Interface lexique-grammaire via des structures de traits

Bruno Guillaume, Guy Perrier

Interface lexique-grammaire via des structures de traits

Bruno Guillaume, Guy Perrier

LORIA, Nancy
{Bruno.Guillaume, Guy.Perrier}@loria.fr

Résumé

Nous proposons un mécanisme pour ancrer une grammaire exprimée dans un formalisme particulier dans un lexique indépendant de tout formalisme. Le principe est d'associer à chaque structure élémentaire de la grammaire une interface donnant les informations nécessaires pour choisir les formes fléchies du lexique qui peuvent ancrer cette structure et pour instancier des paramètres morphologiques ou syntaxiques dans la structure. L'intérêt est qu'à la fois le lexique et le mécanisme d'ancrage sont indépendants du formalisme.

1. Introduction

Les grammaires d'interaction (IG) sont un nouveau formalisme pour l'analyse syntaxique de la langue naturelle (Per03). Le développement des ces grammaires et de leur implémentation dans le prototype LEOPAR (BGP03) conduit naturellement à se poser la question des ressources linguistiques nécessaires. En effet, pour valider le formalisme des IG et pour passer du prototype actuel à un logiciel avec une couverture linguistique plus grande, le problème des ressources est essentiel.

Bien que les IG soient conçues pour faire également de l'analyse sémantique (Per04), nous nous concentrons actuellement essentiellement sur l'analyse syntaxique. Dans ce cadre, les ressources utiles existantes pour la langue française sont de trois types :

- Les *lexiques morphologiques* qui donnent des informations sur la flexion des mots. Parmi les ressources disponibles pour la morphologie, on peut citer par exemple Morphalou [1], ABU [2], Litote [3].
- Les *lexiques syntaxiques descriptifs*. Dans la suite, on appelle ainsi les connaissances lexicales syntaxiques sur la langue qui sont exprimées de façon indépendante de tout formalisme, par exemple les informations de sous-catégorisation des mots prédicatifs. Les ressources les plus intéressantes dans cette catégorie sont les tables du LADL (Gro75) qui donnent en particulier des informations précises pour un grand nombre de verbes du français.
- Les règles syntaxiques générales qui sont exprimées directement dans un formalisme particulier et qu'on appellera *grammaires*. La grammaire d'Anne Abeillé pour les TAG (Abe02) est l'un des exemples de grammaires formalisées pour le français ; cette grammaire a une bonne couverture et fournit un bon point de départ pour développer une grammaire équivalente pour un autre formalisme grammatical.

La réutilisation des lexiques morphologiques ne posent pas de réels problèmes. Pour ce qui est des lexiques syntaxiques descriptifs, les données contenues dans les tables du LADL ne sont pas directement exploitables en TAL et nécessite une transformation difficile techniquement comme conceptuellement. Nous participons à un projet en cours qui vise à une telle transformation des tables du LADL (GGPF05). La structure de données proposée dans ce travail est très proche de celle que nous présentons ici.

Enfin, pour ce qui est des grammaires, le travail d'Anne Abeillé est très lié au formalisme des TAG, il est donc impossible de traiter automatiquement ses données pour les rendre utilisables dans notre cadre. Par ailleurs, les grammaires sont de grande taille et donc difficile à écrire et à maintenir. Pour pallier ces défauts, des travaux récents utilise un langage de haut niveau qui est compilé pour décrire ces grammaires (Can99; GCR02). Ce langage de haut niveau permet d'exprimer des généralisations linguistiques : on décrit une hiérarchie de classes (appelée *méta-grammaire*) qui après compilation donne la grammaire attendue (sans l'ancrage au mot). Dans notre cas, nous utilisons le compilateur XMG (DLP04). Ce compilateur et le langage qu'il utilise en entrée sont suffisamment génériques pour qu'ils puissent être utilisés pour différents formalismes (actuellement TAG et IG).

Lorsque toutes ces ressources sont disponibles, il faut les faire collaborer pour construire les entrées de l'analyseur. La combinaison d'un lexique morphologique avec un lexique syntaxique descriptif ne pose pas de problème. En revanche, la grammaire produite par compilation d'une méta-grammaire fournit les structures syntaxiques élémentaires (des descriptions d'arbre dans le cas des IG) utiles pour l'analyse mais ces structures ne sont pas encore reliées à une forme fléchie. Il faut donc un mécanisme d'*ancrage* qui permettent de croiser ces informations liées au formalisme avec les informations indépendantes du formalisme fournies par les lexiques morphologiques et syntaxiques descriptifs.

Comme la méta-grammaire utilise des classes et de l'héritage entre ces classes pour décrire les entrées de la grammaire, une possibilité d'ancrage est alors d'utiliser la trace d'une entrée grammaticale (*i.e.* l'ensemble des classes utilisées pour produire cette entrée). Dans ce cas, l'ancrage dépend de la façon dont la méta-grammaire est conçue et notamment du nom des classes utilisées, ce qui rend l'ancrage difficile et dépendant du formalisme. Cette possibilité est utilisée dans (Kin00) : les hypertags sont des représentations sous-spécifiées qui font référence aux trois dimensions, aux noms des classes et aux noms des

[1] http://www.atilf.fr/morphalou

[2] http://abu.cnam.fr

[3] http://www.loria.fr/equipes/calligramme/litote

familles d'arbres de la méta-grammaire de (Can99). Dans l'architecture de XTAG (XTA01), l'ancrage des arbres est également dépendant du formalisme TAG en faisant référence aux noms des arbres ou des familles d'arbre TAG utilisés.

Pour éviter cette subordination du lexique au formalisme qui sert à écrire la grammaire, nous proposons que la phase d'ancrage soit prévue dans la conception même de la méta-grammaire. La grammaire produite se présente alors comme un ensemble de couples dont l'une des composantes est une structure propre au formalisme grammatical considéré et l'autre une interface dans un format imposé (des structures de traits dans notre proposition). Cette solution est plus contraignante pour l'écriture de la méta-grammaire qui doit explicitement construire l'interface ; cependant cela nous semble un bon compromis pour fournir une méthode simple d'ancrage avec un lexique qui est indépendant du formalisme.

Notons que même si cet article et ses exemples se concentrent sur la syntaxe, on pourrait utiliser le même mécanisme pour la sémantique.

Dans la suite, nous définissons les formats des lexiques, le format de la grammaire, puis la façon de faire l'ancrage en l'illustrant sur un exemple.

2. Format des lexiques

Après combinaison du lexique morphologique avec le lexique syntaxique descriptif, les entrées du lexique résultant apparaissent comme des structures de traits associées à des formes fléchies du français. Ces structures de traits font apparaître à la fois des informations syntaxiques et des informations morphologiques.

Dans notre lexique, les structures de traits utilisent deux niveaux de profondeur. Le niveau externe contient un trait pour la tête de l'entrée et un trait par argument de l'entrée dans le cas où elle est de nature prédicative ; chacun de ces traits a pour valeur une structure de traits interne qui décrit les propriétés morpho-syntaxiques de l'élément correspondant. Pour les structures de traits internes, on suppose fixée une liste de nom de traits (cat, fonct, ...) et pour chaque trait un domaine fini de valeurs atomiques (pour cat : v, n, ...). Une valeur de trait est alors une disjonction de valeurs atomiques prises dans le domaine associé au nom de trait correspondant.

Par exemple pour la forme fléchie ***vendu***, l'entrée de notre lexique résultant correspondant au verbe transitif est donnée par la figure 1 où v est la tête de l'entrée (notée par l'encadrement de v) et a_0, a_1 sont les deux arguments. Les structures de traits internes indiquent que le sujet et l'objet doivent être des groupes nominaux (traits a_0 et a_1) et que le verbe est passivable et réflexivable (trait v).

Sans entrer dans les détails, nous donnons (figure 2) les deux entrées des lexiques syntaxique descriptif et morphologique qui sont fusionnées pour obtenir l'entrée de la figure 1. Pour produire le lexique final, l'entrée du lexique morphologique est unifiée avec la structure de traits interne de la tête de l'entrée du lexique syntaxique descriptif.

$$\textbf{\textit{vendu}} : \left[\begin{array}{ll} a_0 = & \left[\begin{array}{c} \text{cat} = \text{n} \\ \text{fonct} = \text{sujet} \end{array}\right] \\ \boxed{\text{v}} = & \left[\begin{array}{c} \text{cat} = \text{v} \\ \text{mode} = \text{part\_pass} \\ \text{aux} = \text{avoir} \\ \text{pass} = + \\ \text{refl} = + \\ \text{num} = \text{sing} \\ \text{gen} = \text{masc} \end{array}\right] \\ a_1 = & \left[\begin{array}{c} \text{cat} = \text{n} \\ \text{fonct} = \text{objet} \end{array}\right] \end{array}\right]$$

FIG. 1 – Entrée du lexique : forme fléchie ***vendu*** correspondant à l'usage transitif du lemme ***vendre***.

$$\textbf{\textit{vendre}} : \left[\begin{array}{ll} a_0 = & \left[\begin{array}{c} \text{cat} = \text{n} \\ \text{fonct} = \text{sujet} \end{array}\right] \\ \boxed{\text{v}} = & \left[\begin{array}{c} \text{cat} = \text{v} \\ \text{aux} = \text{avoir} \\ \text{pass} = + \\ \text{refl} = + \end{array}\right] \\ a_1 = & \left[\begin{array}{c} \text{cat} = \text{n} \\ \text{fonct} = \text{objet} \end{array}\right] \end{array}\right]$$

$$\textbf{\textit{vendu}} \{\text{lemme} = \text{vendre}\} \left[\begin{array}{c} \text{mode} = \text{part\_pass} \\ \text{num} = \text{sing} \\ \text{gen} = \text{masc} \end{array}\right]$$

FIG. 2 – Une entrée du lexique morphologique pour la forme fléchie ***vendu*** et du lexique syntaxique descriptif pour le lemme ***vendre***

3. Format de la grammaire

Pour la grammaire, il est impossible de définir un format qui soit indépendant du formalisme utilisé. La solution que nous proposons est de représenter chaque entrée de la grammaire par un couple formé d'une part d'une structure syntaxique spécifique au formalisme et, d'autre part d'une structure de traits (que nous appelons interface) qui est utilisée pour décrire les entrées du lexique qui peuvent ancrer la structure syntaxique associée. Les interfaces des entrées de la grammaire suivent le même format que les structures de traits du lexique.

En plus de leur rôle de sélection lexicale, ces interfaces vont permettre aussi d'instancier des paramètres morphologiques et syntaxiques présents dans la structure spécifique au formalisme. Cette instanciation va se faire à partir des données issues du lexique au moment de l'ancrage. Pour cela, nous utilisons un mécanisme de co-indexation entre les traits de l'interface et les données dans la structure spécifique associée. Il est important de noter que cette structure ne doit pas nécessairement utiliser des structures de traits identiques à celles des interfaces (*e.g.* pour les IG, les traits sont polarisés alors qu'ils ne le sont pas dans les interfaces) que la co-indexation peut se faire partout (pas seulement sur l'ancre) et que l'information co-indexée n'est pas forcément morphologique (on peut par exemple préciser la préposition pour un complément

oblique).

Dans la figure 3, nous donnons un exemple de description d'arbre de IG avec son interface. Cet exemple correspond à la construction avec sujet impersonnel d'un verbe transitif à la voix moyenne (la phrase 8 de la figure 5 en fournit une illustration).

4. Ancrage de la grammaire dans le lexique des formes fléchies

À partir d'une forme fléchie, on sélectionne toutes les entrées du lexique associées. Puis, pour chacune de ces entrées, on sélectionne toutes les structures de la grammaire dont l'interface s'unifie avec l'entrée.

Le mécanisme d'unification n'est pas le même selon le niveau (externe ou interne) auquel on se place :

- au niveau externe, les structures de traits sont fermées ; l'unification n'est donc possible que si les deux structures de traits définissent le même ensemble de traits ;
- pour chacun de ces traits communs, la structure de traits interne est obtenue par l'unification ouverte des deux structures internes de traits correspondant au lexique et à la grammaire. Cette opération consiste à faire l'union des traits des deux structures, et pour les traits qui apparaissent des deux côtés, de faire l'intersection des valeurs atomiques.

Si l'unification réussie, l'entrée de la grammaire est sélectionnée pour la forme fléchie, les valeurs de traits coréférencées sont mises à jour et on peut faire apparaître la forme phonologique de la forme fléchie dans la structure grammaticale (dans les IG, c'est le rôle du trait **phon** qui est ajouté sur l'ancre de l'arbre).

L'ancrage de l'entrée de grammaire de la figure 3 avec l'entrée du lexique de la figure 1 donne la description d'arbre de la figure 4.

5. Implantation en cours

L'implantation pour les IG de ce mécanisme est en cours. La mise sous forme de structures de traits du lexique morphologique ne pose pas de problème. Pour le lexique syntaxique descriptif (construit à partir des tables du LADL), la tâche est de plus grande ampleur et fait l'objet d'un travail parallèle (GGPF05).

Une méta-grammaire est en cours de développement (on utilise alors le compilateur XMG pour construire une grammaire). La méta-grammaire a été surtout développé pour le verbe avec la couverture suivante : traitement des auxiliaires avec les différentes formes d'accord du participe passé, voix active, passive et moyenne, construction personnelle et impersonnelle du sujet, inversion du sujet. Tous les cas de sous-catégorisation du verbe sont pris en compte et notamment le cas de compléments qui sont des infinitives ou des complétives. Si ces compléments sont des infinitives, les différents cas de contrôle sont traités (sujet, objet direct, objet indirect…). La construction causative a été laissée de côté pour l'instant.

Des exemples de phrases pour chacune des huit constructions de la grammaire déjà construite compatibles avec l'interface de la figure 1 sont donnés dans la figure 5.

6. Références

A. Abeillé. *Une grammaire électronique du français*. CNRS Editions, Paris, 2002.

G. Bonfante, B. Guillaume, and G. Perrier. Analyse syntaxique électostatique. *Traitement Automatique des Langues*, 44(3), 2003.

M.-H. Candito. *Organisation modulaire et paramétrable de grammaires électroniques lexicalisées*. Phd thesis, Université Paris 7, 1999.

D. Duchier, J. Le Roux, and Y. Parmentier. The metagrammar compiler : An NLP application with a multiparadigm architecture. In *2nd International Mozart/Oz Conference*, 2004.

B. Gaiffe, B. Crabbé, and A. Roussanaly. A new metagrammar compiler. In *TAG+6,Venice*, 2002.

C. Gardent, B. Guillaume, G. Perrier, and I. Falk. Le lexique-grammaire de Maurice Gross et le traitement automatique des langues. In *Journée d'étude de l'ATALA (Interface lexique-grammaire et lexiques syntaxiques et sémantiques)*, 2005. soumis.

M. Gross. *Méthodes en syntaxe*. Hermann, 1975.

A. Kinyon. Hypertags. In *Proceedings of COLING*, pages 446–452, 2000.

G. Perrier. *Les grammaires d'interaction.* Habilitation à diriger les recherches en informatique, Université Nancy 2, 2003.

G. Perrier. La sémantique dans les grammaires d'interaction. In *Proceedings of 11ième Conférence annuelle sur le Traitement Automatique des Langues (TALN'04)*, pages 351–360, 2004.

XTAG Research Group. A lexicalized tree adjoining grammar for english. Technical Report IRCS-01-03, IRCS, University of Pennsylvania, 2001.

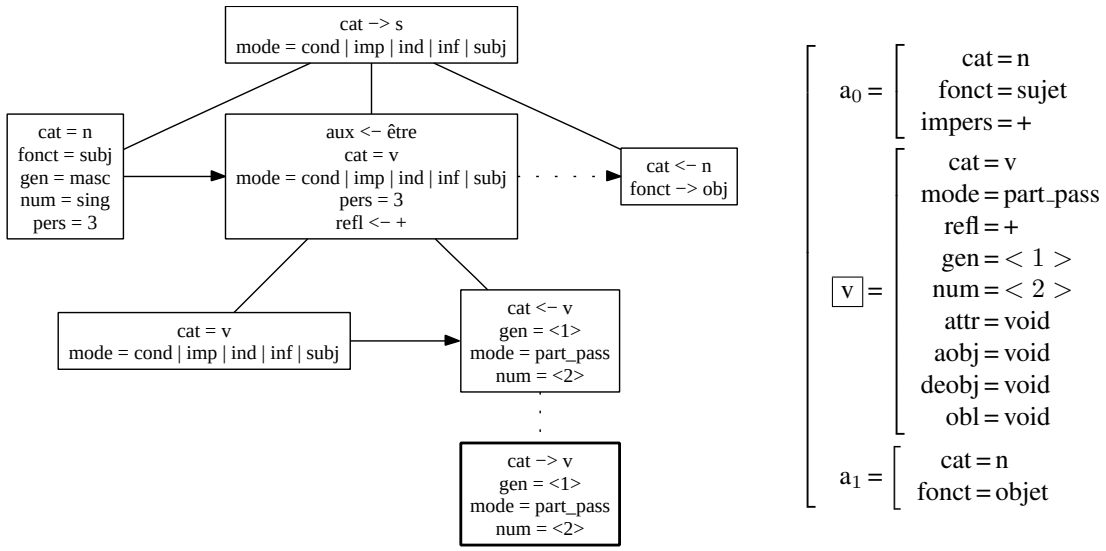

FIG. 3 – Exemple d'entrée de la grammaire : verbe transitif, au participe passé, à la voie moyenne et avec sujet impersonnel

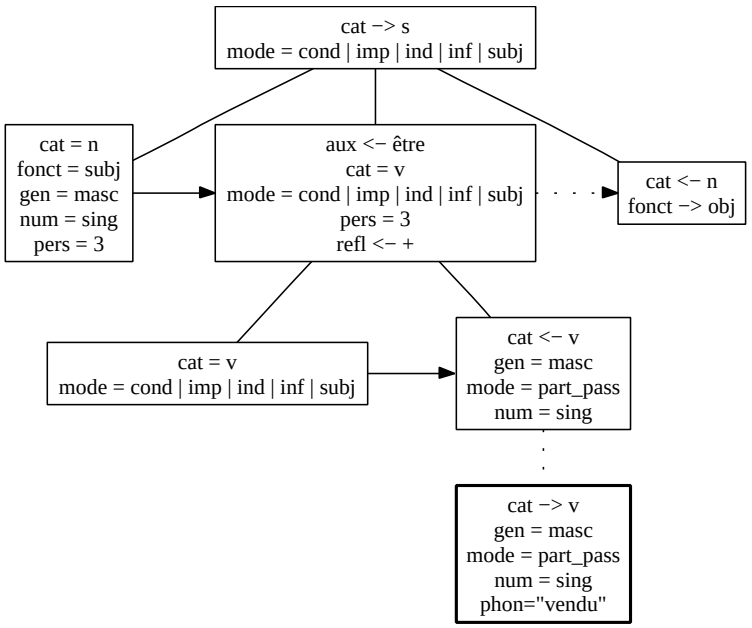

FIG. 4 – Entrée de la grammaire après ancrage

Voix	Impers.	Refl.	Exemples
active	–	–	(1a) *Paul a **vendu** un livre* (1b) *Paul va avoir **vendu** un livre* (1c) *Ayons tout **vendu***
active	–	+	(2) *Paul s'est **vendu***
passive	–	–	(3) *Un livre est **vendu*** (4) *Un livre est **vendu** par Paul*
passive	+	–	(5) *Il a été **vendu** beaucoup de livres* (6) *Il a été **vendu** beaucoup de livres par Paul*
moyenne	–	+	(7) *Ce livre s'est **vendu***
moyenne	+	+	(8) *Il s'est **vendu** beaucoup de livres*

FIG. 5 – Exemple d'utilisation de la forme fléchie ***vendu***