TALN INF 478 CH. RETORÉ 2009/2010

MÉMENTO: GRAMMAIRES FORMELLES ET HIÉRARCHIE DE CHOMSKY

Etant donné un ensemble E on note E^* les suites finies d'éléments de E, dont la suite vide notée ε. La longueur d'une suite $W \in E^*$, notée $|W|$, est le nombre d'éléments de E qui la constituent.

Monoïdes, langages Un **monoïde** est un ensemble M muni d'une loi de composition $\cdot : M \times M \mapsto M$ souvent notée par simple juxtaposition), qui est associative ($\forall w,w',w'' \in M\ w.(w'.w'') = (w.w').w''$) et dotée un élément neutre ε ($\forall w \in M\ w.\varepsilon = w = \varepsilon.w$). En raison de l'associativité, on peut omettre les parenthèses.

Un cas particulier est le monoïde libre sur un ensemble Σ, appelé alphabet (ou lexique), dont les éléments sont appelés terminaux (ou lettres ou mots suivant le contexte). Les éléments de ce monoïde sont les éléments de Σ^* et la loi de composition est la *concaténation* opération qui à deux suites finies $a_1...a_n$ et $b_1...b_p$ de Σ^* associe la suite $a_1...a_nb_1...b_p$ de Σ^* — la concaténation est bien associative et la suite vide ε est bien son élément neutre.

Un **langage** est une partie, finie ou infinie, de Σ^*. On va en donner une description finie par un processus appelé grammaire formelle qui engendre toutes les suites du langage et rien qu'elles.

Grammaires formelles La description d'une grammaire formelle comporte plusieurs ingrédients.

- Il faut se donner deux alphabets disjoints, $\mathscr{N}$ et Σ dont la réunion est notée $\mathscr{V} = \mathscr{N} \cup \Sigma$.
  - $\mathscr{N}$ (dont les éléments sont appelés non terminaux) contenant un symbole distingué S (symbole de départ, S pour *start* ou *sentence*). Les éléments de $\mathscr{N}$ sont généralement notés par des majuscules.
  - Σ un alphabet de terminaux, dont les symboles sont appelés des lettres (si on veut décrire des langages qui soient des ensembles de mots), ou des mots (si on veut décrire des langages qui soient des ensembles de phrases). Les éléments de Σ sont généralement notés en minuscules.
- Un ensemble de règles de production qui sont de la forme $W \to W'$ avec $W \in \mathscr{V}^*\mathscr{N}\mathscr{V}^*$ et $W' \in \mathscr{V}^*$ (W et W' sont des suites de terminaux et de non terminaux, et W contient au moins un non terminal). [Certains auteurs n'utilisent pas la restriction que W contienne au moins un non terminal, c'est-à-dire $W \in \mathscr{V}^*$, ce qui ne change pas grand chose]

On définit la relation $\to$ sur $\mathscr{V}^*$ ainsi: une suite M donne une suite M', ce qu'on notera aussi $M \to M'$ si M est de la forme NWP avec $N,P,W \in \mathscr{V}^*$ et si M' est de la forme $NW'P$ avec $W' \in \mathscr{V}^*$ et que $W \to W'$ est l'une des règles de production.

La réécriture est une relation sur $\mathscr{V}^*$ notée $\longrightarrow$ (remarquez la différence avec $\to$) qui est la clôture réflexive et transitive de $\to$. En d'autres termes, pour $M,M' \in \mathscr{V}^*$ on a $M \longrightarrow M'$ s'il existe un entier n et des suites $M_0,M_1,M_2,\ldots,M_n \in \mathscr{V}^*$ telles que $M = M_0 \to M_1 \to M_2 \cdots \to M_n = M'$ (si $n = 1$ cela revient à $M \to M'$).

Etant donnée une grammaire G, le langage $L(G)$ engendré par G est la partie de Σ^* définie par $L(G) = \{M \in \Sigma^* | S \longrightarrow M\}$: ce sont les suites de *terminaux* que l'on peut obtenir par réécriture à partir de S. Le mot *appartenance* utilisé ci-après désigne l'appartenance d'une phrase $m_1 \cdots m_n \in \Sigma^*$ au langage $L(G)$, et sa complexité est le nombre d'étapes de calcul pour répondre à cette question en fonction de n le nombre de terminaux de la phrase. C'est une question qui permet de classer diverses classes de complexité algorithmique.

La hiérarchie de Chomsky On définit la hiérarchie de Chomsky par des restrictions sur la forme des règles, de sorte que les grammaires de type i contiennent strictement les grammaires de type $i+1$, pour $i = 0,1,2$.

type 0 Aucune restriction. La classe de langages décrite par ces grammaires est celle des langages récursivement énumérables. L'appartenance est une question indécidable (semi-décidable, pour être plus précis).

type 1 Grammaires contextuelles (*context sensitive*): toutes les règles doivent être de la forme $M_1XM_2 \to M_1WM_2$ avec $X \in \mathscr{N}$ et $M_1,M_2,W \in \mathscr{V}^*$ et $S \longrightarrow \varepsilon$ est autorisé si S ne figure jamais dans le membre droit d'une règle. [On utilise souvent une autre caractérisation de cette classe en terme de croissance de la réécriture donnée ci-après]. L'appartenance au langage d'une telle grammaire est un problème décidable (grâce à l'autre caractérisation).

type 2 Grammaires algébriques, non contextuelles, hors-contexte (*context-free*): toutes les règles doivent être de la forme $X \to W$ avec $X \in \mathscr{N}$, $W \neq \varepsilon$ sauf au cas où $X = S$ et S ne figure jamais dans le membre droit d'une règle. [On peut omettre cette restriction, voir le commentaire ci-après]. L'appartenance se décide en $O(n^3)$ pour une phrase à n mot, et les machines associées à ce type de grammaires sont les automates non déterministe à piles — et comme ceux-ci ne sont pas "déterminisables" l'appartenance n'est en générale pas linéaire.

type 3 Grammaire régulière, linéaire (on peut distinguer droite ou gauche). Les règles sont soit toutes de la forme $X \to mY$ ou $X \to n$ soit toutes de la forme $X \to Ym$ ou $X \to n$ avec $X,Y \in \mathscr{N}$ et $m,n \in \Sigma$. Comme précédemment $S \longrightarrow \varepsilon$ est autorisé si S ne figure jamais dans le membre droit d'une règle. Les automates à états finis sont les machines associés à ces grammaires, et l'appartenance est décidable en temps linéaire (en utilisant un automate déterminisé).

La **classe d'un langage** L est la classe de grammaires la plus simple qui contient une grammaire G telle que $L(G) = L$.

Deux grammaires sont dites **équivalentes** lorsqu'elles engendrent le même langage. Hormis pour deux grammaires de type 3 (ou, mais c'est bien plus difficile, pour deux grammaires de type 2 déterministes, résultat de Sénizergues), l'équivalence de deux grammaires n'est pas décidable. Deux grammaires de type différents peuvent très bien être équivalentes.

Variantes et raffinements Une grammaire est dite **croissante** (*length increasing*) lorsque ses productions sont de la forme $W \to W'$ avec $|W'| \geq |W|$ un résultat de Kuroda montre que toute grammaire contextuelle est équivalente à une grammaire croissante. Il est assez aisé de voir que l'appartenance est un problème décidable lorsque la grammaire est croissante.

Les dérivations des grammaires de type 2 se représentent aisément par des arbres.

Une grammaire dont les règles sont de la forme $X \to W$ avec $X \in \mathscr{N}$ mais avec W possiblement égal à ε peut être transformée en une grammaire hors-contexte équivalente. On calculant $\mathscr{N}_\varepsilon = \{X \in \mathscr{N} | X \longrightarrow \varepsilon\}$ et on remplace de toutes les manières possibles les $Y \in \mathscr{N}_\varepsilon$ par ε dans les membres droits des règles. C'est pourquoi certains auteurs autorisent les productions vides dans la définition des grammaires de type 2.

Pour les grammaires de type 2, peut aussi éliminer les règles $X \to Y$ avec $X,Y \in \mathscr{N}$. En introduisant un non terminal $\underline{x}$ par terminal x, peut aussi n'avoir que des règles $X \to W$ avec $W \in \mathscr{N}^*$ et $|W| \geq 2$ ou $X \to a$ dont les $\underline{x} \to x$. Finalement, en introduisant un non terminal $\langle XY \rangle$ pour les couples de non terminaux consécutifs dans un membre droit, avec la règle $\langle XY \rangle \to XY$ et en itérant le processus, on peut transformer toute grammaire de type 2 en une grammaire équivalente dont les règles sont de la forme $X \to YZ$ ou $X \to a$ avec $X,Y,Z \in \mathscr{N}$ et $a \in \Sigma$. Une telle grammaire de type 2 est dire en **forme normale de Chomsky**.