



Création d'un langage interprété

Rapport d'élève ingénieur

Projet de 2^{ème} année

Filière F2 : Génie Logiciel et Systèmes Informatiques

Présenté par : **Franck ALONSO** et **Rémi CHASSAGNOL**

Responsable ISIMA :

Mardi 31/01/2023

Projet de 60h

Campus des Cézeaux. 1 rue de la Chébarde. TSA 60125. 63178 Aubière CEDEX

Remerciements

Nous tenons à exprimer notre profonde gratitude à :

- Notre encadrant et cher professeur M. Loïc YON pour la qualité de son encadrement, et pour nous avoir guidés durant toute la période du projet.
- Mme Murielle MOUZAT, notre professeur de communication pour son aide précieuse et indispensable pour la réussite de notre projet de 2^{ème} année.

Finalement, nous exprimons nos vifs remerciements à toute personne ayant participé de près ou de loin au bon déroulement de ce projet.

Table des matières

Résumé	2
Abstract	2
Introduction	3
Résumé des références	4
1 Conception d'un langage informatique	5
Grammaire de notre langage	5
2 Le lexeur	6
Définition	6
Implémentation	6
3 Le parseur	8
Définition	8
Implémentation	8

Résumé

L'objectif de ce projet est la création d'un langage **interprété** en **C++** dans le but de faire découvrir l'informatique et la programmation à des collégiens et lycéens. Ce projet rentre dans le cadre du sujet commun de la filière F2, qui concerne la conception d'outils et plus généralement en développement logiciel.

Le projet, réalisé sous Visual Studio Code, utilisera un **lexeur** et **parseur** pour pouvoir reconnaître notre langage de programmation voulu.

Mots-clés : **C++**, **interpréteur**, **lexeur**, **parseur**, **langage de programmation**

Abstract

The objective of this project is the creation of an **interpreted** language in **C++** in order to introduce computer science and programming to middle and high school students. This project falls within the framework of the common subject of the F2 major, which relates to the design of tools and more generally in software development.

The project, coded with Visual Studio Code, uses a **lexer** and **parser** to be able to recognize our desired programming language.

Keywords : **C++**, **interpreter**, **lexer**, **parser**, **programming language**

Introduction

Dans le cadre du projet de 2^{ème} année à l'ISIMA, nous avons choisi de réaliser un travail concernant le sujet commun de la filière F2, génie logiciel et systèmes d'informations. Le but du projet est de pouvoir montrer à des élèves de collèges et lycées ce que la filière ingénieur permet de faire.

Le domaine étant vaste, nous avons choisi de concevoir un langage informatique simple en C++ avec comme principal objectif de faire comprendre aux élèves la notion de fonction. Ce langage devait posséder une interface graphique, mais cette tâche s'est avérée trop ambitieuse pour un travail de 60 heures.

Le projet, en plus de sa valeur pédagogique, nous permet de nous familiariser avec les concepts de compilateur, interpréteur et arbre syntaxique.

Pour présenter ce projet, nous commencerons par la forme du langage à créer souhaité, puis nous détaillerons sa structure. Une fois familier avec les différentes étapes de son implémentation, nous introduirons les concepts et outils informatiques qui permettent la réalisation de notre langage.

Résumé des références

- La partie consultable de [1] présentent les bases de flex.
- [2] présente la construction d'un compilateur avec flex et bison. Le compilateur présenté utilise une **table des symboles** ainsi qu'une sorte de **byte code**. Nous avons choisi l'autre méthode qui consiste à utiliser un object-ABS plutôt que directement du byte code. Article très utilisé au départ pour la mise en place du parseur/lexeur.
- [3] : manuel d'utilisation de Flex.
- [4] : nous a permis d'avoir un exemple de code qui allie flex et bison en C++ et non en C.
- [5] explication du fonctionnement d'un compilateur.
- [6] première version de l'article précédent.
- [7] création d'un analyseur syntaxique pour du C/C++ : ASTROLOG. L'article par d'analyse syntaxique et de la construction d'**ABS**.
- [8] L'objectif de l'article est de présenter l'utilisation des **ABS** pour de la méta programmation. Il comporte pas mal d'exemples sur les **ABS** donc je le trouve pertinent.
- [9] : **ABS** en java.

1 Conception d'un langage informatique

Grammaire de notre langage

Avant d'implémenter notre langage informatique, il faut avoir une idée de sa grammaire, c'est-à-dire son fonctionnement. Un de ces objectifs est d'initier les élèves de collèges et lycées à la notion de fonction informatique.

C'est la raison pour laquelle les opérations permises par notre langage ne seront que des fonctions.

Exemple

Nous disposons de 2 variables **a** et **b**.

La variable **a** est lue par commande de l'utilisateur tandis que la valeur 4 est assignée à **b**.

Nous cherchons ensuite à additionner les 2 variables avec la valeur 5. Nous avons alors la syntaxe suivante :

- une fonction *add3* qui additionne les valeurs de ses 3 paramètres
- une fonction *affiche* qui affiche sur l'écran le nombre passé en paramètre
- une fonction *main*, où se trouve toutes les commandes voulues de notre programme

```
fn add3(int a, int b, int c) -> int {  
    return add(a, add(b, c));  
}  
  
fn affiche(int n) {  
    print("nombre : ");  
    print(n);  
}  
  
fn main() {  
    int a;  
    int b;  
    read(a);  
    set(b, 4);  
    affiche(add3(a, b, 5));  
}
```

La particularité de cette grammaire est que les opérations arithmétiques de base sont remplacées par des fonctions :

$\mathbf{a + b}$ devient <i>add(a, b)</i>
$\mathbf{a - b}$ devient <i>mns(a, b)</i>
$\mathbf{a \times b}$ devient <i>tms(a, b)</i>
$\mathbf{a \div b}$ devient <i>div(a, b)</i>

2 Le lexeur

Définition

Le lexeur, ou encore appelé analyseur lexical, a pour but de transformer le texte du code source en des unités lexicales, appelées *tokens*, comme expliqué par la partie consultable de [1].

Exemple

Pour l'expression simple **a = 2 * b**

Les tokens apparaissant sont :

Token	Sa nature
a	Identificateur de variable
=	Symbole d'affectation
2	Valeur entière
*	Opérateur de multiplication
b	Identificateur de variable

Le lexeur a également pour rôle de supprimer les informations inutiles, généralement du caractère espace et des commentaires.

Implémentation

L'outil utilisé pour générer un lexeur à partir du code précédent est Flex (Fast LEXical analyser generator). Au lieu d'écrire un lexeur à partir de zéro, Flex permet d'avoir un lexeur en donnant seulement les modèles des expressions régulières ainsi que le langage de travail (c++ dans notre cas).

[2] a fourni un squelette pour la réalisation du fichier nécessaire à Flex.

Le fichier **main_cpp.l** contient le code qui permet de générer le lexeur avec Flex. Les token doivent être définis dans **main_cpp.y** au préalable. À noter que l'on peut utiliser la variable `yylval` pour transmettre des éléments au parser.

La structure du fichier **main_cpp.l** est la suivante :

```
C and parser declaration

%%
Grammar rules and actions

%%
C subroutings
```

On peut définir des règles dans les déclarations du lexeur :


```
%option c++ interactive noyywrap noyylineno nodefault outfile="lexer.cpp"

alpha [a-zA-Z]
digit [0-9]
int  [+]?{digit}+
float  [+]?{digit}+\.{digit}+
char  '{alpha}'
identifiant  [a-z]({alpha}|{digit}|_)*
```

Concernant la ligne d'option :

- **c++** : indique qu'on travaille avec du c++ et non du c
- **interactive** : utile quand on utilise **std::in**. Le scanner interactif regarde plus de caractères avant de générer un token (plus lent mais permet de lutter contre les ambiguïtés)
- **noyywrap** : ne pas appeler **yywrap()** qui permet de parser plusieurs fichiers
- **noyylineno** : désactive l'enregistrement des lignes (**yylineno**)
- **nodefault** : pas de scanner par défaut (=> on doit tout implémenter)
- **outfile : "file.cpp"** : permet de définir le fichier de sortie

Dans les règles, on suit toujours le même principe, on indique les caractères à reconnaître puis on exécute du code :

```
for      { AFFICHE("L_for"); return Parser::token::FOR; }
{identifiant} {
    AFFICHE("L_id");
    yyval->build<std::string>(yytext);
    return Parser::token::IDENTIFIER;
}
```

Ici on a accès à la variable **yyval** qui est un **Parser::semantic_type*** et qui possède une méthode **build** qui nous permet de transmettre des valeurs à bison.

Ces valeurs sont accessibles via les variables de bison : **\$2**. La variable **yytext** contient le texte traité par Flex. De plus, dans le code, on retourne les *tokens*. Ces **tokens** sont définis dans le fichier de bison.

La fonction appelée par défaut est **yylex**, cependant, pour pouvoir travailler avec bison, nous devons fournir nos propres fonctions, pour ce faire on utilise la macro **YY_DECL**, comme expliqué dans la partie *9 The Generated Scanner* du manuel pour Flex [3].

```
#define YY_DECL int interpreter::Scanner::lex(interpreter::Parser::semantic_type *yyval)
```

3 Le parseur

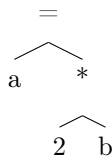

Définition

Également appelé analyseur syntaxique, son rôle principal est la vérification de la syntaxe du code en regroupant les tokens selon une structure suivant des règles syntaxiques.

Exemple

Pour l'expression simple **a = 2 * b**

Les tokens apparaissant sont :

Arbre syntaxique	Évaluation de 2 * b	Affectation de a
		a = 2 * b

Implémentation

À l'instar de Flex pour le lecteur, Bison est un générateur de grammaire qui convertit une description de grammaire en un programme C++ qui analyse cette même grammaire.

L'article [2] s'est encore une fois révélé très utile pour la réalisation du fichier nécessaire à Bison

Le fichier **main_cpp.y** contient le code qui permet de générer le parseur avec Bison. Toutes les règles syntaxiques qui définissent la grammaire du langage y sont comprises. Chaque règle va contenir des blocks de code qui seront exécutés au moment où le parseur la reconnaît, ce code permet de créer des objets qui formeront l'ABS (Abstract Syntactic Tree) du programme.

La structure du fichier **main_cpp.y** est identique au lecteur :

```
C and parser declaration

%%
Grammar rules and actions

%%
C subroutines
```

Les tokens sont définis en début de fichier avec la syntaxe suivante :

```
%token IF ELSE FOR WHILE FN INCLUDE IN
%token <long long> INT
%token <double> FLOAT
%token <char> CHAR
%token <std::string> IDENTIFIER
```

À noter que l'on peut spécifier le type de l'élément, ce qui sera utile pour récupérer les valeurs retournées par le lecteur.

Bison permet de construire le parser, qui va reconnaître des éléments de syntaxe et non pas juste des mots clés. Par exemple, on peut définir une règle pour reconnaître une suite d'inclusion de fichiers :

```
includes: %empty
        | INCLUDE IDENTIFIER SEMI includes
        ;
```

Ici, on définit une règle **includes** qui décrit la syntaxe des *includes*. Selon cette règle, une suite d'inclusions est soit vide, soit elle comporte une inclusion, suivit d'une suite d'inclusion ('|' signifie "ou"). Il faut noter que la syntaxe est "récursive", ce qui nous permet de définir une suite d'éléments. Enfin, les mots en majuscule sont les tokens retournés par le lexeur.

On peut ajouter des blocks de codes qui seront exécutés au moment où le parseur atteint l'élément qui précède le block. Dans l'exemple ci-dessous, le block sera appelé une fois que Bison aura parser le `;`. À noter que l'on peut accéder aux éléments retournés par Flex; ici, `$2` fait référence au second élément de la règle qui est **IDENTIFIER**. Le type de **IDENTIFIER** a été défini comme étant une `std::string`. Le block de code nous permet donc de créer une nouvelle inclusion et de récupérer le nom de la bibliothèque.

```
includes: %empty
        |
        INCLUDE IDENTIFIER SEMI
        {
            std::cout << "new include id: " << $2 << std::endl;
            pb.addInclude(std::make_shared<Include>($2));
        }
        includes
        ;
```

Comme dit plus haut, on peut définir des types pour les tokens, ce qui permet de récupérer des valeurs :

```
value: INT {
    std::cout << "new int: " << $1 << std::endl;
    lastValue.i = $1;
    lastValueType = INT;
} | FLOAT {
    std::cout << "new double: " << $1 << std::endl;
    lastValue.f = $1;
    lastValueType = FLT;
} | CHAR {
    std::cout << "new char: " << $1 << std::endl;
    lastValue.c = $1;
    lastValueType = CHR;
}
;
```

Pour la génération du code, on a deux options :

- utiliser des instructions très simples => sorte de bytecode
- créer un code objet où tous les éléments sont des objets.

Choix de la représentation objet :

- plus simple à comprendre et à visualiser
- plus compliqué à générer : on peut générer du bytecode au fil de l'exécution du parseur, en utilisant des `goto` pour sauter de block d'instruction en block d'instruction. Pour le code objet, les éléments à l'intérieurs des blocks doivent être créés avant le block, et le block est détecté avant les instructions, il faut donc stocker les instructions.

Bibliographie

- [1] J. LEVINE, *Flex & Bison : Text Processing Tools*. " O'Reilly Media, Inc.", 2009. adresse : https://books.google.fr/books?hl=fr&lr=&id=nYUkAAAAQBAJ&oi=fnd&pg=PR3&dq=flex+bison+interpreter&ots=VX9xrg4D9l&sig=p95S6sNhMxdIIl-7u00nK_1u-fM&redir_esc=y#v=onepage&q=flex%20bison%20interpreter&f=false.
- [2] A. A. AABY, "Compiler construction using flex and bison," *Walla Walla College*, 2003. adresse : <http://penteki.web.elte.hu/compiler.pdf>.
- [5] H.-P. CHARLES et C. FABRE, "Compilateur," *Techniques de l'ingénieur Technologies logicielles Architectures des systèmes*, t. base documentaire : TIP402WEB. N° ref. article : h3168, 2017, fre. DOI : 10.51257/a-v2-h3168. eprint : basedocumentaire:TIP402WEB.. adresse : <https://www.techniques-ingenieur.fr/base-documentaire/technologies-de-l-information-th9/systemes-d-exploitation-42305210/compilateur-h3168/>.
- [6] B. LORHO, "Compilateurs," *Techniques de l'ingénieur Technologies logicielles Architectures des systèmes*, t. base documentaire : TIP402WEB. N° ref. article : h3168, 1996, fre. DOI : 10.51257/a-v2-h3168. eprint : basedocumentaire:TIP402WEB.. adresse : <https://www.techniques-ingenieur.fr/base-documentaire/technologies-de-l-information-th9/systemes-d-exploitation-42305210/compilateur-h3168/>.
- [7] R. F. CREW et al., "ASTLOG : A Language for Examining Abstract Syntax Trees," t. 97, p. 18-18, 1997. adresse : https://www.usenix.org/legacy/publications/library/proceedings/ds197/full_papers/crew/crew.pdf.
- [8] E. VISSER, "Meta-programming with concrete object syntax," p. 299-315, 2002. adresse : https://dspace.library.uu.nl/bitstream/handle/1874/23952/visser_02_metaprogramming.pdf?sequence=2.
- [9] E. M. GAGNON et L. J. HENDREN, *SableCC, an object-oriented compiler framework*. IEEE, 1998. adresse : https://central.bac-lac.gc.ca/.item?id=MQ44169&op=pdf&app=Library&oclc_number=46811936.

Webographie

- [3] E. M. GAGNON et L. J. HENDREN. "Lexical Analysis With Flex, for Flex 2.6.2." (2016), adresse : https://westes.github.io/flex/manual/index.html#SEC_Contents.
- [4] CPPTUTOR. "Generating C++ programs with flex and bison." (2020), adresse : <https://learnmoderncpp.com/2020/12/18/generating-c-programs-with-flex-and-bison-3/>.