ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Αναφορά Εξαμηνιαίας Εργασίας

στο μάθημα 8ου εξαμήνου (ροής Λ)

Μεταγλωττιστές

Θέμα εργασίας: “*Ανάπτυξη ενός μεταγλωττιστή για τη γλώσσα προγραμματισμού Alan*”

Στοιχεία ομάδας

Ονοματεπώνυμο:	ΜΑΘΙΟΥΛΑΚΗ ΕΛΕΝΗ
Α.Μ.:	03114040
E-mail:	el.mathioulaki@gmail.com
Ονοματεπώνυμο:	ΝΙΑΡΧΟΣ ΣΩΤΗΡΙΟΣ
Α.Μ.:	03114076
E-mail:	sot.niarchos@gmail.com

Εισαγωγή

Αντικείμενο της εργασίας μας είναι η ανάπτυξη ενός μεταγλωττιστή για τη γλώσσα προγραμματισμού Alan, μία απλή προστακτική γλώσσα προγραμματισμού που περιγράφεται αναλυτικά στην εκφώνηση της εργασίας.

Η παρούσα αναφορά χωρίζεται σε αρκετά μέρη, ένα για κάθε τμήμα του μεταγλωττιστή. Συγκεκριμένα, η παρουσίαση θα γίνει ως εξής:

1. Εποπτεία της συνολικής αρχιτεκτονικής του μεταγλωττιστή
2. Λεκτικός αναλυτής (lexer)
3. Συντακτικός αναλυτής (parser) και κατασκευή αφαιρετικού συντακτικού δέντρου (AST)
4. Σημασιολογικός αναλυτής
5. Παραγωγή ενδιάμεσου κώδικα και βελτιστοποίηση
6. Παραγωγή τελικού κώδικα και βελτιστοποίηση
7. alanc – ο τελικός μεταγλωττιστής

Οι τεχνολογίες που χρησιμοποιήθηκαν σε κάθε τμήμα του μεταγλωττιστή είναι οι εξής:

- Για τον λεκτικό αναλυτή χρησιμοποιήθηκε το εργαλείο flex (έκδοση 2.6.0)
- Για τον συντακτικό αναλυτή χρησιμοποιήθηκε το εργαλείο bison (έκδοση 3.0.4)
- Ο σημασιολογικός αναλυτής και η κατασκευή του AST γράφτηκαν σε C++14
- Η παραγωγή ενδιάμεσου κώδικα έγινε σε C++14 με χρήση των αντίστοιχων LLVM bindings (ο μεταγλωττιστής ελέγχθηκε με τις εκδόσεις 3.8, 6.0 και 10.0.0svn του LLVM. Ανάλογα την έκδοση μπορεί να δημιουργηθούν κάποια warnings κατά το make, τα οποία μπορούν με ασφάλεια να αγνοηθούν. Επιπλέον, για να δουλέψει η μεταγλώττιση με LLVM 10.0.0svn, απαιτείται η αλλαγή της γραμμής 96 του αρχείου **src/codegen.cpp** από:

  ```
  TheModule = llvm::make_unique<llvm::Module>(filename, TheContext);
  ```

  σε:

  ```
  TheModule = std::make_unique<llvm::Module>(filename, TheContext); )
  ```

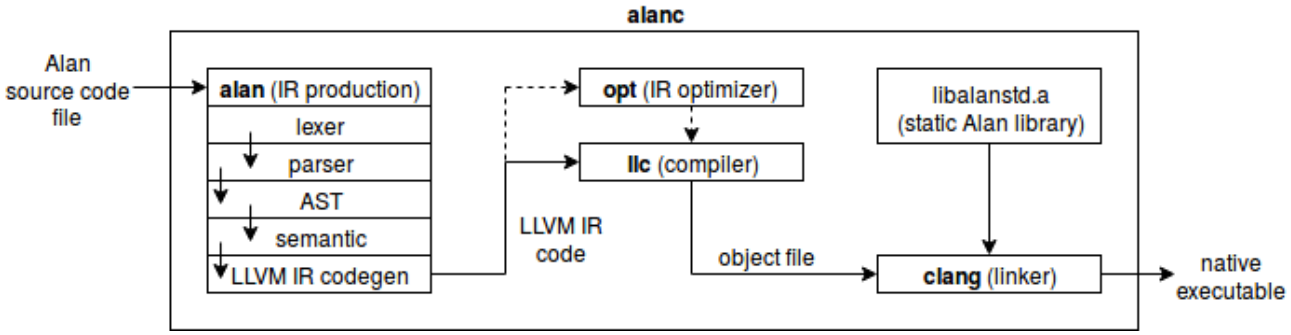
- Η βελτιστοποίηση του ενδιάμεσου κώδικα, καθώς και η παραγωγή του τελικού κώδικα και η βελτιστοποίησή του έγιναν με χρήση κάποιων command line utilities που προσφέρονται από το LLVM (αναλυτικότερη αναφορά σε αυτά θα γίνει στη συνέχεια)
- Η standard βιβλιοθήκη της Alan γράφτηκε από την αρχή σε C, ακολουθώντας όσο πιο πιστά γινόταν αφενός την περιγραφή των συναρτήσεων στην εκφώνηση της εργασίας, αφετέρου τη συμπεριφορά των αντίστοιχων συναρτήσεων της C. Αυτό επιλέξαμε να το κάνουμε στα πλαίσια της προσπάθειας για μία πιο cross-platform προσέγγιση
- Το τελικό script που αλληλεπιδρά με τον χρήστη και συνενώνει όλα τα παραπάνω γράφτηκε σε Python 3.7. Η Python προτιμήθηκε ως “γλώσσα συνένωσης” (glue language) έναντι της αρχικής μας υλοποίησης σε bash με στόχο το τελικό script να είναι όσο το δυνατόν πιο εύχρηστο, ευανάγνωστο, επεκτάσιμο και cross-platform

Για το build του project αρκεί η εκτέλεση της εντολής **make** στο root directory (αυτό αποτελεί και το μόνο σημείο της εργασίας μας που, δυστυχώς, δεν συμμορφώνεται με την ιδέα ενός πλήρως cross-platform project). Τέλος, ο μεταγλωττιστής μας μπορεί να εγκατασταθεί στο σύστημα του χρήστη (δεδομένου του ότι αυτό χρησιμοποιεί κάποιο Unix-based OS) με την εκτέλεση της εντολής **make install**, η οποία προφανώς απαιτεί superuser privileges. Η απεγκατάσταση μπορεί να γίνει με την εκτέλεση της εντολής **make uninstall**.

1. Εποπτεία της συνολικής αρχιτεκτονικής του μεταγλωττιστή

Αρχικά, θα αναφερθούμε στην συνολική αρχιτεκτονική του μεταγλωττιστή, ώστε να είναι ξεκάθαρα η θέση και ο ρόλος του κάθε τμήματος το οποίο θα αναλύουμε στη συνέχεια της αναφοράς.

Μία αφαιρετική οπτική της αρχιτεκτονικής του μεταγλωττιστή που υλοποιήσαμε φαίνεται στο διάγραμμα που ακολουθεί. Πριν προχωρήσουμε στην επιμέρους ανάλυση των τμημάτων του μεταγλωττιστή, θα αναφερθούμε συνοπτικά στον τρόπο που αυτά αλληλεπιδρούν και ποιες είναι οι αρμοδιότητες του κάθε τμήματος, όπως αυτές προκύπτουν από τον σχεδιασμό μας.



Η καρδιά του μεταγλωττιστή μας είναι το πρόγραμμα **alan**. Αυτό είναι υπεύθυνο για την ανάγνωση του πηγαίου κώδικα Alan, την επεξεργασία του και, εν τέλει, την παραγωγή του αντίστοιχου ενδιάμεσου κώδικα, σε LLVM IR. Για να το επιτύχει αυτό, αποτελείται μεταξύ άλλων από τον λεκτικό, τον συντακτικό και τον σημασιολογικό αναλυτή. Ο λεκτικός αναλυτής αναγνωρίζει και προωθεί στο pipeline του alan τα lexemes, τα οποία στη συνέχεια ο συντακτικός αναλυτής χρησιμοποιεί για να κατασκευάσει το abstract syntax tree (AST) που αντιπροσωπεύει το αρχικό πρόγραμμα. Αυτό επιτυγχάνεται, προφανώς, στην περίπτωση που δεν εντοπιστούν ούτε λεκτικά ούτε συντακτικά λάθη. Στη συνέχεια, ο σημασιολογικός αναλυτής ελέγχει ολόκληρο το AST για σημασιολογικά λάθη. Εάν τέτοια δεν εντοπιστούν, ένα ακόμα τελευταίο πέρασμα του AST λαμβάνει χώρα, κατά το οποίο παράγεται ενδιάμεσος κώδικας στη γλώσσα του LLVM (LLVM IR), με χρήση των κατάλληλων C++ bindings.

Σε αυτό το σημείο, έχει χαθεί οτιδήποτε σχετίζεται με την Alan: έχουμε αυτή τη στιγμή στα χέρια μας το ισοδύναμο του αρχικού προγράμματος σε LLVM IR, το οποίο σημαίνει πως τώρα μπορούμε να το αντιμετωπίσουμε ως τέτοιο, απολύτως ανεξάρτητα από την αρχική (πηγαία) γλώσσα. Συνεπώς, η πορεία από εδώ και πέρα καθορίζεται σχεδόν αποκλειστικά από το LLVM.

Πιο συγκεκριμένα, εάν ο χρήστης έχει ζητήσει να γίνουν βελτιστοποιήσεις, ο ενδιάμεσος κώδικας περνάει από το **opt**, ένα command line utility του LLVM το οποίο δέχεται ως είσοδο LLVM IR κώδικα και παράγει (πιθανώς) βελτιστοποιημένο LLVM IR κώδικα. Δεν θελήσαμε να κάνουμε υποχρεωτικό το στάδιο της βελτιστοποίησης του ενδιάμεσου κώδικα, ώστε να μπορούμε να πειραματιστούμε περισσότερο με τις διαφορές που προκύπτουν (με ή χωρίς βελτιστοποίηση) στη συνέχεια του pipeline, αλλά και για να μπορεί ο χρήστης να δει τον ενδιάμεσο κώδικα ακριβώς όπως τον παράγουμε εμείς μέσω των C++ bindings, χωρίς καμία αλλαγή (με χρήση του flag -i, όπως περιγράφεται στην εκφώνηση της εργασίας).

Στη συνέχεια, το **llc**, ο compiler του LLVM, αναλαμβάνει να μετατρέψει τον ενδιάμεσο κώδικα σε object file, να το μεταγλωττίσει, δηλαδή, στη native assembly. Ο llc είναι υπεύθυνος και για την βελτιστοποίηση του τελικού κώδικα, εάν αυτό έχει ζητηθεί από τον χρήστη.

Τέλος, ο **clang** αναλαμβάνει να συνδέσει (link) το object file που έχει παραχθεί με τη standard βιβλιοθήκη της Alan (libalanstd.a), παράγοντας έτσι το τελικό native εκτελέσιμο πρόγραμμα.

2. Λεκτικός αναλυτής (lexer)

Ο λεκτικός αναλυτής (lexer) είναι το πρώτο τμήμα του μεταγλωττιστή μας. Η υλοποίησή του έγινε με τη βοήθεια του εργαλείου **flex**. Ο lexer είναι υπεύθυνος για την ανάγνωση του source code του προγράμματος εισόδου και της παραγωγής των κατάλληλων lexemes, τα οποία και στη συνέχεια προωθούνται στον συντακτικό αναλυτή (parser). Ο lexer υλοποιήθηκε αποκλειστικά στο αρχείο **src/lexer.l**.

Η υλοποίηση του lexer αποτελεί μία αρκετά standard διαδικασία και δεν έχουμε να επισημάνουμε κάποια ιδιαίτερη σχεδιαστική επιλογή.

Ακολουθώντας την εκφώνηση, γράψαμε τα regexes που ορίζουν τις μεταβλητές, τις σταθερές (ακέραιες και δεκαεξαδικές), τους χαρακτήρες διαφυγής, και ούτω καθεξής.

Για τα keywords της Alan χρησιμοποιήσαμε τον parser (βλ. επόμενο μέρος) όπου και τα ορίσαμε και στη συνέχεια τα χρησιμοποιούμε από τον lexer μέσω του header parser.hpp που παράγεται κατά την εκτέλεση του bison.

Για κάθε κανόνα που αντιστοιχεί σε κάποιο lexeme, αυτό προωθείται στον parser μέσω του struct **yylval**, για το οποίο θα μιλήσουμε στο επόμενο μέρος.

Επιπλέον, για τα σχόλια (και κυρίως επειδή αυτά επιτρεπόταν να είναι εμφωλευμένα), χρησιμοποιήσαμε τον μηχανισμό των καταστάσεων (states) που προσφέρει το flex, δημιουργώντας μία ξεχωριστή κατάσταση για την περίπτωση των σχολίων πολλών γραμμών (<MULLINE_COMMENT>). Από αυτήν την κατάσταση φεύγουμε μόνο στην περίπτωση που συναντήσουμε τόσες αρχές σχολίων (“(*”) όσα και τέλη (“*)”). Ενδιαμέσως, δεν αναγνωρίζουμε κανένα άλλο lexeme.

Τέλος, ο lexer χειρίζεται τον μετρητή linecount προσαυξάνοντάς τον κάθε φορά που συναντάει χαρακτήρα αλλαγής γραμμής (“\n”), ο οποίος έχει οριστεί ως εξωτερική (extern) μεταβλητή στον parser.

3. Συντακτικός αναλυτής (parser) και κατασκευή abstract syntax tree (AST)

Ο συντακτικός αναλυτής (parser) αποτελεί την καρδιά του προγράμματος **alan** (βλ. διάγραμμα 1ου μέρους), αφού είναι αυτός στον οποίο ορίζεται η συνάρτηση main και ενορχηστρώνει τόσο την ανάγνωση του προγράμματος εισόδου (χρησιμοποιώντας τον lexer), όσο και την κατασκευή του AST, τον σημασιολογικό έλεγχο αλλά και την παραγωγή του ενδιάμεσου κώδικα σε LLVM IR. Ο parser υλοποιήθηκε αποκλειστικά στο αρχείο **src/parser.ypp**, με χρήση του εργαλείου **bison**.

3.1 Η υλοποίηση του συντακτικού αναλυτή

Αρχικά, θα αναφερθούμε στην υλοποίηση του parser. Όπως και στον lexer, η διαδικασία κατασκευής του parser είναι σε μεγάλο βαθμό standard. Το κύριο μέλημα εδώ είναι η μετάφραση της γραμματικής της Alan σε κατάλληλους κανόνες που θα οδηγήσουν στην κατασκευή ενός σαφούς και εύχρηστου AST (θα αναφερθούμε εκτενώς σε αυτό στη συνέχεια).

Πρώτα απ' όλα, στον parser ορίζουμε ποια lexemes αντιστοιχούν σε keywords της Alan, πληροφορία που γίνεται exported στον lexer μέσω του header parser.hpp, που δημιουργείται αυτόματα.

Στη συνέχεια, ορίζουμε το struct **yylval**, το οποίο λειτουργεί ως ένας "buffer" μεταξύ lexer και parser, όπου ο πρώτος γράφει τα lexemes που βρίσκει στη κατάλληλη θέση ώστε να τα καταναλώσει ο δεύτερος. Εξαίρεση αποτελούν τα keywords, για τα οποία ο lexer επιστρέφει στον parser απλά τον αντίστοιχο αριθμό, όπως ορίστηκε αυτόματα στο parser.hpp. Το yylval δεν είναι παρά ένα union, στο οποίο ορίζονται όλοι οι δυνατοί τύποι δεδομένων για τα αντικείμενα που είτε έρχονται από τον lexer, είτε δημιουργούνται από τον parser:

- κόμβοι του AST (ASTNodes, δημιουργούνται από τον parser)
- χαρακτήρες (από τον lexer)
- strings (από τον lexer)
- ακέραιοι αριθμοί (από τον lexer)
- τύποι (Types, ορισμένοι στο δοσμένο πίνακα συμβόλων, δημιουργούνται από τον parser)

Έπειτα, ορίζεται η προτεραιότητα και η προσεταιριστικότητα των πράξεων/τελεστών με βάση την εκφώνηση της άσκησης, καθώς και οι τύποι (με βάση το yylval) των δεδομένων που δημιουργούνται όταν κάποιος από τους κανόνες του parser ενεργοποιείται.

Σε αυτό το σημείο, ορίζονται οι κανόνες της γραμματικής της Alan. Θεωρούμε ότι δεν υπάρχει λόγος να μπούμε σε εκτενείς λεπτομέρειες, καθώς ακολουθήσαμε πιστά την γραμματική όπως αυτή παρουσιάζεται στην εκφώνηση. Για κάθε κανόνα που ενεργοποιείται, δημιουργείται μία αντίστοιχη δομή (είτε ένα καινούργιο ASTNode, είτε ένα καινούργιο Type, ανάλογα με τον κανόνα), και συνδέεται με την δομή – γονιό, με τη βοήθεια των γνωστών τελεστών $$ και $1, $2, ... του bison. Αφού αναλυθεί συντακτικά το σύνολο του προγράμματος εισόδου, έχουμε στα χέρια μας το συνολικό AST που το αναπαριστά.

3.2 Το αφαιρετικό συντακτικό δέντρο (abstract syntax tree – AST)

Σε αυτό το σημείο θα μιλήσουμε πιο αναλυτικά για τον τρόπο που υλοποιήσαμε το AST που θα αναπαριστά το εκάστοτε πρόγραμμα εισόδου, μετά την επιτυχή συντακτική ανάλυσή του. Το AST υλοποιήθηκε στο αρχείο **include/ast.hpp**, σε C++14.

Στο ast.hpp ορίζεται ο κόμβος (ASTNode) του AST, ως μία κλάση. Τα private πεδία της κλάσης είναι όλες οι πληροφορίες που μπορεί να χρειάζεται να συγκρατήσει *οποιοδήποτε είδος κόμβου του AST*. Με άλλα λόγια, έχουμε ένα και μόνο είδος ASTNode κλάσης. Για παραπάνω πληροφορίες για το κάθε πεδίο της ASTNode παραπέμπουμε στο αντίστοιχο αρχείο, όπου περιγράφεται αναλυτικά σε σχόλια ό,τι χρειάζεται να γνωρίζει κανείς για την κλάση.

Η διαφοροποίηση μεταξύ ASTNodes που επιτελούν άλλη λειτουργία (για παράδειγμα, μεταξύ ενός κόμβου που αναπαριστά δήλωση συνάρτησης και ενός που αναπαριστά μία άθροιση) επιτυγχάνεται με τη χρήση διαφορετικών **constructors**, οι οποίοι ορίζονται ως protected μέθοδοι της κλάσης ASTNode. Στη συνέχεια, ορίζονται πολλές υποκλάσεις της ASTNode, τόσες όσα και τα διαφορετικά είδη κόμβων που χρειαζόμαστε για το AST. Κάθε μία από αυτές τις υποκλάσεις αποτελείται από τις εξής τρεις μεθόδους, και μόνον αυτές (μαζί, προφανώς, με τα private πεδία που κληρονόμησαν από την ASTNode):

- Έναν constructor, ο οποίος καλεί τον αντίστοιχο constructor της γονεϊκής κλάσης ASTNode, που σημαίνει πως γεμίζει μόνο τα private πεδία που αυτός ο συγκεκριμένος κόμβος χρειάζεται,
- την void sem(), η οποία χρησιμοποιείται κατά τον σημασιολογικό έλεγχο (βλ. επόμενο μέρος) και ορίζεται εκ νέου από κάθε κλάση – παιδί (είναι virtual μέθοδος της ASTNode), και
- την llvm::Value * codegen(), η οποία χρησιμοποιείται κατά την παραγωγή ενδιάμεσου κώδικα σε LLVM IR (βλ. Μέρος 5) και ορίζεται εκ νέου από κάθε κλάση – παιδί (είναι virtual μέθοδος της ASTNode).

Τα παραπάνω ορίζουν ένα πολύ απλό interface του AST για τον parser, το οποίο και χρησιμοποιεί εάν χρειαστεί να δημιουργηθεί ένας καινούργιος κόμβος του AST ως αποτέλεσμα ενεργοποίησης κάποιου κανόνα. Για παράδειγμα, εάν ο parser συναντήσει την εντολή “return 42;”, δεν δημιουργεί ένα αντικείμενο ASTNode, αλλά ένα αντικείμενο ASTRet, δίνοντας στον κατασκευαστή του ως όρισμα μόνο την έκφραση “42” (η οποία θα έχει γίνει και αυτή αναδρομικά ένας κόμβος του AST τύπου ASTInt). Έπειτα, ο κατασκευαστής της κλάσης ASTRet θα καλέσει τον κατασκευαστή της ASTNode που του αντιστοιχεί, δηλαδή εκείνον που δέχεται ως όρισμα μόνο μία έκφραση (και αφήνει τα υπόλοιπα πεδία κενά).

3.3 Η λειτουργία του συντακτικού αναλυτή (η συνάρτηση main)

Όλα τα παραπάνω εκκινούν από την main συνάρτηση, η οποία επίσης ορίζεται στο parser.ypp και αποτελεί τη main του προγράμματος alan γενικότερα.

Συγκεκριμένα, πρώτα αρχικοποιείται το όνομα του αρχείου εισόδου (filename) και ο μετρητής γραμμής (linecount). Έπειτα καλείται η **yyparse()**, η οποία κατασκευάζει το συνολικό AST και αποθηκεύει τον ριζικό κόμβο σε μία μεταβλητή t. Εάν η yyparse() αποτύχει, το πρόγραμμα τερματίζει με κατάλληλο μήνυμα και κωδικό εξόδου (1). Σε αντίθετη περίπτωση, δημιουργείται ο πίνακας συμβόλων (βλ. επόμενο μέρος) στον οποίο προστίθενται οι συναρτήσεις βιβλιοθήκης της Alan με μία δική μας συνάρτηση (initLibFunctions()), ώστε η ύπαρξη και το signature τους να είναι γνωστά κατά τον σημασιολογικό έλεγχο του AST. Σε αυτό το σημείο, ελέγχεται εάν η main συνάρτηση του προγράμματος εισόδου (δηλαδή η πιο εξωτερική συνάρτηση στο .alan αρχείο) δέχεται ορίσματα. Επειδή αυτό είναι κάτι που δεν επιτρέπουμε, επιστρέφουμε αντίστοιχο μήνυμα λάθους και διαγράφουμε τους κόμβους του AST που αντιστοιχούν στα ορίσματα της main συνάρτησης.

Στη συνέχεια, καλείται η συνάρτηση sem() του ριζικού κόμβου του AST (μεταβλητή t), η οποία εκκινεί τον αναδρομικό σημασιολογικό έλεγχο του AST. Εάν αυτός αποτύχει σε κάποιον κόμβο, το πρόγραμμα *δεν τερματίζει ακαριαία*, με στόχο να βρεθούν όσο περισσότερα λάθη γίνεται. Αυτό επιλέξαμε να το κάνουμε μιμούμενοι τη συμπεριφορά άλλων compilers, όπως του gcc/g++. Παρόλα αυτά, το γεγονός πως υπήρξε λάθος καταχωρείται σε μία μεταβλητή και, όταν ο σημασιολογικός έλεγχος διατρέξει όλο το AST, ο πίνακας συμβόλων καταστρέφεται και η μεταβλητή αυτή ελέγχεται και, εάν έχει όντως προκύψει σημασιολογικό σφάλμα, το πρόγραμμα τερματίζει επιστρέφοντας κωδικό εξόδου 1. Σε αντίθετη περίπτωση, δεν μπορούν να υπάρξουν λάθη από εδώ και πέρα – έχουμε στα χέρια μας ένα ορθό, τόσο συντακτικά όσο και σημασιολογικά, AST. Έτσι, το τελευταίο βήμα είναι η κλήση της συνάρτησης codegen(), η οποία διατρέχει και πάλι

όλο το AST και αναδρομικά παράγει τον αντίστοιχο LLVM IR κώδικα, τον οποίο και τυπώνει απευθείας στο stdin όταν ολοκληρωθεί. Αυτός, έπειτα, συλλέγεται από τα επόμενα τμήματα του συνολικού μεταγλωττιστή προς τη πιθανή βελτιστοποίηση και, εν τέλει, την παραγωγή τελικού κώδικα και ενός native εκτελέσιμου.

4. Σημασιολογικός αναλυτής

Σε αυτό το μέρος της αναφοράς θα αναφερθούμε σύντομα στον σημασιολογικό αναλυτή του μεταγλωττιστή μας. Ο σημασιολογικός αναλυτής υλοποιείται στο αρχείο **src/ast.cpp**, σε C++14, και επί της ουσίας δεν είναι τίποτα περισσότερο από τον ορισμό της συνάρτησης void sem(), για κάθε κλάση – παιδί της ASTNode.

Η διαδικασία είναι αναδρομική και ξεκινά από την κλήση της sem() στον ριζικό κόμβο του AST, από την main του parser. Σε κάθε κόμβο που φτάνει ο σημασιολογικός έλεγχος, ελέγχεται πλήθος συνθηκών που σχετίζεται με τις ιδιότητες και τη λειτουργία του συγκεκριμένου κόμβου, συναρτήσει του τι αυτός αντιπροσωπεύει στο αρχικό πρόγραμμα. Ταυτόχρονα, προκειμένου να πραγματοποιηθούν οι έλεγχοι αυτοί, είναι απαραίτητη η καταγραφή, για κάθε κόμβο, στοιχείων όπως για παράδειγμα ο τύπος των διαφόρων μεταβλητών, καθώς και η συλλογή πληροφοριών σχετικά με τα scopes, τις συναρτήσεις και τις παραμέτρους τους. Κατά τη διάρκεια της διάσχισης του AST, λοιπόν, συμπληρώνονται επιπλέον πεδία των υπαρχόντων κόμβων και ενημερώνεται ο πίνακας συμβόλων.

Λόγω του πλήθους συνθηκών που ελέγχονται σε κάθε κόμβο, αλλά και των μεγάλων διαφορών μεταξύ των κόμβων, κρίναμε πως δεν χρειάζεται να τα παραθέσουμε στην αναφορά. Ενδεικτικά αναφέρουμε τον έλεγχο τύπων, όπως για παράδειγμα στην περίπτωση type mismatch μεταξύ των εκφράσεων ενός δυαδικού τελεστή ή και της χρήσης μίας έκφρασης int στη συνθήκη ενός while loop, καθώς και τον έλεγχο κατά το πέρασμα παραμέτρων. Για το σύνολο των συνθηκών που ελέγχονται, παραπέμπουμε στον πηγαίο κώδικα, στον οποίο μπορεί κανείς να αποκτήσει μία πλήρη εικόνα των ελέγχων που λαμβάνουν χώρα, τόσο από τα σχόλια όσο και από τα αντίστοιχα μηνύματα λάθους που δίνονται ως ορίσματα στη συνάρτηση error().

Αξίζει παρόλα αυτά να αναφέρουμε 4 σχεδιαστικές μας επιλογές:

- Για τον πίνακα συμβόλων χρησιμοποιήθηκε η έτοιμη υλοποίηση όπως μας δόθηκε από τον διδάσκοντα (αρχεία **include/symbol.hpp** και **src/symbol.cpp**). Χρειάστηκε να γίνουν κάποιες μικρές αλλαγές ώστε ο κώδικας να μετατραπεί από C σε C++ (για παράδειγμα, τα enumerations δεν μπορούσαν πλέον παρά να ορίζονται globally).
- Χρησιμοποιήσαμε ένα stack από SymbolEntry* αντικείμενα για να μοντελοποιήσουμε τις συναρτήσεις του προγράμματος, ώστε να γνωρίζουμε ανά πάσα στιγμή μέσα σε ποια συνάρτηση βρισκόμαστε κατά τον σημασιολογικό έλεγχο, κάτι πολύ χρήσιμο, για παράδειγμα, στους ορισμούς επικεφαλίδων συναρτήσεων (για τη χρήση της endFunctionHeader() του πίνακα συμβόλων) ή για τον έλεγχο κατά την εντολή “return”.
- Καθώς στην αναλυτική περιγραφή της γλώσσας Alan στην εκφώνηση δεν αναφέρεται τίποτα για την περίπτωση η βασική συνάρτηση του προγράμματος να έχει τυπικά ορίσματα, αποφασίσαμε η ύπαρξη αυτών να θεωρείται σημασιολογικό λάθος. Αντίθετα, δε θεωρήθηκε απαραίτητο να υπάρχει κάποιος περιορισμός ως προς τον τύπο επιστροφής της κύριας συνάρτησης. Σε κάθε περίπτωση, το εκτελέσιμο πρόγραμμα που παράγεται από τον μεταγλωττιστή μας επιστρέφει στο λειτουργικό σύστημα την τιμή επιστροφής της κύριας συνάρτησης, εάν αυτή υπάρχει, αλλιώς 0.
- Όσον αφορά την ύπαρξη return instruction στις άλλες συναρτήσεις, προφανώς σε όσες έχουν τον ειδικό τύπο επιστροφής proc δεν είναι απαραίτητη, ενώ στις υπόλοιπες επιλέξαμε να πραγματοποιείται ένας απλός έλεγχος, μιμούμενοι την αντίστοιχη συμπεριφορά του gcc. Συγκεκριμένα, σε περίπτωση που δεν είναι απόλυτα βέβαιο ότι θα εκτελεστεί κάποια εντολή return (δεν υπάρχει δηλαδή κάπου στον κώδικα της συνάρτησης *εκτός* if, if/else και while statements), τυπώνεται ένα προειδοποιητικό μήνυμα (warning), το οποίο απλώς ενημερώνει το χρήστη για την *πιθανότητα* εμφάνισης unexpected behavior (σε σχέση με την αρχική πρόθεσή του όταν έγραφε τον κώδικα), χωρίς να επηρεάζει την ολοκλήρωση του σημασιολογικού ελέγχου. Αξίζει να αναφέρουμε σε αυτό το σημείο πως *πάντα* προσθέτουμε ένα επιπλέον return statement στο τέλος κάθε συνάρτησης (χωρίς τιμή ή με τιμή 0, ανάλογα τον τύπο επιστροφής), για να εξασφαλίσουμε την ομαλή συμπεριφορά του προγράμματος.

5. Παραγωγή ενδιάμεσου κώδικα και βελτιστοποίηση

Σε αυτό το στάδιο του μεταγλωττιστή, το οποίο πραγματοποιείται μόνο στην περίπτωση επιτυχούς ολοκλήρωσης του σημασιολογικού ελέγχου, λαμβάνει χώρα ένα ακόμα, τελευταίο αναδρομικό πέρασμα του AST, με στόχο την παραγωγή ενδιάμεσου κώδικα. Συγκεκριμένα, παράγεται κώδικας στη γλώσσα του LLVM (LLVM IR), από τον οποίο στη συνέχεια θα προκύψει ο τελικός κώδικας.

Η παραγωγή του ενδιάμεσου κώδικα πραγματοποιείται στο αρχείο **src/codegen.cpp**, σε C++14, με χρήση της συνάρτησης **void codegen()** και αναδρομική κλήση της αντίστοιχης virtual μεθόδου για κάθε κλάση – παιδί της ASTNode.

Πέρα από τη χρήση των μεταβλητών και των συναρτήσεων που παρέχει το LLVM, υλοποιήθηκε και μία επιπλέον δομή δεδομένων, ο Logger, ο οποίος είναι υπεύθυνος για την αποθήκευση των πληροφοριών που αφορούν όλες τις συναρτήσεις, όπως για παράδειγμα τις παραμέτρους και τις τοπικές μεταβλητές τους, καθώς και τους τύπους κάθε μεταβλητής. Συγκεκριμένα, πρόκειται για έναν vector από δομές scopeLog, η κάθε μία εκ των οποίων αφορά μία συνάρτηση (που ταυτίζεται στην alan με ένα scope) και περιλαμβάνει πληροφορίες για:

- τον τύπο και τη διεύθυνση στη στοίβα των μεταβλητών και παραμέτρων της συνάρτησης
- τις συναρτήσεις που ορίζονται στο εσωτερικό της

Με χρήση, λοιπόν, των εργαλείων του LLVM και του Logger που σταδιακά ενημερώνεται κατά τη διάρκεια της διάσχισης του AST, υλοποιήσαμε για κάθε κλάση-παιδί της ASTNode τη μέθοδο codegen(), η οποία ανάλογα με τη λειτουργία και τις ιδιαιτερότητες του κάθε κόμβου παράγει τον αντίστοιχο ενδιάμεσο κώδικα. Για αναλυτική επεξήγηση της διαδικασίας παραγωγής IR σε κάθε κόμβο, παραπέμπουμε στον πηγαίο κώδικα, ο οποίος συνοδεύεται από επαρκή σχολιασμό. Αξίζει, παρόλα αυτά, να αναφερθούν τα παρακάτω σημεία:

- Όταν χρησιμοποιείται η διεύθυνση των μεταβλητών, ιδιαίτερη προσοχή απαιτεί ο χειρισμός των πινάκων, των τυπικών παραμέτρων, των μεταβλητών που περνιούνται by reference ως παράμετροι συναρτήσεων κτλ. Συγκεκριμένα, είναι εξαιρετικά σημαντικό να διακρίνουμε τις περιπτώσεις που είναι απαραίτητο το dereferencing ώστε να αποκτήσουμε την πραγματική διεύθυνση της μεταβλητής. Για το λόγο αυτό χρησιμοποιείται η συνάρτηση **calcAddr()**, η οποία ανάλογα με τον τύπο της μεταβλητής επιστρέφει την ανάλογη διεύθυνση.
- Καθώς το LLVM δεν υποστηρίζει εμφωλευμένες συναρτήσεις, στις οποίες βασίζεται η δομή της Alan, είναι απαραίτητο να βρεθεί ένας εναλλακτικός τρόπος να είναι ορατές σε κάθε scope οι μεταβλητές που ορίζονται στο εξωτερικό του. Η λύση που επιλέξαμε ως την πιο αποτελεσματική και κατανοητή, είναι αυτό να γίνεται περνώντας όλες τις εξωτερικές μεταβλητές ως “κρυφές” παραμέτρους σε κάθε συνάρτηση.
- Όπως ζητήθηκε στην εκφώνηση, η αποτίμηση των συνθηκών που χρησιμοποιούν τους τελεστές & και | γίνεται με **βραχυκύκλωση (short-circuit)**. Συγκεκριμένα, αν το αποτέλεσμα της συνθήκης μπορεί να κριθεί μόνο από την αποτίμηση του πρώτου operand, όλα τα επόμενα operands δεν αποτιμώνται καθόλου. Για την υλοποίηση της βραχυκύκλωσης για 2 ή και περισσότερα τελούμενα, απαραίτητη ήταν η δημιουργία ενός ξεχωριστού BasicBlock για την αποτίμηση κάθε τελούμενου, και η χρήση **Phi Node** για την τον καθορισμό της τελικής τιμής της συνθήκης.

Τέλος, όσον αφορά στη βελτιστοποίηση του ενδιάμεσου κώδικα, αυτή γίνεται με χρήση του command line utility **opt** του LLVM, με το flag **-O3**, όπως εξηγείται αναλυτικά στο τελευταίο μέρος της αναφοράς.

6. Παραγωγή τελικού κώδικα και βελτιστοποίηση

Βρισκόμαστε, πλέον, στο σημείο όπου ο ενδιάμεσος κώδικας έχει παραχθεί (και ενδεχομένως βελτιστοποιηθεί). Για την παραγωγή του τελικού κώδικα, δεδομένου του ότι ο ενδιάμεσος κώδικας είναι σε LLVM IR, αρκεί να χρησιμοποιήσουμε το command line utility **llc**, τον compiler του LLVM. Με την εντολή “llc -filetype=obj” και είσοδο τον ενδιάμεσο κώδικα από το προηγούμενο τμήμα του μεταγλωττιστή (συγκεκριμένα, το stdout του) παράγουμε ένα object file σε native assembly. Εάν ζητήθηκε βελτιστοποίηση, τότε προσθέτουμε και το flag -O3 στην παραπάνω εντολή. Διαφορετικά, σε περίπτωση που δεν είναι επιθυμητή η βελτιστοποίηση του τελικού κώδικα, προσθέτουμε το flag -O0, ώστε να αποφύγουμε το ενδεχόμενο της by default βελτιστοποίησης από το llc.

Για να έχουμε στα χέρια μας ένα τελικό εκτελέσιμο, απαιτείται ένα ακόμα βήμα – η σύνδεση (linking) του παραπάνω object file με τη βιβλιοθήκη της Alan, ώστε να γίνουν resolve τα ονόματα συναρτήσεων που έχουν δηλωθεί μέσα στο object file, αλλά που το τελευταίο δεν γνωρίζει πού βρίσκονται οι υλοποιήσεις τους. Αυτό επιτυγχάνεται με την εντολή “clang <object file> libalanstd.a -o <outname>”, οπότε και παράγεται απευθείας το τελικό εκτελέσιμο πρόγραμμα.

7. alanc – ο τελικός μεταγλωττιστής

Σε αυτό το μέρος της παρούσας αναφοράς θα περιγράψουμε τον τρόπο χρήσης και λειτουργίας του τελικού εκτελέσιμου του μεταγλωττιστή, του **alanc**. Στο πρώτο μέρος παρουσιάσαμε συνοπτικά την γενική αρχιτεκτονική του script, ενώ σε αυτό το μέρος θα αναφερθούμε πιο αναλυτικά στην υλοποίησή του καθώς και στις σχεδιαστικές επιλογές που αποφασίσαμε να κάνουμε.

Αρχικά, παρατίθεται το μήνυμα που λαμβάνει ο χρήστης με την εκτέλεση της εντολής **./alanc -h**:

```
usage: alanc [-h] [-O] [-x] [-i] [-f] [-c] [-o OUTNAME] [infile]

ALANC - the Alan Limitless and Amazingly Neat Compiler

positional arguments:
  infile      (if -f or -i not given) the Alan source code to compile

optional arguments:
  -h, --help    show this help message and exit
  -O            optimize IR and final (assembly) code
  -x, --no-dump do not emit IR and assembly code in two seperate files (switched
                on by default)
  -i            read source code from stdin, print IR code to stdout, then exit
                (no executable is produced)
  -f            read source code from stdin, print final code to stdout, then
                exit (no executable is produced)
  -c            create object file and stop, skipping the linking phase
  -o OUTNAME    (if -f or -i not given) the name of the produced executable
                (default: a.out)
```

Όπως φαίνεται παραπάνω, υπάρχει μία σειρά από flags που προσφέρονται στον χρήση με σκοπό την παραμετροποίηση της λειτουργίας του μεταγλωττιστή.

Ως default συμπεριφορά του μεταγλωττιστή (όταν δεν χρησιμοποιείται κανένα από τα τρία flags -x, -i, -f), ορίσαμε, όπως υπονοείται από την εκφώνηση της εργασίας, την παραγωγή δύο επιπλέον αρχείων, ένα που περιέχει τον ενδιάμεσο και ένα τον τελικό κώδικα που παράγεται από τον μεταγλωττιστή, σε human-readable μορφή. Και τα δύο αυτά αρχεία κρίναμε κατάλληλο να αποθηκεύονται **στο ίδιο directory** με αυτό του προγράμματος εισόδου (σε αντίθεση με το τελικό εκτελέσιμο το οποίο, μιμούμενοι την συμπεριφορά του gcc, τοποθετούμε στο τρέχον directory).

Παρόλα αυτά, κρίθηκε σκόπιμη η ύπαρξη του flag **-x (--no-dump)** με τη χρήση του οποίου αποφεύγεται η δημιουργία των δύο επιπλέον αρχείων σε περίπτωση που ο χρήστης θέλει να μεταγλωττίσει κάποιο πρόγραμμα και να δημιουργηθεί μόνο το τελικό εκτελέσιμο.

Τα flags **-i** και **-f** ακολουθούν τις προδιαγραφές της εκφώνησης, όπως φαίνεται και στις περιγραφές τους στο μήνυμα παραπάνω. Αξίζει να σημειωθεί πως αποφασίσαμε να κάνουμε τη χρήση τους mutually exclusive, προς αποφυγή συγχύσεων. Επίσης, όπως φαίνεται και στα παραπάνω μηνύματα, αποφασίσαμε να μην παράγεται το τελικό εκτελέσιμο σε αυτές τις περιπτώσεις, δίνοντας την δυνατότητα στον μεταγλωττιστή μας να λειτουργήσει και ως μέρος κάποιου toolchain που θέλει να κατασκευάσει ο χρήστης, όπου και δεν περιμένει ή επιθυμεί τέτοιου είδους side effects.

Το flag **-O** ενεργοποιεί την βελτιστοποίηση τόσο του ενδιάμεσου (LLVM IR) όσο και του τελικού κώδικα (native assembly). Εξηγήσαμε στο πρώτο μέρος της αναφοράς την απόφασή μας να μην είναι υποχρεωτική η βελτιστοποίηση του ενδιάμεσου κώδικα. Σε αυτό το σημείο θεωρούμε σημαντικό να αναφέρουμε πως **επιλέξαμε να επιτρέψουμε το inlining optimization pass του opt**. Αυτό σημαίνει πως ο βελτιστοποιημένος LLVM IR κώδικας που παράγεται από τον μεταγλωττιστή μας έχει πολύ συχνά επαναλήψεις κώδικα και συναρτήσεων. Αν και αυτό μεγαλώνει ανούσια τον όγκο του ενδιάμεσου και του τελικού κώδικα, θεωρήσαμε πως δεν μας δημιουργεί πρόβλημα,

αφενός διότι η Alan είναι μία αρκετά απλή γλώσσα και τα προγράμματα που γράφουμε για αυτήν δεν ξεπερνούν στην γενική περίπτωση τις μερικές δεκάδες γραμμές, αφετέρου διότι θεωρήσαμε πιο σημαντικό το κέρδος της αφαίρεσης κάποιων εντολών call (που γενικά είναι από τις πιο επιβαρυντικές κατά την εκτέλεση ενός προγράμματος) από την ζημία που προκύπτει από την μικρή αύξηση στον όγκο του τελικού προγράμματος και το redundancy του ενδιάμεσου και τελικού κώδικα.

Το flag **-c** έχει ακριβώς την ίδια σημασία με το αντίστοιχο flag του gcc: σταματάει στη φάση της μεταγλώττισης του πηγαίου προγράμματος, χωρίς να προχωρήσει στο στάδιο της σύνδεσης (linking) με τη standard βιβλιοθήκη της Alan και, τελικά, την παραγωγή του εκτελέσιμου προγράμματος. Τελικό προϊόν σε περίπτωση που δοθεί αυτό το flag είναι το αντίστοιχο object file, με κατάληξη ".o". Σημειώνεται πως αυτό το αρχείο πάντα παράγεται, απλά σε περίπτωση που δεν έχει δοθεί το εν λόγω flag, διαγράφεται μετά την παραγωγή του εκτελέσιμου προγράμματος.

Επιπλέον, δίνεται στο χρήστη η δυνατότητα να δώσει κάποιο όνομα που επιθυμεί στο τελικό εκτελέσιμο με τη χρήση του flag **-o** (όπως και στον gcc, το default όνομα για το παραγόμενο εκτελέσιμο πρόγραμμα είναι a.out).

Τέλος, ο χρήστης μπορεί να δώσει το όνομα ενός αρχείου προς μεταγλώττιση, ακριβώς όπως στον gcc και στον clang. Προφανώς, εάν δώσει κάποιο όνομα αρχείου, δεν μπορεί να έχει δώσει και κάποιο από τα flags **-i** ή **-f**.

Τα παραπάνω flags ορίστηκαν και υλοποιήθηκαν με βάση το module **argparse** της Python. Όταν το πρόγραμμα ξεκινά, γίνονται κάποιοι έλεγχοι σχετικοί με τα flags:

1. Ο χρήστης επιτρέπεται είτε να χρησιμοποιήσει ένα εκ των flags **-i** ή **-f** είτε να δώσει ένα όνομα αρχείου προς μεταγλώττιση.
2. Ένα από τα δύο παραπάνω πρέπει να έχουν δοθεί (δηλαδή, δεν επιτρέπεται να λείπουν και τα δύο)
3. Μόνο ένα εκ των flags **-i** ή **-f** επιτρέπεται να χρησιμοποιηθεί
4. Δεδομένου του ότι η χρήση των flags **-i** ή **-f** συνεπάγεται μη παραγωγή τελικού εκτελέσιμου προγράμματος, δεν επιτρέπεται να χρησιμοποιούνται μαζί με το flag **-o**.

Σε περίπτωση που οποιαδήποτε από τις παραπάνω συνθήκες δεν πληρούται, η εκτέλεση του μεταγλωττιστή τερματίζεται με αντίστοιχο μήνυμα λάθους από το argparse.

Στη συνέχεια, ορίζεται το όνομα του προγράμματος (αποκόπτεται η κατάληξη εάν αυτή είναι ".alan", αλλιώς μένει ως έχει), του object file που θα παραχθεί, καθώς και των αρχείων .imm και .asm κατά τις προδιαγραφές της εκφώνησης, σε περίπτωση που αυτά ζητήθηκαν από τον χρήστη.

Έπειτα, ορίζονται οι εντολές που εκτελούνται για τα 4 βήματα της μεταγλώττισης που ακολουθούν μαζί με τα αντίστοιχα flags τους. Τα 4 αυτά βήματα φαίνονται εποπτικά στο διάγραμμα του πρώτου μέρους της παρούσας αναφοράς:

1. Για την **παραγωγή του ενδιάμεσου κώδικα** χρησιμοποιείται το πρόγραμμα **alan** με μοναδικό όρισμα το όνομα του προγράμματος ώστε να χρησιμοποιηθεί εσωτερικά από το LLVM αλλά και για πιθανά μηνύματα λάθους. Δέχεται το πηγαίο κώδικα από το stdin και παράγει τον ενδιάμεσο κώδικα στο stdout.
2. Για την **βελτιστοποίηση του ενδιάμεσου κώδικα** χρησιμοποιείται το command line utility **opt** του LLVM, με το flag **-O3**, το οποίο καλύπτει τα ζητούμενα της πρώτης bonus μονάδας της εργασίας. Δέχεται τον ενδιάμεσο κώδικα από το stdin και παράγει (πιθανώς) βελτιστοποιημένο ενδιάμεσο κώδικα στο stdout. Αυτό το βήμα παραλείπεται σε περίπτωση που δεν έχει δοθεί από τον χρήστη το αντίστοιχο flag. Επιπλέον, μετά από αυτό το βήμα (είτε έχει εκτελεστεί είτε όχι) και εάν αυτό έχει ζητηθεί από τον χρήστη, τότε ο ενδιάμεσος κώδικας είτε αποθηκεύεται σε κατάλληλο .imm αρχείο και η εκτέλεση του μεταγλωττιστή συνεχίζεται είτε τυπώνεται στην οθόνη και το πρόγραμμα τερματίζει επιτυχώς.
3. Για την **παραγωγή του τελικού κώδικα** χρησιμοποιείται το command line utility **llc**, ο

μεταγλωττιστής του LLVM, μαζί με το flag **-filetype=obj**. Αποφασίσαμε σε αυτό το στάδιο να παράγεται object file και όχι κάποια άλλη μορφή assembly ώστε να μπορούμε να υποστηρίξουμε ευκολότερα τη λειτουργία του flag **-c**, να είναι ευκολότερο το linking στη συνέχεια, αλλά και γιατί αποτελεί standard συμπεριφορά πολλών μεταγλωττιστών η παραγωγή ενός object file σε αυτή τη φάση, κάτι που μπορεί ο χρήστης να επιθυμεί. Σε περίπτωση που δόθηκε το flag βελτιστοποίησης, προστίθεται στα flags του llc το **-O3**, το οποίο καλύπτει τα ζητούμενα της δεύτερης bonus μονάδας της εργασίας. Η εντολή αυτή δέχεται τον (πιθανώς βελτιστοποιημένο) ενδιάμεσο κώδικα από το stdin και παράγει ένα object file το οποίο αποθηκεύεται στο file system, στο τρέχον directory. Επιπλέον, μετά από αυτό το βήμα και εάν αυτό έχει ζητηθεί από τον χρήστη, τότε ο τελικός κώδικας είτε αποθηκεύεται σε κατάλληλο .asm αρχείο και η εκτέλεση του μεταγλωττιστή συνεχίζεται είτε τυπώνεται στην οθόνη, διαγράφεται το object file και το πρόγραμμα τερματίζει επιτυχώς. Για να συμβούν αυτά τα δύο εφόσον ζητηθούν, ξανακαλούμε τον llc, αλλά αυτή τη φορά χωρίς το flag **-filetype=obj**, ώστε να παραχθεί human-readable assembly αντί για ένα object file.

4. Το τελευταίο βήμα είναι **η σύνδεση του object file με τη standard βιβλιοθήκη της Alan (linking)**. Για αυτό το βήμα, εκτελείται απλώς η εντολή:
   `clang <objectfile> libalanstd.a -o <outname>`
   Μετά από αυτήν την εντολή, έχει παραχθεί το τελικό εκτελέσιμο πρόγραμμα με όνομα outname. Εάν έχουμε φτάσει σε αυτό το σημείο, διαγράφεται και το object file από το file system.

Σημειώνεται πως σε κάθε ένα από τα παραπάνω βήματα ελέγχεται στο stderr κάθε εντολής. Σε περίπτωση που δεν είναι κενό, το μήνυμα προωθείται στον χρήστη και ο μεταγλωττιστής τερματίζει, ανεπιτυχώς (με αρνητικό κωδικό εξόδου).

Για την εκτέλεση όλων των παραπάνω εντολών χρησιμοποιήθηκε το module **subprocess** ενώ για τον χειρισμό των directories, των paths και των αρχείων χρησιμοποιήθηκε το module **os**, και πάλι σε μία προσπάθεια για cross-platform υλοποίηση.