Vers la construction d'une bibliothèque en-ligne de grammaires linguistiques

Mathieu Constant

Vers la construction d'une bibliothèque en-ligne de grammaires linguistiques

Matthieu Constant

Université de Marne-la-Vallée

mconstant@univ-mlv.fr

URL : http://ladl.univ-mlv.fr

ABSTRACT. Local grammars efficiently recognize local syntactic constraints in texts. As their number is exploding and the places where they are stored are spread all over the world, we plan to build a system that manages an on-line library of local grammars. We describe their formalism and give an overview of where they are used within the informal European network of RELEX laboratories. Finally, we describe briefly the on-line library we have implemented.

KEYWORDS: lexicon-grammar; local grammars; natural language processing

RÉSUMÉ. Les grammaires locales sont un moyen simple et efficace de repérer et d'analyser des contraintes syntaxiques locales dans des textes. L'explosion de leur nombre et leur éparpillement géographique nous pousse à implanter un outil de gestion : une bibliothèque en-ligne de grammaires locales. Après avoir décrit leur formalisme, nous faisons un large état des lieux de l'utilisation des grammaires locales dans le cadre du réseau informel de laboratoires européens RELEX. Nous insistons principalement sur les travaux réalisés sur le français. Enfin, nous décrivons brièvement notre système de gestion de grammaires locales.

MOTS-CLÉS : grammaires locales ; lexique-grammaire ; traitement automatique des langues

1. Introduction

Le développement des moyens de communication et plus particulièrement d'Internet a fait exploser le nombre de textes électroniques disponibles, rendant ainsi le traitement automatique des langues naturelles et ses applications incontournables depuis quelques années. La plupart des outils implémentés utilisent des approches statistiques [ABNEY 96 ; CHARNIAK 97]. Cependant, depuis longtemps, les chercheurs connaissent l'intérêt d'intégrer à ces systèmes de vastes bases de données de descriptions linguistiques fines [ABEILLÉ 00]. Dans cette optique, le Laboratoire d'Automatique Documentaire et Linguistique puis le réseau de laboratoires européens RELEX accumulent depuis les années soixante-dix une large variété de composants linguistiques où le lexique joue un rôle fondamental. Avec l'aide d'une méthodologie claire et rigoureuse et de la technologie à états finis [ROCHE 97 ; MOHRI 97], de larges dictionnaires et grammaires ont été créés et appliqués à des textes avec les logiciel Intex [SILBERZTEIN 93, 94] et Unitex [PAUMIER 02] et leurs extensions [LAPORTE 99 ; PAUMIER 00]. Actuellement, nous assistons à une augmentation spectaculaire du nombre de ressources notamment des grammaires sous la forme de

graphes. Nous proposons d'implémenter un outil de gestion de grammaires : une bibliothèque en-ligne.

Dans cet article, nous décrivons dans un premier temps le cadre théorique de travail du réseau RELEX[1], puis nous dressons un état des lieux des différents types de grammaires existants. Enfin, nous décrivons notre projet de bibliothèque en-ligne de grammaires et les services implémentés.

2. Le cadre théorique et méthodologique

2.1 Le lexique-grammaire

Le lexique-grammaire est un ensemble de méthodes linguistiques dû à M. Gross (1975), largement inspiré par les travaux de Z. Harris (1968). L'unité minimale d'étude est la phrase élémentaire comprenant un prédicat et des arguments. Chaque prédicat (verbe, nom, adjectif) est classé selon sa structure de surface de base (nombre et forme des arguments). Par exemple, les quatre prédicats *dire*, *demander*, *procès* et *donner* appartiennent à trois classes. Dans les phrases ci-dessous, *Ni* signifie le *i*ème argument d'un prédicat (*i* : entier) et *P* désigne une phrase :

N0 (dire + demander) à N1 que P

N0 donner N1 à N2

N0 faire un procès à N1

Dans les années soixante-dix, M. Gross et l'équipe du LADL ont entamé une étude systématique transformationnelle et distributionnelle, pour chaque prédicat du français. Cette méthodologie a été reprise par les différents laboratoires du réseau RELEX (C. Leclère et al. 1991). Ces études exhaustives ont montré que chaque prédicat avait un comportement quasi-unique. Par exemple, les deux verbes *obéir* et *penser* qui ont la même structure syntaxique de surface *N0 V Prep N1* se comportent différemment lors de la pronominalisation :

Luc obéit à Max = Luc lui obéit

*Luc pense à Max = * Luc lui pense*

Un autre volet fondamental de cette approche est l'étude et la classification systématique des expressions figées. [GROSS 84] Environ 25,000 expressions ont été répertoriées en Français selon la méthode décrite précédemment. Cette liste sert notamment de base à un travail de comparaison entre différentes variantes du français (BFQS = Belge Français Québécois Suisse) [LECLÈRE 91].

---

1. Le réseau RELEX est un ensemble informel de laboratoires européens travaillant dans les domaines de la linguistique et du traitement automatique des langues naturelles. Les différentes équipes travaillent sur un nombre important de langues comme le français, l'anglais, le portugais, l'allemand, l'espagnol, le norvégien, le coréen, le thaï,… Elles utilisent une méthodologie commune : le lexique-grammaire. Le lexique y occupe une place fondamentale, ce qui se traduit par la construction de bases de données linguistiques à large couverture. Le logiciel INTEX sert de plate-forme linguistique commune pour appliquer ces ressources à des textes réels. Des réunions formelles sont organisées tous les ans sous la forme d'une conférence, le Colloque international « grammaires et lexiques comparés », et sous la forme d'un atelier de travail, les journées INTEX [FAIRON 99 ; DISTER 00]. Par ailleurs, il existe un service de veille de corpus journalistiques sur Internet, Glossanet, très utile pour les linguistes [FAIRON 00].

2.2 Les données linguistiques

Le traitement automatique des langues requiert de larges bases de données linguistiques avec un grand degré de précision. Dans le réseau RELEX, il existe trois grands types de données : les dictionnaires électroniques, les grammaires et les tables de lexique-grammaire.

D'abord, les dictionnaires électroniques de formes fléchies de mots simples (DELAF) et de mots composés (DELACF) permettent de reconnaître des unités lexicales et de réaliser un étiquetage lexical précis des textes. [COURTOIS 90]. Ces formes fléchies sont automatiquement générées à partir de leur forme canonique et d'une classe flexionnelle associée. Au départ sous la forme de listes simples, les dictionnaires sont compressés sous la forme de transducteurs à états finis minimaux [REVUZ 91]. Par exemple, le DELAF français comprenant 900,000 mots a une taille d'environ 1 MO dans sa forme compressée. Le codage des entrées lexicales suit le même format dans toutes les langues RELEX. Nous donnons ci-dessous un exemple d'une entrée ambiguë du DELAF *avions* et d'une entrée du DELACF *pommes de terre*. Nous notons qu'il existe aussi des dictionnaires phonétiques DELAP [LAPORTE 88].

avions,avoir.V:I1p (verbe *avoir* conjugué à l'imparfait à la troisième personne du pluriel)

avions,avion.N:mp (nom *avion* au masculin pluriel)

pommes de terre, pomme de terre.N+NDN:fp (nom composé *pomme de terre* au féminin pluriel)

Ensuite, les tables de lexique-grammaire sont un moyen simple et efficace de représenter le comportement distributionnel et transformationnel des prédicats dans les phrases simples [GROSS 75]. Elles ont la forme de matrices. Chaque ligne correspond à une entrée lexicale (ou un prédicat). Chaque colonne correspond à une propriété. A l'intersection, il y a un signe + si l'entrée lexicale accepte cette propriété ; un signe – si elle ne l'accepte pas ; une information lexicale si besoin est.

Enfin, les grammaires sous forme de graphes sont au départ des extensions des dictionnaires de mots composés puis ont évolué vers des niveaux d'analyse supérieurs (voir section 3). Elles peuvent, par exemple, être éditées à l'aide des éditeurs de graphe d'Intex et d'Unitex. Elles présentent de nombreux avantages indéniables comme la représentation compacte de descriptions linguistiques fines [GROSS 97].

3. Grammaires : état des lieux

Dans cette section, nous regardons en détail un type de données linguistiques : les grammaires. Nous établissons un état des lieux dans le réseau RELEX. Nous décrivons d'abord les grammaires locales formellement, puis nous regardons les différents niveaux d'analyse pouvant être faites à l'aide des grammaires. Enfin, nous listons différentes applications utilisant ce type de ressources. Nous parlons essentiellement des travaux réalisés dans la langue française car c'est la langue qui a le niveau le plus avancé. Pour des précisions spécifiques sur d'autres langues, les lecteurs sont priés de se référer à la bibliographie générale.

3.1 Formalisme

Les grammaires que nous considérons ont la forme de réseaux récursifs de transitions [WOODS 1970 ; SILBERZTEIN 93]. Chaque grammaire g possède un alphabet *N*

d'éléments non terminaux, un alphabet *T* d'éléments terminaux (avec $N \cap T = \emptyset$), un ensemble de règles *G* sous la forme de graphes (terme gauche : nom du graphe, soit un élément de *N* ; termes droits : factorisés sous la forme d'un graphe sur $N \cup T$) et un axiome de départ (ou graphe principal) g_0. Par exemple, prenons $N = \{X,Y,Z\}$, $T = \{a,b,c,d\}$ et *X* correspondant au nom de g_0. *G* est représenté par l'ensemble des graphes ci-dessous (fig. 1, 2 et 3). Les graphes se lisent de gauche à droite. Ils sont équivalents à des automates finis. Les états ne sont pas représentés (sauf l'état initial et l'état final). Les étiquettes des transitions se trouvent dans des boîtes. Chaque étiquette grisée est un élément non-terminal (c'est-à-dire un appel à un sous-graphe). Ainsi, notre grammaire reconnaît des expressions telles que *aaccbb* ou *bcdab*.

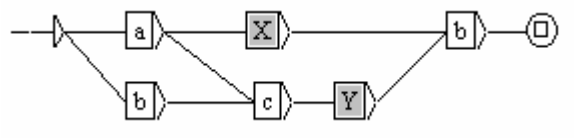

Figure 1 : X

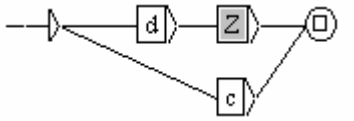

Figure 2 : Y

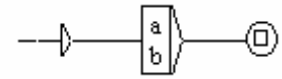

Figure 3 : Z

Pour l'instant, nous avons défini les éléments des alphabets comme de simples symboles. Nous précisons maintenant leur forme réelle dans nos grammaires linguistiques. Les symboles non-terminaux sont des noms arbitraires donnés à des graphes. Sous Intex et Unitex, ces noms sont toujours précédés du caractère « *:* ». Les symboles terminaux représentent, dans la majorité des cas, des mots au sens linguistique et sont donc très variés. Afin de limiter le nombre de transitions dans les graphes, nous utilisons des abréviations pour désigner des ensembles d'éléments terminaux. Par exemple,

– <*station*> désigne toutes les formes fléchies de la forme canonique *station* : {*station, stations*}
– <*V*> désigne n'importe quel verbe codé dans nos dictionnaires, c'est équivalent au OU logique de tous les verbes codés dans le dictionnaire
– <*N:ms*> désigne n'importe quel nom masculin singulier du dictionnaire (noms simples et composés)
– <*NB*> désigne n'importe quel nombre représenté par l'expression régulière : *(0+1+2+3+4+5+6+7+8+9) (0+1+2+3+4+5+6+7+8+9)** où le symbole * représente le symbole de Kleene et le signe + symbolise le OU logique

Ainsi, l'exemple 5 représente une classe de mots composés sémantiquement proches de *station de ski*, et reconnaît des expressions telles que *station de sports d'hiver* ou *station*

de haute montagne. Dans certains cas, l'alphabet des symboles terminaux n'est pas composé de mots linguistiques mais de caractères typographiques de la langue de travail. L'exemple 4 représente les différentes variantes orthographiques du toponyme *Vietnam* (O. Piton, D. Maurel 1997). Notons également qu'un mélange des deux niveaux est possible, notamment pour des formes telles que *re#<V>* où # est un symbole qui interdit l'espace entre le préfixe *re* et le verbe *<V>*.

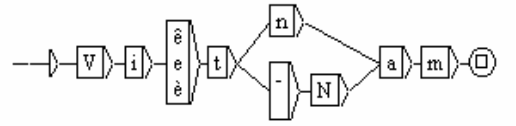

Figure 4: Vietnam

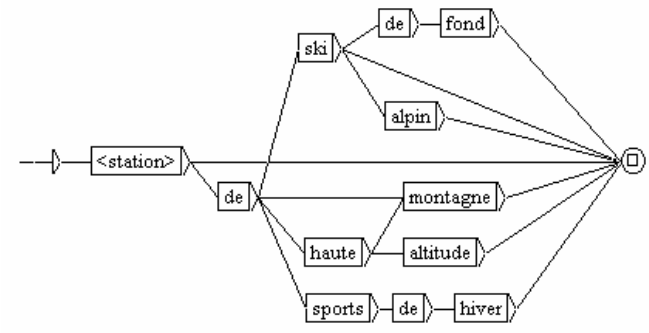

Figure 5: Station

Il est possible d'ajouter des informations en sortie des graphes. Ainsi, nos grammaires peuvent se comporter comme des transducteurs à états finis [SILBERZTEIN 99]. Par exemple, le graphe 6 qui décrit des adverbes de temps tels que *à l'aube* ou *en fin de matinée* peut servir à étiqueter les expressions qu'il reconnaît comme des adverbes de temps à l'aide des informations de sortie écrites en gras sous les boites du graphe. Ainsi, après application de cette grammaire, le texte *Marie est arrivée en fin de matinée* peut être étiqueté *Marie est arrivée <ADV+Time> en fin de matinée<\ADV+Time>*.

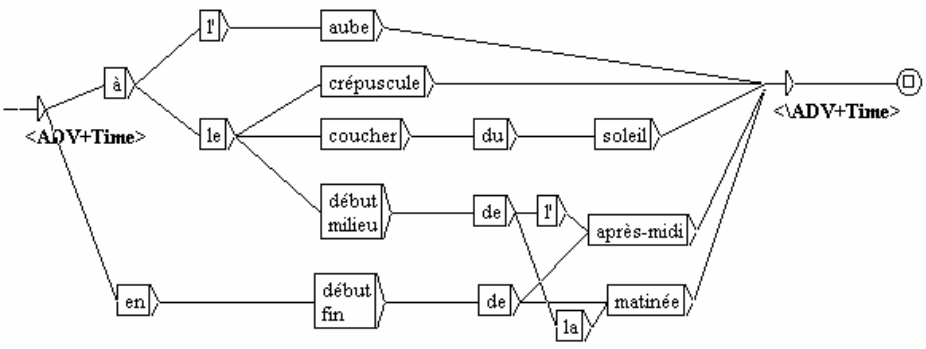

Figure 6 : Adverbes de temps

Par ailleurs, il est possible de construire des graphes décrivant des règles de réécriture : ce sont des graphes à variables. Le graphe 7 permet de décrire la phrase *Paul est à 10 km de la ville* et de l'interpréter sémantiquement : la distance *d* entre *Paul* et la *ville* est égale à *10 km* ou *d(Paul,la ville)=10 km*. En effet, les séquences reconnues par les morceaux de graphes entre parenthèses indexés par le nombre *i* (*i* : entier) sont stockées dans les variables *$i* et sont réécrites en sortie de l'application du graphe quand elles apparaissent en sortie dans le transducteur.

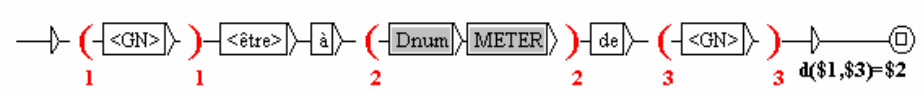

Figure 7: Interprétation sémantique

Enfin, les graphes-patrons représentent des sur-ensembles de séquences [ROCHE 93 ; SENELLART 99] et permettent de transformer les informations codées sous la forme de tables de lexique-grammaire en graphes. Pour chaque entrée lexicale (ou chaque ligne), on crée automatiquement un graphe associé à partir des informations de la table. Nous utilisons un système de variables qui sont placées dans les boîtes des graphes. Soient *i* et *j* deux entiers, *TLG* une table de lexique-grammaire et *g* le graphe-patron associé à *TLG*. Etant donné une ligne *i* de *TLG*, la variable *@j* (placée dans *g*) correspond au contenu de l'intersection de la ligne *i* et de la colonne *j* de *TLG*. Ainsi, chaque variable correspond à une colonne des tables, soit une propriété, c'est-à-dire un ensemble de séquences ou/et des informations lexicales. Pour chaque ligne *i*, si *TLG(i,j)* = -, alors on supprime la boite contenant *@j*, on supprime un ensemble de séquences non désirées. Si *TLG(i,j)* = +, on remplace *@j* par l'élément vide. Par défaut, on remplace *@j* par le contenu de *TLG(i,j)*. Par exemple, à partir de la table 8 et de son graphe-patron associé (figure 9), on génère automatiquement les deux graphes de la figure 10.

C1	C2
X	+
Y	-

Figure 8: Table

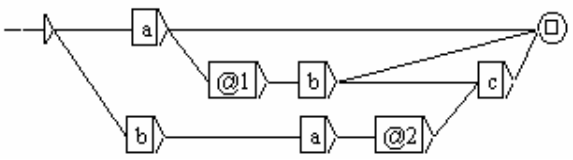

Figure 9 : Graphe de référence

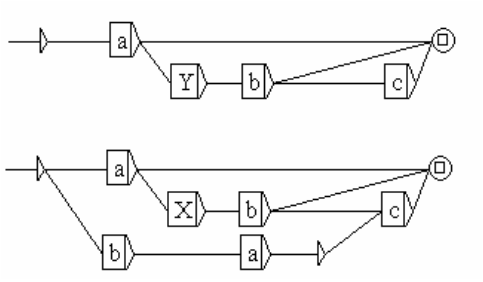

Figure 10: Graphes générés à partir de la table 8 et le graphe 9

3.2 Les différents niveaux d'analyse

Le gros avantage des grammaires sous la forme de graphes est qu'elles permettent différents niveaux d'analyse des textes. Dans la précédente section, nous avons déjà distingué deux niveaux selon l'unité minimale utilisée (caractère ou mot). Lorsque l'unité minimale est le caractère, nous pouvons parler de traitement morphologique. Dans ce cas, les graphes utilisés servent à décrire des variantes orthographiques de manière compacte et donc à alimenter les dictionnaires électroniques. Nous nous intéressons maintenant au cas où l'unité minimale est le mot. Les niveaux d'analyse y sont plus nombreux. Tout d'abord, les graphes peuvent être assimilés à des extensions des dictionnaires des mots composés. Par exemple, la description des dates sous la forme d'automates factorise de façon significative un ensemble d'expressions quasiment impossible à traiter sous forme de listes [MAUREL 90 ; BAPTISTA 99]. Ensuite, il est possible de décrire les contraintes locales autour d'un mot de manière très fine. Ainsi, nous pouvons constituer des classes de mots composés ayant un sens proche, comme le graphe **Station**. Ce dernier graphe permet notamment de distinguer deux entrées lexicales de *station* : (1) *station (E + de ski)* et (2) *station (E + de métro)*. J. Senellart (1998) a construit des grammaires pour des noms d'activités telles que *ministre (E+ de l'intérieur)*. L'étape suivante est de construire des constituants de phrases comme les groupes nominaux ou les groupes verbaux comme le montre M. Salkoff (1973) pour construire une grammaire en chaîne du français. Afin de reconnaître automatiquement des expressions figées dans les corpus, J. Senellart (1999) a élaboré quelques grammaires de groupes nominaux simples comme montré dans le graphe 11 ci-dessous. C. Dominguès (2001) a regardé le comportement de groupes nominaux contenant une coordination. La constitution de grammaires complètes reconnaissant les GN est l'un des futurs enjeux du réseau RELEX. Par ailleurs, il existe des grammaires de groupes verbaux composés en anglais [GROSS 99].

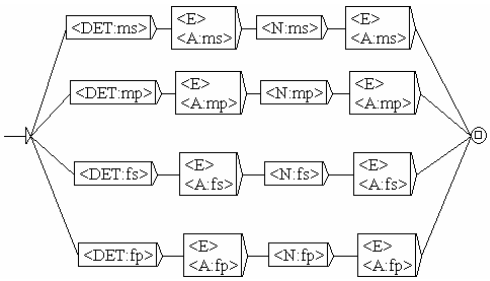

Figure 11: GN

A partir de l'étape précédente, il est possible de décrire des phrases simples libres contenant un prédicat (verbe, nom, adjectif) et des arguments comme l'a fait E. Roche (1993, 1999) à l'aide de tables de lexique-grammaires et de transducteurs à états finis. Un travail de grande envergure dans la continuité de cette étude est actuellement mené au sein de l'université de Marne-la-Vallée. J. Senellart (1999), à l'aide de graphes, a décrit les expressions figées du français, à partir des tables de M. Gross (1983) et montré par la même occasion leur présence en grand nombre dans les textes français (de l'ordre de 30% dans des textes journalistiques). Les phrases simples avec prédicats nominaux et adjectivaux peuvent être réduites en GN ou GA. Leur description permet d'affiner celle des GN généraux. La reconnaissance de phrases complexes, combinaisons de phrases simples sous toutes les formes, d'adverbes de temps (M. Gross 2002), de lieux (M. Constant 2002), d'incises [FAIRON 00] et de conjonctions, est un objectif majeur du réseau RELEX. Enfin, quelques études ont été menées dans le domaine spécifique de la bourse [GROSS 97 ; NAKAMURA 01) et ont montré la limitation du lexique et des structures syntaxiques employés. Des grammaires entièrement lexicalisées ont ainsi pu être construites.

3.3 Les applications

Nous décrivons, dans cette section, quelques applications qui ont été développées par les membres de la communauté RELEX. Elles sont nombreuses ; nous listons les principales :

Découpage : Une des premières étapes du traitement automatique des textes est la segmentation de ces derniers en phrases à l'aide de la ponctuation. Des transducteurs ont été construits à cet effet, insérant à la fin des phrases un symbole comme *{S}* [DISTER 01]. Les résultats sont encourageants et dépendent beaucoup des corpus.

Morphologie : Un problème crucial dans la construction de dictionnaires à large couverture est la génération automatique de toutes les formes fléchies d'un lemme. A chaque lemme est associé une classe de flexion qui représente un transducteur. L'application de ce transducteur permet de résoudre une grande majorité des problèmes rencontrés [SILBERZTEIN 97].

Étiquetage : La consultation des dictionnaires permet de faire un étiquetage lexical des textes. Le résultat est représenté sous la forme d'un transducteur permettant par la même occasion de montré l'impressionnante ambiguïté de la langue [SILBERZTEIN 97]. Les grammaires lexicalisées permettent de reconnaître des séquences composées et d'étiqueter ces dernières de manière très satisfaisante. La prochaine étape est l'analyse syntaxique complète de textes. Pour cela, il est nécessaire de trouver, en tenant compte de l'énorme ambiguïté de la langue, de nouveaux formalismes et des algorithmes associés. Des travaux sont en cours à l'université de Marne-la-Vallée.

Extraction : Un des sujets à la mode actuellement est l'extraction d'information. Le plus brûlant d'entre eux est l'extraction de noms propres. De nombreuses études ont été menées et sont en cours. J. Senellart (1998) extrait automatiquement des noms de personnalités en leur associant une fonction politique ou professionnelle à l'aide de grammaires sous forme de graphes. N. Friburger et al. (2001) extraient des noms propres de personne à l'aide de cascades de transducteurs. D'autres sujets ont aussi été abordés comme l'extraction de noms de gènes dans les corpus en génomique [POIBEAU 01].

Filtrage : Le filtrage d'information est aussi un sujet majeur de ces dernières années, notamment pour la distribution personnalisée des dépêches AFP. A. Balvet (2000) montre l'intérêt d'utiliser des graphes linguistiques pour repérer les textes adéquats à une requête donnée.

Ambiguïté : La levée d'ambiguïté des textes est fondamentale pour le traitement automatique. De nombreuses études pointues ont montré l'intérêt d'utiliser des batteries de transducteurs [SILBERZTEIN 97 ; DISTER 99 ; CARVALHO 02]. C'est un sujet très important dans la communauté : un module de levée d'ambiguïté (ELAG) a même été implémenté par E. Laporte et al. (1999) qui permet de supprimer les mauvais chemins dans le transducteur du texte.

Traduction : La traduction automatique des langues est sûrement l'objectif le plus difficile du TAL [GROSS 92]. Devant la difficulté de la tâche, les études dans ce domaine s'orientent vers l'aide automatique à la traduction. C. Fairon et J. Senellart (1998) ont construit un ensemble de transducteurs lexicalisés traduisant des adverbes de temps du français à l'anglais. Dans le même esprit, J. Baptista et D.Catala (2002) ont réalisé quelques grammaires autour de trois mots permettant de traduire des adverbes de temps du portugais vers l'espagnol et vice-versa.

Génération : La génération n'est pas un sujet très porteur dans la communauté RELEX plus attirée par l'analyse. Cependant, il existe un projet de génération automatique de sujets d'examen à l'aide de graphes à mémoires ou graphes de réécriture (FAIRON 01).

4. Gestion de grammaires : une bibliothèque en-ligne

4.1 Vers une centralisation des grammaires

Dans la section précédente, nous avons vu la diversité des grammaires que nous utilisons. Nous avons avant tout exposé les travaux sur le français qui reste la langue la plus avancée dans les descriptions linguistiques. Mais les autres langues auront atteint le même niveau dans quelques années. Ainsi, nous prédisons une future explosion du nombre de grammaires disponibles. Par ailleurs, la dispersion géographique des laboratoires est source de redondance. Le formalisme utilisé (la modularité des grammaires) est une qualité essentielle pour le travail en équipe. En effet, un chercheur construisant une grammaire peut facilement insérer une sous-grammaire déjà construite par lui-même ou par un autre, au moyen d'un appel à un sous-graphe, aussi faut-il que ces sous-graphes soient facilement accessibles.

Actuellement, il n'existe pas de gestion commune des grammaires, ce qui est préjudiciable pour la collaboration et est source de redondance. Les seuls moyens de s'informer ou d'informer les autres de l'existence d'une grammaire particulière sont les articles, les communications aux différents colloques ou les discussions orales ou par courrier électronique. Le meilleur moyen pour se transmettre les grammaires est le courrier électronique. Nous proposons de créer un outil de gestion des grammaires permettant de centraliser les grammaires dans un même endroit et de donner un moyen simple aux utilisateurs de stocker leurs données et de consulter le catalogue de grammaires disponibles.

4.2 Organisation de la bibliothèque

Le système que nous proposons a une architecture client-serveur. Du côté serveur, est disposée la base de données contenant les grammaires et différentes informations, et y

sont aussi réunis un certain nombre de modules traitant cette base de données. Du côté client, est fournie une interface au moyen de laquelle l'utilisateur envoie ses requêtes au serveur.

La base de données contient deux principales entités : les grammaires et les utilisateurs. L'entité utilisateur contient quelques données sur les utilisateurs du système. L'entité Grammaire contient les grammaires sous forme de graphes (physiquement un fichier pour chaque graphe), des informations techniques relatives aux graphes (langue, auteur, type…) et de la documentation écrite sur les graphes.

Chaque utilisateur a un compte personnel auquel il accède au moyen d'un nom d'utilisateur et d'un mot de passe. Il l'organise librement à l'aide d'un système d'arbre de répertoires comme dans tous les systèmes d'exploitation. Tous les utilisateurs ont accès en lecture à toutes les grammaires. Les auteurs ont accès en écriture à leurs propres graphes uniquement.

4.3 Les différentes opérations : stockage et recherche d'information

Pour l'instant, il existe deux types d'opérations qui ont été implémentées : le stockage de grammaires et la recherche simple d'informations.

Une caractéristique de cette bibliothèque est que les graphes n'existent pas seulement physiquement, ils sont aussi documentés. Ainsi, lors de l'insertion d'un graphe dans la bibliothèque, son auteur doit fournir un certain nombre d'informations dont une description linguistique de celui-ci. Côté client, il existe un éditeur de documentation facilitant la tâche de l'utilisateur. Les graphes d'une grammaire sont automatiquement calculés et fournis sous la forme d'une liste. Ainsi, pour éditer la documentation d'un graphe, il lui suffit de cliquer sur un graphe de la liste et un formulaire simple à remplir apparaît pour ce graphe. Par ailleurs, certaines caractéristiques des graphes (comme le type) sont automatiquement calculées réduisant le nombre de champs à remplir dans le formulaire.

Le stockage de grammaires comprend deux opérations essentielles : insertion et suppression. Contrairement à des bases de données classiques qui contiennent des objets simples, ces opérations ne sont pas évidentes du fait de la complexité du formalisme des grammaires. Les algorithmes utilisés mettent en jeu des objets mathématiques complexes comme les composantes fortement connexes lors de la suppression. Nous n'entrons pas dans les détails dans cet article principalement dédié à une présentation générale du système. Par ailleurs, l'utilisateur peut créer dans son compte personnel son arbre de travail personnel en insérant ou supprimant des répertoires. Notre outil de stockage permettra d'insérer un dictionnaire personnel de l'utilisateur quand il est utile dans certains graphes. Exemple : supposons que l'on insère une grammaire sur les noms de ville, il est nécessaire d'utiliser un dictionnaire de toponymes, non implanté dans le dictionnaire général du système. De même, il sera possible d'insérer la table de lexique-grammaire associée à un graphe patron.

Un utilisateur peut être tenté de rechercher des informations sur le catalogue de grammaires puis de télécharger les graphes qui l'intéressent. Nous avons implémenté un explorateur avancé de la bibliothèque de graphe. Cet explorateur effectue des tris et des filtres sur des critères simples comme l'auteur, le type, la langue… D'autres outils plus complexes ont été implémentés. Par exemple, il existe des modules qui recherchent toutes les grammaires dans lesquelles certains mots existent ou toutes les grammaires qui reconnaissent ou comprennent une séquence de mots donnée. Il est également

possible de faire des recherches sur les descriptions textuelles des graphes ou les mots-clés.

Il peut être intéressant de vérifier si une grammaire n'existe pas déjà, ce qui revient à faire une intersection de grammaires. Or, l'intersection de grammaires hors contextes est indécidable (GROSS 67). Dans l'avenir, il conviendra de réaliser des intersections approximatives.

5. Conclusions et perspectives

Dans cet article, nous avons décrit le cadre de travail de la communauté RELEX, puis fait un large état des lieux sur un type de données linguistiques : les grammaires. Nous avons montré la nécessité de construire un outil de gestion de cette ressource : une bibliothèque en-ligne de grammaires. Nous avons décrit son organisation et les services implémentés (stockage et recherche). Dans le futur, nous espérons améliorer les services déjà existants notamment en permettant des recherches sur des critères complexes (cf. 4.3). Nous souhaitons aussi, permettre aux utilisateurs d'échanger des informations (critiques et remarques sur les grammaires notamment). Par ailleurs, il faudrait étendre cette bibliothèque aux autres données linguistiques (tables de lexique-grammaire et dictionnaires) de façon plus large.

Références

[ABEILLÉ 00] ABEILLÉ A., BLACHE P., Grammaires et analyseurs syntaxiques, In: PIERREL J.M., *Ingénierie des langues*, Hermès science publications, Paris, pp. 51-76, 2000

[ABNEY 96] ABNEY S., Statistical Methods and Linguistics, In: Judith KLAVANS and Philip RESNIK (eds.), *The Balancing Act: Combining Symbolic and Statistical Approaches to Language.* The MIT Press, Cambridge, MA, 1996

[BALVET 00] BALVET A., Evaluation de stratégies linguistiques pour le filtrage d'information, In A. Dister (ed.), *Actes des Troisièmes Journées Intex*, Revue Informatique et Statistique dans les Sciences Humaines, Liège, Université de Liège, pp. 29—52, 2000

[BAPTISTA 99] BAPTISTA J., Manhã, Tarde, Noite: analysis of temporal adverbs using local grammars, *Seminarios de Linguistica 3*, Faro, Universidade do Algarve, pp. 5–31, 1999

[BAPTISTA 02] BAPTISTA J., D. CATALÀ, Compound temporal adverbs in Portuguese and in Spanish, In: Ranchhod E., Mamede N. (Eds), *Advances in Natural Language Processing*, Lecture Notes in Artificial Intelligence (LNAI), 2389, Springer, pp., 2002

[CHARNIAK 97] CHARNIAK E., Statistical techniques for natural language parsing, *AI Magazine*, 1997.

[CARVALHO 02] CARVALHO P., MOTA C., RANCHHOD E., Complex lexical units and automata, In: Ranchhod E., Mamede N. (Eds), *Advances in Natural Language Processing*, Lecture Notes in Artificial Intelligence (LNAI), 2389, Springer, pp. 229-238, 2002

[CONSTANT 02] CONSTANT M., On the analysis of locative phrases with graphs and lexicon-grammar: the classifier/proper noun pairing, In: Ranchhod E., Mamede N.

(Eds), *Advances in Natural Language Processing*, Lecture Notes in Artificial Intelligence (LNAI), 2389, Springer, pp. 33-42, 2002

COURTOIS B., SILBERZTEIN M., *Les dictionnaires électroniques du français*, Langue Française 87, Paris: Larousse, 1990

[DISTER 00] DISTER A., *Actes des Troisièmes Journées Intex*, Revue Informatique et Statistique dans les Sciences Humaines, Liège, Université de Liège, 2000

[DOMINGUES 01] DOMINGUES C., *Etude d'outils informatiques et linguistiques pour l'aide à la recherche d'information dans un corpus documentaire*, Thèse de doctorat en informatique, Université de Marne-la-Vallée, 2001

[FAIRON 99] FAIRON, C., SENELLART, J., Classes d'expressions bilingues gérées par des transducteurs finis, dates et titres de personnalité (anglais-français). Linguistique contrastive et traduction, Approches empiriques. Louvain-la-Neuve, 1999

[FAIRON 99] FAIRON C., *Analyse lexicale et syntaxique: le système INTEX*, Lingvisticae Investigationes, John Benjamins Publishing Company, Amsterdam-Philadelphia, 1999

[FAIRON 00] FAIRON C., *Structures non-connexes : Grammaires des incises en français : description linguistique et outils informatiques*, thèse de doctorat en informatique, Université de Marne-la-Vallée, 2000

[FAIRON 01] FAIRON C., INTEX dans un système de génération automatique de tests de raisonnement analytique, http://www.nyu.edu/pages/linguistics/intex/, 2001

[FRIBURGER 01] FRIBURGER N., MAUREL D., Elaboration d'une cascade de transducteurs pour l'extraction de motifs : l'exemple des noms de personnes, *Actes de la 8ème conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Tours, pp.183 – 192, 2001

[GROSS 67] GROSS M., LENTIN A., *Introduction to formal grammars*, Springer-Verlag, Berlin-Heidelberg-New York, 1967

[GROSS 75] GROSS M. , *Méthodes en syntaxe*, Paris: Hermann, 1975

[GROSS 84]GROSS M., Une classification des phrases figées du français, In: ATTAL P. et MULLER C. (Eds), *De la syntaxe à la pragmatique*, Lingvisticae Investigationes Supplementa, John Benjamins publishing company, pp. 141-180, 1984

[GROSS 92] GROSS M., Quelques réflexions sur le domaine de la traduction automatique, TAL, Paris, pp., 1992

[GROSS 97] GROSS M., The Construction of Local Grammars, In: E. ROCHE and Y. SCHABES (Eds.), *Finite State Language Processing*, The MIT Press, Cambridge, Mass. pp.329–352, 1997

[GROSS 99] GROSS M., Lemmatization of compound tenses in English, In : FAIRON C. (Ed), *Analyse lexicale et syntaxique: le système INTEX*, Lingvisticae Investigationes, John Benjamins publishing company, Amsterdam-Philadelphia, pp. 71-122, 1999

[GROSS 02] GROSS M., Les déterminants numéraux, un exemple : les dates horaires, *Langages* n°145, Paris : Larousse, pp. 21-37, 2002

[HARRIS 68] HARRIS Z.S., *Mathematical Structures of Language*. New York: John Wiley and sons, 1968

[LAPORTE 88] LAPORTE E., *Méthodes algorithmiques et lexicales de phonétisation de textes: applications au français*, thèse de doctorat en informatique, Université Paris 7, 1988

[LAPORTE 99] LAPORTE E. & MONCEAUX A., Elimination of lexical ambiguities by grammars: the ELAG system, In: FAIRON C. (Ed), *Analyse lexicale et syntaxique: le système INTEX*, Lingvisticae Investigationes, John Benjamins publishing company, Amsterdam-Philadelphia, pp. 341-368, 1999

[LECLÈRE 91] LECLÈRE C., SUBIRATS-RÜGGEBERG, C., A bibliography of studies on lexicon-grammar. Lingvisticæ Investigationes, XV:2, 347-409, 1991

[MAUREL 90] MAUREL D., Adverbes de date: étude préliminaire à leur traitement automatique, *Lingvisticae Investigationes*, Vol. XIV:1, John Benjamins, Amsterdam Philadelphia, pp. 31–63, 1990

[MOHRI 97] MOHRI M., Finite-State Transducers in Language and Speech Processing, *Computational Linguistics* 23:2, pp. 269-312, 1997

[NAKAMURA 01] NAKAMURA T., Analyse du discours économique, *Communication au colloque international des lexiques et grammaires comparés*, Londres, 2001

[PAUMIER 00] PAUMIER S. (2000), Nouvelles méthodes pour la recherche d'expressions dans de grands corpus, In A. Dister (ed.). *Actes des Troisièmes Journées Intex*, Revue Informatique et Statistique dans les Sciences Humaines. Liège, Université de Liège, pp. 289—296

[PAUMIER 02] PAUMIER S., manuel d'utilisation d'Unitex, http://www-igm.univ-mlv.fr/~unitex/, 2002

[PITON 97] PITON O., MAUREL D., le traitement informatique de la géographie politique internationale, *Cahiers Eco & Maths* 97.68, Université Paris 1, Paris, 1997

[POIBEAU 01] POIBEAU T., Extraction d'information dans les bases de données textuelles en génomique au moyen de transducteurs à états finis, *Actes de la 8ème conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Tours, pp. 295-304, 2001

[REVUZ 91] REVUZ, D., *Dictionnaires et lexiques : méthode et algorithmes*, Thèse de doctorat en informatique, Paris, Université Paris 7, 1991

[ROCHE 93] ROCHE E., *Analyse syntaxique transformationnelle du français par transducteurs et lexique-grammaire*, Thèse de doctorat en informatique, Paris, Université Paris 7, 1993

[ROCHE 97] ROCHE E., SCHABES Y., *Finite State Language Processing*, The MIT Press, Cambridge, Mass, 1997

[ROCHE 99] ROCHE E., Finite state transducers: parsing free and frozen sentences, In A. Kornai (ed.), *Extended finite state models of language* (pp. 108--121). Cambridge Press, 1999

[SALKOFF 73] SALKOFF M., *Une grammaire en chaîne du français : Analyse distributionnelle*, Paris: Dunod, 1973

[SENELLART 98] SENELLART J., Locating noun phrases with finite state transducers. In: *Proceedings of the 17th International Conference on Computational Linguistics (COLING98)*. Montréal, pp. 1212–1219, 1998

[SENELLART 99] SENELLART J., Reconnaissance automatique des entrées du lexique-grammaire des expressions figées, In : Lamiroy B. (Ed.), *Le lexique-grammaire*, Travaux de linguistique, Bruxelles, pp. 109-121, 1999

[SILBERZTEIN 93] SILBERZTEIN M., *Dictionnaires électroniques et analyse automatique de textes : Le système INTEX*, Masson, Paris, 1993

[SILBERZTEIN 94] SILBERZTEIN M., INTEX: a corpus processing system. In: *Proceedings of the 15th International Conference on Computational Linguistics (COLING94)*, Kyoto, Japan, pp. 579–582, 1994

[SILBERZTEIN 97] SILBERZTEIN M., The Lexical Analysis of Natural Languages, In: E. Roche and Y. Schabes (Eds.), *Finite State Language Processing*, The MIT Press, Cambridge, Mass., pp. 329–352, 1997

[SILBERZTEIN 99] SILBERZTEIN M., Transducteurs pour le traitement automatique des textes, In : Lamiroy B. (Ed.), *Le lexique-grammaire*, Travaux de linguistique, Bruxelles, pp. 127-142, 1999

[WOODS 70] WOODS W.A., Transition Network Grammars for Natural Language Analysis, *Communications of the ACM*, Vol. 13:10, 1970