Automates d’arbre
TD n°5bis : Document Type Definition (DTD)

**Definition** (DTD). A *DTD* is a tuple \((\Sigma, s, \delta)\) with \(s \in \Sigma\) (the start symbol) and \(\delta : \Sigma \rightarrow \text{Rat}_\Sigma\), where \(\text{Rat}_\Sigma\) is the set of rational expressions on \(\Sigma\).

**Definition** (Language of a DTD). The language of a DTD \(D = (\Sigma, s, \delta)\) is the language of the DHTA \(A = (Q_D, \Sigma, Q_{D,f}, \Delta_D)\) where:
- the set of states \(Q_D = \Sigma\)
- the set of final states \(Q_{D,f} = \{s\}\)
- the transitions are the \(a(\delta(a)) \rightarrow a\) for \(a \in \Sigma\) (the first ’a’ is the letter a and the second one is the state a)

**Exercise 1:**
Give a recognizable language which is not the language of a DTD.

**Definition** (Marking, 1-unambiguous rational expression). The marking \(m(E)\) of a rational expression on \(\Sigma\) is the rational expression on \(\Sigma \times \mathbb{N}\) obtained by replacing the \(i\)th occurrence of \(a\) in \(E\) by \((a, i - 1)\) for all \(a \in \Sigma\). For example:
\[
m((a + b)^* b(ab)^*) = ((a, 0) + (b, 0))^*(b, 1)((a, 1)(b, 2))^*
\]
We say that \(E\) is 1-unambiguous if for all words \(u, v\) and \(w\) in \(\Sigma \times \mathbb{N}\), for all symbols \(a, b\) in \(\Sigma\) and all integers \(i, j\), if \(u(a, i)v\) and \(u(b, j)w \in L(m(E))\) then \((a, i) \neq (b, j)\) implies \(a \neq b\).

We say that a DTD is 1-unambiguous if for all \(a \in \Sigma, \delta(a)\) is 1-unambiguous.

*Remark.* There are languages that are not definable by a 1-unambiguous rational expression but it is decidable in polynomial time whether a language is definable by such an expression. See A. Brüggemann-Klein, *Formal Models in Document Processing.*

**Exercise 2:**
Let \(E\) be a regular expression on \(\Sigma\). Define the following sets:
- \(\text{first}(E) = \{ a \in \Sigma | \exists w \in \Sigma^*, aw \in L(E) \}\)
- \(\text{last}(E) = \{ a \in \Sigma | \exists w \in \Sigma^*, wa \in L(E) \}\)
- \(\text{follow}(E, a) = \{ b \in \Sigma | \exists v, w \in \Sigma^*, vabw \in L(E) \}\)
- \(\text{sym}(E) = \{ a \in \Sigma \text{ appearing in } E \}\)

1) Prove that for all \(n \geq 1, x_1 x_n \in L(m(E))\) iff the three following conditions hold:
- \(x_1 \in \text{first}(m(E))\)
- \(x_n \in \text{last}(m(E))\)
- for all \(1 \leq i < n, x_{i+1} \in \text{follow}(m(E), x_i)\)

2) Prove that if \(E\) is 1-unambiguous then:
- for all \((a, i), (b, j) \in \text{first}(m(E))\), \((a, i) \neq (b, j)\) implies \(a \neq b\)
- for all \((a, i) \in \text{sym}(m(E)), (b, j), (c, k) \in \text{follow}(m(E), (a, i))\), \((b, j) \neq (c, k)\) implies \(b \neq c\)

3) Let \(E\) be 1-unambiguous. Construct in polynomial time on the size of \(E\) a deterministic automaton that recognizes \(L(E)\).

4) What can we deduce on 1-unambiguous DTD?