Analyse lexicale

Cyril Rabat
cyril.rabat@univ-reims.fr

Licence 3 Informatique - Info0602 - Langages et compilation

2020-2021

Cours n°2

Qu'est-ce qu'un analyseur lexical ?

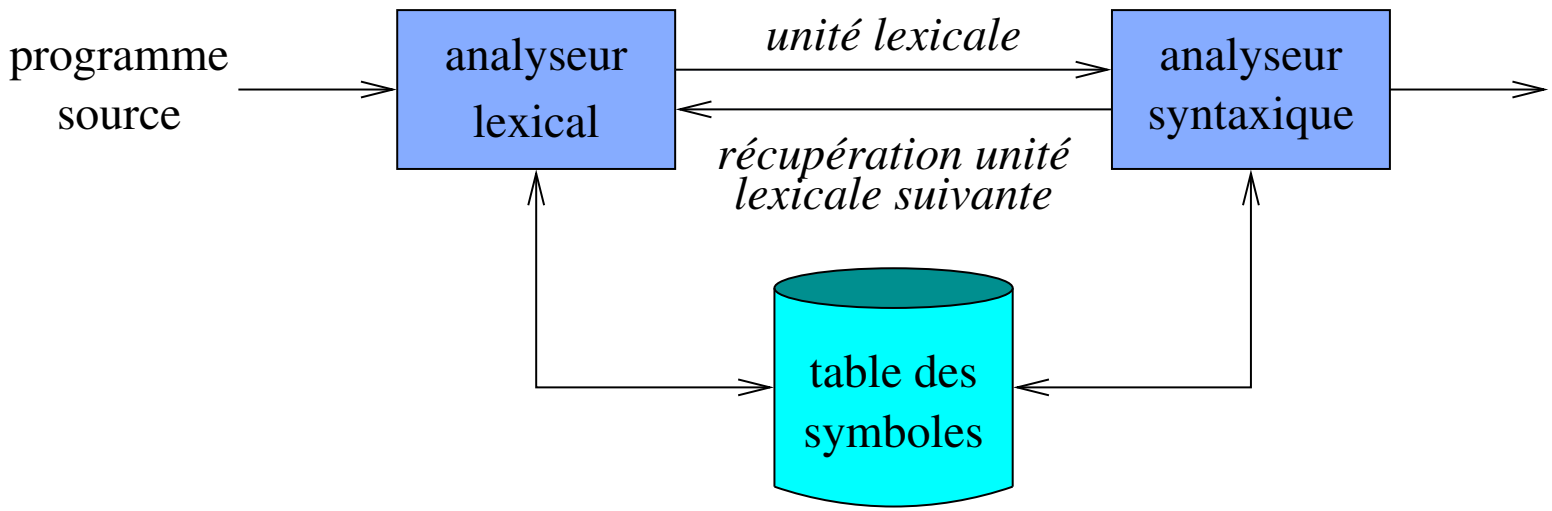
Langages, expressions régulières, automates finis

Version 14 décembre 2020

Table des matières

Un analyseur lexical

- Lecture des caractères en entrée
- Production d'unités lexicales
  $\hookrightarrow$ Analysées par l'analyseur syntaxique
- Interactions entre les deux analyseurs (lexical et syntaxique) :



Tâches secondaires de l'analyseur lexical

- Éliminer :
  - Les commentaires
  - Les caractères "inutiles" (espaces, tabulations, lignes vides. . .)
- Faciliter la gestion des erreurs :
  - Conservation/calcul du numéro de ligne
  - Associer les messages d'erreur à une ligne
- Le principal intérêt de l'analyseur lexical est de simplifier l'analyseur syntaxique

Modèle et unité lexicale

Définition : modèle

Règle qui décrit un ensemble de chaînes

Exemples

- $[0-9]^*$
- $a^* b$

Définition : unité lexicale

Éléments produit par l'ensemble des chaînes du modèle

Exemples

- mots-clés, opérateurs, identificateurs, constantes, chaînes littérales...

Lexème et attribut

Définition : lexème

Suite de caractères du programme source qui correspond au modèle

Exemples

- Modèle : $[0-9]^*$
- Unité lexicale : 100, 001, 123

Définition : attribut

Données liées aux unités lexicales

Exemple

- L'entrée dans la table des symboles pour un identificateur

Alphabet et mots

Définition : alphabet

*Un **alphabet** est un ensemble fini de symboles appelés caractères. Il est noté $\mathcal{A}$.*

- Exemples de symboles : lettres et caractères
- Exemples d'alphabets : $\{0, 1\}$ (l'alphabet binaire), l'ASCII

Définition : mot (ou chaîne)

*Un **mot** sur un alphabet est une séquence finie de symboles de cet alphabet. La **longueur du mot** w (notée $|w|$) est le nombre de symboles dans ce mot. Le **mot vide**, noté ϵ, est un mot de longueur 0.*

- Exemples de mots sur l'alphabet $\{a, b, c\}$: a, $baba$

Partie de mots

- **Préfixe** de w : mot obtenu en supprimant un nombre quelconque de symboles en fin de w (voire aucun)
- **Suffixe** de w : mot obtenu en supprimant un nombre quelconque de symboles en début de w (voire aucun)
- **Sous-mot** de w : mot obtenu en supprimant un préfixe et un suffixe de w
- **Préfixe propre** de w : tout mot non vide x, préfixe de w tel que $x \neq w$
- Idem pour **suffixe propre** et **sous-chaîne propre** de w
- **Sous-suite** de w : tout mot obtenu en supprimant un nombre quelconque de symboles de w, éventuellement aucun, pas nécessairement consécutifs

Opérations sur les mots

- **Concaténation de mots** : si x et y sont des mots, la concaténation xy est la chaîne formée en joignant x et y
  $\hookrightarrow$ Exemple : pour $\mathcal{A} = \{a, b\}$, si $x = aa$ et y $= bb$, alors $xy = aabb$
- **Exponentiation** : $s^0 = \epsilon$; $s^i = s^{i-1}s$
  $\hookrightarrow$ Exemple : pour $\mathcal{A} = \{a, b\}$, si $x = ba$ alors $x^3 = bababa$

Langage

Définition : langage

Un langage est un ensemble de mots définis sur un même alphabet.

- Soit $\mathcal{A} = \{1, 2, 3\}$, l'ensemble $\{1, 11, 12, 21\}$ est un langage sur $\mathcal{A}$
- Le langage vide est noté $\emptyset$
- Le langage $\{\epsilon\}$ ne contient que le mot vide

$$\emptyset \neq \{\epsilon\}$$

Opérations sur les langages (1/2)

Soit deux langages L_1 et L_2 définis respectivement sur les alphabets $\mathcal{A}_1$ et $\mathcal{A}_2$.

Définition : union de deux langages

L'union de L_1 et L_2 définie sur $\mathcal{A}_1 \cup \mathcal{A}_2$ est le langage contenant tous les mots de L_1 et L_2 :

$$L_1 \cup L_2 = \{w \mid w \in L_1 \vee w \in L_2\}$$

Définition : intersection de deux langages

L'intersection de L_1 et L_2 définie sur $\mathcal{A}_1 \cap \mathcal{A}_2$ est le langage contenant tous les mots qui sont à la fois dans L_1 et L_2 :

$$L_1 \cap L_2 = \{w \mid w \in L_1 \wedge w \in L_2\}$$

Opérations sur les langages (2/2)

Définition : complément d'un langage

Le complément de L_1 est le langage défini sur $\mathcal{A}_1$ contenant tous les mots qui ne sont pas dans L_1 :

$$\mathcal{C}(L_1) = \{w \mid w \in \mathcal{A}_1 \wedge w \notin L_1\}$$

Définition : différence de deux langages

La différence de L_1 et L_2 est le langage défini sur $\mathcal{A}_1$ contenant tous les mots de L_1 qui ne sont pas dans L_2 :

$$L_1 - L_2 = \{w \mid w \in L_1 \wedge w \notin L_2\}$$

Produit et puissances

Définition : produit de deux langages

Le produit ou ***concaténation*** *de L_1 et L_2 est le langage défini sur $\mathcal{A}_1 \cup \mathcal{A}_2$ contenant tous les mots formés d'un mot de L_1 suivi d'un mot de L_2 :*

$$L_1.L_2 = \{w_1 w_2 \mid w_1 \in L_1 \wedge w_2 \in L_2\}$$

Définition : puissances d'un langage

Les puissances successives de L_1 définies sur $\mathcal{A}_1$ sont définies récursivement :

- $L_1^0 = \{\epsilon\}$
- $L_1^n = L_1.L_1^{n-1}$ *pour* $n \geq 1$

Fermeture itérative

Définition : fermeture de Kleene de deux langages

La fermeture de Kleene de L_1, appelée également la ***fermeture itérative****, définie sur $\mathcal{A}_1$, est l'ensemble des mots formés par une concaténation finie des mots de L_1 :*

$$L_1^* = \{w \mid \exists k \geq 0 \wedge w_1 \dots w_k \in L_1 \ tels \ que \ w = w_1 w_2 \dots w_k\}$$

- On définit également L_1^+ :
  $L_1^+ = \{w \mid \exists k > 0 \wedge w_1 \dots w_k \in L \text{ tels que } w = w_1 w_2 \dots w_k\}$

Langage fini et infini

Définition : langage fini

Un langage fini peut être décrit par l'énumération des mots qui le compose. Ce qui n'est pas le cas pour un langage infini.

- Certains langages infinis peuvent être décrits à l'aide d'opérations sur des langages simples
- Certains langages infinis peuvent être décrits à l'aide de règles (grammaires)
- Les langages qui ne peuvent être décrits ni par des opérations, ni par des grammaires sont des langages **indécidables**.

Expressions régulières

Définition : expression régulière

Les expressions régulières pour un alphabet $\mathcal{A}$ sont les expressions formées par les règles suivantes :

- *$\emptyset$, ϵ et les symboles de $\mathcal{A}$ sont des expressions régulières*
- *Si α et β sont des expressions régulières sur $\mathcal{A}$, $(\alpha|\beta)$, $(\alpha.\beta)$ et $(\alpha)^*$ sont des expressions régulières*

- On note indifféremment $\alpha.\beta$ et $\alpha\beta$
- On définit une priorité décroissante sur les opérateurs : $*$, . et $|$

Langage décrit par une expression régulière

Définition : langage décrit par une expression régulière

Le langage $L(E)$ où E est une expression régulière définie sur $\mathcal{A}$, est défini comme suit :

- $L(E) = \emptyset$ *si* $E = \emptyset$
- $L(E) = \{\epsilon\}$ *si* $E = \epsilon$
- $L(E) = \{a\}$ *si* $E = a$ *pour tout* $a \in \mathcal{A}$
- $L(E) = L(E_1) \cup L(E_2)$ *si* $E = E_1|E_2$
- $L(E) = L(E_1).L(E_2)$ *si* $E = E_1.E_2$
- $L(E) = L(E)^*$ *si* $E = E_1^*$