I Miroir

On note $\widetilde{u} = u_n...u_1$ le miroir d'un mot $u = u_1...u_n$ et $\widetilde{L} = \{\widetilde{u} \mid u \in L\}$ le miroir d'un langage L.
Soit L un langage hors-contexte. Montrer que $\widetilde{L}$ est un langage hors-contexte.

Solution : Soit $G = (\Sigma, V, R, S)$ une grammaire hors-contexte engendrant L. On construit une grammaire $G' = (\Sigma, V, R', S)$ engendrant $\widetilde{L}$ en remplacant chaque règle $X \to u \in (V \cup \Sigma)^*$ par $X \to \widetilde{u}$.

II Trouver une grammaire

Montrer que les langages suivants sont non-contextuels :

1. $L_1 = (ab)^*$
2. $L_2 = \{a^n b^p \mid n \geq p\}$.
3. L_3 : représentations des multiples de 3 en base 2.
4. $L_4 = \{a^i b^j c^k \mid i = j + k\}$.

Solution :

1. $X \to abX \mid \varepsilon$
2. $$S \to aSb \mid aS \mid \varepsilon$$
3. Similaire à l'automate reconnaissant les multiples de 5 en base 2 (voir TD automates). On utiliser des variables X_0, X_1 et X_2 pour les restes modulo 3 et les égalités $\overline{u0}^2 = 2\overline{u}^2$, $\overline{u1}^2 = 2\overline{u}^2 + 1$. On utilise aussi S pour que les mots reconnus commencent par 1.
$$\begin{aligned} S &\to 1X_1 \mid 0 \\ X_0 &\to 0X_0 \mid 1X_1 \mid \varepsilon \\ X_1 &\to 0X_2 \mid 1X_0 \\ X_2 &\to 0X_1 \mid 1X_2 \end{aligned}$$
4. On ajoute d'abord les c puis les b :
$$S \to aSc \mid X$$
$$X \to aXb \mid \varepsilon$$

III Trouver le langage engendré

Déterminer les langages engendrés par les grammaires suivantes avec S comme symbole initial, en le prouvant :

1. G_1 :
$$\begin{aligned} S &\to X \mid Y \\ X &\to aX \mid aZ \\ Y &\to Yb \mid Zb \\ Z &\to \varepsilon \mid aZb \end{aligned}$$
2. G_2 :
$$\begin{aligned} S &\to 0A1 \mid \varepsilon \\ A &\to 1S0 \mid \varepsilon \end{aligned}$$
3. G_3 :
$$\begin{aligned} S &\to X \mid XaS \\ X &\to aXbX \mid bXaX \mid \varepsilon \end{aligned}$$
4. G_4 :
$$\begin{aligned} S &\to X \mid Y \\ X &\to Z0X \mid Z0Z \\ Y &\to Z1Y \mid Z1Z \\ Z &\to \varepsilon \mid 1Z0Z \mid 0Z1Z \end{aligned}$$

Solution :

1. Clairement, $L_Z(G_1) = \{a^n b^n \mid n \in \mathbb{N}\}$. Comme X y ajoute au moins un a, $L_X(G_1) = \{a^n b^p \mid n > p\}$. De même, $L_Y(G_1) = \{a^n b^p \mid n < p\}$. Donc $L(G_1) = \{a^n b^p \mid n \neq p\}$.
2. Soit $L = (01)^*$.
Soit H_n : « $(01)^n \in L(G_2)$ ».
H_0 est vrai car $\varepsilon \in L(G_2)$. H_1 est vrai car $S \Rightarrow 0A1 \Rightarrow 01$.
Soit $n \geq 2$ et supposons H_k vrai pour $k < n$. Soit $u = (01)^n$.

D'après l'hypothèse de récurrence, $v \in (01)^{n-2}$: $S \Rightarrow^* (01)^{n-2}$.
Alors $S \Rightarrow 0S1 \Rightarrow 01A01 \Rightarrow^* 01(01)^{n-2}01 = u$. Donc H_n est vrai.

3. On a montré dans le cours que l'ensemble des mots générés à partir de X est $L_X(G_3) = \{w \in \{a,b\}^* \mid |w|_a = |w|_b\}$. Comme S peut ajouter des a : $L(G_3) = \{u \in \{a,b\}^* \mid |u|_a > |u|_b\}$.
4. De même, $L_Z(G_4) = \{w \in \{0,1\}^* \mid |w|_0 = |w|_1\}$. S donne soit X qui ajoute au moins un 0, soit Y qui ajoute au moins un 1. Donc $L(G_4) = \{w \in \{0,1\}^* \mid |w|_0 \neq |w|_1\}$.

IV Régulier $\implies$ hors-contexte

Montrer par induction structurelle que tout langage régulier est un langage hors-contexte (ce qui est une preuve alternative passant par un automate, donnée en cours).

V CCP 2023

On considère la grammaire algébrique G sur l'alphabet $\Sigma = \{a,b\}$ et d'axiome S dont les règles sont : $S \to SaS \,|\, b$

1. Cette grammaire est-elle ambiguë ? Justifier.
2. Déterminer (sans preuve pour cette question) le langage L engendré par G. Quelle est la plus petite classe de langages à laquelle L appartient ?
3. Prouver que $L = L(G)$.
4. Décrire une grammaire qui engendre L de manière non ambiguë en justifiant de cette non ambiguité.
5. Montrer que tout langage dans la même classe de langages que L peut être engendré par une grammaire algébrique non ambiguë.

Solution :

1. Cette grammaire est ambiguë car le mot $babab \in L(G)$ admet les deux arbres syntaxiques suivants :

Or ces derniers sont différents.

2. On constate que $L(G)$ est rationnel car dénoté par l'expression rationnelle $(ba)^\star b$.
3. Montrons par récurrence forte sur $n \in \mathbb{N}^\star$ la propriété $H(n)$ suivante : si $u \in L$ est un mot de taille n alors u est engendré par la grammaire G. C'est bien sûr le cas pour $n = 1$ puisque le seul mot de L de cette taille est b, engendré par la deuxième règle de G. Soit donc $n \in \mathbb{N}^\star$ et u un mot de L de taille $n+1$. Comme $|u| \geq 2$, ba est nécessairement préfixe de u et il existe donc $v \in (ba)^\star b = L$ tel que $u = bav$. Par hypothèse, ce mot v est engendré par G : il existe une dérivation telle que $S \Rightarrow^\star v$. On en déduit que

$$S \Rightarrow SaS \Rightarrow baS \Rightarrow^\star bav = u$$

est une dérivation licite et donc que $u \in L(G)$. Cette récurrence montre que $L \subset L(G)$. Montrons réciproquement que $L(G) \subset L$ en montrons par récurrence forte sur $n \in \mathbb{N}^\star$ la propriété $H(n)$ suivante : si $u \in \Sigma^\star$ se dérive de S en n dérivations alors $u \in L$. C'est acquis pour $n = 1$: le seul mot de $\Sigma^\star$ qu'on peut obtenir en une dérivation est $b \in L$. Soit donc $n \in \mathbb{N}^\star$ et u un mot dans $L(G)$ tel que $S \Rightarrow^{n+1} u$. Comme ce mot est obtenu en au moins 2 dérivations, les règles de G nous informent que la première est nécessairement $S \to SaS$ (sans quoi ce serait $S \to b$ et dans ce cas u serait obtenu en une seule dérivation). Donc la dérivation permettant d'obtenir u se décompose en :

$$S \Rightarrow SaS \Rightarrow^n u$$

On en déduit qu'il existe $v, w \in \Sigma^\star$ et $k_1, k_2 \in [\![1, n]\!]$ tels que $u = vaw$, $S \Rightarrow^{k_1} v$, $S \Rightarrow^{k_2} w$ et $k1 + k2 = n$. L'hypothèse

de récurrence (forte) s'applique à v et w et on en déduit que ces deux mots appartiennent au langage dénoté par $(ba)^\star b$ donc qu'il existe $r_1, r_2 \in \mathbb{N}$ tels que $v = (ba)^{r_1} b$ et $w = (ba)^{r_2} b$. Par conséquent, $u = (ba)^{r_1} ba (ba)^{r_2} b = (ba)^{r_1+r_2+1} b \in L$.

4. On sait à présent que $L(G) = (ba)^\star b$; il s'agit donc de trouver une grammaire non ambiguë engendrant ce langage. On peut proposer par exemple la grammaire dont les règles sont :

$$S \to Tb \qquad T \to baT \,|\, \varepsilon$$

les règles sur T permettant de générer le facteur dans $(ba)^\star$ et la première de rajouter le b final. Cette grammaire G' est non ambiguë car pour tout mot dans $L(G')$, il existe une unique dérivation permettant de le construire (donc évidemment un seul arbre syntaxique) ; cette unicité découlant du fait que dans cette grammaire un non terminal se dérive toujours en un mot qui contient au plus un seul non terminal.

5. La question demande de montrer que tout langage rationnel peut être engendré par une grammaire non ambiguë. Soit donc L un langage rationnel. Par le théorème de Kleene, il existe un automate fini $A = (\Sigma, Q, \{q_0\}, F, \delta)$ qui reconnaît L qu'on peut loisiblement supposer déterministe. Considérons alors la grammaire dont les non terminaux sont $\{V_q \,|\, q \in Q\}$, l'axiome est V_{q_0}, les terminaux sont les lettres de Σ et dont les règles sont données par :

   - Pour toute transition $q \xrightarrow{a} q'$ dans A, on ajoute la règle $V_q \to aN_{q'}$.
   - Pour tout $q \in F$, on ajoute la règle $V_q \to \varepsilon$.

   Cette grammaire engendre L de manière non ambiguë graĉe au déterminisme de A.

VI Forme normale de Chomsky

Une grammaire $G = (\Sigma, V, R, S)$ est en forme normale de Chomsky si toutes ses règles sont de la forme $X \longrightarrow YZ$ (où $Y, Z \in V$), $A \longrightarrow a$ (où $a \in \Sigma$) ou $S \longrightarrow \varepsilon$.
Soit G une grammaire qui n'engendre pas ε. Montrer qu'il existe une grammaire G' en forme normale de Chomsky telle que $L(G') = L(G)$.

Solution : Pour chaque $a \in \Sigma$, on ajoute une variable X_a et une règle $X \longrightarrow a$.
On remplace chaque lettre apparaissant dans un membre droit de règle par des variables.
On remplace chaque règle $X \longrightarrow X_1 X_2 ... X_n$ avec $n > 2$ par $X \longrightarrow X_1 Y_1$, $Y_1 \longrightarrow X_2 Y_2$, ..., $Y_{n-2} \longrightarrow X_{n-1} X_n$, où les Y_i sont de nouvelles variables.

VII Mots de Dyck

Soit $\Sigma = \{a, b\}$. Un mot u sur Σ est un mot de Dyck (ou : mot bien parenthésé) si :

a) u contient autant de a que de b

b) chaque préfixe de u contient au moins autant de a que de b

On note D l'ensemble des mots de Dyck.

1. Montrer que D n'est pas un langage régulier.
2. Montrer que tout mot de Dyck non-vide se décompose de manière unique sous la forme $aubv$, où u et v sont des mots de Dyck.
3. Soit G la grammaire donnée par $S \to SS \,|\, aSb \,|\, \varepsilon$. Montrer que $L(G) = D$.
4. Montrer que G est ambigüe.
5. Donner une grammaire non-ambigüe engendrant D.
6. Donner une bijection entre les mots de Dyck et les arbres binaires stricts (tels que tout nœud possède 0 ou 2 fils).
7. Soit C_n le nombre de mots de Dyck de longueur $2n$. Trouver une équation de récurrence sur C_n.
8. Après avoir fait le cours de mathématiques sur les séries entières, montrer que $C_n = \dfrac{1}{n+1}\dbinom{2n}{n}$.
9. Dans cette question, on peut utiliser des parenthèses différentes (par exemple, `{}` et `[]`). Décrire un algorithme en complexité linéaire pour savoir si un mot est bien parenthésé.

10. Décrire un algorithme en complexité linéaire pour trouver la longueur du plus long facteur bien parenthésé d'un mot sur Σ.

On pourra résoudre les deux dernières questions sur LeetCode : Valid Parentheses et Longest Valid Parentheses.

Solution :

1. Supposons que L soit régulier. Soit $n \in \mathbb{N}$ donné par le lemme de l'étoile.
Soit $m = a^n b^n$. Comme $m \in L$ et $|m| \geq n$, il existe x, y, z tels que $|xy| \leq n$, $y \neq \varepsilon$, $m = xyz$ et $xy^*z \subseteq L$.
Comme $|xy| \leq n$, y ne contient que de a. Comme $y \neq \varepsilon$, y contient au moins un a. Donc xy^2z contient strictement plus de a que de b, ce qui contredit le fait que $xy^2z \in L$: absurde.
Donc L n'est pas régulier.

2.

Existence : Soit $m = m_1...m_n$ un mot de Dyck non vide.
L'ensemble $\{i \in [\![1, n]\!] \mid m_1...m_i$ contient autant de a et $b\}$ est non vide (car m contient autant de a que de b) et minoré donc admet un minimum k.
Soit $u = m_2...m_{k-1}$ et $v = m_{k+1}...m_n$. Comme $m_1 = a$ (sinon m aurait a comme préfixe), $m_k = b$ (sinon $m_1...m_{k-1}$ aurait plus de b que a) et $m_1...m_k$ vérifie a), u et v vérifient aussi a). Par minimalité de k, u vérifie b) et on en déduit que v vérifie b). Ainsi, $m = aubv$ où u et v sont des mots de Dyck.

Unicité : Supposons que $m = au_1bv_1 = au_2bv_2$ avec u_1, u_2, v_1, v_2 des mots de Dyck et $|u_1| < |u_2|$.
Alors $u_2 = u_1w$. De plus, comme $au_1bv_1 = au_2bv_2$, la première lettre de w est un b. u_1b est alors un préfixe de u_2 qui contient plus de b que de a, ce qui est absurde.

3.

$D \subset L(G)$: Montrons par récurrence forte sur $n \in \mathbb{N}^\star$ la propriété P_n : si $u \in D$ est un mot de taille n alors $u \in L(G)$.
P_0 est vrai car $\varepsilon \in L(G)$. P_1 est vrai $D = \emptyset$.
Soit $n \geq 2$. Supposons P_k vrai pour $k < n$. Soit $w \in D$ de taille n.
D'après la question précédente, il existe a, b tels que $u = aubv$ avec $u, v \in D$.
Par hypothèse de récurrence, $u, v \in L(G)$. Donc $S \Rightarrow^* u$ et $S \Rightarrow^* v$.
Alors $S \Rightarrow SS \Rightarrow SS \Rightarrow aSbS \Rightarrow^* aubv = w$. Donc $w \in L(G)$.

$L(G) \subset D$: Montrons par récurrence forte sur $n \in \mathbb{N}^\star$ la propriété P_n : si $S \Rightarrow^n u \in \Sigma^*$ alors $u \in D$.
P_1 est vrai car si $S \Rightarrow u$ alors $u = \varepsilon \in D$. P_2 est vrai car la seule dérivation de longueur 2 est $S \Rightarrow aSb \Rightarrow ab$ et $ab \in D$.
Soit $n > 2$. Supposons P_k vrai pour $k < n$.
Soit $u \in \Sigma^*$ tel que $S \Rightarrow^n u$. Considérons la première de cette dérivation :

- Si $S \Rightarrow SS \Rightarrow^{n-1} u$ alors $u = u_1u_2$ avec $S \Rightarrow^k u_1$ et $S \Rightarrow^{n-1-k} u_2$.
Par hypothèse de récurrence, $u_1, u_2 \in D$. On vérifie alors que $u = u_1u_2$ vérifie la définition de mot de Dyck.
- Si $S \Rightarrow aSb \Rightarrow^{n-1} u$ alors $u = au_1b$ avec $S \Rightarrow^{n-2} u_1$.
Par hypothèse de récurrence, $u_1 \in D$. On vérifie alors que $u = au_1b$ vérifie la définition de mot de Dyck.

Ainsi, $u \in D$.

4. ε possède deux dérivations gauches : $S \Rightarrow \varepsilon$ et $S \Rightarrow SS \Rightarrow \varepsilon S \Rightarrow \varepsilon\varepsilon = \varepsilon$. Donc G est ambiguë.
5. On montre que G' définie par $S \to aSbS \mid \varepsilon$ engendre D et est non ambiguë, d'après la question 2.
6. Soit V l'arbre vide et $N(g, d)$ l'arbre de sous-arbre gauche g et sous-arbre droit d. On définit par induction un arbre $f(m)$ associé à un mot m :

$$f(\varepsilon) = V$$
$$f(aubv) = N(f(u), f(v))$$

D est bien bijective puisqu'on peut définir son inverse :

$$h(V) = \varepsilon$$
$$h(N(g, d)) = ah(g)bh(d)$$

7. $c_0 = 1$ car ε est le seul mot de Dyck de longueur 0.
   Soit $n \geqslant 0$. D'après la question 2, un mot de Dyck de longueur $2n$ se décompose de manière unique sous la forme $aubv$ où u et v sont des mots de Dyck de tailles $2k$ et $2(n-k)$. Comme k peut prendre toutes les valeurs entre 0 et $n-1$, on a $C_{n+1} = \sum_{k=0}^{n} C_k C_{n-k}$.

8. Comme $C_n \leq 2^{2n} = 4^n$, le rayon de convergence de $f(x)$ est $R \geqslant \frac{1}{4}$ et, pour $x < R$:

$$f(x) = \sum_{n=0}^{\infty} C_n x^n = 1 + x \sum_{n=1}^{\infty} C_{n+1} x^n = 1 + x \sum_{n=1}^{\infty} \sum_{k=0}^{n} c_k C_{n-k} x^n = 1 + x f(x)^2$$

   Où on a reconnu un produit de Cauchy.

   On en déduit $f(x) = 1 + x f(x)^2$, qui est une équation du second degré en $f(x)$: $f(x) = \frac{1 \pm \sqrt{1-4x}}{2x}$.

   Comme $\frac{1+\sqrt{1-4x}}{2x} \underset{x \longrightarrow +0}{\longrightarrow} \infty$ alors que $f(0) = 1$, on a $f(x) = \frac{1-\sqrt{1-4x}}{2x}$.

   Après simplification, $\sqrt{1-4x} = -\sum_{n=0}^{\infty} \frac{\binom{2n}{n}}{2n-1} x^n$. Donc :

$$f(x) = \frac{1}{2x} \sum_{n=1}^{\infty} \frac{\binom{2n}{n}}{2n-1} x^n = \sum_{n=0}^{\infty} \frac{\binom{2(n+1)}{n+1}}{2(2n+1)} x^n$$

   De plus, $\binom{2(n+1)}{n+1} = \frac{(2n+2)(2n+1)}{(n+1)^2} \binom{2n}{n}$ donc :

$$f(x) = \sum_{n=0}^{\infty} \frac{(2n+2)(2n+1)}{2(2n+1)(n+1)^2} \binom{2n}{n} x^n = \sum_{n=0}^{\infty} \frac{1}{n+1} \binom{2n}{n} x^n$$

   Par unicité du développement en série entière : $c_n = \frac{1}{n+1} \binom{2n}{n}$.

VIII Lemme de l'étoile algébrique, intersection et complémentaire

On admet la version suivante du lemme de l'étoile pour les langages non-contextuels :

Théorème : Lemme d'Ogden

Si L est un langage hors-contexte alors il existe un entier n tel que, pour tout mot $t \in L$ tel que $|t| \geq n$, on peut écrire $t = uvwxy$ avec :
- $|vwx| \leq n$;
- $vx \neq \varepsilon$;
- $\forall i \in \mathbb{N}, uv^i wx^i y \in L$.

Soient $L_1 = \{a^n b^n c^p \mid n, p \in \mathbb{N}\}$ et $L_2 = \{a^n b^n c^n \mid n \in \mathbb{N}\}$.

1. Montrer que L_1 est un langage hors-contexte.
2. Montrer que L_2 n'est pas un langage hors-contexte.
3. Montrer que l'ensemble des langages hors-contextes n'est pas stable par intersection ni par passage au complémentaire.