                        <meta name="something" content="something">
                                        <form id="f1" style="color:black" input_count=2>
                                                           <!--this is a comment-->
<input id="i1" type="text" value="something">
<input id="i2" type="checkbox" value="something">
<input id="i5" type="checkbox" value="something">
<input id="i3" type="submit" value="something">
<label for="i1"></label>
<label for="i2"></label></label></label></label>
                                                              <!--this is a comment-->
                                           </form>
<!--this is a comment-->
cp id="p1" style="font_size:12%">
                                           <!--this is a comment-->

<div style="color:red; font_size:12px;">
<a id="a1" href="./something.com">
<a id
                                                                              An image
                                                                              <img src="http://image.png" alt="example">
                                                             33:
                                             </div>
         34:
        35: </body>
36: </MYHTML>
          37:
        ERROR: Title contains more than 60 characters at line 3 loukas@MacBookPro build %
```

Ερώτημα 2.a:

Ο τίτλος περιέχει πάνω από 60 χαρακτήρες

```
in build — -zsh — 117×48
loukas@MacBookPro build % ./myhtml example_2b.myhtml
 ====== INPUT ==
1: <MYHTML>
2: <head>
3: <tit
          <title>Test Title</title>
4:
5:
          <!--this is a comment-->
<meta charset="test">
          <meta name="something" content="something">
7: </head>
7: </head>
8: <body>
9: <fc
10:
11:
12:
13:
           <form id="f1" style="color:black" input_count=2>
                  <!--this is a comment-->
<input id="i1" type="text" value="something">
<input id="i2" type="checkbox" value="something">
<input id="i5" type="checkbox" value="something">
<input id="i5" type="submit" value="something">
                  <label for="i1"></label>
<label for="i2"></label>
<!--this is a comment-->
16:
17:
18:
19:
20:
21:
22:
23:
24:
25:
26:
27:
28:
29:
30:
            </form>
            <!--this is a comment-->
            <!--this is a comment-->
            Hello
            An image
                        <img src="http://image.png" alt="example">
31:
32:
33:

</div>
34:
35: </body>
36: </MYHTML>
ERROR: ID is not unique at line 15 loukas@MacBookPro build %
```

Ερώτημα 2.b:

Δύο id είναι ίδια οπότε εμφανίζεται ανάλογο μήνυμα.

Ερώτημα 2.c:

Το χαρακτηριστικό href δεν ακολουθεί την μορφή url, ούτε τη μορφή #element_id.

Ο επιτυχής έλεγχος της μορφής #element_id φαίνεται στην γραμμή 32.

```
build — -zsh — 117×48
loukas@MacBookPro build % ./myhtml example_2d.myhtml
 ===== INPUT ====
1: <MYHTML>
2: <head>
          <title>Test Title</title>
3: <ii
4: <!--
5: <me*
6: <me*
7: </head>
8: <body>
9: <fo:
10:
11:
12:
13:
14:
15:
16:
17:
          <!--this is a comment-->
<meta charset="test">
<meta name="something" content="something">
           <form id="f1" style="color:black" input_count=2>
                  <!--this is a comment-->
<input id="i1" type="text" value="something">
<input id="i2" type="checkbox" value="something">
<input id="i5" type="checkbox" value="something">
<input id="i3" type="submit" value="something">
<label for="i1"></label>
<label for="i2"></label></label>
17:
18:
19:
20:
21:
22:
23:
24:
25:
26:
27:
28:
29:
30:
31:
                   <!--this is a comment-->
             </form>
             <!--this is a comment-->
cp id="p1" style="font_size:12%">
             <!--this is a comment-->
             Hello
             An image
                         <img src="not_a_url" alt="example">
32:
                   </div>
34:
35: </body>
36: </MYHTML>
37:
 ERROR: SRC value is not valid at line 30
 loukas@MacBookPro build %
```

Ερώτημα 2.d:

Το χαρακτηριστικό src δεν ακολουθεί την μορφή url.

```
uild — -zsh — 117×48
 loukas@MacBookPro build % ./myhtml example_2e.myhtml
  ====== INPUT ===
 1: <MYHTML>
1: <MYHTML:
2: <head>
3: <ti'
4: <!--
5: <me*
6: <me*
7: </head>
             <title>Test Title</title>
             <!--this is a comment--> <meta charset="test">
             <meta name="something" content="something">
7: </heads
8: <body>
9: <fc
10:
11:
12:
13:
14:
15:
16:
17:
18:
19:
              <form id="f1" style="color:black" input_count=2>
                     <!--this is a comment-->
<input id="i1" type="text" value="something">
<input id="i2" type="checkbox" value="something">
<input id="i3" type="submit" value="something">
<input id="i6" type="checkbox" value="something">
<input id="i6" type="checkbox" value="something">
<input id="i4" type="something_else" value="something">
<input id="i4" type="something_else" value="something">
                      <label for="i1"></label>
                      <label for="i2"></label>
 19:
20:
21:
22:
23:
24:
25:
26:
27:
28:
29:
30:
31:
32:
                      <!--this is a comment-->
               <!--this is a comment-->

               <!--this is a comment-->
               Hello
               An image
                             <img src="http://image.png" alt="example">
                      </a>
 33:
                      34:
 35:
 36: </body>
 37: </MYHTML>
 38:
 ERROR: Type attribute is not valid at line 16
ERROR: Type attribute is not valid at line 15
loukas@MacBookPro build %
```

Ερώτημα 2.e:

Το στοιχείο τύπου submit δεν είναι τελευταίο.

Το στοιχείο input στην γραμμή 16 δεν έχει επιτρεπτή τιμή.

```
uild — -zsh — 117×48
 loukas@MacBookPro build % ./myhtml example_2f.myhtml ======= INPUT ======
 1: <MYHTML>
 2: <head>
 3:
            <title>Test Title</title>
4:
5:
             <!--this is a comment-
             <meta charset="test">
5: <meta charset="test">
6: <meta name="something" content="something">
7: </head>
8: <body>
9: <form id="f1" style="color:black" input_count=2>
10:
11: <!--this is a comment-->
12: <input id="i1" type="text" value="something"
13: <input id="i2" type="checkbox" value="something"
14: <input id="i3" type="checkbox" value="something"
15: <input id="i3" type="submit" value="something"
                     <!--this is a comment-->
<input id="i1" type="text" value="something">
<input id="i2" type="checkbox" value="something">
<input id="i5" type="checkbox" value="something">
<input id="i3" type="submit" value="something">
 15:
16:
17:
18:
19:
20:
                     <label for="i1"></label>
                     <label for="i10"></label>
                     <label for="i1"></label>
                     <!--this is a comment-->
 21:
              </form>
22:
23:
24:
25:
26:
27:
28:
29:
30:
              <!--this is a comment-->
               <!--this is a comment-->
              Hello
              An image
 31:
                            <img src="http://image.png" alt="example">
 32:
 33:
                     34:
               </div>
35:
36: </body>
 37: </MYHTML>
 38:
 ERROR: For label is not valid at line 18 ERROR: For label is not valid at line 17 loukas@MacBookPro build %
```

Ερώτημα 2.f:

Το χαρακτηριστικό for στην γραμμή 18 συνδέεται σε input που έχει συνδεθεί ήδη ένα άλλο for χαρακτηριστικό.

Το χαρακτηριστικό for στην γραμμή 17 δεν αντιστοιχεί σε κάποιο id στοιχείου input.

```
build — -zsh — 117×48
loukas@MacBookPro build % ./myhtml example_2g.myhtml
====== INPUT ======
1: <MYHTML>
2: <head>
       <title>Test Title</title>
4:
5:
        <!--this is a comment-
        <meta charset="test">
        <meta name="something" content="something">
6: <me1
7: </head>
8: <body>
9: <fo1
10:
11:
        <form id="f1" style="color:black" input_count=2>
             <!--this is a comment-->
<input id="i1" type="text" value="something">
<input id="i2" type="checkbox" value="something">
13:
14:
              <input id="i5" type="checkbox" value="something">
15:
              <input id="i3" type="submit" value="something">
              <label for="i1"></label>
16:
17:
18:
19:
20:
21:
              <label for="i2"></label>
              <!--this is a comment-->
         </form>
         <!--this is a comment-->
22:
         24:
         <!--this is a comment-->
25:
         Hello
26:
27:
28:
29:
30:
         An image
                  <img src="http://image.png" alt="example">
31:
              </a>
32:
              33:
         </div>
34:
35:
         <div style="something_else:red; font_size:12px;">
37: </body>
38: </MYHTML>
39:
ERROR: Style properties are not valid at line 35 ERROR: Style properties are not valid at line 27 loukas@MacBookPro build % ■
```

Ερώτημα 2.g:

Η τιμή του χαρακτηριστικού style στην γραμμή 27 δεν έχει επιτρεπόμενο value.

Η τιμή του χαρακτηριστικού style στην γραμμή 35 δεν έχει επιτρεπόμενο property.

```
in build — -zsh — 117×48
loukas@MacBookPro build % ./myhtml example_3.myhtml
 ====== INPUT =:
1: <MYHTML>
2: <head>
3: <t:
4: <!-
          <title>Test Title</title>
          <!--this is a comment-->
<meta charset="test">
6: <met
7: </head>
8: <body>
9: <for
          <meta name="something" content="something">
          <form id="f1" style="color:black" input_count=3>
                 <!--this is a comment-->
<input id="i1" type="text" value="something">
<input id="i2" type="checkbox" value="something">
<input id="i5" type="checkbox" value="something">
<input id="i3" type="submit" value="something">
<label for="i1"></label>
</abel>
11:
12:
14:
15:
16:
17:
18:
                  <label for="i2"></label>
                  <!--this is a comment-->
19:
20:
21:
22:
23:
24:
25:
26:
27:
28:
29:
30:
31:
32:
33:
            </form>
            <!--this is a comment-->
cp id="p1" style="font_size:12%">
            <!--this is a comment-->
            Hello
            An image
                       <img src="http://image.png" alt="example">
                  </a>
                 35: </body>
36: </MYHTML>
37:
ERROR: Input checkbox count does not match the input_count attribute at line 20 loukas@MacBookPro build %
```

Ερώτημα 3:

Η τιμή του χαρακτηριστικού input_count δεν αντιστοιχεί στο ακριβές πλήθος των στοιχείων input τύπου checkbox

```
uild — -zsh — 117×48
 loukas@MacBookPro build % ./myhtml example_3_2.myhtml
 ====== INPUT ===
 1: <MYHTML>
3: <ti1
4: <!--
5: <met
7: </head>
8: <body>
9: <for
10:
11:
12:
13:
14:
15:
16:
17:
18:
19: </r
20: <!
21: <p>22:
23: <!
24: Hei
25: </i>26: <d:
27:
28:
29:
30:
          <title>Test Title</title>
         <!--this is a comment-->
<meta charset="test">
<meta name="something" content="something">
          <form id="f1" style="color:black" input_count=2>
                 <!--this is a comment-->
<input id="i1" type="text" value="something">
                <input id="i3" type="submit" value="something">
<label for="i1"></label>
<label for="i3"></label>
                 <!--this is a comment-->
           <!--this is a comment-->

            <!--this is a comment-->
            Hello
           An image
                      <img src="http://image.png" alt="example">
 31:
                 32:
            </div>
 34: </body>
35: </MYHTML>
 36:
 ERROR: No input checkbox types exist so no input_count attribute should be used at line 19 loukas@MacBookPro build %
```

Ερώτημα 3:

Η χρήση του χαρακτηριστικού input count δεν επιτρέπεται, αφού δεν υπάρχει κανένα στοιχείο input τύπου checkbox.