Représentation ontologique du LVF et son utilisation en traitement automatique de la langue

Radia Abdi, Guy Lapalme
RALI-DIRO, Université de Montréal
C.P. 6128, Succ Centre-Ville
Montréal, Québec, Canada, H3C 3J7
abdi.radia@gmail.com, lapalme@iro.umontreal.ca

Résumé. Nous présentons une version ontologique du dictionnaire LVF (Les Verbes Français) de J. Dubois et F. Dubois-Charlier. Elle a été obtenue par une transformation automatique de la version XML du LVF. Nous en démontrons l'utilisation dans le domaine du traitement automatique de la langue avec une application d'annotation sémantique développée dans l'environnement GATE.

Abstract. We present an ontological version of the LVF dictionary (Les Verbes Français) by J. Dubois and F. Dubois-Charlier. It was produced automatically by transforming the XML version of the LVF. We illustrate its use in the field of natural language processing with a semantic annotation application developed in the GATE environment.

Mots-clés : LVF, Les Verbes Français, peuplement d'ontologies, ressource lexicale, web sémantique, extraction d'information, OWL. .
Keywords: LVF, Les Verbes Français, ontology population, lexical resource, semantic web, information extraction, OWL.

1 Introduction

Des ressources lexicales riches et disponibles en accès libre en langue anglaise ont facilité le développement des recherches en traitement automatique de cette langue, tels que *WordNet*, *VerbNet* ou *FrameNet*. Il n'existe malheureusement que peu d'équivalents en français disponibles en accès libre, ce qui complique la recherche et les travaux dans le traitement automatique du français.

L'ouvrage « Les Verbes Français » (LVF), réalisé par Jean Dubois et Françoise Dubois-Charlier est une ressource lexicale qui fournit une description linguistique et sémantique détaillée des verbes français. À cause de problèmes de diffusion et de distribution, le LVF n'a malheureusement pas pu être exploité par les chercheurs et les linguistes qui, pour plusieurs, en ignoraient même l'existence. Certains travaux ont rendu le LVF plus accessible en termes d'encodage et de format de données : Denis Le Pesant en a créé une version sous format Excel pour faciliter sa consultation manuelle, mais ce mode d'accès ne s'est pas avéré pratique pour les applications informatiques ; Guy Lapalme en a alors proposé une version XML qui en facilite l'exploitation par les applications de traitement automatique de la langue et Hadouche et Lapalme (2010) l'ont comparé à d'autres ressources lexicales.

Ces dernières années, il y a eu un regain d'intérêt pour la notion d'ontologie, sous l'impulsion du web sémantique. La recherche sur le web, étant devenue une activité à haute valeur ajoutée, a poussé un développement rapide de modèles, de langages et d'outils permettant d'expliciter la sémantique des données issues du web et de raisonner sur ces données. La représentation du LVF en format XML, considéré comme un des standards de base du web sémantique, nous a incités à développer une représentation du LVF en un standard plus puissant, en l'occurrence OWL pour obtenir une ontologie des verbes. L'intérêt de l'application des ontologies au traitement automatique de la langue a été démontré par de nombreuses recherches dans le traitement automatique de la langue anglaise. Des versions OWL de WordNet ou FrameNet ont été développées afin de désambigüiser le sens des mots ou de les intégrer dans une autre ontologie.

Cet article présente tout d'abord la structure du LVF et sa version XML; la section 3 décrit le processus de transformation de LVF en une ontologie OWL et son application dans le cadre d'une application d'annotation sémantique.; nous présentons enfin l'application d'annotation sémantique, développée dans GATE, qui sert à annoter les verbes à partir de l'ontologie LVF. Nous concluons en évoquant quelques problèmes rencontrés ainsi que des travaux futurs. Nous montrons les apports mutuels entre le web sémantique et le TAL et jusqu'à quel point la représentation ontologique du LVF peut améliorer son exploitation et utilisation en TAL et en web sémantique.

2 Organisation du LVF

Les Verbes Français (LVF) est une base de données numérique réalisée par J. Dubois et F. Dubois-Charlier dont le but est de classifier les verbes selon leur syntaxe et leurs interprétations sémantiques. Le principe de la classification repose sur l'adéquation entre les schèmes syntaxiques et la sélection distributionnelle dans la construction, et l'interprétation sémantique (François et coll. 2007). Les schèmes sont regroupés en classes et sous-classes : 248 sous-classes syntaxiques, 54 classes sémantico-syntaxiques et 14 classes génériques.

LVF comprend plus de 25 610 entrées représentant 12 310 verbes différents avec 4 188 verbes ayant plusieurs entrées. Une entrée est représentée par 11 rubriques, par exemple, le verbe `chercher` présente 10 entrées ou emplois différents (`chercher 01, …, chercher 10`) correspondant à des schèmes syntaxiques différents. Une entrée est définie par un schème syntaxique et un opérateur codé pour faciliter le traitement automatique, mais aussi par d'autres informations linguistiques telles que le sens, des exemples de phrases, le lexique (entier entre 1 et 6 correspondant au type de lexique, du plus élémentaire au plus spécialisé, où on trouve cette entrée), la conjugaison …etc. Un schème syntaxique est une suite de caractères alphanumériques (tels que `T1318` ou `P3008`) qui indiquent la nature du verbe (*transitif direct, indirect, intransitif, pronominal*), le type du constituant sujet et objet (*humain, animal ou chose, complétive*) et aussi la nature des compléments (*locatif, prépositionnel, instrumental, à modalité,* etc.). Un opérateur est une étiquette interprétative du sens et de l'emploi du verbe ; par exemple `loq AV` veut dire *parler avec*. Pour notre travail, nous avons utilisé la version XML du LVF qui est plus structurée et qui offre une facilité de manipulation et d'exploitation automatique des données avec des feuilles de transformation XSLT.

3 Ontologies et standard OWL

Le web sémantique répond à certains problèmes et limitations du web actuel : (i) aucune sémantique n'est attribuée au contenu web, (ii) les métadonnées utilisées sont non structurées et limitées dans leur usage, (iii) l'absence de modèle de représentation de connaissances et de données publiées sur le web rend le processus de raisonnement et d'inférence pratiquement impossible.

Le web sémantique propose des solutions à ces problèmes. Une d'entre elles est de mettre en œuvre des formalismes et des langages standardisés de représentation de données et de connaissances pour représenter et modéliser la sémantique des ressources web. On fournit ainsi des ontologies qui sont des ressources conceptuelles représentées par ces langages modélisant les domaines des connaissances et on facilite leur accès et leur partage. Les ontologies représentent des ressources de modélisation et de conceptualisation très importantes (Noy et McGuinness, 2000). Elles constituent en soi un modèle de données représentatif d'un ensemble de concepts dans un domaine, ainsi que des relations entre ces concepts. Les ontologies sont employées pour raisonner à propos des objets du domaine concerné. OWL (Motik et coll. 2012) est un langage de représentation des connaissances, développé par le W3C. Il fournit les moyens pour définir des ontologies web structurées et riches. Il permet de décrire des ontologies complexes de domaines concrets. Le vocabulaire OWL est constitué d'un ensemble de notions qui spécifient des concepts (classes) et des propriétés telles que : la hiérarchie des classes et des relations (propriétés), l'équivalence des classes, la symétrie et la transitivité des relations, la notion de cardinalité de classes…etc.

OWL est basé essentiellement sur le formalisme des logiques des descriptions. Avec l'aide de *reasoners* qui traitent la logique de description, il devient possible de se doter d'une capacité d'inférence et de raisonnement déductif sur les concepts de l'ontologie. OWL est le langage le plus utilisé pour la description d'ontologies.

4 Représentation OWL du LVF

L'ontologisation de la ressource lexicale LVF fournit un format relationnel aux données du LVF comme certains chercheurs l'ont déjà fait pour WordNet (Niles et Pease 2003) (Van Assem et coll. 2006). Deux principales raisons nous ont motivés pour le développement d'une ontologie OWL du LVF :

- la représentation du LVF en standard formel du W3C nous permet de répondre à des besoins des applications du web sémantique telles que l'annotation sémantique, l'extraction d'informations…etc. Cette ontologie permettra d'ouvrir de nouveaux horizons pour différents champs d'applications sémantiques et de traitement automatique de la langue française qui utilisent le LVF.
- les standards W3C du web sémantique représentent des langages sophistiqués qui offrent un niveau de qualité supérieur d'application et d'interopérabilité entre les applications.

Dans cette section, on présente le processus de transformation du LVF à partir du format XML vers OWL.

4.1 Conception et définition du schéma général

Dans la représentation OWL du LVF, nous nous intéressons à traduire les fichiers XML de ce dictionnaire en une ontologie OWL donc en un modèle de données ou en graphe de concepts reliés par des relations sémantiques. Nous avons défini uns structure générale de ce modèle à partir de la structure hiérarchique des fichiers XML du LVF pour en extraire automatiquement les concepts de base ainsi que leurs relations. Dans un premier temps, nous avons énuméré tous les termes de ce dictionnaire pour définir par la suite les classes (concepts) et leur hiérarchie taxinomique ; par la suite, nous avons déterminé les attributs de chaque classe ainsi que les relations sémantiques possibles entre les différentes classes. Ce processus nous a permis d'obtenir une idée générale du schéma et de la nature des constituants de l'ontologie LVF générée automatiquement à partir des fichiers XML à l'étape suivante.

4.2 Transformation automatique du XML à OWL

Plusieurs stratégies pour la transformation de XML en OWL ont été proposées (Bohring et Auer 2007, Ferdinand et coll. 2004). Certaines approches proposent une méthode générique de transformation XML en modèle OWL à partir d'un schéma XML et des données du fichier XML, d'autres pensent qu'il est impossible de proposer une approche automatique convenable pour une transformation automatique complète de XML vers OWL, car XML ne définit aucune contrainte sémantique. Contrairement à cela, d'autres approches considèrent qu'il y a une sémantique dans les documents XML qui peut être découverte à partir de la structure des documents, en l'occurrence l'approche de Melnik (1999).

Même si XML n'est pas censé représenter d'informations sémantiques ou de sémantique entre les données, les balises imbriquées peuvent représenter une relation `is-a` ou `part-of` ou `subType-of`. On peut considérer la structure XML comme relationnelle et se baser sur celle-ci pour obtenir le modèle OWL. Le processus de transformation est divisé en deux étapes : la génération du modèle de l'ontologie et la génération des instances (individus) de l'ontologie.

4.2.1 Génération du modèle de l'ontologie

Le modèle de données XML décrit un arbre de nœuds, par contre le modèle OWL est représenté à base de triplets RDF sujet-prédicat-objet. Nous exploiterons donc la structure d'arbre XML pour générer la hiérarchie de classes correspondante. Le schéma XML est un fichier qui permet de décrire la structure d'un document XML, plus précisément, il définit les éléments/nœuds et les attributs XML ainsi que leurs types de données, il permet aussi de définir l'ordre d'imbrication des nœuds XML c'est-à-dire quel élément est l'élément parent ou l'élément fils. Un document XML est validé par son schéma XML dans le but de vérifier la consistance des données dans le document. Comme le schéma XML définit la structure et les facettes des données d'un fichier XML, on va l'utiliser pour générer automatiquement la structure de notre ontologie. On suppose que le document XML contient une structure relationnelle entre les données et on déterminera la signification et les relations possibles entre les éléments du document XML. Les nœuds du document XML peuvent représenter des classes car ils représentent des concepts dans la ressource LVF tels que `verbe`, `entree`, `domaine`, `sens`, `opérateur` …etc. L'imbrication des nœuds peut dans certains cas indiquer la présence d'une relation de type `is-a` ou `part-of` mais dans notre cas, on considère la relation de type `has-` dans les cas suivants :

- Un verbe a des entrées
- Une entrée a un domaine, une classe, un opérateur, des phrases, un sens, une construction, un lexique, un nom ...etc.

Le document XML du LVF définit les données sur les verbes et les entrées. Cependant, Il existe d'autres fichiers XML qui apportent des informations supplémentaires sur les classes, les schèmes, les codes de conjugaison, les codes des opérateurs et de dérivation. Ces fichiers XML décrivent certains détails importants pour la compréhension des codes utilisés dans le LVF tels que les codes des opérateurs, les schèmes syntaxiques, les codes de conjugaisons ainsi que les codes des différentes classes. Les schémas XML de ces fichiers ont été aussi exploités dans le processus de transformation afin de compléter le modèle de données de l'ontologie LVF pour une représentation plus complète du LVF.

Nous avons utilisé une feuille de transformation XSLT pour définir des règles d'extraction des classes, de leur hiérarchie et de leurs propriétés. Cette feuille de style prend en entrée le schéma XML du document LVF pour produire un modèle de l'ontologie LVF écrit en OWL. Le fichier résultant va contenir la définition des classes et de la hiérarchie des classes, la définition des *Object Properties* qui relient deux classes et des *Data Properties* qui relient une classe et une constante (chaîne de caractère, nombre, valeur booléenne, etc.). Le résultat de cette transformation à partir des fichiers de schéma et des fichiers XML décrivant les codes des opérateurs est appelé le **modèle** de l'ontologie.

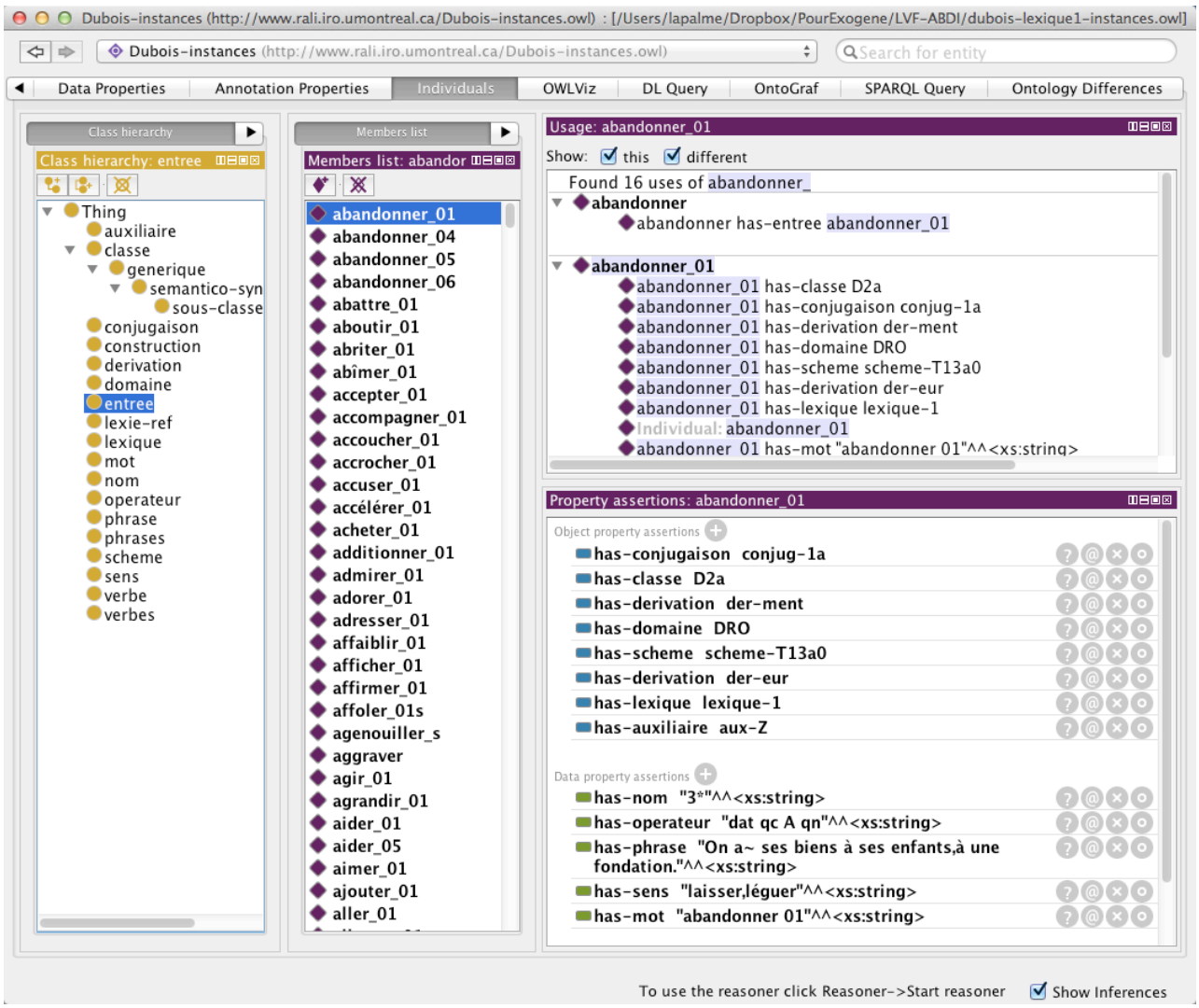

Figure 1 : Ontologie LVF dans l'éditeur Protégé. Le panneau de gauche montre la hiérarchie des classes générées à partir du schéma XML. Le panneau du milieu montre quelques verbes tirés du lexique de niveau 1 du LVF. Les panneaux de droite donnent de l'information à propos du verbe «abandonner 01».

Il correspond aux noms et à la structure des classes de l'ontologie illustrée à la figure 1. Il comprend aussi la définition des noms de domaine et de portée des propriétés dont on retrouve quelques exemples dans la partie en bas à droite de la figure.

4.2.2 Génération automatique des instances de l'ontologie

Le peuplement de l'ontologie LVF a été effectué à l'aide d'une deuxième feuille de style XSLT qui sert à transformer le document XML du LVF en un document OWL en peuplant l'ontologie avec des instances de classes et à les relier par leurs propriétés et à affecter des valeurs aux attributs. L'ontologie résultante importe l'ontologie du modèle décrite à la section précédente. Cette ontologie peut ensuite être chargée dans un éditeur d'ontologie comme Protégé ou dans l'environnement GATE comme nous le verrons plus loin. Il n'y a aucun problème à traiter le 25 000 entrées du LVF avec ce processus, mais afin de limiter l'espace mémoire nécessaire pour les traitements subséquents, nous nous sommes limités aux 867 entrées marquées comme étant du lexique de niveau 1 (dictionnaire fondamental). Cette expérimentation montre donc la faisabilité du principe de l'approche générale sur les verbes considérés les plus fréquents.

Dans un fichier XML, les nœuds XML représentent les classes OWL et leur hiérarchie représente les relations entre les classes que nous avons déjà définies à l'aide de la première feuille de style. De ce fait, nous avons parcouru les fichiers XML qui contiennent les instances, en respectant le modèle de classe qui a été généré précédemment pour générer les instances de chaque classe et de chaque propriété.

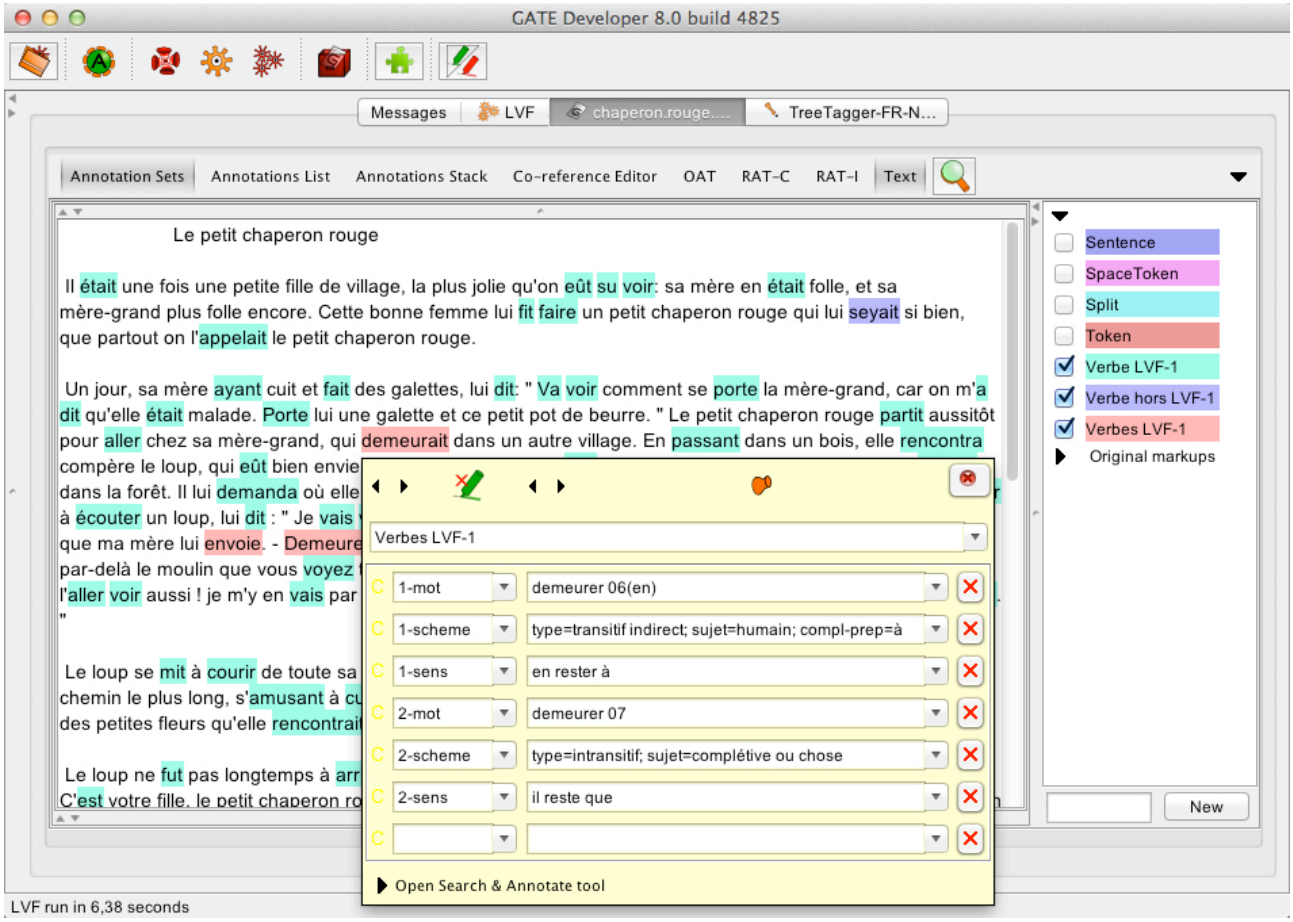

Figure 2 : Annotation semi-automatique de verbes du LVF (lexique de niveau 1) dans GATE. Trois types d'annotation sont mise en évidence: les verbes reconnus, ceux qui ne sont pas dans le lexique, les verbes avec une ambiguïté sur l'acception. En cliquant sur un verbe on obtient le mot, l'information sur le schème et les sens tirés de l'ontologie.

5 Annotation sémantique à partir du LVF

Nous avons développé, dans la plateforme GATE (Cunningham et coll. 2011), une application d'annotation sémantique des verbes à partir de l'ontologie LVF ou les verbes sont annotés avec leurs entrées tout en sachant que chaque entrée est en relation avec d'autres concepts de l'ontologie tels que le sens, le schème syntaxique, l'opérateur sémantique, le domaine, la conjugaison …etc. Dans GATE, le pipeline (une suite de *processing resources* où les sorties de l'un servent d'entrée au suivant) comprend les ressources suivantes : *French Tokenizer*, *RegEx Sentence Splitter*, *Adapt Tokenizer to Tagger* et le *TreeTagger,* un étiqueteur grammatical testé avec succès sur plusieurs langues dont le français.

Idéalement, chaque verbe devrait être annoté avec son entrée correspondante dans le LVF. La détermination automatique de l'entrée pourrait se faire grâce à l'analyse de son schème syntaxique ou sémantique. Le schème syntaxique d'une entrée est codé selon un modèle bien défini, tel que `T1318`, `P3008` ou `A16`, qui indique la nature du verbe (transitif direct ou indirect, intransitif ou pronominal), le type du sujet et de l'objet ainsi que la nature des compléments.

Dans ce travail, nous avons opté pour une approche semi-automatique pour la réalisation de cet annotateur. Elle consiste à annoter les verbes avec les différentes entrées possibles accompagnées par leurs sens et leurs schèmes syntaxiques, présentées sous forme d'une liste dans laquelle l'utilisateur peut supprimer les entrées non pertinentes. Pour y arriver, nous avons développé une nouvelle ressource GATE de type *JAPE Transducer* qui s'ajoute au pipeline décrit plus haut. Ce module est chargé de l'annotation sémantique des verbes à partir de l'ontologie LVF en suivant les étapes suivantes :

– **Extraction automatique des verbes et lemmatisation** : elle est effectuée grâce aux étiquettes grammaticales créées par le *TreeTagger* dans la phase de prétraitement, et qui sont sauvegardées dans la structure d'annotation de GATE. Nous avons utilisé le langage JAPE pour accéder à ces annotations en définissant les règles qui permettent de récupérer toutes les annotations/étiquettes qui commencent par `VER:`. Les lemmes des verbes ont été récupérés à l'aide du *TreeTagger*, qui fournit les lemmes des mots traités en plus de leurs étiquettes grammaticales.

– **Recherches des entrées dans LVF** : les entrées de chaque verbe à partir de l'ontologie LVF sont présentées sous forme d'une liste à l'utilisateur. Comme une liste d'entrées n'est peut-être pas toujours significative pour un utilisateur, on a ajouté à chaque entrée d'autres informations à partir de l'ontologie LVF: le sens et le/les schèmes syntaxiques de chaque entrée. Pour pouvoir déterminer l'entrée correspondante, l'utilisateur pourra soit, apparier le schème syntaxique avec le verbe, soit, procéder à l'élimination des entrées selon leur sens.
– **Création des annotations GATE** : la liste des entrées est affichée lorsque l'usager clique sur un verbe reconnu aux étapes précédentes, comme on peut le voir à la figure 2. Les entrées sont numérotées pour pouvoir lier leur sens et schème. Il est possible de supprimer les entrées qui ne correspondent pas au contexte du verbe à partir de la liste et de sauvegarder par la suite le document avec les annotations pertinentes.

Ce travail illustre la possibilité d'utiliser le LVF dans le contexte d'une application de TAL. Même si elle reste relativement simpliste, requérant une grande implication de la part de l'usager, cette expérience est prometteuse. L'automatisation de la prise en compte des informations des schémas aurait été intéressante à explorer, mais elle aurait impliqué l'utilisation d'un parseur ce qui dépassait l'ampleur de ce travail exploratoire que nous comptons poursuivre. On pourrait aussi imaginer l'utilisation d'heuristiques simples basées sur l'étiquetage du TreeTagger combinées avec la présence de pronoms personnels devant le verbe (sujet humain), ou la présence de certaines prépositions.

6 Conclusion

On a présenté dans ce travail une version OWL du dictionnaire LVF qui a été le résultat d'une transformation automatique à partir de ses fichiers XML. Par la suite, on a démontré l'intérêt et l'utilisation de cette version dans une application d'annotation sémantique qui sert à annoter les verbes français à partir des concepts et instances de l'ontologie LVF, plus précisément à partir des instances de la classe « Entree », « Sens » et « Schème » tout en sachant que l'entrée d'un verbe définit son schème syntaxique et sémantique et donc l'emploi du verbe. Le processus d'annotation des verbes est basé sur une approche semi-automatique qui propose une liste d'entrées possibles pour chaque verbe à l'aide de leurs sens et schème syntaxique correspondants.

Dans le futur, nous envisageons d'intégrer un module qui automatiserait la sélection de l'entrée correspondante à l'emploi du verbe parmi l'ensemble des entrées possibles. Pour y arriver il faudrait implanter un processus d'analyse automatique du schème syntaxique de chaque verbe. En effet si on arrivait à déterminer le schème syntaxique d'un verbe, on pourrait en déduire automatiquement l'entrée correspondante ainsi que sa nature sémantique et syntaxique.

Références

Bohring, H. et S. Auer. Mapping XML to OWL Ontologies, In *Leipziger Informatik-Tage,* vol. 72, 2005, pp. 147–156. Society, Washington, DC, USA, 2007.

Ferdinand, Matthias, Christian Zirpins, and David Trastour. Lifting XML Schema to OWL. *Web Engineering*. Springer Berlin Heidelberg, 2004. 354-358.

François, Jacques, Denis Le Pesant et Danielle Leeman. Présentation de la classification des Verbes Français de Jean Dubois et Françoise Dubois-Charlier. *Langue française* 1, 2007, p 3-19.

Hadouche, Fadila et Guy Lapalme. Une version électronique du LVF comparée avec d'autres ressources lexicales. *Langages* 3 (2010): 193-220.

H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damljanovic, T. Heitz, M. A. Greenwood, H. Saggion, J. Petrak, Y. Li, and W. Peters. *Text Processing with GATE (Version 6)*. 2011.

Melnik, S. *Bridging the gap between XML and RDF*. http://wwwdb.stanford.edu/~melnik/rdf/fusion.html , 1999.

Motik, Boris, Peter F. Patel-Schneider et Bijan Parsia (eds), *OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax* (Second Edition), W3C Recommendation, 11 December 2012.

Niles, Ian et A. Pease. Mapping WordNet to the SUMO ontology. *Teknowledge Technical Report*, 2003.

Noy, Natalya F. et Deborah L. McGuinness. *Développement d'une ontologie 101: Guide pour la création de votre première ontologie*. Stanford, University Traduit de l'anglais par Anila Angjeli. http://www.bnf.fr/pages/infopro/normes/pdf/no-DevOnto.pdf (2000).

Van Assem, Mark, Aldo Gangemi et Guus Schreiber. Conversion of WordNet to a standard RDF/OWL representation. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. 2006.