GraalWeb ou accéder à une bibliothèque décentralisée de grammaires locales

Matthieu Constant

GraalWeb ou accéder à une bibliothèque décentralisée de grammaires locales

Matthieu Constant
IGM - Université de Marne-la-Vallée
5, bd Descartes
Champs-sur-Marne
77454 Marne-la-Vallée Cedex
mconstan@univ-mlv.fr

Résumé

Cet article présente un système en-ligne de partage de grammaires locales de descriptions linguistiques : la bibliothèque Graal. Celle-ci a pour caractéristique d'être décentralisée : chaque auteur gère ses propres grammaires sur son propre serveur. L'interface GraalWeb permet de visualiser, explorer et télécharger les grammaires disponibles dans cette bibliothèque à partir d'un index précalculé, de manière transparente.

Mots-clefs :

diffusion, grammaires locales, recherche d'information, ressources linguistiques

1 Introduction

Le partage et la diffusion de ressources informatiques est en plein essor depuis une dizaine d'années avec la démocratisation d'Internet. La naissance de Linux et la création de la licence GPL[1] a également généré un élan extraordinaire de diffusion de ressources libres. Dans la communauté TAL, la diffusion de ressources linguistiques est, depuis peu, une composante valorisée du fait de la prise de conscience collective du réel besoin de telles ressources (Romary, 2000). Les corpus et les lexiques sont les ressources les plus usuellement diffusées et partagées. De telles entreprises n'existent pas ou très peu pour les grammaires. Cela s'explique peut-être en partie par le foisonnement des formalismes et des formats. Pourtant, depuis quelques années, avec la lexicalisation des grammaires comme (Abeillé, 2002), le développement de grammaires de plus en plus précises rend nécessaires des collaborations dans des cadres formels unifiés. Ainsi, on assiste à la création

[1] GNU Public License

de "réseaux" dédiés à de telles tâches[2].

Nous nous intéressons spécifiquement aux grammaires locales (Gross, 1993; Gross, 1997), formalisme de description linguistique partagé par la communauté RELEX formée d'une trentaine d'équipes. De ce fait, elles rentrent particulièrement bien dans le cadre d'un projet de diffusion.

Dans cet article, nous décrivons un système de bibliothèque décentralisée de telles grammaires, Graal. Ce système, ouvert à tous, a pour ambition de
– proposer un support simple à des chercheurs isolés pour diffuser leurs grammaires locales,
– faire office, à terme, d'état-de-l'art dans le domaine des grammaires locales,
– permettre une utilisation intensive des grammaires locales dans des applications du TAL au moyen d'outils d'importation.

Cette bibliothèque est accessible en-ligne de manière transparente au moyen de l'applet GraalWeb (`http://igm.univ-mlv.fr/~mconstan/library/`). Elle est pour l'instant limitée aux grammaires au format Unitex (Paumier, 2003).

Dans un premier temps (section 2), nous présenterons assez brièvement les grammaires locales. Nous décrirons ensuite de manière détaillée le système Graal (section 3). Nous développerons enfin quelques aspects pratiques avec la description d'un outil donnant accès à ce système : l'applet GraalWeb, une vue en-ligne de Graal (section 4).

2 Grammaires locales

Les formalismes de grammaires foisonnent en TAL et apparaissent à différents niveaux d'analyse allant de la morphologie à la syntaxe. Les modèles à états finis tels que les expressions régulières, les automates ou les transducteurs ont montré une efficacité certaine, notamment pour l'analyse morphologique et l'analyse lexicale (Mohri, 1997; Karttunen, 2001). L'analyse syntaxique nécessite des formalismes un peu plus évolués, même si plusieurs études ont montré l'intérêt des transducteurs pour cette opération (Roche, 1993; Abney, 1996). Utilisés historiquement, les grammaires algébriques puis les réseaux récursifs de transitions (Woods, 1970) ont montré des limites pratiques et ont évolués vers des grammaires algébriques décorées de contraintes d'unification telles que LFG (Bresnan & Kaplan, 1982). La famille des grammaires d'arbres adjoints (Joshi, 1987) est également très largement utilisée dans la communauté, ainsi que les grammaires de contraintes telles que HPSG (Pollard & Sag, 1994) qui prennent de plus en plus d'ampleur.

Entre analyse lexicale et analyse syntaxique, il existe un niveau intermédiaire aux limites floues formé d'un ensemble de phénomènes locaux figés ou semi-figés. Les modèles à états finis y ont été appliqués avec un certain succès (Maurel, 1990). Le formalisme des grammaires locales (Gross, 1993; Gross, 1999), une extension de ces modèles, est apparu comme une évolution très intéressante du fait de sa simplicité et sa modularité.

[2]On citera, par exemple, le projet LinGO (grammaires HPSG), le réseau RELEX (lexiques et grammaires) ; projet PAPILLON (lexiques multilingues) ; projet lexsynt (lexiques syntaxiques).

2.1 Représentation

Les grammaires locales sont équivalentes à des réseaux récursifs de transitions. Elles comportent deux alphabets disjoints : un alphabet de symboles terminaux et un alphabet de symboles non-terminaux. A chaque symbole non-terminal, est associé un automate sur les deux alphabets. Il existe un symbole non-terminal particulier jouant le rôle d'axiome, soit le point d'entrée de la grammaire. Ces grammaire reconnaissent théoriquement des langages algébriques. Elles ont également une singularité pratique : l'utilisation de masques lexicaux complexes comme étiquettes des automates. Les masques lexicaux définissent des sous-ensembles d'items lexicaux eux-même définis dans des lexiques. Par exemple, l'étiquette *<N+Conc :ms>* correspond à l'ensemble des noms concrets au masculin singulier. Cette singularité ne change rien au niveau formel car un masque lexical peut être remplacé par une disjonction d'items lexicaux. Chaque automate d'une telle grammaire est représenté sous la forme d'un graphe orienté dont les étiquettes sont sur les sommets. Les symboles non-terminaux sont des appels à d'autres automates.

Le formalisme des grammaires locales est partagé par un réseau informel d'une trentaine d'équipes de recherche en informatique linguistique. Les plate-formes Intex (Silberztein, 1993) et Unitex (Paumier, 2003) offrent un cadre unifié de travail (formalisme et formats utilisés).

2.2 Analyse de texte et applications

L'intérêt majeur des grammaires locales est de représenter de manière simple et compacte des contraintes lexico-syntaxiques définissant des classes syntaxiques comme les déterminants nominaux (Silberztein, 2003), les complexes verbaux (Gross, 1999) et même des classes syntaxico-sémantiques comme les adverbes de dates (Maurel, 1990), les prépositions locatives (Constant, 2003). À un moindre niveau, les grammaires locales sont aussi utilisées pour l'analyse locale de surface basée sur des contraintes grammaticales ou graphiques : ex. chunking (Blanc *et al.*, 2007), reconnaissance d'entités nommées (Friburger & Maurel, 2004), etc. Elles servent aussi à l'analyse de textes spécialisés comme les bulletins boursiers (Nakamura, 2005).

L'intégration de grammaires locales dans des processus industriels est de plus en plus courante comme le montre le projet Outilex (Blanc & Constant, 2006) financé par le Ministère français de l'Industrie. Ce projet rassemblant une dizaine de partenaires dont la moitié d'industriels est basé sur la technologie des grammaires locales.

3 Graal, un système décentralisé de catalogue de grammaires locales

Les grammaires locales se développent de manière anarchique dans la communauté et il est difficile d'avoir une vue précise de l'ensemble des grammaires locales existantes. Pour palier à ce problème, nous proposons un système de partage de telles ressources : la bibliothèque Graal[3] qui consiste en

[3]Graal signifie "Grammar and automata library".

un ensemble de serveurs HTTP de grammaires locales comme le montre la figure 1. Ces serveurs jouent l'unique rôle de "dépôt". Un utilisateur ou une application souhaitant avoir accès à leur contenu passent par un serveur d'accès. Ce serveur comporte un index à partir duquel toutes les requêtes sont traitées. L'architecture décentralisée est ainsi transparente pour l'utilisateur.

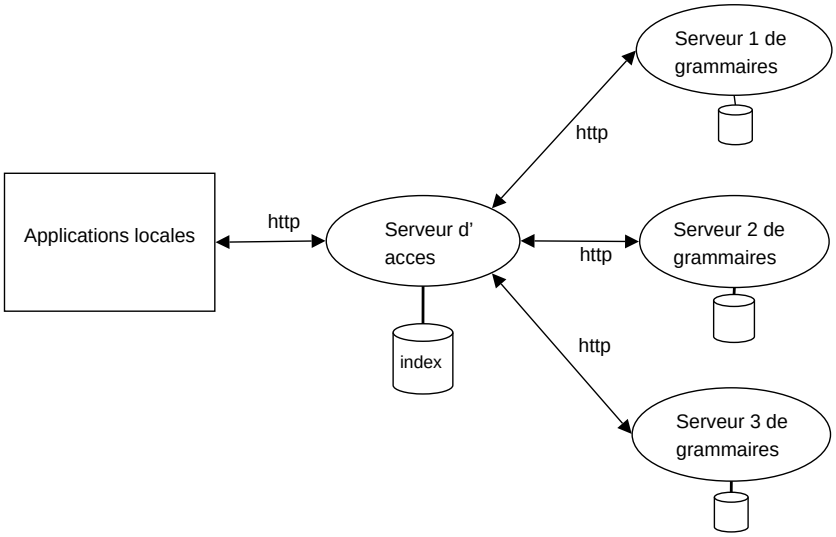

FIG. 1 – Architecture de Graal

3.1 Un ensemble de dépôts indépendants

Le système Graal est donc constitué d'un ensemble de dépôts (ou serveurs de grammaires locales) gérés indépendamment les uns des autres par leur propriétaires respectifs. Un dépôt est défini par une URL de base et un propriétaire. Par exemple, il peut être situé sur le propre site web d'un auteur de grammaires. Cet ensemble de dépôts est connu à l'avance par le système. Tout auteur souhaitant avoir son propre dépôt référencé doit donc en informer l'administrateur de Graal.

Chaque dépôt est composé d'un ensemble de paquetages de grammaires. Un paquetage de grammaires est une archive comprenant une collection d'automates (ou "graphes"), une licence et des documentations XML et HTML. La documentation n'est pas obligatoire mais elle est fortement conseillée car le document XML comprend des informations utiles telles que les auteurs, la langue, une description linguistique, les points d'entrées (les automates principaux), des exemples de séquences reconnues, etc. Notons qu'un automate d'un paquetage peut faire référence à des automates du même paquetage, mais aussi à des automates d'un autre paquetage pouvant être situé dans un autre dépôt[4]. Un paquetage est défini par un dépôt et un chemin relatif dans ce dépôt. Pour être connus du système, les chemins relatifs des paquetages d'un dépôt doivent être listés par son propriétaire dans un fichier défini à l'avance. Ainsi, un auteur est libre de rajouter ou supprimer des paquetages de Graal quand il ou elle le souhaite.

[4]Un système de référencement des automates a été mis en place à cet effet et a été rendu compatible avec Unitex.

3.2 Un serveur d'accès et un index

Cette architecture décentralisée est rendue transparente pour l'utilisateur grâce à un index situé sur le serveur d'accès. Cet index détient des informations précises sur la bibliothèque, notamment sur le contenu linguistique des différents dépôts ce qui permet aux utilisateurs de définir des requêtes spécifiques sur le contenu, lexical notamment (cf. section 4). Graal n'est donc pas seulement un répertoire organisé de liens vers des ressources, avec pour chacune d'elles des descriptions méta-linguistiques, comme la plupart des catalogues en-ligne. Dans son état actuel, l'index comprend les informations suivantes :

– la liste des paquets de grammaires (langue, dépôt, chemin),
– les termes utilisés dans certains champs-clés de la documentation des paquetages,
– les termes utilisés dans les automates,
– la dépendance entre les automates : quels automates appellent quels automates,
– les automates principaux (soit donnés dans la documentation ; soit calculés automatiquement par l'indexeur)

L'indexation des différents dépôts est lancée périodiquement (pour l'instant manuellement). Le référencement d'un paquetage n'est donc pas instantané après son placement dans un dépôt : il faut attendre la prochaine indexation. C'est d'ailleurs le cas avec les moteurs de recherche du type Google. Dans un futur proche, nous souhaitons mettre en place un système de veille afin de "réagir" au plus vite aux mises à jour de la bibliothèque. Par ailleurs, à chaque indexation, les paquetages sont téléchargés pour obtenir une sauvegarde de la bibliothèque.

Notre système ressemble un peu par son architecture à celui proposé dans (Romary, 2000). Cependant, la décentralisation de ce dernier est plus poussée : le serveur d'accès est limité à la gestion des flux de requêtes ; ce sont les différents serveurs de ressources (pour nous, grammaires locales) qui traitent les requêtes elles-mêmes. Nous n'avons pas souhaité reprendre ce système par souci de simplicité.

3.3 La question de la qualité

La décentralisation de notre système engendre un certain nombre de problèmes. En particulier, comment garantir la qualité du contenu de la bibliothèque ? Une trop grande liberté donnée aux auteurs ne risque-t-elle pas de conduire à la construction d'une bibliothèque au contenu linguistique médiocre ? En effet, un tel système offre moins de contrôle qu'un système centralisé. Par exemple, dans un système centralisé tel que celui du projet Papillon (Mangeot-Lerebours *et al.*, 2003) pour la construction de lexiques multilingues, l'analyse préalable des nouvelles entrées par un collège d'experts permet de garantir la qualité du lexique. Ce n'est pas le cas dans notre bibliothèque. Cependant, quelques solutions existent, comme :

– *un contrôle a priori :*
  Lors de la demande d'une personne d'ajouter son site à l'ensemble des dépôts, si le propriétaire n'apparaît pas fiable, l'administrateur est libre de ne pas l'ajouter.
– *un contrôle a posteriori :*
  Une commission d'évaluation de la bibliothèque pourrait être mise en place et rédigerait des recommandations pour chaque dépôt. En cas de non-respect des recommandations par le pro-

priétaire, son dépôt pourrait simplement être exclu de Graal.

4 GraalWeb, le "google des grammaires locales"

GraalWeb est une applet Java qui donne une vue en ligne de Graal au moyen d'un moteur de recherche et d'un explorateur. Elle permet aussi de télécharger les paquetages de grammaires disponibles dans la bibliothèque. Le moteur de recherche utilise des techniques classiques du domaine de la recherche d'informations. L'explorateur permet d'avoir une vue d'ensemble de la bibliothèque (l'ensemble des paquetages, les dépendances entre les grammaires) et une vue détaillée du contenu au moyen d'un visualisateur avancé d'automates.

Pour l'instant, les requêtes tournent en local, c'est-à-dire que l'index est chargé au niveau du client à chaque chargement de l'applet, ce qui nécessite un index de petite taille (i.e. avec un nombre limité d'informations). Dans l'avenir, avec l'augmentation de la taille de Graal, il sera nécessaire de faire évoluer la bibliothèque vers un système où les requêtes sont traitées sur le serveur d'accès (et non en local). Une telle évolution a été prévue dans l'implantation actuelle de GraalWeb.

4.1 Un moteur de recherche

GraalWeb comporte un moteur de recherche permettant de trouver les grammaires utilisant un certain nombre de termes soit dans leur lexique, soit dans leur documentation. En entrée, chaque requête est définie par un ensemble de mots-clés. Son traitement produit une liste d'automates triés selon leur degré de pertinence par rapport à la requête. Cet outil est basé sur des techniques de recherche d'informations classiques avec une représentation des documents et des requêtes au moyen de vecteurs de termes, et des mesures de similarité entre ces vecteurs (Baeza-Yates & Ribeiro-Neto, 1999). Dans l'état actuel de nos travaux, les vecteurs utilisés sont des vecteurs binaires de termes : l'élément associé à un terme est égal à 1 si ce terme apparaît dans le document ; il est égal à 0 dans le cas contraire.

Nous avons développé trois techniques pour la recherche de grammaires selon leur contenu lexical. La première technique consiste à considérer que les termes d'un automate sont ses symboles terminaux. Le degré de similarité entre un automate et une requête est le cosinus de l'angle de leurs vecteurs respectifs. La deuxième technique (indépendante de la requête) consiste à ne tenir compte que de la dépendance entre les différents automates de la bibliothèque et donner un degré d'importance à chaque automate en utilisant la technique du PageRank de Google (Page *et al.*, 1998) : plus un automate est appelé par des automates importants, plus il est important. Nous appelons ce calcul *GrammarRank* en hommage à son illustre inspirateur. La troisième technique consiste à combiner lexique et dépendance. Elle est basée sur le fait qu'un terme utilisé dans un automate est aussi utilisé indirectement par un automate qui l'appelle. Notre algorithme consiste simplement à propager les termes dans le graphe de dépendance inverse de la bibliothèque. Des expériences récentes relativement similaires ont été testées avec succès pour la recherche documentaire sur le Web comme dans (Qin *et al.*, 2005). Le score final d'un automate combine ses trois techniques auxquelles on affecte des coefficients : 0.1 pour la première technique, 0.05 pour le GrammarRank,

0.85 pour la technique hybride.

Lorsque l'on effectue une recherche sur la documentation des paquetages, on utilise une variante de la première technique. Si un paquetage est jugé pertinent, seuls les automates principaux sont listés.

4.2 Un explorateur de grammaires locales

GraalWeb comporte aussi un explorateur de grammaires locales permettant d'avoir une vue à la fois globale et détaillée du système Graal. Tout d'abord, il est possible d'avoir la liste des paquetages disponibles et, pour chacun d'eux, voir la documentation HTML associée (si disponible) pour avoir une idée de son contenu linguistique. La structure de dépendance entre les automates d'un paquetage est également visualisable sous la forme d'un arbre (qui se déploie au fur et à mesure de son exploration) comme pour les explorateurs de systèmes de fichiers. Les noeuds enfants de la racine de l'arbre sont les graphes principaux du paquetage. Chaque automate est également visualisable à l'aide d'un visualiseur de graphes implémenté à partir du code source d'Unitex. Il permet de voir chaque automate de manière détaillée. Un appel à un automate est considéré comme un lien hypertexte que l'on peut suivre pour le visualiser. Notre explorateur permet aussi de suivre la dépendance inverse de la bibliothèque, c'est-à-dire avoir la liste des automates parents de l'automate courant par un simple clic droit de souris, et de les visualiser en les sélectionnant.

5 Conclusion et perspectives

La bibliothèque Graal a l'ambition d'être un système de partage de grammaires locales dans la communauté TAL. Elle a la particularité d'être décentralisée ; plus précisément, chaque auteur dispose ses grammaires sur son propre site. Un indexeur se charge de centraliser les informations sur un serveur d'accès. L'applet GraalWeb permet aux utilisateurs d'avoir une vue en-ligne de cette bibliothèque de manière transparente.

La bibliothèque est pour l'instant limitée aux grammaires Unitex, mais nous projetons de l'étendre à d'autres formats. Par ailleurs, nous souhaitons améliorer nos fonctionnalités de recherche d'informations au moyen de techniques plus évoluées, comme, par exemple, la recherche des grammaires qui incluent une séquence donnée de mots (Constant, 2003). Enfin, il sera prochainement possible de réaliser une projection de tout ou d'une partie de Graal sur un système local intégrant des modules de traitements de textes alimentés par des grammaires locales.

À l'heure actuelle, il existe huit paquetages de grammaires tous fournis par des chercheurs de l'Institut Gaspard Monge. L'ensemble contient environ 1 700 automates de descriptions linguistiques. Nous espérons que ce petit ensemble de grammaires servira de pompe d'amorçage, en incitant les chercheurs du domaine à partager leurs grammaires au moyen de notre système et ainsi permettre de nouvelles avancées significatives dans le domaine.

Remerciements

Cette recherche a été en partie financée par le CNRS et le projet Outilex du Ministère de l'Industrie.

Références

ABEILLÉ A. (2002). *Une grammaire électronique du français*. Paris : CNRS Editions.

ABNEY S. (1996). Partial parsing via finite-state cascades. *Natural Language Engineering*, **2**(4), 337–344.

BAEZA-YATES R. & RIBEIRO-NETO B. (1999). *Modern Information Retrival*. Addison-Wesley.

BLANC O. & CONSTANT M. (2006). Outilex, a linguistic platform for text processing. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, p. 73–76.

BLANC O., CONSTANT M. & WATRIN P. (2007). Segmentation en super-chunks. In *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse : ATALA.

BRESNAN J. & KAPLAN R. (1982). *Lexical-functional grammar : A formal system for grammatical representation*, In J. BRESNAN, Ed., *The Mental Representation of Grammatical Relations*, p. 173–281. The MIT Press : Cambridge, Mass.

CONSTANT M. (2003). *Grammaires locales pour l'analyse automatique de textes : Méthodes de construction et outils de gestion*. PhD thesis, Université de Marne la Vallée.

FRIBURGER N. & MAUREL D. (2004). Finite-state transducer cascades to extract named entities in texts. *Theoretical Computer Science*, **313**, 94–104.

GROSS M. (1993). Local grammars and their representation by finite automata. In M. HOEY, Ed., *Data, Description, Discourse, Papers on the English Language in honour of John McH Sinclair*, p. 26–38. Berlin/New York : Springer Verlag.

GROSS M. (1997). *The Construction of Local Grammars*, In E. ROCHE & Y. SCHABES, Eds., *Finite State Language Processing*, p. 329–352. The MIT Press : Cambridge, Mass.

GROSS M. (1999). Lemmatization of compound tense in english. *Linguisticae Investigationes*, **22**.

JOSHI A. K. (1987). *An introduction to tree adjoining grammars*, In MANASTER-RAMER, Ed., *Mathematics of Language*, p. 329–352. John Benjamins : Amsterdam.

KARTTUNEN L. (2001). Applications of finite-state transducers in natural language processing. In S. YU & A. PAUN, Eds., *Implementation and Application of Automata*, volume 2088 of *Lecture Notes in Computer Science*, p. 34–46. Heidelberg : Springer Verlag.

MANGEOT-LEREBOURS M., SÉRASSET G. & LAFOURCADE M. (2003). Construction collaborative d'une base lexicale multilingue - le projet papillon. *Traitement Automatique des Langues (TAL)*, **44**(2), 151 – 176.

MAUREL D. (1990). Description par automate des dates et des adverbes apparentés. *Mathématiques et Sciences Humaines*, **109**, 5–16.

MOHRI M. (1997). Finite-state transducers in language and speech processing. *Computational Linguistics*, **23**(2), 269–312.

NAKAMURA T. (2005). *Analysing texts in a specific domain with local grammars : The case of stock exchange market reports*, In Y. KAWAGUCHI, S. ZAIMA, T. TAKAGAKI, K. SHIBANO & M. USAMI, Eds., *Linguistic Informatics - State of the Art and the Future*, p. 76–98. Benjamins : Tokyo University of Foreign Studies, Amsterdam/Philadelphia.

PAGE L., BRIN S., MOTWANI R. & WINOGRAD T. (1998). *The PageRank Citation Ranking : Bringing Order to the Web*. Stanford Digital Technologies.

PAUMIER S. (2003). *De la reconnaissance de formes linguistiques à l'analyse syntaxique*. PhD thesis, Université de Marne la Vallée.

POLLARD C. & SAG I. A. (1994). *Head-Driven Phrase Structure Grammar*. Chicago : CSLI Publications.

QIN T., LIU T.-Y., ZHANG X.-D., CHEN Z. & MA W.-Y. (2005). A study of relevance propagation for web search. In *The 28th Annual International ACM SIGIR Conference*, New York, NY : ACM Press.

ROCHE E. (1993). *Analyse syntaxique transformationnelle du français par transducteurs et lexique-grammaire*. PhD thesis, Université Paris 7.

ROMARY L. (2000). *Outils d'accès à des ressources linguistiques*, In J.-M. PIERREL, Ed., *Ingénierie des langues*. Série Informatique et systèmes d'information. Hermès Science : Paris.

SILBERZTEIN M. (2003). Finite-state description of the french determiner system. *Journal of French Language Studies*, **13**(2).

SILBERZTEIN M. D. (1993). *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX*. paris : Masson.

WOODS W. A. (1970). Transition network grammars for natural language analysis. *Communications of the ACM*, **13**(10), 591–606.