Devoir 2 : Langages et flux balisés

À rendre le 3 mai 2009 à minuit au plus tard.

April			1	2	3	4	5
	6	7	8	9	10	11	12
	13	14	15	16	17	18	19
	20	21	22	23	24	25	26
	27	28	29	30			
May					1	2	3

Ce devoir est l'occasion d'étudier une classe restreinte des langages algébriques préfixes déterministes qui peut servir à modéliser des documents XML. On considère dans tout le reste de l'énoncé les trois alphabets

1. $A_n = \{a_1, \ldots, a_n\}$ de n symboles distincts,
2. $\bar{A}_n = \{\bar{a}_1, \ldots, \bar{a}_n\}$ une copie disjointe de A_n, et
3. $\Sigma_n = A_n \uplus \bar{A}_n$.

Définition 1 (Langage enraciné de Dyck)**.** Soit D_n^* l'ensemble des mots de Dyck sur Σ_n. L'ensemble E_n des *mots enracinés de* Dyck sur Σ_n est l'union

$$E_n = \bigcup_{a_i \in A_n} a_i D_n^* \bar{a}_i \ .$$

1 Langages balisés

Définition 2 (Grammaires balisées)**.** Une grammaire algébrique $G = \langle N \uplus \{S\}, \Sigma_n, P, S\rangle$ est une *grammaire balisée* sur Σ_n, avec l'alphabet non terminal $N \uplus \{S\}$, si les règles de P sont de l'une des trois formes

i. $S \to a_i A \bar{a}_i$, avec $A \in N$ et $a_i \in A_n$,

ii. $A \to a_i B \bar{a}_i C$, avec $A, B, C \in N$ et $a_i \in A_n$,

iii. $A \to \varepsilon$, $A \in N$.

On appelle *langage balisé* sur Σ_n un langage $L \subseteq \Sigma_n^*$ tel qu'il existe une grammaire balisée G sur Σ_n avec $L = L(G)$.

Exercice 1.

1. Montrer que si L est un langage balisé sur Σ_n, alors $L \subseteq E_n$.
2. Montrer que la famille des langages balisés est fermée par union et intersection.

Exercice 2 (Ambiguïté)**.**

1. Montrer qu'une grammaire algébrique réduite $\langle N, \Sigma, P, S\rangle$ est ambiguë si et seulement si au moins l'une des deux conditions suivantes est remplie :
   - il existe deux productions $A \to \alpha_1$ et $A \to \alpha_2$ avec α_1 et α_2 dans $(N \uplus \Sigma)^*$ telles que $L(\alpha_1) \cap L(\alpha_2) \neq \emptyset$ ou
   - il existe une production $A \to \alpha_1\alpha_2$ avec α_1 et α_2 dans $(N \uplus \Sigma)^*$ telle qu'il existe u et v dans Σ^* et w dans Σ^+ tels que u et uw appartiennent à $L(\alpha_1)$ et v et wv appartiennent à $L(\alpha_2)$.
2. Montrer que l'on peut décider si une grammaire balisée est ambiguë.
3. Proposer une grammaire balisée non ambiguë mais néanmoins non LR(k).

Exercice 3 (Déterminisme)**.** Montrer que tout langage balisé L sur Σ_n est un langage déterministe préfixe (i.e. $L\Sigma_n^+ \cap L = \emptyset$), c'est-à-dire que L est reconnu par un automate à pile déterministe acceptant par pile vide.

2 Flux balisés

Définition 3 (Flux balisé)**.** Un *flux balisé* L sur Σ_n s'écrit comme l'intersection du langage enraciné de Dyck E_n et d'un langage rationnel R sur Σ_n :

$$L = E_n \cap R\ .$$

On représente un flux balisé à l'aide d'un automate fini (non déterministe) $\mathcal{A}$ tel que $L(\mathcal{A}) = R$. Un flux balisé est *local* si R est un langage local, c'est-à-dire s'il existe $R_p \subseteq \Sigma_n$, $R_f \subseteq \Sigma_n^2$ et $R_s \subseteq \Sigma_n$ tels que

$$R\backslash\{\varepsilon\} = (R_p\Sigma_n^* \cap \Sigma_n^* R_s)\backslash\Sigma_n^* R_f \Sigma_n^*$$

(voir aussi l'exercice 5 du TD 1 sur les langages locaux).

Par le théorème de Chomsky-Schützenberger, on sait que tout langage algébrique peut s'écrire comme la projection d'un flux balisé local.

Exercice 4 (Langages balisés)**.**

1. Montrer que tout flux balisé est un langage balisé.
2. Montrer (par un argument de pompage) que le langage balisé généré par la grammaire

$$\begin{aligned}
S &\to aA\bar{a}\\
A &\to aA\bar{a}B \mid cE\bar{c}C \mid \varepsilon\\
B &\to bE\bar{b}E\\
C &\to aA\bar{a}E\\
E &\to \varepsilon
\end{aligned}$$

n'est pas un flux balisé. On pourra s'intéresser aux deux chaînes de E_3

$$a(c\bar{c}a)^m a(c\bar{c}a)^m a\bar{a}^m\bar{a}b\bar{b}\bar{a}^m\bar{a}$$
$$a(c\bar{c}a)^{m+k} a(c\bar{c}a)^{m+k} a\bar{a}^m\bar{a}b\bar{b}\bar{a}^{m+2k}\bar{a}$$

pour des valeurs de m et k bien choisies.

Exercice 5 (Langages simples)**.**

1. Montrer que tout flux balisé local est un langage simple, par exemple en montrant qu'il est reconnu par un automate à pile déterministe, sans transition ε, à un seul état, et acceptant par pile vide.
2. Proposer un langage simple qui est aussi un flux balisé mais qui n'est pas un flux balisé local.
3. Proposer un flux balisé qui n'est pas un langage simple (on pourra s'inspirer des langages vus au cours du TD 6).

Exercice 6 (Opérations booléennes)**.**

1. Montrer que l'union de deux flux balisés est un flux balisé.
2. Montrer que l'intersection de deux flux balisés est un flux balisé.
3. Montrer que l'intersection du complément d'un flux balisé avec E_n (soit encore le complémentaire dans E_n d'un flux balisé) est un flux balisé.

Exercice 7 (Problèmes de décision)**.** Montrer que si L_1 et L_2 sont des flux balisés arbitraires sur Σ_n, alors les problèmes suivants sont décidables :

1. $L_1 = E_n$,
2. $L_1 \subseteq L_2$.

Quelle complexité obtenez-vous ?

3. Montrer que si L est un langage algébrique et L' est un flux balisé local sur Σ_n, alors $L \cap L' = \emptyset$ est indécidable en général.