Mécanismes de contrôle pour l'analyse en Grammaires de Propriétés

Philippe Blache, Stéphane Rauzy

Université de Provence & CNRS, Laboratoire Parole et Langage
{pb ; stephane.rauzy}@lpl.univ-aix.fr

Résumé

Les méthodes d'analyse syntaxiques hybrides, reposant à la fois sur des techniques statistiques et symboliques, restent peu exploitées. Dans la plupart des cas, les informations statistiques sont intégrées à un squelette context-free et sont utilisées pour contrôler le choix des règles ou des structures. Nous proposons dans cet article une méthode permettant de calculer un indice de corrélation entre deux objets linguistiques (catégories, propriétés). Nous décrivons une utilisation de cette notion dans le cadre de l'analyse des Grammaires de Propriétés. L'indice de corrélation nous permet dans ce cas de contrôler à la fois la sélection des constituants d'une catégorie, mais également la satisfaction des propriétés qui la décrivent.

Mots-clés : analyseur syntaxique, modèle des patrons, indice de corrélation, Grammaires de Propriétés, technique d'analyse hybride.

Abstract

Hybrid parsing techniques based both on statistical and symbolic methods remain rare. In general, they integrate statistical information into a context-free skeleton, in order to control the selection of rules and structures. We propose a statistical method which allows the evaluation of a correlation index between two linguistic objects, namely, category, and property. We describe how to integrate this statistical information into the framework of Property Grammars. The correlation index is used for controling both the selection process of category constituents and the evaluation of properties satisfaction.

Keywords: parsing, patterns model, correlation index, Property Grammars, hybrid parsing techniques.

1. Introduction

L'analyse syntaxique automatique doit aujourd'hui être en mesure de fournir des résultats sur du matériel tout-venant. Après la couverture, les questions de robustesse et d'efficacité redeviennent ainsi des problèmes majeurs du domaine. Plusieurs réponses sont aujourd'hui apportées en ce sens. Les techniques numériques sont la solution généralement choisie dans ce cas. À condition de disposer des corpus annotés adéquats, il est ainsi possible de mettre en oeuvre des techniques probabilistes offrant l'avantage d'être rapides et pouvant s'adapter à des données non canoniques pour peu qu'elles aient été identifiées dans le corpus d'apprentissage.

Les approches symboliques ne sont cependant pas en reste, comme l'indiquent les résultats de la campagne d'évaluation des analyseurs syntaxiques Easy (Vilnat, 2004). Les analyseurs symboliques superficiels sont ainsi très rapides tout en étant plus tolérants aux constructions non canoniques. Les analyseurs profonds quant à eux, s'ils offrent une bonne couverture et une bonne résistance aux données mal formées (tout en offrant une analyse détaillée), présentent

toutefois une efficacité plus limitée, voire franchement mauvaise en termes de rapidité d'exécution. Les solutions proposées pour pallier ce type de problème reposent sur la combinaison de plusieurs techniques consistant par exemple à contrôler les analyseurs profonds à l'aide d'analyseurs superficiels (Blache, 2005b). D'une façon plus générale, cette question de l'association de techniques différentes est régulièrement proposée pour améliorer l'efficacité des systèmes. Se pose ainsi naturellement la question de l'association de techniques numériques et symboliques. Cette idée est déjà ancienne (Klavans, 1996), elle consiste à contrôler le choix d'une structure (généralement un sous-arbre) par l'intermédiaire de probabilités. Il peut s'agir de sélection de structures entières (Bod, 1998), et l'utilisation des probabilités relève de la stratégie d'analyse, ou de contrôle de sélection de règles (Johnson, 1998 ; Manning, 1997). Dans ce cas, les probabilités sont intégrées à la représentation de la grammaire, offrant ainsi une architecture symbolique contrôlée par une technique numérique. Dans ces approches, le contrôle porte sur une partie de la structure syntaxique et se situe donc à un niveau élevé, ce qui peut rendre l'approche quelquefois peu efficace.

Nous proposons d'intégrer l'utilisation de techniques numériques dans le cadre d'une approche permettant de contrôler l'information à un niveau plus fin. Il s'agit plus précisément de tirer parti des capacités de contrôle de ces stratégies à la fois pour ce qui concerne la détermination des unités syntaxiques, mais également de leurs propriétés linguistiques. Nous décrivons dans cet article les possibilités d'exploitation d'une information probabiliste, la corrélation, répondant à cet objectif double. Après une présentation de cette technique, nous en proposons une application dans le cadre des Grammaires de Propriétés offrant ainsi la possibilité d'introduire au sein d'une architecture symbolique un niveau de contrôle probabiliste.

Les propositions faites dans cet article sont, en l'état, théoriques. Leur validation est en cours d'expérimentation.

2. Probabilités sur l'espace des séquences de catégories

Il s'agit ici de proposer un mécanisme pour probabiliser l'espace $\mathcal{S}$ composé de toutes les séquences de catégories pouvant être produites. Nous présentons une nouvelle approche, le modèle des patrons, qui permet de réaliser cet objectif. Le modèle des patrons est caractérisé par deux propriétés :
- une extraction optimale de l'information contenue dans le corpus d'apprentissage,
- une faible complexité de calcul des probabilités des séquences de catégories.

2.1. Le modèle des patrons

On considère dans la suite que l'ensemble des catégories $\mathcal{C}$ est de taille N, $\mathcal{C} = \{c_1, ..., c_j, ..., c_N\}$. Nous disposons d'un corpus d'apprentissage constitué par la réalisation d'une séquence de catégories de taille importante, contenant de l'information sur la distribution des catégories et sur leurs interdépendances. L'objectif est de donner une estimation de la probabilité d'une séquence de catégories donnée, par exemple la séquence $(c_3, c_{14}, c_{12}, c_3, c_5, c_{14})$ représentant la séquence $(Det, Nc, Cc, Det, Adj, Nc)$ de l'énoncé *"la maison et le grand chêne"*. La probabilité de la séquence est donnée par la formule canonique des probabilités conditionnelles, *i.e.*

$$P(c_3, c_{14}, c_{12}, c_3, c_5, c_{14}) = \pi_1 \times \pi_2 \times \pi_3 \times \pi_4 \times \pi_5 \times \pi_6 \quad (1)$$

avec $\pi_1 = P(c_3) = P(c_3|\emptyset)$; $\pi_2 = P(c_{14}|c_3)$, $\pi_3 = P(c_{12}|c_3, c_{14})$, $\pi_4 = P(c_3|c_3, c_{14}, c_{12})$, $\pi_5 = P(c_5|c_3, c_{14}, c_{12}, c_3)$ et $\pi_6 = P(c_{14}|c_3, c_{14}, c_{12}, c_3, c_5)$. Dans le terme π_4 par exemple

$P(c_3|c_3, c_{14}, c_{12})$ est la probabilité de c_3 conditionnée par la séquence (c_3, c_{14}, c_{12}).

Un *patron* σ_i est caractérisé par deux informations :
- son identifiant s_i, c'est-à-dire la séquence de catégories qui le constitue, *e.g.* $s_i = (c_3, c_{14}, c_{12})$
- un vecteur de taille N, $(P_{i,1}, ..., P_{i,j}, ..., P_{i,N})$ donnant la probabilité de chaque catégorie c_j conditionnée par l'identifiant du patron σ_i, $P_{i,j} = P(c_j|s_i)$. On a par définition $\sum_{j=1}^{N} P_{i,j} = 1$.

La liste des patrons est extraite du corpus d'apprentissage en prenant garde aux problèmes d'échantillonnage. Un patron sera inclus dans le modèle si le nombre d'occurrences de sa séquence identifiante dans le corpus d'apprentissage permet une évaluation fiable des probabilités conditionnelles $P_{i,j}$. Les problèmes standards liés à l'estimation des paramètres d'une loi multinomiale (*zero-frequency problem*, etc.) seront traités par des techniques de lissage du type back-off model (Katz, 1987).
Contrairement aux modèles du type *n-grammes*, la stratégie n'est pas ici de fixer une taille commune aux patrons, puis d'appliquer des méthodes d'interpolation pour estimer les paramètres $P_{i,j}$ pour les patrons qui ne sont pas présents dans le corpus d'apprentissage (information manquante). Le modèle des patrons est composé de patrons de taille variable. Le critère d'inclusion du patron σ_i dans le modèle repose sur la fiabilité de l'estimation statistique des paramètres $P_{i,j}$ à partir du corpus d'apprentissage. L'objectif est ici d'extraire une information optimale du corpus, en conservant notamment les patrons de taille importante possédant une séquence identifiante fréquemment rencontrée dans le corpus.

On considère dans la suite que nous avons extrait M patrons de taille variable définissant le modèle des patrons $\mathcal{M} = \{\sigma_1, ..., \sigma_i, ..., \sigma_M\}$. Pour chaque patron σ_i et pour chaque catégorie c_j, nous introduisons la notion de *patron cible* $\sigma_k = succ(\sigma_i, c_j)$: c'est le patron appartenant au modèle dont l'identifiant est le plus proche (du point de vue du conditionnement) lorsqu'on ajoute la catégorie c_j à l'identifiant s_i du patron σ_i. Par exemple, si $s_i = (c_{14}, c_{12}, c_3)$ et $c_j = c_5$, on recherchera dans le modèle si le patron d'identifiant $(c_{14}, c_{12}, c_3, c_5)$ existe, puis (c_{12}, c_3, c_5), puis (c_3, c_5), etc., jusqu'à rencontrer le patron cible dans la liste des patrons du modèle.

Le modèle des patrons $\mathcal{M}$ est finalement caractérisé par la matrice des probabilités conditionnelles $P_{i,j}$ de taille $M \times N$, et la matrice $I_{i,j}$ de taille $M \times N$ qui donne pour chaque patron σ_i et pour chaque catégorie c_j, l'indice du patron cible dans la liste des patrons du modèle.

Pour calculer la probabilité d'une séquence, la méthode consiste à approximer la formule donnée équation (1) par un produit des probabilités conditionnelles des patrons du modèle. Par exemple, imaginons que le modèle contienne 9 patrons d'identifiants $s_1 = \emptyset$, $s_2 = c_3$, $s_3 = c_{12}$, $s_4 = c_5$, $s_5 = c_{14}$, $s_6 = (c_3, c_{14})$, $s_7 = (c_3, c_5)$, $s_8 = (c_{12}, c_3)$ et $s_9 = (c_{12}, c_3, c_5)$. Les valeurs de la matrice $I_{i,j}$ utilisées pour calculer la probabilité de la séquence $S = (c_3, c_{14}, c_{12}, c_3, c_5, c_{14})$ sont successivement $I_{1,3} = 2$, $I_{2,14} = 6$, $I_{6,12} = 3$, $I_{3,3} = 8$, $I_{8,5} = 9$ et la probabilité de la séquence est approximée par :

$$P(c_3, c_{14}, c_{12}, c_3, c_5, c_{14}) \approx P_{1,3} \times P_{2,14} \times P_{6,12} \times P_{3,3} \times P_{8,5} \times P_{9,14} \quad (2)$$

La comparaison termes à termes des équations (1) et (2) nous donne les approximations effectuées : $P(c_{14}|c_3, c_{14}, c_{12}, c_3, c_5) \approx P(c_{14}|c_{12}, c_3, c_5)$, $P(c_5|c_3, c_{14}, c_{12}, c_3) \approx P(c_5|c_{12}, c_3)$ et $P(c_3|c_3, c_{14}, c_{12}) \approx P(c_3|c_{12})$. Cette approximation représente la meilleure estimation que l'on peut faire de la probabilité de la séquence, compte tenu des informations fournies par le corpus d'apprentissage.

Cet exemple montre que le modèle des patrons peut être vu comme un système à M états, les patrons σ_i, qui change d'état à chaque ajout d'une catégorie à la séquence. Nous généralisons

cette interprétation dans la section suivante.

2.2. Calcul de la probabilité d'une séquence d'observations

L'objectif est de calculer la probabilité d'une séquence composée de la succession de T *observations* $S_T = (\mathbf{o}_1, ..., \mathbf{o}_t, ..., \mathbf{o}_T)$. Deux cas peuvent être distingués :
- *Mélange* : L'information que l'on a sur la catégorie à la position t est ambiguë et propose plusieurs solutions possibles. Dans ce cas $\mathbf{o}_t = (\alpha_{t,1}, ..., \alpha_{t,j}, ..., \alpha_{t,N})$ est caractérisée par une distribution de probabilité d'observation sur l'ensemble des catégories $\mathcal{C}$, avec par définition $\sum_{j=1}^{N} \alpha_{t,j} = 1$. C'est le cas par exemple lorsque l'information sur $\mathbf{o}_t$ est extraite d'un lexique de formes proposant plusieurs catégories pour une forme observée (*e.g.* la forme *"montres"* en entrée donne deux catégories syntaxiques en sortie (Nc, Vm)).
- *Cas pur* : L'information de la catégorie à la position t n'est pas ambiguë et donne c_{12} par exemple. Dans ce cas, on a pour les coefficients de $\mathbf{o}_t$: $\alpha_{t,12} = 1$ et $\alpha_{t,j} = 0$ pour $j \neq 12$.
Cette notation permet de prendre en compte le cas particulier où l'on ne possède aucune information sur la catégorie à la position t. La distribution des coefficients sur les catégories est alors équiprobable, *i.e.* $\mathbf{o}_t = \mathbf{o}_{NI} = (1/N, ..., 1/N, ..., 1/N)$.

Le système est décrit par M états, les patrons σ_i du modèle $\mathcal{M}$, et un vecteur *densité* qui caractérise la distribution de probabilité des états du système à la position t de la séquence, $\rho_t = (\rho_{t,1}, ..., \rho_{t,i}, ..., \rho_{t,M})$ avec $\sum_{i=1}^{M} \rho_{t,i} = 1$ par définition. L'évolution du système est régie par l'ajout successif des observations à la séquence. On peut montrer que la probabilité de la séquence d'observation $S_T = (\mathbf{o}_1, ..., \mathbf{o}_t, ..., \mathbf{o}_T)$ est donnée par :

$$P(\mathbf{o}_1, ..., \mathbf{o}_t, ..., \mathbf{o}_T) = P(\mathbf{o}_1, ..., \mathbf{o}_t, ..., \mathbf{o}_{T-1}) \times \sum_{i=1}^{M} \rho_{T,i} \sum_{j=1}^{N} \alpha_{T,j} P_{i,j} = \prod_{t=1}^{T} \sum_{i=1}^{M} \rho_{t,i} \sum_{j=1}^{N} \alpha_{t,j} P_{i,j}$$

avec une équation d'évolution du système gouvernant la densité ρ_t de la forme :

$$\rho_{t+1,k} = \frac{1}{A} \sum_{i=1}^{M} \rho_{t,i} \sum_{j=1}^{N} \alpha_{t,j} P_{i,j} \delta[k, I_{i,j}]$$

où $\delta[k, I_{i,j}]$ est la distribution de Kroeneker (*i.e.* $\delta[k, I_{i,j}] = 1$ si $k = I_{i,j}$, $\delta[k, I_{i,j}] = 0$ sinon) et A un facteur de normalisation guarantissant $\sum_{k=1}^{M} \rho_{t+1,k} = 1$. En pratique, on doit initialiser les coefficients de ρ_1 en choisissant par exemple un coefficient égal à 1 pour le patron défini par l'identifiant $\emptyset$ (patron non-conditionné) et des coefficients égaux à 0 pour les autres patrons.

Le modèle des patrons permet de calculer la probabilité de toutes séquences d'observations. La complexité du calcul est faible, de l'ordre de $M \times N \times T$ opérations, pour une séquence de longueur T, dans un modèle à M patrons et N catégories. La probabilité de toutes les séquences d'une longueur T donnée est normalisée, *i.e.* $\sum_{S_T \in \mathcal{S}_T} P(S_T) = 1$.

2.3. La notion de corrélation entre catégories

Pour une séquence donnée $S_T = (\mathbf{o}_1, ..., \mathbf{o}_t, ..., \mathbf{o}_T)$, nous souhaitons former une quantité qui permet de mesurer l'influence, au sens statistique, de l'observation à la position t_1 sur l'observation à la position $t_2 > t_1$. Pour ce faire, nous définissons 2 quantités calculables à partir du modèle des patrons :

$$f_C(c_j) = P(c_j | \mathbf{o}_1, ..., \mathbf{o}_{t_1}, ..., \mathbf{o}_{t_2-1}) \;\; ; \;\; f_{NC}(c_j) = P(c_j | \mathbf{o}_1, ..., \mathbf{o}_{t_1} = \mathbf{o}_{NI}, ..., \mathbf{o}_{t_2-1})$$

La quantité $f_C(c_j)$ est la distribution de probabilités prédites sur les N catégories à la position t_2 de la séquence, conditionnée par l'observation de $\mathbf{o}_{t_1}$ et influencée par les autres observations. La quantité $f_{NC}(c_j)$ est la distribution de probabilités prédites sur les N catégories à la position t_2, non-conditionnée par l'observation de $\mathbf{o}_{t_1}$ (on a remplacé l'observation par le vecteur $\mathbf{o}_{NI} = (1/N, ..., 1/N, ..., 1/N)$ indiquant qu'aucune information n'est disponible à la position t_1) mais influencée par les autres observations. Nous nous attendons, si l'observation $\mathbf{o}_{t_1}$ n'a pas d'influence sur l'observation à la position t_2, à ce que les distributions $f_C(c_j)$ et $f_{NC}(c_j)$ soient identiques. Pour chacune des N catégories, nous formons la quantité :

$$\Delta_j(\mathbf{o}_{t_1}, \mathbf{o}_{t_2}) = \frac{f_C(c_j) - f_{NC}(c_j)}{f_C(c_j) + f_{NC}(c_j)}$$

Chaque Δ_j peut varier de 1 à -1. Les Δ_j mesurent l'influence statistique en contexte de l'observation $\mathbf{o}_{t_1}$ sur l'observation à la position t_2. Nous devons maintenant prendre en compte la valeur de l'observation $\mathbf{o}_{t_2} = (\alpha_{t_2,1}, ..., \alpha_{t_2,j}, ..., \alpha_{t_2,N})$, les coefficients $\alpha_{t_2,j}$ venant pondérer les quantités Δ_j. Nous retiendrons finalement comme critère de corrélation, entre deux observations $\mathbf{o}_{t_1}$ et $\mathbf{o}_{t_2}$ de la séquence S_T :

$$\mathbf{o}_{t_1} \text{ et } \mathbf{o}_{t_2} \text{ sont corrélées si } \Delta(\mathbf{o}_{t_1}, \mathbf{o}_{t_2}) = \sum_{j=1}^{N} \alpha_{t_2,j} \times \Delta_j(\mathbf{o}_{t_1}, \mathbf{o}_{t_2}) > \Delta_{seuil}$$

avec Δ_{seuil} une quantité seuil à préciser.

3. L'analyse syntaxique en Grammaires de Propriétés

3.1. Les caractéristiques des Grammaires de Propriétés

Les Grammaires de Propriétés permettent de représenter l'information syntaxique de façon décentralisée et locale. Là où les approches classiques manipulent des structures, les GP permettent de spécifier des propriétés sur des ensembles de catégories, voire de traits, indépendamment de toute structure. Cette caractéristique est essentielle pour le traitement de données incomplètes, partielles ou non canoniques. Nous sommes ainsi en mesure d'exprimer des relations entre deux objets indépendamment de leur position dans la chaîne ou dans une structure. De plus, la description d'une unité syntaxique - nous parlons de *construction* - se fait en tenant compte aussi bien des propriétés satisfaites que de celles qui ne le sont pas. Nous sommes ainsi en mesure de proposer une description très fine de l'information, y compris dans le cas de données non canoniques ou non grammaticales.

Sans entrer dans les détails des Grammaires de Propriétés, rappelons-en malgré tout les grandes lignes. La caractéristique essentielle des GP repose sur l'idée qu'il est possible de représenter toute l'information syntaxique à l'aide de propriétés (**?**). Celles-ci sont des relations entre deux ou plusieurs catégories. Les propriétés expriment différents types d'information : l'ordre linéaire, la cooccurrence impérative ou impossible, la dépendance, la répétition, etc. Cette liste est évolutive, il est toujours possible d'ajouter de nouveaux types de propriétés, à condition bien entendu d'en préciser la sémantique. Une catégorie syntaxique est ainsi décrite dans la grammaire par un ensemble de propriétés mettant en relation ses constituants. Voici par exemple quelques propriétés décrivant la construction *SV* :

Linéarité :	V $\prec$ SN
Cooccurrence :	Aux $\Rightarrow$ V[ppas]
Restriction de cooccurrence :	V[intrans] $\not\Rightarrow$ SP

L'analyse d'un énoncé consiste donc à vérifier pour chaque construction de l'énoncé l'ensemble des propriétés qui lui correspondent dans la grammaire. Certaines de ces propriétés peuvent être satisfaites, tandis que d'autres peuvent être enfreintes. Nous obtenons ainsi un ensemble de propriété évaluées (satisfaites ou pas) que nous appelons "*caractérisation*". Une telle approche permet donc de décrire n'importe quel type de construction : partielle, discontinue, non canonique, etc., répondant ainsi à l'objectif exprimé plus haut.

Cette souplesse d'utilisation a cependant un coût : le problème de l'analyse en GP est en effet d'une complexité exponentielle (VanRullen, 2005). Cette complexité, nous y reviendrons, vient tout d'abord de la possibilité offerte de prendre en considération toutes les unités, indépendamment de leur position. Ceci nous permet de prendre en compte des relations distantes ou non projectives entre deux unités (dépendance à distance, constituants discontinus, dépendances croisées). Par ailleurs, l'analyse d'entrées non canoniques repose sur la possibilité de construire des descriptions à l'aide de propriétés plus ou moins satisfaites. En termes d'implantation, une propriété étant une contrainte, il faut proposer une possibilité de relâchement de contrainte. Ces deux facteurs sont essentiels dans la complexité du problème.

3.2. Les mécanismes d'analyse en GP

Nous décrivons dans cette section un mécanisme théorique - et naïf - d'analyse syntaxique en GP. Il ne s'agit pas d'une stratégie d'analyse, mais d'un schéma permettant de décrire les facteurs de complexité.

Le processus d'analyse consiste à instancier l'ensemble des constructions décrivant un énoncé. Ce mécanisme revient en fait à produire les caractérisations leur correspondant. Ce processus nécessite l'identification des constituants entrant en jeu dans la construction, afin d'en vérifier les propriétés pertinentes. Les constructions pouvant être discontinues, il convient donc d'effectuer cette vérification sur toutes les combinaisons de constituants, en d'autres termes l'ensemble des sous-ensembles possibles de catégories prises en considération. Nous appelons *affectation* chaque combinaison de constituants. L'ensemble des catégories de départ est ainsi formé par l'ensemble des catégories correspondant aux mots de l'énoncé à analyser avec leur position dans l'énoncé. Toutes les affectations, construites à partir de cet ensemble, sont évalués. Ceci revient pour chaque affectation à parcourir la grammaire, autrement dit, l'ensemble des propriétés en les évaluant lorsque c'est possible. Pour certaines affectations, aucune propriété n'est pertinente et la caractérisation construite est un ensemble vide et l'affectation est non productive. En revanche, pour d'autres, une caractérisation peut être construite. Au premier niveau de l'analyse, toutes les constructions ont pour constituants des catégories lexicales, comme dans les exemples suivants :

Construction	*Constituants*	*Caractérisation*
SA	{Adv, Adj}	{Adv $\prec$ Adj ; Adv $\rightsquigarrow$ Adj ; ...}
SN	{Det, N}	{Det $\prec$ N ; Det $\rightsquigarrow$ N ; N $\not\Rightarrow$ Pro ; ...}

Une construction instanciée a pour conséquence d'ajouter son étiquette à l'ensemble des catégories à prendre en considération. Dans les exemples précédents, les catégories *SA* et *SN* sont ainsi ajoutées à l'ensemble des catégories lexicales de départ. Un nouvel ensemble d'affectations peut alors être construit, incluant ces nouvelles catégories, permettant ainsi d'identifier de nouvelles constructions.

```
Initialisation
   Pour chaque mot à une position i de l'énoncé
      créer l'ensemble des c_i, ses catégories possibles
   K ← {c_i | 1<i<nombre de mots}
   S ← ensemble des sous-ensembles de K
Répéter
   Pour chaque S_i ∈ S
      si S_i a une caractérisation productive
         ajouter k_i l'étiquette de la caractérisation à K
   S ← ensemble des sous-ensembles de K
Tant que de nouvelles caractérisations sont produites
```

Ce processus d'analyse illustre la complexité liée au nombre d'affectations à prendre en compte, celles-ci devant être reconstruites à chaque étape, c'est à dire chaque fois qu'une nouvelle construction est identifiée. À chaque niveau, il faut donc régénérer l'ensemble des sous-ensembles de catégories.

Précisons maintenant l'opération de *caractérisation*. Le processus consiste, étant donné une affectation, à vérifier toutes les propriétés évaluables de la grammaire. Une propriété *p* est évaluable lorsque l'affectation $\mathcal{A}$ contient les catégories nécessaires au calcul de *p*. Dans le cas de propriétés unaires portant sur une catégorie c il suffit donc que $c \in \mathcal{A}$. Les propriétés binaires ont une évaluabilité différente selon qu'elles sont positives ou négatives. Les premières mettent en relation deux catégories réalisées (par exemple la nécessité de cooccurrence). Dans ce cas, si c_1 et c_2 sont ces catégories, on a $\{c_1, c_2\} \subset \mathcal{A}$. Dans le cas de propriétés binaires négatives, si c_1 et c_2 sont les catégories de p, on a soit $c_1 \in \mathcal{A}$, soit $c_2 \in \mathcal{A}$.

Lorsqu'une propriété est évaluable pour un $\mathcal{A}$ donné, sa satisfaisabilité est calculée, en fonction de sa propre sémantique. Chaque propriété est ainsi associée à un solveur, leur détail n'est pas exposé ici. Le processus global s'écrit :

```
Soit G l'ensemble des propriétés, soit A une affectation
  ∀p_i ∈ G
     Si p_i est évaluable
        Calculer la satisfaction de p_i pour A
        Ajouter p_i et le résultat de son évaluation à la caractérisation C de A
  Décider si C est productive
```

Ce processus signifie que pour toutes les affectations, toutes les propriétés de la grammaire doivent être parcourues pour en vérifier la satisfaction.

Le dernier volet du processus concerne l'évaluation de la caractérisation elle-même. Une caractérisation productive permet d'instancier la catégorie correspondante ou, en d'autres termes, de considérer que cette catégorie est réalisée. Il est alors possible de l'ajouter à l'ensemble des catégories $\mathcal{K}$ à prendre en considération. Une caractérisation est bien entendu productive si elle est entièrement composée de propriétés satisfaites. Mais il est également possible de la considérer comme étant productive même si elle contient des propriétés enfreintes. Ceci revient à relâcher les contraintes. Ce processus doit bien entendu être contrôlé : définition d'un seuil de contraintes à satisfaire, spécification d'une hiérarchisation des contraintes, etc. Ce processus de décision est effectué sur la base de caractérisations complètes.

La construction des affectations, le calcul de leur évaluabilité et de leur productivité constituent les trois processus fondamentaux en GP. Chacun représente un facteur de complexité que nous proposons de contrôler grâce à la corrélation.

4. La corrélation au service des GP

La question est de savoir s'il est possible d'intégrer des informations numériques au cœur de l'architecture symbolique décrite dans la section précédente. Nous proposons pour cela d'appliquer la corrélation aux deux processus fondamentaux de l'analyse en GP : la construction des affectations et celle de la caractérisation.

4.1. Le contrôle des affectations

La génération des affectations, si elle n'est pas contrôlée, est un processus coûteux compte tenu de la quantité de calculs qu'elle entraîne à la fois pour leur construction et leur évaluation. L'utilisation de la corrélation permet d'en réduire nombre.

Ainsi que nous l'avons vu, la corrélation indique une dépendance existant entre deux catégories pour un contexte donné. Toutes les catégories appartenant à une unité syntaxique ont donc une influence mutuelle et sont corrélées.

L'intérêt de la méthode proposée ici réside dans le fait qu'il est possible pour chaque couple de catégories de donner une indication sur leur indice de corrélation. La proposition consiste à limiter la construction des affectations aux seules catégories corrélées. Remarquons que ce calcul est peu coûteux, il est fonction du nombre de patrons, du nombre de catégories et du nombre d'éléments dans la séquence.

```
Soit 𝒫 l'ensemble des affectations
(initialisé aux singletons des catégories de l'énoncé concerné)
  Répéter
     ∀c_i ∈ 𝒦
        ∀𝒜_n ∈ 𝒫
           si ∃c_j ∈ 𝒜_n  tq  Δ(c_i, c_j) > Δ_seuil
              𝒫 ← 𝒜_n ∪ {c_j}
  Tant que de nouvelles affectations sont créées
```

Le principe consiste à créer les affectations par concaténation de nouvelles catégories à une affectation support. La concaténation n'est effectuée que si la nouvelle catégorie est corrélée à au moins l'une des catégories de l'affectation support.

4.2. Le contrôle de la satisfaction

Nous proposons dans cette section de généraliser cette méthode aux propriétés. De la même façon que les probabilités ont été appliquées sur l'ensemble des catégories, nous suggérons de probabiliser l'espace formé de toutes les caractérisations (autrement dit des séquences de propriétés). Une telle information peut être calculée sur la base d'un corpus annoté à l'aide des grammaires de propriétés. A ce jour, les corpus dont nous disposons ont été créés automatiquement et n'ont pas été corrigés. Il peuvent néanmoins former une base d'expérimentation pour l'acquisition de données statistiquement significatives.

L'observation des caractérisations permet, comme pour les catégories, d'indiquer des ensembles de propriétés dont on observe une cooccurrence. Chaque propriété prend valeur dans l'ensemble $\{vrai, faux\}$ indiquant sa satisfaction. Il est possible, comme pour les catégories, d'indiquer des probabilités de satisfaction pour les propriétés, compte tenu des autres propriétés satisfaites. Il est donc possible de proposer des ensembles de propriétés corrélées, cette information étant obtenue comme précédemment pour les catégories.

Cette technique présente l'avantage d'identifier dans l'ensemble total des propriétés formées par la grammaire des sous-ensembles. La vérification de la caractérisation d'une affectation donnée pourra donc porter seulement sur ces sous-ensembles.

Nous sommes ainsi de plus en mesure d'identifier directement des propriétés de plus haut niveau, dont la satisfaction dépend de celle d'autres propriétés. L'exemple suivant illustre ce phénomène :

(1) a. *Elle est longtemps partie*
b. *Elle est partie longtemps*

La différence d'interprétation entre ces deux énoncés dépend de la position de l'adverbe (**?**) : dans le premier cas, il s'agit d'un adjoint, dans le second un complément. Nous avons donc les corrélation suivantes :

$$Adv \prec V \Rightarrow Adv \rightsquigarrow_{adj} V$$
$$V \prec Adv \Rightarrow Adv \rightsquigarrow_{comp} V$$

Cette corrélation permet d'identifier directement des propriétés et les insérer dans la caractérisation sans avoir besoin d'interpréter la caractérisation. D'une façon plus générale, l'utilisation de corrélation entre propriétés est particulièrement utile pour établir les relations existant entre propriétés linguistiques venant de domaines différents (prosodiques, syntaxiques, sémantiques, etc.) et constituera ainsi un mécanisme de contrôle important pour l'analyse de données multimodales.

5. Conclusion

Le développement de techniques d'analyse hybrides, tirant parti à la fois de stratégies numériques et symboliques, est souvent limité à l'utilisation de coefficients de pondération permettant de guider le choix des règles. La notion de *corrélation* décrite dans cet article permet un usage plus fin et généralisable à plusieurs types d'information. Une telle notion est très utile dans le cadre de l'analyse des Grammaires de Propriétés : il devient en effet possible de contrôler à la fois les constituants, mais également les propriétés caractérisant une entrée.

L'expérimentation de la validité de cette proposition et de son efficacité est bien entendu indispensable. Elle est actuellement en cours, sur la base des corpus annotés dans le cadre de la campagne d'évaluation Easy.

Références

ABEILLÉ A. & GODARD D. (2003). « The Syntactic Flexibility of French degree adverbs », in Proceedings of *HPSG 2003*, CSLI publications, pp. 26-46.

BÈS G. & BLACHE P. (1999). « Propriétés et analyse d'un langage », in proceedings of *TALN'99*.

BLACHE P. (2005). « Property Grammars : A Fully Constraint-Based Theory », in *Constraint Satisfaction and Language Processing*, H. Christiansen *et al.* (eds), Springer-Verlag LNAI 3438.

BLACHE P. (2005). « Combiner analyse superficielle et profonde : bilan et perspectives », in actes de *TALN-05*.

BOD R. (1998). *Beyond Grammar*, CSLI.

JOHNSON M. (1998). « PCFG Models of Linguistic Tree Representations », in *Computational Linguistics* 24(4).

KATZ S. (1987). « Estimation of probabilities from sparse data for the language model component of a speech recognizer », *IEEE Trans. ASSP* 35 : 400-401.

KLAVANS J., RESNIK P. (1996). *The Balancing Act. Combining Symbolic and Statisctical Approaches to Language*, MIT Press.

MANNING C., CARPENTER B. (1997). « Probabilistic Parsing Using Left Corner Language Models », in proceedings of *IWPT'97*.

VILNAT A., MONCEAUX L., PAROUBEK P. , ROBBA I., GENDNER V., ILLOUZ G., JARDINO M. (2004). « Annoter en constituants pour évlauer des analyseurs syntaxiques », in actes de *TALN-04*.

VAN RULLEN T. (2005). *Vers une analyse syntaxique à granularité variable*, Thèse de l'Université de Provence.