TALN 2005, Dourdan, 6–10 juin 2005

Les Méta-RCG, un formalisme linguistique non-linéaire : description et mise en œuvre

Benoît Sagot
INRIA - Projet Atoll
Domaine de Voluceau, B.P. 105, 78153 Le Chesnay Cedex, France
benoit.sagot@inria.fr

Mots-clefs : Analyse syntaxique, interface syntaxe-sémantique, grammaires non-linéaires, Grammaires à Concaténation d'Intervalles (RCG)

Keywords: Parsing, syntax-semantics interface, non-linear grammars, Range Concatenation Grammars (RCG)

Résumé Nous présentons dans cet article un nouveau formalisme linguistique qui repose sur les Grammaires à Concaténation d'Intervalles (RCG), appelé *Méta-RCG*. Nous exposons tout d'abord pourquoi la non-linéarité permet une représentation adéquate des phénomènes linguistiques, et en particulier de l'interaction entre les différents niveaux de description. Puis nous présentons les Méta-RCG et les concepts linguistiques supplémentaires qu'elles mettent en œuvre, tout en restant convertibles en RCG classiques. Nous montrons que les analyses classiques (constituants, dépendances, topologie, sémantique prédicat-arguments) peuvent être obtenues par *projection partielle* d'une analyse Méta-RCG complète. Enfin, nous décrivons la grammaire du français que nous développons dans ce nouveau formalisme et l'analyseur efficace qui en découle. Nous illustrons alors la notion de projection partielle sur un exemple.

Abstract In this paper, we present a novel linguistic formalism based on Range Concatenation Grammars (RCG), called *Meta-RCG*. We first expose why non-linearity allows a satisfying representation of linguistic phenomena, and in particular of the interaction between the different levels of description. Then we introduce Meta-RCGs and the extra linguistic concepts they manipulate, while remaining compilable into classical RCGs. Moreover, we show that classical analyses (consituency, dependency, topology, predicate-arguments semantics) can be obtained by *partial projection* of a full Meta-RCG parse. Finally, we describe the grammar for French we develop in this new formalism and the associated efficient parser. We illustrate the notion of partial projection with an example.

1 Introduction

Dans (Sagot & Boullier, 2004), les auteurs montrent que les Grammaires à Concaténation d'Intervalles (Range Concatenation Grammars, ci-après RCG) sont bien adaptées à la description du langage naturel. Toutefois, et malgré la disponibilité d'un analyseur (Boullier, 2004), ils ne présentent pas de système d'analyse complet incluant une grammaire linguistique.

L'objectif de cet article est de décrire un tel système d'analyse. Mais cela ne se limite pas à une grammaire : en réalité, les RCG ne sont pas directement utilisables pour représenter des phénomènes linguistiques. Si le principe qui leur est sous-jacent, celui de la concaténation d'intervalles de la chaîne d'entrée, est adapté à la description du langage naturel, d'autres concepts linguistiques de base sont à prendre en considération, tels les phénomènes d'homonymie, de syntagmes à têtes, de traits, d'extraction ou d'interface syntaxe-sémantique.

Les langages définis par les RCG couvrent tout *PTIME*, c'est-à-dire l'ensemble des langages analysables en temps polynomial. Ceci est dû à une propriété fondamentale des RCG, leur non-linéarité (ou clôture par intersection). Dans la première section, nous montrons que cette non-linéarité permet une description appropriée des réalités linguistiques, alors que la plupart des formalismes développés jusqu'à présent ont un squelette linéaire (car clos par homomorphisme, voir plus bas). En particulier, cette non-linéarité permet la définition d'un formalisme qui étend les RCG en prenant en compte les concepts linguistiques de base cités plus haut, tout en restant convertible en RCG standard. La deuxième section donne un aperçu de ce formalisme, appelé Méta-RCG. Enfin, nous présentons dans la troisième section une grammaire du français écrite en Méta-RCG. Ainsi, nous montrons que la grammaire que nous avons écrite fait interagir intimement tous les niveaux d'analyse linguistique, de la morphologie à la sémantique lexicale en passant par la syntaxe. De ce fait, les analyses classiques (en constituants, en dépendances, en boîtes topologiques, en prédicats sémantiques) peuvent être extraites de nos analyses Méta-RCG par *projection partielle*.

2 Non-linéarité et grammaires pour les langues naturelles

La complexité des langues naturelles est au-delà de celle des Grammaires Non-Contextuelles (CFG). Pour des raisons pratiques et théoriques, tous les formalismes envisagés pour décrire les langues naturelles étendent les CFG, même si la façon dont ils les étendent varie. La plupart reposent sur une architecture à deux niveaux. Ils utilisent tout d'abord un *squelette syntaxique* qui étend lui-même les CFG tout en étant clos par homomorphisme[1]. Au dessus de ce squelette, ils mettent en jeu des structures, dites *décorations*, qui sont calculées sur les analyses syntaxiques (souvent des arbres) fournies par le squelette, et ce le plus souvent par des mécanismes reposant sur l'unification. C'est par exemple le cas des Grammaires Fonctionnelles-Lexicales (LFG) ou des Grammaires d'Arbres Adjoints avec décorations (*Feature-based TAG*).

Ce n'est pourtant pas le seul choix possible. En réalité, un formalisme qui étend les CFG et qui est de complexité raisonnable[2] ne peut être clos à la fois par intersection et par homomorphisme. Et choisir, comme cela se fait généralement, la clôture par homomorphisme a une grande influence sur la nature des formalismes. En effet, de nombreux faits linguistiques de différentes

[1]Un homomorphisme étant un cas particulier de substitution, la clôture par homomorphisme découle de la clôture par substitution, plus classiquement utilisée et décrite.

[2]Plus précisément, qui définit un sous-ensemble strict des langages récusivement énumérables.

natures peuvent concerner les mêmes parties d'une phrase. La linéarité induite par la clôture par homomorphisme impose alors de faire un choix : une seule classe de faits linguistiques est décrite par le squelette syntaxique et les autres ne peuvent effectivement qu'être relégués dans des décorations à la complexité mal contrôlée. C'est la raison d'être des architectures à deux niveaux (TAG et traits pour les FTAG, CFG et équations fonctionnelles pour les LFG, etc.), qui ont certains inconvénients[3]. Il est donc intéressant d'explorer l'autre piste, celle des formalismes clos par intersection, ou *non-linéaires*, puisqu'ils permettent de se débarasser des décorations et ont de nombreux avantages computationnels et linguistiques.

Les formalismes non-linéaires à un seul niveau ont la puissance d'expression nécessaire pour décrire le langage naturel, même si l'on fait comme ici l'hypothèse que l'on peut se limiter à des formalismes polynomiaux[4]. En prenant les RCG comme exemple d'un tel formalisme, (Boullier, 2003) présente ainsi des grammaires pour les langages $\{a^{2^n}\}$ et $\{a^n b a^{n-1} b \ldots bab\}$ (nombres chinois, génitifs en géorgien ancien,...) et pour une version abstraite du *scrambling*. De nombreux langages modélisant des phénomènes purement syntaxiques non représentables dans un formalisme syntaxique linéaire peuvent être décrits par des grammaires non-linéaires. Mais cette augmentation de la complexité ne se fait pas au détriment de l'efficacité d'analyse[5].

Par ailleurs, la non-linéarité est appropriée pour décrire les langues car celles-ci reposent elles-mêmes sur des mécanismes non-linéaires : si dans les formalismes à deux niveaux il faut reporter dans les décorations le traitement de certains phénomènes, c'est bien parce que la non-linéarité des langues ne peut être représentée par un formalisme linéaire. On peut citer, comme exemples de phénomènes linguistiques non-linéaires, les mécanismes non-linéaires purement syntaxiques[6], comme les verbes à contrôle ou à montée, et l'interaction entre différents types de contraintes linguistiques, et en particulier la syntaxe et la sémantique lexicale. De fait, un formalisme non-linéaire permet par exemple de contraindre trivialement la construction d'une dépendance à des contraintes syntaxiques (accord, ...) et à des contraintes sémantiques (restrictions de sélection, ...) de manière simultanée, voire à toutes sortes de contraintes[7].

Outre une meilleure représentation des faits linguistiques, la non-linéarité permet une meilleure efficacité d'analyse. En effet, la multiplication des contraintes de diverses natures induit un abandon précoce des analyses invalides : dès qu'une contrainte échoue, l'analyse échoue. Ainsi, plus il y a de contraintes, et plus elle a de chances d'échouer vite. C'est l'antithèse des formalismes à deux niveaux comme LFG, où l'on peut être amené à effectuer un nombre colossal d'analyses de premier niveau (CFG) et à calculer sur les arbres produits de coûteuses décorations (que ce soit fait en une seule phase ou en deux phases).

Un formalisme non-linéaire étendant les CFG couvre au moins *PTIME*. Or nous faisons l'hypothèse que l'on peut se cantonner aux formalismes polynomiaux. Suivant ainsi (Sagot & Boullier, 2004), nous sommes donc intéressés par les formalismes couvrant exactement *PTIME*.

[3]Ces inconvénients sont à la fois computationnels et linguistiques : l'unification induit une complexité algorithmique trop élevée (analyseurs exponentiels), des mécanismes d'interaction (successifs ou simultanés) entre le squelette et les décorations, et un arbitraire linguistique dans la position exacte de la séparation entre squelette et décorations. De plus, il existe des phénomènes purement syntaxiques qui dépassent strictement la puissance d'expression des squelettes linéaires (CFG, TAG, MC-TAG), etc.

[4]Dont les grammaires définissent des langages analysables en temps polynomial en la longueur de l'entrée.

[5]Ainsi, on peut convertir une CFG ou une TAG en une RCG fortement équivalente qui s'analyse respectivement en $O(n^3)$ ou en $O(n^6)$. On peut analyser en $O(n^3)$ toute intersection de CFG. Le langage $\{a^{2^n}\}$, quoiqu'ayant peu de rapport avec la linguistique, montre la puissance de la non-linéarité : il ne respecte pas la propriété de croissance constante (CGP) mais est reconnaissable en temps logarithmique et donc sublinéaire avec une RCG.

[6]Bien qu'aucun formalisme ne les reconnaisse comme tels, puisqu'ils ne sont pas en mesure de les traiter ainsi

[7]Des expériences sont en cours concernant la structuration du discours.

Deux de ces formalismes ont été plus particulièrement étudiés : les *simple Literal Movement Grammars* (sLMG, (Groenink, 1996)) et les Grammaires à Concaténation d'Intervalles (RCG, (Boullier, 2004)). Leur définition formelle et leurs propriétés ont été déjà publiées à maintes reprises, et ne seront pas répétées ici. Rappelons simplement que dans les deux cas une grammaire est un ensemble de clauses à la Horn qui mettent en jeu des prédicats sur des portions de la chaîne d'entrée. Pour plusieurs raisons, les RCG sont plus adaptées aux descriptions linguistiques[8]. Mais ce choix n'est pas véritablement déterminant, et nous l'avons fait en partie aussi en raison de la disponibilité d'un analyseur extrêmement efficace pour les RCG (Boullier, 2004).

3 Les Méta-RCG : une brève description

Si la notion d'intervalle est bien adaptée à la description linguistique, il manque aux RCG un certain nombre de concepts primaires pour pouvoir prétendre au titre de formalisme linguistique. Nous avons donc défini un formalisme qui étend la syntaxe des RCG pour rendre accessibles ces concepts. Ce formalisme, décrit dans (Sagot, 2005), s'appelle Méta-RCG. Il n'introduit pas de décorations formant un second niveau, comme dans les formalismes au squelette syntaxique linéaire : la non-linéarité permet à ces extensions d'être incluses à l'intérieur de la grammaire, par *compilation* de toute Méta-RCG en une RCG classique. Nous avons donc écrit un compilateur qui effectue cette transformation, ainsi qu'un convertisseur permettant de traduire l'analyse d'une phrase obtenue à l'aide de cette RCG en une analyse Méta-RCG.

3.1 Extensions linguistiques de la syntaxe des RCG

Une Méta-RCG, comme une RCG, est constituée de clauses manipulant des prédicats. Et toute RCG est une Méta-RCG. Cependant, les prédicats et les arguments Méta-RCG sont des extensions des prédicats et des arguments RCG. Nous allons donc passer en revue successivement les trois familles d'extensions : les têtes de syntagmes, les traits et numéros d'homonymes, et les contextes.

La notion de tête d'un syntagme est répandue dans la littérature linguistique. Dans le cas des Méta-RCG, l'idée est formulée de la façon suivante : il est impossible, en raison de la non-linéarité des RCG et donc des Méta-RCG, de rendre visible une portion d'analyse à différents endroits de l'analyse, car cela dépasserait *PTIME*. Or on a parfois besoin de rendre visible plus d'informations qu'un simple intervalle (ce qui est la version basique de la non-linéarité). De ce fait, on essaie de construire des représentations partielles de portions d'analyses. On fait alors l'hypothèse suivante : on n'a jamais besoin de connaître à propos d'un syntagme et à l'extérieur de celui-ci plus que ses têtes (il y en a plusieurs en cas de coordination) et un certain nombre de traits (voir ci-dessous). Outre un intervalle simple, un argument Méta-RCG peut donc représenter en plus la liste de ses têtes[9]. On appelle *argument syntagmatique* un tel argument à têtes.

[8] Il est toujours possible de convertir une sLMG en une RCG et inversement. La différence entre RCG et sLMG réside dans ce que l'on entend par « portion » de la chaîne d'entrée. Pour les sLMG, il s'agit de sous-chaînes, alors que pour les RCG il s'agit d'intervalles (c'est-à-dire d'occurrences de sous-chaînes). En conséquence, et à titre d'exemple, il n'y a qu'une seule chaîne vide pour les sLMG, alors que pour les RCG il y a $n+1$ intervalles vides distincts entre les mots d'une phrase de longueur n, ce qui semble plus satisfaisant.

[9] On dispose naturellement d'opérateurs d'empilement d'une tête dans une telle liste.

Les prédicats Méta-RCG étendent les prédicats RCG. Tout d'abord, un prédicat Méta-RCG est décoré par des *traits* définis sur des domaines finis qui permettent de représenter de manière élégante des propriétés de syntagmes, c.-à-d. des propriétés de portions d'analyses et non seulement d'intervalles. Ensuite, on associe à chaque nom de prédicat une liste de positions d'arguments dits *à numéros d'homonymes* : ces arguments, remplis par un intervalle vide ou par un « mot », sont associés à des *numéros* qui différencient les homonymes. Enfin, à chaque nom de prédicat est associée une liste de *contextes* (des intervalles syntagmatiques et/ou des traits) qui peuvent être traités comme des intervalles (ou des traits) standard. Sauf indication contraire, un contexte présent dans deux prédicats d'une même clause a une valeur unique[10].

3.2 Projection partielle d'une analyse globale en analyses classiques

La non-linéarité des Méta-RCG permet de traiter au même niveau les phénomènes de dépendance, de constituance, de sémantique lexicale, et autres. De plus, le fondement même des RCG, à savoir la concaténation d'intervalles, rend la vision topologique de la langue inhérente à toute grammaire Méta-RCG. Aussi l'analyse d'une phrase obtenue à partir d'une grammaire Méta-RCG regroupe-t-elle des faits linguistiques participant de ces diverses classes de phénomènes.

Il est ainsi aisé d'extraire d'une analyse Méta-RCG des *vues partielles* qui constituent des analyses classiques (en constituants, en dépendances, en boîtes topologiques, en relations prédicats-arguments sémantiques, etc.). Cette extraction se fait par *projection partielle* : aucune information n'est calculée à partir de l'analyse Méta-RCG, elles en sont extraites par projection.

4 Une Méta-RCG du français

Le développement du formalisme Méta-RCG s'est fait en parallèle avec celui d'une grammaire du français. Faute de place, nous ne pouvons en donner ici de fragment. Actuellement, elle comporte 370 clauses grammaticales (non lexicales) se compilant en une RCG à 625 clauses, d'arité 73 et de degré maximal 40. Elle couvre déjà un grand nombre de phénomènes linguistiques.

À titre d'illustration, considérons les deux phrases suivantes :

(1) Les entreprises dans lesquelles le Japon veut que la Commission accepte que l'Europe investisse fabriquent des ordinateurs.

(2) Cette idée pose un problème à Nancy.

La phrase (1) inclut une relative qui modifie un verbe enchâssé à travers un contrôle sujet et une complétive. L'efficacité du système permet d'obtenir une analyse unique en 0,38s[11], dont la projection partielle pour obtenir un graphe de dépendances produit la figure 1. La phrase (2) illustre l'interaction entre syntaxe et sémantique au niveau de la grammaire. En effet, *Nancy* étant ambigu (prénom ou lieu), et *pose* (verbe transitif direct ou verbe support à complément d'objet indirect), seules les contraintes sémantiques permettent de n'obtenir que deux analyses, ce qui est bien le cas ici (en 0.01s)[12]. Nous montrons dans la figure 2 l'arbre de constituance (commun aux deux analyses) obtenu par projection partielle de notre analyse.

[10] La notion de *barrière* est alors facilement implémentée : en n'associant pas un certain contexte à un certain prédicat, on l'empêche d'être visible dans la portion d'analyse dont ce prédicat est la racine. D'où un traitement élégant de divers phénomènes, tels que dépendances à longue distance, contrôle, montée, etc.

[11] L'architecture utilisée est un Apple Powerbook avec processeur G4 à 1,5 GHz, et `gcc` 3.3.

[12] L'ambiguïté sur *pose* est levée, mais pas celle sur *Nancy*. Naturellement, si l'on remplace *un problème* par *un vase*, on n'obtient qu'une seule analyse, où *Nancy* est un lieu.

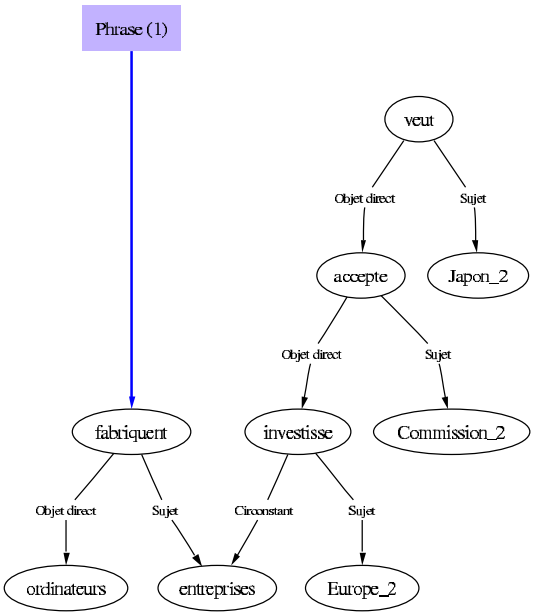

Figure 1. Graphe de dépendance produit par projection partielle de l'analyse de (1).

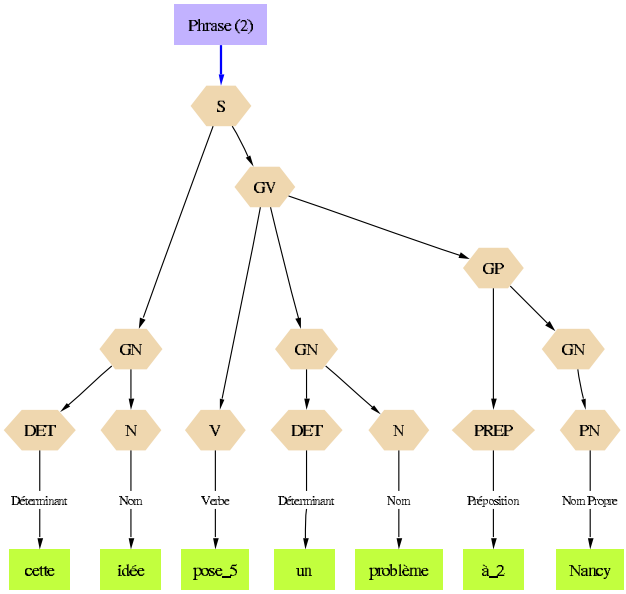

Figure 2. Arbre des constituants produit par projection partielle de l'analyse de (2).

5 Conclusion

Dans cet article, nous avons montré l'importance du concept de non-linéarité pour la description formelle des langues. Nous avons introduit en particulier un formalisme, les Méta-RCG, fondé sur les Grammaires à Concaténation d'Intervalles (RCG). Nous avons ensuite décrit très rapidement la grammaire du français que nous développons, montrant ainsi la pertinence de la prise en compte simultanée de faits linguistiques de natures différentes, et le fait que les analyses classiques peuvent être obtenues à partir de nos analyses par simple projection.

Nous envisageons désormais d'augmenter la couverture de la grammaire, de développer des méthodes automatiques pour la construction du lexique très riche sous-jacent à un tel type de grammaires, et d'étendre la portée de notre grammaire en y incluant des faits linguistiques généralement négligés dans les grammaires, tels que la structuration du discours.

Références

BOULLIER P. (2003). Counting with range concatenation grammars. *Theoretical Computer Science*, **293**, 391–416.

BOULLIER P. (2004). Range concatenation grammars. In *New developments in parsing technology*, p. 269–289. Kluwer Academic Publishers.

GROENINK A. (1996). Mild context-sensitivity and tuple-based generalizations of context-free grammar. In D. JOHNSON & L. MOSS, Eds., *Actes de MoL 4* : Linguistics and Philosophy.

SAGOT B. (2005). Linguistic facts as predicates over ranges of the sentence. In *Proceedings of LACL 05*, Bordeaux, France. To appear.

SAGOT B. & BOULLIER P. (2004). Les RCG comme formalisme grammatical pour la linguistique. In *Actes de TALN 04*, p. 403–412, Fès, Maroc.