Vers une ressource prédicative pour l'extraction d'information

Aurélien BOSSARD
LIPN – Université Paris 13, 93200 Villetaneuse
aurelien.bossard@lipn.univ-paris13.fr

Résumé. Cet article présente une méthode pour construire, à partir d'une ressource lexicale prédicative existante, une ressource enrichie pouvant servir à une tâche d'extraction. Nous montrons les points forts et les lacunes de deux ressources existantes pour le Français : les Tables du LADL et Volem. Après avoir montré pourquoi nous avons sélectionné Volem, nous listons les données nécessaires à la tâche d'extraction d'information. Nous présentons le processus d'enrichissement de la ressource initiale et une évaluation, à travers une tâche d'extraction d'information concernant des textes de rachats d'entreprise.

Abstract. In this article, we present a method aiming at building a resource for an information extraction task, from an already existing French predicative lexical resource. We point out the weaknesses and strengthnesses of two predicative resources we worked with : Les tables du LADL and Volem. We present why we select Volem as the most interesting resource for the task. Thereafter, we make a list of the needs an information extraction task implies, and how we include missing information in the resource we selected. We evaluate the resource completed by those missing informations, using it in an information extraction task.

Mots-clés : ressource prédicative, extraction d'information, patrons lexico-syntaxiques.

Keywords: predicative resource, information extraction, lexico-syntactic patterns.

1 Introduction

Cet article vise à montrer la façon dont nous avons procédé pour contribuer à une ressource lexicale pour le français, utile à des fins d'extraction d'information, en nous servant de ressources déjà existantes. La création d'une ressource pour le français assez complète pour couvrir de larges champs d'application en TAL apparaît aujourd'hui comme un enjeu majeur. En effet, en raison du manque de ressources, seules deux approches sont aujourd'hui possibles afin de réaliser des applications du type extraction d'information : l'apprentissage automatique et/ou la création de règles à la main.

L'approche que nous défendons ici est fondée sur la notion de schéma prédicatif. Les éléments pertinents pour l'extraction sont ceux qui se situent autour d'une relation sémantique, portée par un nom prédicatif ou, plus souvent, par un verbe. L'étude des schémas prédicatifs permet d'attribuer un rôle à chacun des arguments du prédicat ; l'étude de la syntaxe de la phrase permet en outre une mise en relation des arguments d'un prédicat avec leur rôle thématique. Nous testons notre approche dans le cadre d'une application d'extraction d'information portant sur des

rachats d'entreprise. La tâche consiste par exemple à donner une représentation (sémantique) identique pour les trois phrases suivantes :

– CPI rachète Fulmar.
– CPI a racheté Fulmar à son PDG pour 50 millions d'euros.
– CPI a indiqué avoir racheté Fulmar.

Dans chacune de ces trois phrases, nous pouvons identifier un acheteur : *CPI*, une entreprise acheté : *Fulmar*, auxquelles peuvent s'ajouter des données sur le montant de la transaction, le vendeur, la date de la transaction... Ce type d'applications a déjà été développé, y compris pour le français (Poibeau, 2003). Notre but est ici de l'envisager avec un nouveau regard, en nous focalisant sur des ressources sémantiquement riches. Plutôt que de développer des ressources de manière *ad hoc*, nous cherchons à caractériser l'intérêt des données déjà existantes pour le français.

Les questions que nous nous sommes posé sont les suivantes :
– Quelles sont les informations qu'une ressource lexicale syntaxique doit encoder pour que l'on puisse arriver à un tel résultat ?
– Est-il possible, avec les ressources existantes, de créer une telle ressource ?
– Quel réel intérêt aurait une telle ressource (précision de l'extraction, rappel, automatisation...) ?

Dans un premier temps, nous présentons un état de l'art traitant des ressources lexicales syntaxique existantes. Dans une seconde partie, nous caractérisons plus en détail notre sujet d'étude avant d'aborder, dans une troisième partie, les expériences réalisées. Nous présentons ensuite les résultats et leur analyse avant de poser, dans une dernière partie, les conclusions de notre recherche.

2 Choix d'une ressource

L'anglais dispose aujourd'hui de trois ressources à large couverture encodant d'une manière ou d'une autre la notion de schéma prédicatif : VerbNet, PropBank et FrameNet. Ces trois ressources sont fondées sur des approches différentes : approche syntaxique pour VerbNet et PropBank, et approche sémantique pour FrameNet (Pitel, 2006). De nombreuses applications ont été développées autour de FrameNet, comme la désambiguïsation sémantique (Fillmore & Baker, 2001), (Lowe *et al.*, 1997), mais aussi l'extraction d'information. Des recherches ont été menées pour utiliser conjointement ces trois ressources afin d'améliorer l'étiquetage sémantique (Giuglea & Moschitti, 2004).

Il existe beaucoup moins de richesse pour le français. Le Dictionnaire Explicatif et Combinatoire (DEC, `http://www.olst.umontreal.ca/decfr.html`) d'I. Melc'ŭk a été exclu car il n'offrait pas une couverture suffisante pour la tâche. DicoValence (`http://bach.arts.kuleuven.be/dicovalence/`) n'était quant à lui pas disponible au moment de l'étude mais mériterait sinon d'être pris en considération. Nous nous sommes alors focalisés sur deux ressources pour le français : Volem et les tables du LADL. Il s'agit dans cette partie d'expliquer le choix que nous avons fait concernant la ressource que nous avons utilisée.

2.1 Les Tables du LADL

Les Tables du LADL, aussi connues sous le nom de lexique-grammaire, ont été établies sous la direction de Maurice Gross. Elles regroupent 6000 verbes répartis dans des tables construites d'après des similitudes de comportement syntaxique. Chaque table du Lexique-Grammaire contient un certain nombre de propriétés, qui sont validées ou invalidées pour chacun des verbes qui y figure (matrice de + et de −). Les propriétés encodent des informations sur (Gross, 1975) :
– Les réalisations possibles des arguments (restrictions de sélection : arguments à trait « humain » ou « non-humain », argument de type abstrait, objet...) ;
– Les propriétés syntaxiques du verbe ou de ses arguments (pronominalisation des verbes, introduction d'un argument par une préposition...)
– Les sous-catégorisations alternatives ;
– Les possibilités de redistributions (passif long, passif court...).
Les informations contenues dans les tables du LADL sont riches sur le plan syntaxique mais relativement pauvres sur le plan sémantique. Les arguments ne sont typés que par des restrictions de sélection (`humain`, `non-humain`, `objet concret`, `abstrait`...). Le format en colonnes des tables et le fait que l'information soit répartie sur plusieurs colonnes rend les traitements difficiles. Le fait que, selon les tables, les propriétés codées dans les colonnes ne sont pas toujours les mêmes complique encore le traitement. Il est nécessaire d'effectuer un travail important de transformation pour rendre ces tables exploitables directement par des applications de TAL (Gardent *et al.*, 2005).

2.2 Volem

Volem (Saint-Dizier *et al.*, 2002) est une ressource multilingue (français-espagnol-catalan). Les entrées sont des verbes : la ressource décrit leur comportement syntaxique et sémantique à travers la description des arguments et des schémas de sous-catégorisation. Cette ressource décrit à l'heure actuelle 1700 verbes.

Description du verbe : acheter

GRILLE THEMATIQUE :
[[inic(agent),dest],[th],[src]]
LCS :
ALTERNANCES :
caus_2np_pp , anti_pr_np , anti_pr_np_pp , pas_etre_part_np_2pp , pas_etre_part_np_pp , caus_2np , caus_refl_pr_2np , caus_np_pp , caus_support_np
WN :
[13,2,3], [13,3,1] , [13,3,8]
EXEMPLE :
Il a acheté ce livre à un brocanteur

FIG. 1 – L'entrée lexicale du verbe « acheter » dans Volem

Cette ressource est fondée sur une liste de rôles thématiques génériques, mais assez précis. Les différents rôles thématiques peuvent être combinés afin de décrire aux mieux les arguments d'un verbe (*cf.* Figure 1).

Les principaux inconvénients de cette ressource sont :

– L'absence de gestion de la polysémie (les concepteurs de la ressource ont fait le choix de ne coder qu'un sens par verbe, correspondant à l'emploi le plus fréquent) ;
– La faible couverture de la ressource (1700 verbes) ;
– L'absence de description précise des schémas syntaxiques que représentent les différentes alternances utilisées dans Volem.

Volem a une couverture moindre que celle des tables du LADL. L'absence de gestion des rôles thématiques dans les tables du LADL constitue cependant un inconvénient de taille pour une tâche d'extraction d'information, qui demande plus poussée sur la nature des arguments. Nous avons donc choisi de concentrer notre étude sur la ressource lexicale Volem, qui paraît avoir un potentiel de description plus fort que les Tables du LADL pour notre application d'extraction.

3 Méthode d'enrichissement de la ressource

Les informations contenues dans Volem n'étaient pas suffisantes pour réaliser une tâche d'extraction d'information. Le format de Volem au format XML n'était pas non plus directement exploitable. Nous avons donc retravaillé la ressource sur les points suivants afin de pouvoir l'utiliser dans le cadre d'une tâche d'extraction d'information :

1. Codage de la ressource sous la forme d'une table de contraintes
2. Ajout des informations manquantes
3. Codage d'automates patrons

3.1 Codage de la ressource sous la forme d'une table de contraintes

Nous voulons, à partir de Volem, créer des automates d'extraction (format unitex : `http://www-igm.univ-mlv.fr/~unitex/`). Pour cela, nous avons besoin d'une ressource codée à l'aide d'un tableau. Nous avons donc créé un convertisseur pour passer du format de Volem à notre format. Nous utilisons une colonne par alternance de Volem, et validons l'alternance pour un verbe en mettant un « + » dans la case correspondante, et un « - » sinon. Pour encoder les informations sur les rôles thématiques, nous écrivons une colonne par argument du verbe (3 colonnes) et nous remplissons chacune d'elle avec la description enregistrée au sein de Volem du rôle thématique de l'argument. Ainsi, la ressource est directement exploitable par des graphes unitex, qui exploitent les tables de contraintes.

3.2 Ajout des informations manquantes

Cependant, la ressource en l'état n'encode toujours pas assez d'informations pour réaliser une extraction d'information précise. En effet, il lui manque encore :

1. les auxiliaires des verbes
2. les différentes prépositions introduisant éventuellement un argument
3. les noms associés aux verbes (*e.g.* *rachat* pour *racheter*)
4. les adjonctions essentielles à une relation à extraire.

3.2.1 Les auxiliaires

Deux possibilités se sont offertes à nous pour ajouter les auxiliaires d'un verbe à la table de données que nous avions construite. Soit récupérer les auxiliaires depuis un dictionnaire, soit identifier l'auxiliaire d'un verbe grâce aux occurrences de celui-ci en corpus. Nous avons opté pour la deuxième solution. A partir de la table de données que nous avons créée, un automate est créé qui reconnaît les différents verbes ainsi que les auxiliaires qui les accompagnent. Si l'auxiliaire « avoir » apparaît au moins une fois dans le texte pour un verbe donné, un « + » est ajouté à l'intersection de la ligne correspondant à ce verbe et de la colonne correspondant à l'auxiliaire « avoir ». L'inconvénient de cette méthode est qu'elle demande des textes correctement écrits. Mais elle n'est pas dépendante d'une ressource comme l'aurait été la première : en effet, en utilisant les tables du LADL comme référence, nous n'aurions pas pu ajouter automatiquement certains verbes qui ne sont pas encodés dans la ressource. Nous avons procédé au repérage des auxiliaires avec la version « brute » des corpus que nous avons utilisés pour extraire les relations de rachat d'entreprises (*cf.* §4.1).

3.2.2 L'ajout des prépositions

Nous avons déjà mentionné qu'une seule préposition est codée par argument au sein de Volem, alors que plusieurs prépositions peuvent apparaître pour certains verbes (*acheter à* ou *auprès de*). Nous avons donc réalisé un outil permettant d'ajouter à la table de données les prépositions qui introduisent les arguments d'un verbe. Cet outil nécessite une validation des résultats par l'utilisateur.

Le système est fondé sur une série d'automates « à trou » : chaque « trou » correspond à une préposition possible introduisant un argument. Le système renvoie quelques erreurs (soit des groupes de mots qui ne sont pas des prépositions, soit des prépositions rallongées de certains mots qui les suivaient dans le texte). Après validation des résultats par l'utilisateur, les prépositions validées sont ajoutées à la table de données. Cet outil nécessite un corpus annoté par entités nommées. Nous avons travaillé sur les corpus utilisés pour l'extraction de relations de rachats d'entreprise (*cf.* §4.1).

3.2.3 Les adjonctions

Volem ne gère que les arguments clé d'un verbe. Cependant, certains arguments qui ne sont pas nécessaires d'un point de vue syntaxique jouent un rôle extrêmement important dans des relations à extraire. Par exemple, un achat selon Volem ne fait pas intervenir de montant : le montant est quasiment toujours présent quand on se base sur l'analyse en corpus. C'est donc un argument clé, sémantiquement parlant.

L'approche développée par les auteurs de FrameNet est du même type (Fillmore & Baker, 2001). La description des schémas prédicatifs au sein de cette ressource se fonde sur l'étude en corpus des réalisation du verbe. Si un complément intervient fréquemment pour un verbe donné, alors celui-ci sera assimilé à un argument, même s'il est considéré comme un ajout dans la grammaire traditionnelle.

Nous avons alors tenté, en dénombrant ces adjonctions au sein des corpus étudiés, de déterminer quelles adjonctions essentielles pouvaient tenir lieu d'argument et dans quelle mesure celles-ci

pouvaient être repérées par une analyse statistique. La méthode se fonde sur le repérage et le regroupement des compléments circonstanciels (temps, lieu, montant...) pour chaque verbe au sein du corpus.

Après dénombrement des adjonctions, nous ne retenons que celles au-dessus d'un seuil de 10 % (défini manuellement). Cela signifie que nous ne sélectionnons que celles qui sont apparues dans au moins 10 % des phrases contenant un verbe donné. Cette méthode nous a permis de compléter nos informations concernant les données à extraire avec les adjonctions adéquates pour 80 % des verbes sélectionnés (pour les 20 % restant, l'adjonction « montant » était apparue dans moins de 10 % des phrases à extraire, contre 15 % pour l'adjonction « date »). Les résultats doivent toutefois être validés par un expert avant d'ajouter les adjonctions au schéma argumental des verbes concernés.

3.3 Les automates patrons

La dernière étape de l'enrichissement de la ressource a consisté en la création d'automates patrons pour chacune des alternances listée dans Volem. Un automate patron est un automate lexicalement vide, encodant une famille d'alternances ; il est instancié par l'ensemble des verbes correspondant à la famille d'alternances visée. Pour cela, il a fallu dans un premier temps identifier les différentes formes de surface que présentent chacune des alternances de Volem, puis réaliser pour chacune d'entre elles des graphes qui permettent de les reconnaître.

La figure 2 (*cf.* dernière page de l'article) présente un extrait d'un graphe qui permet de reconnaître l'alternance *caus_2np* de Volem.

Ces automates patrons prennent en entrée la ressource que nous avons créée, et produisent en sortie autant d'automates que d'alternances à reconnaître pour chaque verbe. Les automates ainsi créés annotent un texte avec les informations que l'on veut extraire. En l'occurrence, pour notre extraction sur les rachats d'entreprise, ils reconnaissent les fragments de textes correspondant à l'acheteur, au vendeur, à l'élément vendu, au montant et à la date.

4 Expériences

4.1 Données d'évaluation

Nous avons choisi de travailler sur une tâche d'extraction précise : le rachat d'entreprises. Les expériences ont été menées sur plusieurs corpus :
– Un corpus tiré d'un autre site spécialisé : FUSACQ (300ko, 25000 mots) (`http://www.fusacq.com`)
– Un corpus tiré de différents journaux généralistes (400ko)
– Un corpus d'entraînement pour repérer les entités nommées (200ko).

L'utilisation de plusieurs corpus permet d'évaluer les performances en tenant compte (dans la mesure du possible) du genre textuel. On ne trouve pas les mêmes constructions ni les mêmes expressions suivant que l'on a affaire à un corpus journalistique ou à un site web. Nous verrons dans la discussion que cette hypothèse se vérifie lors de l'étude des performances sur les différents corpus.

4.2 Résultats

Nous avons mis en place deux protocoles d'évaluation, afin d'isoler les éventuels problèmes ; dans l'un, nous passons les règles d'extraction sur un corpus dans lequel les entités nommées ont été annotées grâce à un outil pour annoter développé au LIPN (TagEN, `http://www-lipn.univ-paris13.fr/~poibeau/tagen.html`). Dans l'autre, nous utilisons un corpus dans lequel nous avons annoté toutes les entités nommées à la main. Nous pouvons ainsi procéder d'un côté à une évaluation « en conditions réelles », et de l'autre, de nous focaliser sur l'évaluation des schémas prédicatifs, indépendamment des erreurs dues à la mauvaise reconnaissance des entités.

Dans le tableau 1, nous entendons « relations » comme des structures grammaticales comportant un verbe et ses arguments participant à un rachat d'entreprise.

	Protocole 1	Protocole 2	Nombre total de relations
Nombre de relations repérées	101	184	285
% de relations	35	64	100

TAB. 1 – Tableau des résultats de l'extraction sur le corpus FUSACQ

Un peu plus seulement de la moitié des entités nommées correspondant à un acheteur potentiel ou à un vendeur potentiel ont été annotées. L'annotation des entités nommées n'a pas été menée plus avant, étant donnée qu'elle n'est pas au centre de notre étude. Notre outil permet de repérer dans un texte dans lequel l'annotation des entités nommées est correcte, 65 % des relations d'achat.

35 % des relations restent tout de même non repérées. Ceci provient du fait que nous n'avons pas géré certains schémas syntaxiques :
– les subordonnées relatives
– les verbes introducteurs précédés d'un verbe marquant soit le passé, soit le futur (ambiguïté sémantique possible. Ex. : « Bull vient d'annoncer le rachat de CP8 à Schlumberger »)
– les structures complexes (ex. :« COMPANY1 s'était diversifié à travers l'acquisition de COMPANY2 »
– L'alternance passive sans groupe prépositionnel (non encodé dans Volem pour les verbes qui nous intéressent. Ex. : « COMPANY a été racheté »)
– Les structures faisant intervenir un pronom (pas de résolution d'anaphores).
L'outil de repérage des prépositions renvoie des résultats bruités à hauteur de 8 %.

Les adjonctions nécessaires à une tâche d'extraction sont la date et le montant de la transaction. Les corpus sur lesquels nous avons fait nos expériences ont montré ce point, même s'ils sont de taille trop faible pour donner des chiffres significatifs statistiquement.

4.3 Discussion

Nous avons vu que Volem est incomplet du fait qu'il ne gère pas la polysémie et que toutes les alternances n'y sont pas codées.

Est-il possible de rajouter aux entrées de Volem les alternances que cette ressource n'encode

Verbe	Alternances	nombre d'occurrences de l'alternance (FUSACQ)	nombre d'occurrences de l'alternance (Corpus général)
racheter	caus_2np	37	16
	caus_2np_pp	6	12
	pas_etre_part_np_pp	6	12
revendre	caus_2np	1	0
acheter	caus_refl_pr_np	2	0
	caus_2np	0	16
vendre	pas_etre_part_np_2pp	2	0
	caus_2np	2	0
	caus_2np_pp	2	0
acquérir	caus_2np	44	0
	caus_2np_pp	1	4
	pas_etre_part_np_pp	0	16
céder	caus_2np_pp	24	4
	caus_2np	9	0
	pas_etre_part_np_2pp	2	8
fusionner	aucune occurrence	0	0
détenir	caus_2np	5	0
	pas_etre_part_np_pp	1	12
offrir	caus_2np_pp	1	16
	caus_2np	1	0
reprendre	pas_etre_part_np_pp	11	16
	caus_2np	12	8

TAB. 2 – Répartition des alternances selon les verbes dans les phrases extraites des corpus (FUSACQ annoté et extrait du corpus général)

pas ? L'ajout des alternances constitue un réel problème. En effet, il est possible, par des méthodes statistiques, de sélectionner des schémas de sous-catégorisation acceptables pour un verbe (Salmon-Alt & Chesley, 2005),(Briscoe & Carroll, 1997). Mais la sélection d'alternances acceptables constitue un tout autre problème ; il faut en effet pouvoir distinguer une phrase du type : « COMPANY1 a acheté une usine à LIEU. » d'une phrase du type : « COMPANY1 a acheté des terrains à LIEU{la ville de LIEU}. Ce problème mériterait une étude approfondie.

Un point intéressant est la variation dans l'usage du verbe suivant le corpus. On s'aperçoit, même sur des corpus de taille modeste, des variations d'usage, une alternance étant plutôt employée dans un corpus, une autre dans un autre corpus (*cf.* tableau2). Nous faisons l'hypothèse qu'il s'agit de variations dans le style d'écriture propre aux différents genres textuels. Ainsi, l'alternance « caus_2np » du verbe « détenir » constitue 78 % des variations syntaxiques pour le verbe « détenir » dans le corpus FirstInvest, et 80 % dans FUSACQ, mais n'apparaît pas dans le corpus tiré de journaux non spécialisés.

Les résultats obtenus sont satisfaisants. En effet, l'ajout des structures syntaxiques manquantes (*cf.* §4.2) permettrait d'obtenir environ 65 % de rappel sur une tâche d'extraction des rachats d'entreprise, soit 15 % de moins que le rappel obtenu par Thierry Poibeau, sur la même tâche et le même type de corpus (Poibeau, 2003), mais en utilisant une méthode à base de ressources qui se distingue des autres travaux par une certaine généricité.

5 Conclusion et Perspectives

Les ressources pour le français sont beaucoup moins complètes que celles pour l'anglais. La ressource pour le français qui nous a semblé la plus adaptée à l'extraction d'information (Volem), présente des manques (non gestion de la polysémie, couverture faible, alternances non encodées...) qu'il est cependant possible de combler par des méthodes semi-automatiques.

Les résultats obtenus pour une tâche d'extraction pour l'anglais (Giuglea & Moschitti, 2004) montre que l'extraction à base de ressources à large couverture permet d'obtenir de bons résultats et évite de redévelopper de manière *ad hoc* des connaissances pour chaque nouvelle application.

Il reste certes un travail à réaliser dépendant de l'application (ajout d'adjonctions, prépositions), cependant la mise en place d'une méthode utilisant des ressources extérieures permet une réutilisabilité *a contrario* des méthodes purement *ad hoc* tout en garantissant malgré tout une couverture et une certaine généricité.

Cet article a cherché à montrer comment compléter Volem, notamment avec les prépositions et le filtrage des alternances non pertinentes pour un sens donné. Il reste à définir une méthode semi-automatique pour l'ajout des alternances non référencées par Volem.

Références

BRISCOE T. & CARROLL J. (1997). Automatic extraction of subcategorization from corpora.

FILLMORE C. & BAKER C. (2001). Frame semantics for text understanding. In *WordNet and Other Lexical Resources Workshop*, NAACL, Pittsburgh.

GARDENT C., GUILLAUME B., FALK I. & PERRIER G. (2005). Le lexique-grammaire de m. gross et le traitement automatique des langues.

GIUGLEA A.-M. & MOSCHITTI A. (2004). Knowledge discovering using framenet, verbnet and propbank. In *International Workshop on Mining for and from the Semantic Web*, Seattle, USA.

GROSS M. (1975). *Méthodes en syntaxe*. Paris : Hermann.

LOWE J., BAKER C. & FILLMORE C. (1997). A frame-semantic approach to semantic annotation.

PITEL G. (2006). Framenet, théorie, produit, processus, multilingualité et connexions. In *Autour de FrameNet et de la Sémantique Lexicale Multilingue : projets en cours et points de contacts entre les différentes approches*. date de la conférence : 28 Février 2006.

POIBEAU T. (2003). *Extraction automatique d'information, du texte brut au web sémantique*. Paris : Hermes.

SAINT-DIZIER P., FERNANDEZ A., VAZQUEZ G., KAMEL M. & BENAMARA F. (2002). The Volem Project : a Framework for the Construction of Advanced Multilingual Lexicons . In *Language Technology 2002 , Hyderabad,* , p. 123–142 : Springer Verlag, Lecture Notes. Dates de conférence : décembre 2002.

SALMON-ALT S. & CHESLEY P. (2005). Le filtrage probabiliste dans l'extraction automatique de cadres de sous-catégorisation. In *Journée ATTALA du 12/03/2005*.

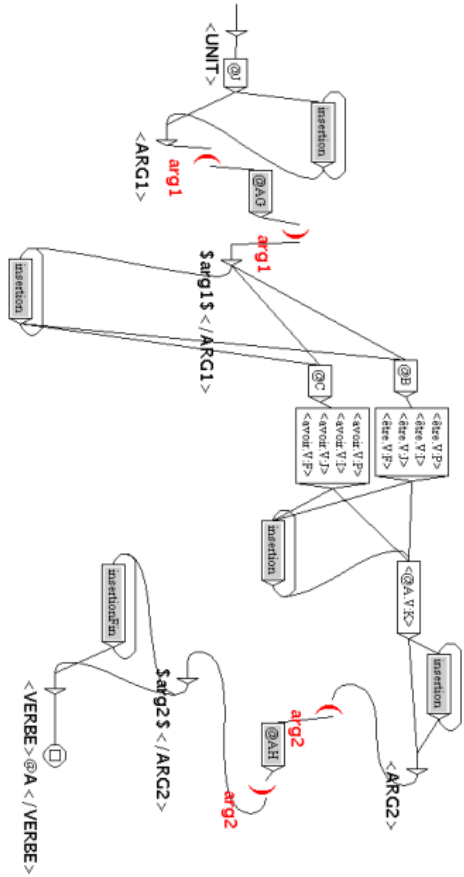

FIG. 2 – Exemple d'un automate patron