



Création d'un langage Transpileur

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'une grammaire	5
Définition de la grammaire	5
Les variables	5
Les fonctions	5
Les structures de contrôle	5
Les opérations	5
Inclusion de fichier	6
Entrée et sortie	6
Exemple de code	6
Arbre syntaxique	7
Les blocs de code	7
Les conteneurs d'opérations	7
Les opérations	7
2 Construction de l'AST	8
Le lexeur	8
Définition	8
Exemple	8
Implémentation	9
Le parseur	11
Définition	11
Implémentation	11
Fabrique à programme	14
3 Analyse des symboles	14
Création d'une table des symboles	14
Traitement des des symboles dans le parseur	14
4 Le transpileur	14
5 Conception d'un langage informatique	14
A Diagramme de classes de l'AST	16

Résumé

L'objectif de ce projet est la création d'un langage **transpilé** 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.

Ce projet s'inscrit dans le cadre du projet commun de la filière 2. L'objectif du projet commun est la création d'outils de démonstration servant à présenter la filière génie logiciel. Notre travail a consisté en la création d'un langage de programmation transpilé. Pour ce faire, nous avons dû étudier des concepts et des outils utilisés pour la création de compilateurs.

Mots-clés : **C++**, **transpileur**, **lexeur**, **parseur**, **flex**, **bison**, **langage de programmation**

Abstract

The objective of this project is the creation of an **transpiled** 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.

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 2, génie logiciel et systèmes d'informations. Le but du projet commun est la création d'outils de démonstration servant à présenter la filière.

Nous souhaitions au départ, créer un langage de programmation interprété ainsi qu'un IDE, cependant, ce projet s'est avéré trop ambitieux pour être réalisé en 60 heures. Pour simplifier, nous avons décidé de concevoir un langage transpilé, déléguant ainsi une partie des tâches complexes comme la gestion de la mémoire à un compilateur existant.

Ce projet permettra de présenter un langage de programmation très simple pour introduire des lycéen à la programmation. De plus, il pourra constituer une maquette pour les élèves de l'ISIMA souhaitant étudier le fonctionnement des compilateurs.

Pour présenter ce projet, nous commencerons par la forme du langage à créer souhaité, puis nous détaillerons sa structure. Une fois familiarisé 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'une grammaire

Définition de la grammaire

Avant d'implémenter notre langage informatique, il faut avoir une idée de sa grammaire. Puisque nous souhaitons utiliser ce langage à des fins pédagogiques, nous avons décidé de simplifier au maximum la syntaxe. Par exemple, nous avons choisi de supprimer tous les opérateurs, ainsi, toutes les opérations arithmétiques et booléennes auront la même syntaxe que les fonctions. Par exemple, $a + b$ s'écrira `add(a, b)`. De plus, pour différer au maximum de python, notre syntaxe s'inspirera de celle des langages **C** et **Rust**. Pour définir la grammaire de manière formelle, nous utiliserons la forme de Backus-Naur.

Les variables

Pour stocker des données, notre langage utilise des variables dont la convention de nommage est la même qu'en **C**. Le langage est à typage statique, ce qui signifie que toutes les variables doivent être déclarées avec leur type avant d'être utilisées. Pour l'instant, nous possédons trois types : les entiers (*int*), les nombre à virgule flottante (*flt*) et les caractères (*chr*). Voici la syntaxe pour déclarer une variable :

```
<identifiant> ::= 'a'-'z'( <alpha> | '0'-'9' )*
<int> ::= "int"
<flt> ::= "flt"
<chr> ::= "chr"
<type> ::= <int> | <flt> | <chr>
<declaration> ::= <type> <identifiant> ";"
```

Les fonctions

Pour définir des fonctions, nous utilisons le mot clé **fn**, suivi du nom de la fonction avec ses paramètres entre parenthèses et enfin le bloque de code qui contient les instructions entre accolade. De plus, il faut spécifier le type de la valeur retour de la fonction en utilisant **-> type** quand c'est nécessaire.

```
<function> ::= fn <identifiant> "("<parameters>)" [ "->" <type> ] "{"<instructions>"}"
```

Pour pouvoir retourner une valeur, il faudra utiliser le mot clé *return* dans la fonction.

Les structures de contrôle

Nous avons aussi ajouté les structures de contrôle présentes dans la plupart des langages de programmation.

```
<if> ::= if "("<condition>)" "{"<instructions>"}" [ else "{" <instructions>"}" ]

<range> ::= range "("<valueSymbol>, <valueSymbol>, <valueSymbol>)"
<for> ::= for <identifiant> in <range> "{"<instructions>"}"

<while> ::= while "("<condition>)" "{" <instructions> "}"
```

Les opérations

Comme dit précédemment, tous le langage ne comporte pas d'opérateur donc toutes les opérations se font à l'aide de fonctions directement intégrées dans le transpileur. Voici une liste des opérations possibles :

opération	syntaxe
$\mathbf{a} + \mathbf{b}$	<i>add</i> (a , b)
$\mathbf{a} - \mathbf{b}$	<i>mns</i> (a , b)
$\mathbf{a} \times \mathbf{b}$	<i>tms</i> (a , b)
$\mathbf{a} \div \mathbf{b}$	<i>div</i> (a , b)
$\mathbf{a} == \mathbf{b}$	<i>eql</i> (a , b)
$\mathbf{a} < \mathbf{b}$	<i>inf</i> (a , b)
$\mathbf{a} > \mathbf{b}$	<i>sup</i> (a , b)
$\mathbf{a} \leq \mathbf{b}$	<i>ieq</i> (a , b)
$\mathbf{a} \geq \mathbf{b}$	<i>seq</i> (a , b)
not a	<i>not</i> (a)

Inclusion de fichier

TODO : Non fonctionnel :/

Entrée et sortie

Le langage possède aussi la possibilité d’afficher et de lire du texte depuis la console. La lecture se fait avec la fonction *read* qui prend en paramètre une variable qui stockera le résultat. Pour la l’affichage, il faudra utiliser la fonction *print* qui peut être utilisée de deux manières :

- `print("Hello, World!")` : affiche du texte
- `print(variable)` : affichage d’une valeur (depuis une variable ou une fonction).

Exemple de code

Dans cet 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));
}
```


Arbre syntaxique

Pour donner un sens aux éléments textuels du langage, nous utiliserons une représentation sous forme d'arbre. On appelle cela un arbre syntaxique ou AST (Abstract Syntax Tree). Les arbres syntaxiques sont nés avec la théorie des langages et comme on peut le voir dans cet article [5], les AST sont très utilisés par les compilateurs car ce sont des structures plus simples à manipuler que du texte. Voici l'arbre syntaxique de la fonction *add3* :

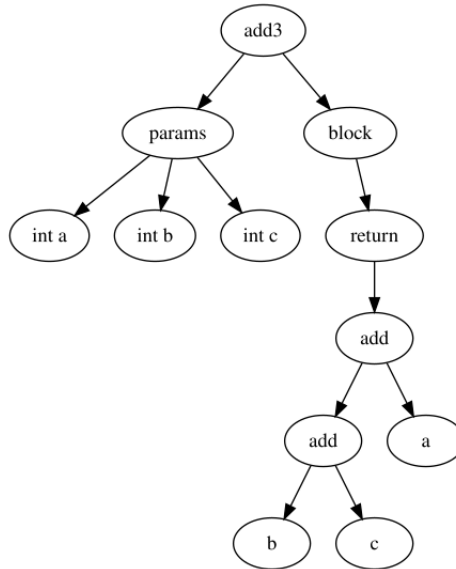


FIGURE 1 – Exemple d'arbre syntaxique

Pour cette partie, nous nous sommes inspiré d'un projet appelé **minijava** [10] ainsi que d'un article [9] qui couvre aussi l'utilisation des AST en java. Nous avons choisi l'approche objet, avec un AST représenté par des classes, où chaque nœud de l'arbre a sa propre classe.

Toutes les classes de l'arbre sont stockées dans le fichier AST/AST.hpp et héritent toutes d'une classe abstraite ASTNode. L'emploi de l'héritage ici est très important car nous profiterons par la suite des avantages du polymorphisme pour stocker des opérations de différents types. La classe ASTNode est abstraite car elle ne doit pas être instanciée, il faut que chaque nœud ait un type concret utilisable dans le transpileur. Le diagramme de classes complet est disponible en annexe A. Détaillons maintenant les parties importantes.

Les blocs de code

La classe Block correspond aux blocs de code entre accolades. Elle possède une liste de nœuds qui sont les opérations contenues dans le bloc.

Les conteneurs d'opérations

La classe Statement correspond aux éléments qui contiennent des blocs de code. Cela comprend les fonctions et les structures de contrôles. La grammaire ne permet pas l'emploi des blocs ailleurs.

Les opérations

Les opérations sont traitées avec la classe OperationBinaire qui correspond à tous les opérateurs.

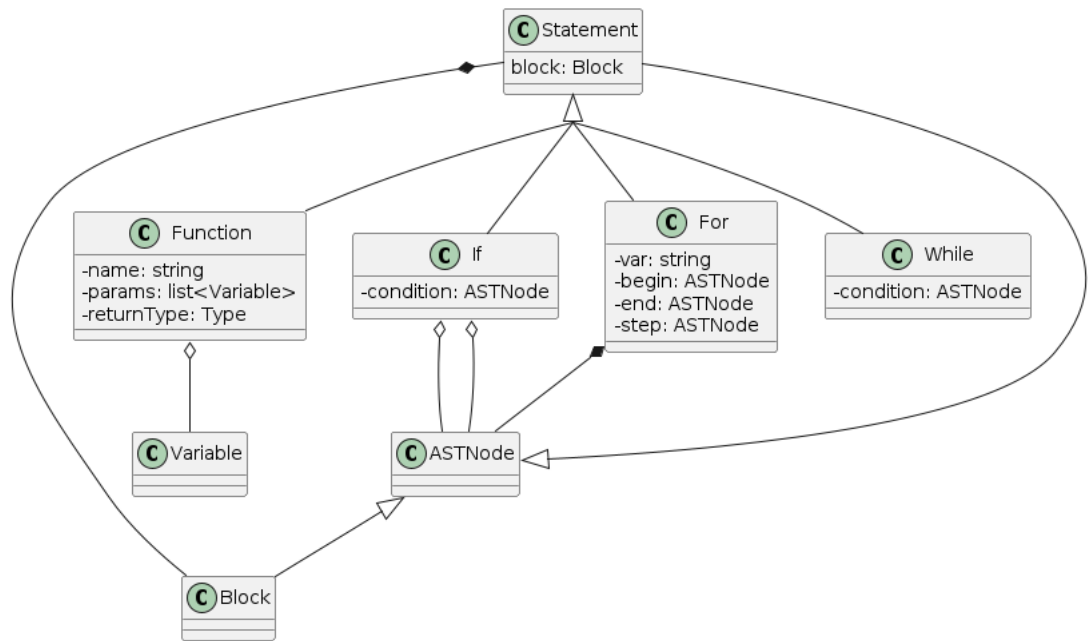


FIGURE 2 – Classe Statement

2 Construction de l'AST

Dans cette partie nous expliquerons la création d'un parseur en utilisant les outils **GNU Flex** et **GNU Bison**. De plus, nous détaillerons la génération de l'AST en utilisant les classes décrites précédemment. Pour cette partie, nous nous sommes aidé d'un article [2] qui présente la construction d'un compilateur avec Flex et Bison en C. Pour porter le code en C++, nous avons utilisé la documentation officielle des deux outils ainsi qu'un article qui nous a permis d'avoir un squelette de base pour le lexeur et le parseur [4].

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* [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 blancs (espaces et tabulations) et des commentaires.

Implémentation

L'outil utilisé pour générer le lexeur est **GNU Flex** (Fast LEXical analyser generator). Il permet de générer le code C++ du lexeur à partir d'un fichier. Dans notre cas, le fichier utilisé sera **main_cpp.l** et possède la structure suivante [2] :

```
// code C++, options et declarations de raccourcis

%%
// Definition des tokens et actions

%%
// Fonctions C++
```

Dans la première partie en haut du fichier, on place les inclusions de bibliothèques C/C++ ainsi que des options pour flex.

```
%{
#include "parser.hpp"
#include "lexer.hpp"
}%

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

Détail des fonctions utilisée :

- **c++** : indique qu'on travail avec du cpp 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
- **nodefault** : pas de scanner par défaut (=> on doit tout implémenter)
- **outfile : "file.cpp"** : permet de définir le fichier de sortie

Après les options on peut définir des raccourcis en utilisant des expressions régulières. Par exemple, dans le code ci-dessous, nous avons défini les règles suivantes :

- **alpha** : un caractère alphabétique est composé d'une lettre minuscule ou d'une lettre majuscule.
- **digit** : les chiffres sont les caractères entre 0 et 9.
- **int** : les entiers correspondes à une suite de chiffres et peuvent être positifs ou négatifs.
- **float** : similaire aux entiers sauf qu'ici on a obligatoirement un point suivit d'une suite de chiffre à la fin.
- **char** : les caractère sont toujours écrit entre ' (par exemple 'a').
- **identifier** : correspond aux noms de fonctions et de variables et suit le standard du C. Un identifiant commence par une lettre minuscule et peut être suivit d'une suite de lettres, de nombre et de _.

```

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

```

Dans la seconde partie du fichier, on définit des règles et des actions. À noter que l'on peut utiliser les raccourcis définis précédemment en mettant leurs noms entre accolades comme fait ci-dessous pour **identifieur**. La définition d'une règle suit le principe suivant, on commence par donner une suite de caractères qui sera consommée. Ensuite met du code entre accolades, et ce code sera exécuté quand le lecteur consommera la chaîne. Par exemple, si on prend la première ligne ci-dessous, quand le lecteur trouvera le mot **for**, il affichera *L_for* dans le terminal puis retournera le token **FOR**.

À noter que les tokens disponibles dans l'espace de nom **Parser : :token** doivent être définis dans le fichier bison que l'on détaillera dans la partie suivante.

```

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

```

Ici on a accès à la variable **yylval** de type **Parser : :semantic_type*** qui possède une méthode **build** permettant de transmettre des valeurs à 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(Parser::semantic_type *yylval, Parser::
    location_type *yylloc)

```

Cette macro permet de définir le type de la fonction **lex**, ici, on a pour paramètre **yylval** (qui comme dit précédemment permet de transmettre des valeurs à bison) et **yylloc** qui doit être fourni quand on utilise les positions dans le parseur (pour avoir le numéro de ligne en cas d'erreur par exemple), mais cela nécessite une option particulière pour bison.

Le parseur

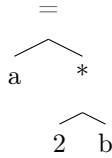

Définition

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, suivie 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 parsé 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.

Fabrique à programme

3 Analyse des symboles

Création d'une table des symboles

Traitement des des symboles dans le parseur

4 Le transpileur

5 Conception d'un langage informatique

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=p95S6sNhMxdI1l-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/dsl97/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/>.
- [10] J. P. JOAO CANGUSSU et V. SAMANTA. “Modern Compiler Implementation in Java : the MiniJava Project.” (2002), adresse : <https://www.cambridge.org/resources/052182060X/#java>.

A Diagramme de classes de l'AST

