

DM COMP : Création d'un parseur d'expressions arithmétiques

Aurèle Barrière & Antonin Garret

29 septembre 2016

1 Introduction

Le but de ce projet est d'implémenter un parseur d'expressions arithmétiques en **OCaml**, en utilisant les **stream parsers** de l'extension **Camlp4**. Le but est ainsi de convertir une chaîne de caractères (écrite sur un alphabet comportant des nombres et des opérateurs arithmétiques) en un arbre syntaxique abstrait donnant une représentation correcte (c'est-à-dire qui respecte les propriétés de priorité et d'associativité voulues) de l'expression.

On effectuera une analyse descendante.

Dans une première partie, nous verrons comment aboutir à une grammaire d'expressions arithmétiques prête à être implémentée. Ensuite, nous verrons comment l'implémenter. Enfin, nous nous intéresserons à la vérification empirique de notre implémentation.

2 Une grammaire pour des expressions arithmétiques

Pour implémenter notre parseur avec analyse descendante avec les **stream parsers**, il faut définir un ensemble de fonctions mutuellement récursives qui décomposent la chaîne de caractère souhaitée. Dans un premier temps, la chaîne de caractère aura été parcourue par un *scanner* (ou **lexer**) qui repère les mot-clés (ou **tokens**), et qui convertit donc une chaîne de caractères en une chaîne de mot-clés.

Les fonctions mutuellement récursives prennent donc en entrée une chaîne de mot-clés, et renvoient un arbre syntaxique abstrait d'expression, qu'il nous faut définir. Une expression sera ainsi : un nombre ou une addition ou soustraction ou multiplication ou division ou puissance de deux expressions. On aura donc en **OCaml** :

```
type expr = Num of int | Add of expr * expr | Sub of expr * expr | Mul
  of expr * expr | Div of expr * expr | Pow of expr * expr
```

2.1 Une première version

Pour parser correctement des expressions arithmétiques dans lesquelles figurent l'addition, la soustraction, la multiplication et des parenthèses, en respectant les propriétés suivantes :

- L'addition, la soustraction et la multiplication sont associatives à gauche.
- Les parenthèses sont prioritaires, puis la multiplication, puis l'addition (ou soustraction).

On dispose d'une première grammaire. Les récursivités gauches témoignent de l'associativité gauche des opérateurs. On a décomposé nos expressions en plusieurs sous-expressions : (E : expressions, F : facteurs, A : atomes) en suivant l'ordre inverse de priorité. Les dépendances cycliques entre ces différentes fonctions récursives nous obligeront à les déclarer comme des fonctions mutuellement récursives.

$$\begin{aligned} E &\rightarrow E + F \mid E - F \mid F \\ F &\rightarrow F * A \mid A \\ A &\rightarrow \text{constante} \mid (E) \end{aligned}$$

2.2 Ajout de la division

La division est un opérateur qui doit être associatif à gauche et avoir la même priorité que la multiplication. On ajoute donc la règle $F \rightarrow F / A$.

$$\begin{aligned} E &\rightarrow E + F \mid E - F \mid F \\ F &\rightarrow F * A \mid F / A \mid A \\ A &\rightarrow \text{constante} \mid (E) \end{aligned}$$

2.3 Récursivités gauches

Cependant, la méthode utilisée est l'analyse descendante. On ne peut donc pas garder nos règles avec récursivité gauche. Il faut donc créer de nouvelles règles pour les transformer en récursivités droites.

$$\begin{aligned} E &\rightarrow F E_aux \\ E_aux &\rightarrow + F E_aux \mid - F E_aux \mid \epsilon \\ F &\rightarrow P F_aux \\ F_aux &\rightarrow * A F_aux \mid / A F_aux \mid \epsilon \\ A &\rightarrow \text{constante} \mid (E) \end{aligned}$$

2.4 Ajout de l'opérateur de puissance

Ensuite, pour ajouter un opérateur de puissance, de telle sorte qu'il soit prioritaire sur la multiplication (mais pas sur les parenthèses) et associatif à droite, on ajoute une fonction supplémentaire entre les facteurs et les atomes. Comme elle utilise la récursivité droite, il n'est pas nécessaire de la transformer comme les précédentes. On obtient donc la grammaire :

$$\begin{aligned} E &\rightarrow F E_aux \\ E_aux &\rightarrow + F E_aux \mid - F E_aux \mid \epsilon \\ F &\rightarrow P F_aux \\ F_aux &\rightarrow * P F_aux \mid / P F_aux \mid \epsilon \\ P &\rightarrow A ^ P \mid A \\ A &\rightarrow \text{constante} \mid (E) \end{aligned}$$

2.5 Changement de la règle de puissance

Cependant, la documentation des **stream parsers** précise que pour une même fonction, on ne peut pas commencer deux schémas par les mêmes mot-clés. Or, P se décompose en $A \wedge P$ ou A , qui commencent tous deux par A . On peut utiliser les paramètres d'exception **Stream.error** pour se rendre compte qu'en effet, l'analyse descendante attendra le reste de la règle ($\wedge P$) même si l'expression devrait utiliser le deuxième schéma de P .

Pour résoudre ce problème, on décompose la règle en deux règles :

$$\begin{aligned} P &\rightarrow A \ P_aux \\ P_aux &\rightarrow \wedge P \mid \epsilon \end{aligned}$$

2.6 Grammaire finale

Ainsi, la grammaire obtenue est :

$$\begin{aligned} E &\rightarrow F \ E_aux \\ E_aux &\rightarrow + \ F \ E_aux \mid - \ F \ E_aux \mid \epsilon \\ F &\rightarrow P \ F_aux \\ F_aux &\rightarrow * \ P \ F_aux \mid / \ P \ F_aux \mid \epsilon \\ P &\rightarrow A \ P_aux \\ P_aux &\rightarrow \wedge \ P \mid \epsilon \\ A &\rightarrow \text{constante} \mid (\ E \) \end{aligned}$$

3 Implémentation

La grammaire obtenue n'a plus qu'à être recopiée avec la syntaxe des **stream parsers**.

On notera tout de même qu'il convient d'envoyer en argument l'expression déjà lue par une fonction principale aux fonctions auxiliaires, qui en auront besoin pour construire l'expression.

4 Vérification

Une fois le parseur implémenté, nous avons souhaité le tester pour vérifier sa correction. Dans un premier temps, nous avons vérifié chaque associativité, chaque priorité en donnant divers exemples. Mais nous avons souhaité automatiser le processus de vérification.

Ainsi, nous avons donc un évaluateur, qui à partir d'une expression (sous forme d'arbre syntaxique abstrait), calcule le résultat.

Dans un second temps, nous avons créé un générateur aléatoire d'expressions arithmétiques (sous forme de chaînes de caractères) en **Ocaml**.

Enfin, un programme en **bash** appelle le générateur et récupère une chaîne de caractères. Puis cette chaîne de caractère est scannée, analysée puis évaluée par notre parseur. D'un autre côté, on évalue également l'expression en **Python** (qui utilise les mêmes priorités et associativité que celles de notre parseur). Enfin, on compare les résultats.

La vérification automatique nous a permis de ...

Comme l'évaluateur de **Python** et notre évaluateur en **Ocaml** ne gèrent pas de la même manière les dépassements mémoire, il nous arrive d'avoir des résultats qui diffèrent lorsque le résultat (ou des sous-résultats) dépassent la capacité maximale de nos entiers.

Il ne s'agit pas d'une vérification exhaustive de toutes les expressions correctes que pourraient analyser notre parseur. En effet, nous avons ajouté des conditions de profondeur maximale dans la génération d'une expression, pour éviter les dépassements mémoires la plupart du temps. Cependant ...

5 Conclusion

Pour réaliser ce parseur, il nous a fallu nous intéresser au fonctionnement de l'analyse descendante, pour adapter nos règles à la méthode utilisée (pas de récursivités gauches par exemple). Il a fallu également s'adapter à l'outil utilisé, les **stream parsers** de **Camlp4** (pas de schémas qui commencent par la même séquence).

Le développement d'un outil de vérification empirique nous a permis de nous rendre compte de certains problèmes dans notre implémentation, et de nous convaincre de sa robustesse.