Université Gustave Eiffel

Projet de Compilation

Licence d'informatique

—2019-2020—

Le but du projet est d'écrire un compilateur en utilisant les outils `flex` et `bison`.

Le langage source est un petit langage de programmation appelé TPC, qui ressemble à un sous-ensemble du langage C. Le langage cible est un sous-ensemble de l'assembleur `nasm` 64 bits. Vous vérifierez le résultat de la compilation d'un programme en exécutant le code obtenu.

Le projet est à faire en binôme ou seul. Si vous préférez le faire en binôme mais que n'avez pas de partenaire, contactez Eric Laporte. Chaque binôme réutilisera son projet d'analyse syntaxique du premier semestre et pourra le modifier [1]. Les dates limite de rendu sont :

- le lundi 27 avril 2020 à 23h55 au plus tard pour une première version du compilateur [2] avec les fonctionnalités décrites ci-dessous dans la section Travail demandé, Rendu intermédiaire (1/4 de la note de projet) ;
- le samedi 13 juin 2020 à 23h55 au plus tard pour le compilateur complet (3/4 de la note de projet).

1 Définition informelle du langage source

Un programme TPC est une suite de fonctions. Chaque fonction est constituée de déclarations de variables (locales à la fonction), et d'une suite d'instructions. Les fonctions peuvent être récursives. Il peut y avoir des variables de portée globale. Elles sont alors déclarées avant les fonctions.

Tout programme doit comporter la fonction particulière `main` par laquelle commence l'exécution. Les types de base du langage sont `int` (entier signé codé sur 4 octets) et `char`. Le mot clé `void` est utilisé pour indiquer qu'une fonction ne fournit pas de résultat ou n'a pas d'arguments. Les arguments d'une fonction sont transmis par valeur : pour passer une adresse, on déclare le paramètre comme étant de type pointeur, donc l'argument transmis est bien la valeur du pointeur.

Le langage TPC utilise `print` pour afficher un entier ou un caractère. Appliquer `print` à un pointeur est une erreur sémantique. Le mot-clé `readc` permet d'obtenir un caractère (`char`) lu au clavier ; `reade` permet de lire un entier (`int`) en notation décimale.

2 Définition des éléments lexicaux

Les identificateurs sont constitués d'une lettre, suivie éventuellement de lettres, chiffres, symbole souligné ("_"). Vous pouvez fixer une longueur maximale pour un identificateur. Il y a distinction entre majuscule et minuscule. Les mots-clés comme `if`, `else`, `return`, etc., doivent être écrits en minuscules. Ils sont reconnus par l'analyseur lexical et ne peuvent pas être utilisés comme identificateurs.

Les éléments lexicaux pour les constantes numériques sont des suites de chiffres.

Les caractères littéraux dans le programme sont délimités par le symbole `'`, comme en C. Dans les caractères littéraux, la barre oblique inverse ("\") est utilisée pour déspécialiser `'` et pour spécialiser `n` et `t` : `\n` et `\t` sont le caractère fin de ligne et la tabulation.

Les commentaires sont délimités par `/*` et `*/` et ne peuvent pas être imbriqués.

Les différents opérateurs et autres éléments lexicaux sont :

`=`	: opérateur d'affectation
`+`	: addition ou plus unaire
`-`	: soustraction ou moins unaire
`*`	: multiplication
`*`	: déréférencement de pointeur
`/` et `%`	: division et reste de la division entière
`!`	: négation booléenne

(1). Si vous avez déjà validé Analyse syntaxique l'an dernier, mais pas Compilation, vous devez quand même faire le projet d'analyse syntaxique (on ne peut pas faire de compilateur sans analyseur syntaxique)

(2). Déposez votre projet sur la plateforme elearning dans la zone prévue à cet effet.

`==`, `!=`, `<`, `>`, `<=`, `>=` : les opérateurs de comparaison
`&&`, `||` : les opérateurs booléens
`&` : adresse de variable
`;` et `,` : le point-virgule et la virgule
`(`, `)`, `{` et `}` : les parenthèses et les accolades

Chacun de ces éléments sera identifié par l'analyse lexicale, qui devra produire une erreur pour tout élément ne faisant pas partie du lexique du langage.

3 Notations et sémantique du langage

Dans ce qui suit,
- **CARACTERE** et **NUM** désignent respectivement un caractère littéral et une constante numérique ;
- **IDENT** désigne un identificateur ;
- **TYPE** désigne un nom de type qui peut être `int` ou `char` ;
- **EQ** désigne les opérateurs d'égalité ('`==`') et d'inégalité ('`!=`') ;
- **ORDER** désigne les opérateurs de comparaison '`<`', '`<=`', '`>`' et '`>=`' ;
- **ADDSUB** désigne les opérateurs '`+`' et '`-`' (binaire ou unaire) ;
- **OR** et **AND** désignent les deux opérateurs booléens '`||`' et '`&&`'.
- Les mots-clés sont notés par des unités lexicales qui leur sont identiques à la casse près.

L'instruction nulle est notée '**;**'.

4 Grammaire du langage TPC

Prog : DeclVars DeclFoncts ;
DeclVars : DeclVars **TYPE** Declarateurs **';'**
| ;
Declarateurs : Declarateurs **','** **IDENT**
| Declarateurs **','** **'*'** **IDENT**
| **IDENT**
| **'*'** **IDENT** ;
DeclFoncts : DeclFoncts DeclFonct
| DeclFonct ;
DeclFonct : EnTeteFonct Corps ;
EnTeteFonct : **TYPE IDENT** **'('** Parametres **')'**
| **TYPE '*' IDENT** **'('** Parametres **')'**
| **VOID IDENT** **'('** Parametres **')'** ;
Parametres : **VOID**
| ListTypVar ;
ListTypVar : ListTypVar **','** **TYPE IDENT**
| ListTypVar **','** **TYPE '*' IDENT**
| **TYPE IDENT**
| **TYPE '*' IDENT** ;
Corps : **'{'** DeclVars SuiteInstr **'}'** ;
SuiteInstr : SuiteInstr Instr
| ;
Instr : LValue **'='** Exp **';'**
| **READE** **'('** **IDENT** **')'** **';'**
| **READC** **'('** **IDENT** **')'** **';'**
| **PRINT** **'('** Exp **')'** **';'**
| **IF** **'('** Exp **')'** Instr
| **IF** **'('** Exp **')'** Instr **ELSE** Instr
| **WHILE** **'('** Exp **')'** Instr
| **IDENT** **'('** Arguments **')'** **';'**
| **RETURN** Exp **';'**
| **RETURN** **';'**
| **'{'** SuiteInstr **'}'**
| **';'** ;
Exp : Exp **OR** TB
| TB ;
TB : TB **AND** FB
| FB ;
FB : FB **EQ** M
| M ;
M : M **ORDER** E
| E ;
E : E **ADDSUB** T
| T ;
T : T **'*'** F
| T **'/'** F
| T **'%'** F
| F ;
F : **ADDSUB** F
| **'!'** F
| **'&'** IDENT
| **'('** Exp **')'**
| **NUM**
| **CHARACTER**
| LValue
| **IDENT** **'('** Arguments **')'**
| **'*'** **IDENT** **'('** Arguments **')'** ;
LValue : **IDENT**
| **'*'** **IDENT** ;
Arguments : ListExp
| ;
ListExp : ListExp **','** Exp
| Exp ;

5 Sémantique

La sémantique de la plupart des expressions et instructions du langage est la sémantique habituelle en langage C.

Tout identificateur utilisé dans un programme doit être déclaré avant son utilisation et dans la partie de déclaration appropriée.

De même, la grammaire n'impose pas que le programme comporte la fonction `main`, mais l'analyse sémantique doit vérifier sa présence.

Le typage des expressions est à peu près comme en C :

- Tout `char` auquel on applique une opération (sauf `print()`) est implicitement converti en `int`.
- Toute valeur peut être interprétée comme booléenne, avec la convention que 0 représente "faux" et tout autre entier "vrai", et toute expression à sens booléen est de type `int`. En particulier, l'opérateur de négation produit comme résultat l'entier 1 quand on l'applique à l'argument 0.
- Utiliser un pointeur comme opérande d'une opération arithmétique est une erreur sémantique.
- Utiliser un pointeur comme opérande d'une opération de comparaison **ORDER** est une erreur sémantique.
- Si on affecte un `int` à une LValue de type `char`, le compilateur émet un avertissement (*warning*).

Pour le typage, les instructions `return Exp`, `reade()` et `readc()` sont considérées comme des affectations ; dans les appels de fonctions, on considère que chaque paramètre effectif correspond à une affectation.

6 Langage cible

Le langage cible est un sous-ensemble de l'assembleur `nasm` 64 bits. Les commandes autorisées sont :

`mov`, `movsx`, `call`, `ret`, `syscall`,
`add`, `sub`, `idiv`, `imul`,
`and`, `or`, `xor`,
`cmp`, `je`, `jg`, `jne`, `jng`, `jmp`,
`push`, `pop`,
`resb`, `resd`, `resq`, `db`, `dd`, `dq`.

Les registres autorisés sont :

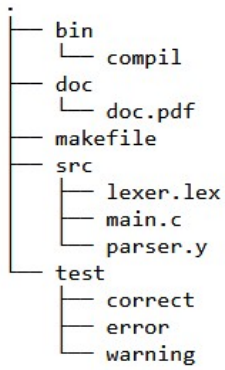
`rax`, `rbx`,
`rdi`, `rsi`, `rdx`, `rcx`, `r8`, `r9`,
`rsp`, `rbp`

et tous leurs sous-registres : `eax`, `ax`, `ah`, `al`, `r8d`, `r8w`, `r8b`...

7 Travail demandé

Écrire un compilateur de ce langage en utilisant `flex` pour l'analyse lexicale et `bison` pour l'analyse syntaxique et la traduction. Les messages d'erreur doivent donner le numéro de ligne et le numéro du caractère dans la ligne, puis reproduire la ligne et indiquer le caractère par une flèche verticale ou un circonflexe. Vous pourrez modifier la grammaire qui vous a été fournie, pour lever des conflits d'analyse ou pour faciliter la traduction, mais vos modifications ne peuvent s'écarter de la définition du langage TPC que si cela l'enrichit.

Le répertoire que vous déposerez doit être organisé correctement : un répertoire `src` pour les sources, un autre nommé `doc` pour la documentation, un autre nommé `test` pour les tests... Le répertoire racine du projet doit contenir un makefile nommé `makefile`. L'analyseur créé avec le makefile doit être nommé `compil`. Merci de respecter ces consignes et le schéma ci-dessous : cela facilitera l'examen de vos projets, augmentant ainsi la probabilité que les évaluateurs soient de bonne humeur, ce qui est toujours si avantageux pour ceux dont le travail va être évalué :



La commande suivante doit exécuter votre analyseur :

./compil [-o *prog***.asm] <** *prog***.tpc**

Le programme devra renvoyer 0 si et seulement si *prog***.tpc** est correct, c'est-à-dire ne contient aucune erreur, ni syntaxique, ni sémantique ; les avertissements ne comptent pas comme des erreurs.

Déposez votre projet sur la plateforme elearning dans la zone prévue à cet effet, sous la forme d'une archive tar compressée de nom "ProjetCompilationL3_NOM1_NOM2.tar.gz", qui, au désarchivage, crée un répertoire "ProjetCompilationL3_NOM1_NOM2" contenant le projet.

Rendu intermédiaire On demande une première version du compilateur qui doit au moins :

- construire la table des symboles et y mettre au moins les variables
- détecter les variables qui sont utilisées sans être déclarées et émettre les messages d'erreur
- typer les expressions et émettre les messages d'avertissement.

On demande aussi d'écrire trois jeux de tests, un pour les programmes TPC corrects sans avertissements, un autre pour les programmes incorrects, et un troisième pour les programmes corrects avec des avertissements. Enfin, on demande un script de déploiement des tests, qui produit un rapport unique donnant les résultats de tous les tests. Pour le rendu intermédiaire, on ne demande pas de documentation.

Pour le rendu final, on demande les fonctionnalités suivantes.

Fonctions Votre compilateur doit pouvoir traiter les programmes avec plusieurs fonctions et des appels de fonctions.

Pointeurs Il doit pouvoir initialiser les pointeurs, utiliser leur valeur comme addresse pour accéder à la mémoire, les passer comme paramètres des fonctions et comme valeur de retour.

Documentation Vous décrirez dans votre documentation vos choix et les difficultés que vous avez rencontrées.