Dictionnaires électroniques et étiquetage syntactico-sémantique

Pierre-André BUVET, Emmanuel CARTIER, Fabrice ISSAC, Salah MEJRI
LDI UMR 7187– Université Paris 13
{prénom.nom}@lli.univ-paris13.fr

Résumé. Nous présentons dans cet article le prototype d'un système d'étiquetage syntactico-sémantique des mots qui utilise comme principales ressources linguistiques différents dictionnaires du laboratoire *Lexiques, Dictionnaires, Informatique* (LDI). Dans un premier temps, nous mentionnons des travaux sur le même sujet. Dans un deuxième temps, nous faisons la présentation générale du système. Dans un troisième temps, nous exposons les principales caractéristiques des dictionnaires syntactico-sémantiques utilisés. Dans un quatrième temps, nous détaillons un exemple de traitement.

Abstract. We present in this paper a syntactico-semantics tagger prototype which uses as first linguistic resources various dictionaries elaborated at LDI. First, we mention several related works. Second, we present the overall sketch of the system. Third, we expose the main characteristics of the syntactico-semantic dictionaries implied in the processes. Last, using an example, we explicit the main stages of the analysis.

Mots-clés : étiqueteur sémantique, dictionnaire, LMF, XML, XPATH;

Keywords : word sense disambiguation (WSD), dictionnary, LMF, XML, XPATH;

1 Introduction

L'une des activités majeures du Laboratoire *Lexiques, Dictionnaires, Informatique* (LDI) est d'élaborer des dictionnaires électroniques à large couverture qui sont dédiés à des systèmes opérant sur des textes numérisés. Les descriptions contenues dans les dictionnaires sont de nature morphologique, d'une part, syntactico-sémantique, d'autre part. Les descriptions du second type sont effectuées dans le cadre théorique du modèle des classes d'objets (Gross, 1995, Le Pesant et Mathieu-Colas, 1998). Nous présentons dans cet article le prototype d'un système d'étiquetage syntactico-sémantique des mots qui utilise comme principales ressources linguistiques différents dictionnaires du LDI. Dans un premier temps, nous mentionnons des travaux sur le même sujet. Dans un deuxième temps, nous faisons la présentation générale du système. Dans un troisième temps, nous exposons les principales caractéristiques des dictionnaires syntactico-sémantiques utilisés. Dans un quatrième temps, nous détaillons un exemple de traitement.

2 État de l'art[1]

Nous rappelons comment l'étiquetage sémantique est généralement défini, puis nous indiquons dans quel cadre ce type d'étiquetage peut être utilisé et nous précisons finalement ses principales caractéristiques.

L'étiquetage sémantique est la tâche qui consiste à attribuer une valeur sémantique à un mot lexical[2]. Il s'agit d'une tâche intermédiaire dans un processus de traitement automatique des langues, puisqu'elle sert de point de départ à d'autres tâches plus directement en rapport avec la finalité du processus. Les étiquetages morphosyntaxiques et syntaxiques sont également des tâches intermédiaires que l'on considère disjointes de l'étiquetage sémantique.

La notion de valeur sémantique est beaucoup moins précise que celles de valeur morphosyntaxique et de valeur syntaxique car la structure et l'étendue des informations de nature sémantique sont beaucoup plus complexes. A l'instar de l'étiquetage morphosyntaxique, l'étiquetage sémantique implique que le choix d'une étiquette ne dépend pas seulement du mot mais aussi de son contexte. Ainsi dans les phrases *La pièce est dans le porte-monnaie* et *Le porte-monnaie est dans la pièce*[3], le calcul du sens du mot *pièce* nécessite des connaissances sur sa combinatoire compte tenu de sa polysémie.

Selon les types d'applications, les étiquettes sémantiques ont plus ou moins d'importance. Elles sont cruciales en traduction automatique ou en traduction assistée par ordinateur du fait que la transposition d'un texte d'une langue cible vers une langue source nécessite, entre autres tâches, d'attribuer la valeur exacte d'une forme donnée dans un contexte donné. Ainsi, pour traduire *abattre* par *cut down* dans *abattre un arbre* et *kill* par *tuer* dans *abattre un criminel*, il faut étiqueter la forme verbale comme un synonyme de *couper* dans un cas, de *tuer* dans l'autre. De même, en recherche d'information, il faut faire appel non seulement à l'étiquetage morphosyntaxique mais aussi à l'étiquetage sémantique pour construire des index et analyser les requêtes. Ainsi, une requête comme *Quelle est la vitesse du jaguar ?* ne sera pas associée à des résultats pertinents si la nature du mot *jaguar* (félin, avion ou voiture) n'est pas précisée.

L'étiquetage sémantique est conçu comme un processus en deux étapes : (i) utiliser une ressource linguistique du type dictionnaire ou du type ontologie pour attribuer un ensemble de sens à tous les mots pleins d'un texte ; (ii) utiliser des techniques symboliques ou numériques faisant appel à des ressources linguistiques pour éliminer les sens incorrects. Ce type de représentation permet de distinguer les ressources linguistiques à utiliser et les traitements à appliquer. La première étape ne présente aucune difficulté particulière pour peu que l'on dispose d'une ressource suffisamment complète ; elle consiste à attribuer des étiquettes sémantiques à des formes. La deuxième étape fait appel à des ressources linguistiques beaucoup plus riches[4] ou bien à un échantillon pré-étiqueté.

Le système utilise dans les deux cas des ressources linguistiques du type dictionnaire pour étiqueter sémantiquement les mots lexicaux. Nous montrons maintenant l'intérêt de faire appel à des dictionnaires électroniques paramétrés de façon syntactico-sémantique.

3 Présentation générale du système

L'élaboration du système est un projet en cours qui s'inscrit dans le programme TAL du LDI visant à construire une plateforme d'analyse syntactico-sémantique des textes dédiée à la mise

[1] On trouvera une description plus complète du domaine dans (Ide et Véronis, 1998) et (Véronis, 2004) ainsi qu'un aperçu des applications les plus récentes dans (Mihalcea et Edmonds, 2004).

[2] En anglais, Word Sense Disambiguation (WSD).

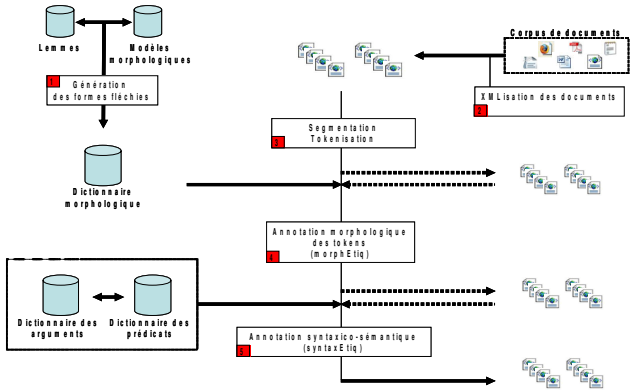
[3] Adaptation au français des phrases « the box is in the pen » et « the pen is in the box » de Yehoshua Bar-Hillel (Sabah, 1996).

[4] Par exemple, WordNet / EuroWordNet (Preiss, Stevenson, 2004) ou EST (Piao et al., 2005)

en place d'applications (veille, traduction, etc.) et à la gestion de corpus de documents par les linguistes du laboratoire .

Nous indiquons quelle est l'architecture générale du système puis nous précisons pour chaque module quelles sont les ressources utilisées. Ensuite, nous présentons le corpus de travail et nous détaillons la chaîne de traitements

3.1 Architecture générale



3.2 Ressources du système

Le système utilise trois **ressources** lexicographiques et un corpus :

- **dictionnaire morphologique** : Il comprend pour l'instant l'ensemble des formes simples du français (actuellement plus de 750 000 formes), qui sont générées à partir d'une table des lemmes et d'une table de modèles morphologiques. Par la suite, il intègrera l'ensemble des mots composés du français[5].
- **dictionnaire des arguments** : *cf. Infra*.
- **dictionnaire des prédicats** : *cf. Infra*.
- **Corpus** : il est composé pour l'essentiel d'articles du journal *Le Monde* au format texte ainsi que d'autres types de documents dans des formats variés (HTML, Open Office, PDF, Word et RTF)

Un pré-traitement permet de normaliser les documents en les convertissant de leur format d'origine vers un format XML commun. Il consiste à récupérer des informations sur la structure des documents. Nous avons opté, en l'état actuel, pour une DTD « minimale », permettant d'annoter les titres, sous-titres, sections et sous-sections ainsi que les paragraphes du texte. Dans une phase ultérieure, des annotations supplémentaires permettront de rendre compte de structures textuelles plus fines, comme les listes et les tableaux. La normalisation XML des documents permet de les décorer de balises `<text><title level='1..N'><section level='1..N'><p>`. Le pré-traitement permet également de convertir les documents de leur encodage d'origine vers l'encodage UTF-8[6].

[5] Ce dictionnaire est issu des travaux de Michel Mathieu-Colas au LLI

[6] Ce travail a été effectué avec la promotion 2006-2007 des étudiants du Master PRO TILDE de l'Université Paris 13.

Le système comprend deux étapes de pré-traitement (génération du dictionnaire morphologique, étape 1, et normalisation XML des documents, étape 2) et trois étapes de traitement. A chaque étape, le système prévoit une sortie afin d'effectuer une évaluation des résultats.

3.3 Étapes du traitement

Etape 3 : segmentation en phrases et en mots des documents XML : la segmentation en phrases et en mots permet de décorer les documents de deux nouvelles balises : `<s>` et `<w>` (respectivement « sentence » et « word »).
La phase de segmentation en phrases (définies du point de vue graphique) résulte de l'analyse et du classement des différents sortes de 'point' dans le corpus « Le Monde ». Des listes d'abréviations et d'acronymes courants ainsi qu'une série d'expressions régulières permettent de repérer les 'points' qui ne sont pas en fin de phrase. Un sous-corpus a été manuellement annoté pour constituer un corpus de référence de 150 000 mots. Les résultats de la segmentation automatique sont un taux de rappel de 98,7% et un taux de précision de 97,6%.
La phase de segmentation en mots a consisté tout d'abord à reconnaître un certain nombre d'unités textuelles, qu'il s'agisse d'entités numériques (*5,7%*, *13,2 millions*), d'entités temporelles à base numérique (*le 12/12/2005*) ou encore d'entités spécifiques (url, mail, etc.). Elle a également consisté à normaliser les signes de ponctuation (réduction de plusieurs espaces en un seul, séparation des signes de ponctuation, normalisation des signes d'élision...), puis à segmenter les phrases en mots sur la base de l'espace.[7]
Etape 4 : annotation morpho-syntaxique des tokens : les chaînes de caractères reconnues comme des mots font l'objet d'une annotation morphosyntaxique. Par projection, les informations contenues dans le dictionnaire morphologique sont associées aux tokens : lemme, catégorie morphosyntaxique, informations de nombre, genre, mode, temps, personne[8].
A ce stade, les mots peuvent être ambigus. Par exemple, *porte* comprend l'ensemble des informations liées aux formes verbales (mode : indicatif, subjonctif ; temps : présent, personne$1 : 1, 3) et à la forme nominale (nom, féminin, singulier). Dans le fichier XML adéquat, cela se traduit par l'annotation suivante :

```
<w                                                              case='lc'>
 <morph  lemma='porter'  cat='v'  mood=''subj'  tense='pres'  pers='1'/>
 <morph  lemma='porter'   cat='v'  mood=''subj'  tense='pres'  pers='3'/>
 <morph  lemma='porter'   cat='v'  mood=''ind'  tense='pres'  pers='1'/>
 <morph  lemma='porter'   cat='v'  mood=''ind'  tense='pres'  pers='3'/>
 <morph      lemma='porte'       cat='n'      nb='plu'      gender='fem'/>
 porte
</w>
```

Etape 5 : annotation syntaxico-sémantique : l'annotation syntaxico-sémantique est l'étape la plus complexe du traitement. Elle fait appel au dictionnaire des arguments et à celui des prédicats. Elle se décompose en plusieurs sous étapes :

- 5a. désambiguisation morpho-syntaxique : le système élimine un certain nombre de solutions morphologiques issues de la projection des dictionnaires sur les mots « hors contexte » en faisant appel à des règles correspondant à des grammaires locales.

[7] Dès cette phase, un certain nombre d'informations typographiques sont incluses sous forme d'attributs : le type de l'élément (valeur de l'élément : w pour word, num pour infos numériques, punct pour signe de ponctuation, symb pour symbole) ainsi que des attributs de casse pour w (case='lowercase,uppercase,titlecase') et un attribut type pour num (afin de reconnaître, entre autres, les dates).
[8] *Cf.* Mathieu-Colas, 2007, à propos des informations morphologiques utilisées.

- 5b. projection du dictionnaire des arguments : le dictionnaire des arguments est projeté sur le fichier XML afin d'ajouter des informations de type sémantique au niveau de certains signes de type mot « w ».
- 5c. Annotation des constructions syntaxiques : le système identifie diverses constructions syntaxiques : groupes nominaux, groupes prépositionnels, etc.. Pour chaque groupe, un attribut « head » correspondant à la « tête sémantique» des différents groupes est également récupéré. Sont également reconnues les séquences relatives à des entités nommées (personnes, organisations, lieux, artefacts, événements).
- 5d. Analyse prédicative : il s'agit enfin d'identifier les structures prédicat-arguments à l'aide du dictionnaire des prédicats. Le système génère automatiquement une expression XPATH à partir des descriptions lexicographiques. Par exemple, pour l'entrée « subir », dans le sens de subir(EVENEMENT,ENTITE CONCRETE), nous aurons l'expression XPATH :
  `group[1][@cat='gn' and @sem='ENTITE'] and group[2][@cat='gv' and @head='subir'] and group[3][@cat='gn' and @sem='EVE'].`

Le XML résultant de l'analyse prédicative se présente comme suit :

```
<s>
       <gn head='12/12/2003'>
              <w case='tc' ><morph lemma='le' cat='det'>Le</morph></w>
              <num type='date'>12/12/2003</num>
       </gn>
              <punct>,</punct>
<pred sem='subir' X0='population' X1='hausse'>
       <gn head='population'>
          ...
      </gn>
      <gv head='subir'>
       ...
      </gv>
      <gn head='hausse'>
       ...
     </gn>
</pred>
              <w typo='punct'>,</w>

    <gadj head='atteindre'>
          ...
      </gadj>
              <punct>.</punct>
</s>
```

4 Les dictionnaires syntactico-sémantiques

Le modèle des classes d'objets subdivise les unités linguistiques à partir du postulat suivant : toute phrase élémentaire est constituée d'un prédicat du premier ordre et de ses éventuels arguments, les autres constituants phrastiques ressortissant à l'actualisation. Il s'ensuit trois sous-catégories majeures : celle des prédicats, celle des arguments élémentaires et celles des actualisateurs. La description des items de la dernière sous-catégorie est subordonnée à celles des items des autres sous-catégories (Buvet, à paraitre). Les descriptions des arguments élémentaires et des prédicats sont formalisées dans des dictionnaires électroniques. Nous discutons de ces dictionnaires du point de vue linguistique puis du point de vue informatique.

4.1 Structuration linguistique

Deux sortes de dictionnaires syntactico-sémantiques sont élaborés au LLI : ARGU-DIC et PREDI-DIC. Le premier dictionnaire décrit les arguments élémentaires, le second les prédicats.

ARGU-DIC : les arguments élémentaires sont des substantifs qui ne peuvent jamais occuper une position prédicative dans une construction à support (Gross et Vives, 1986). D'un point

de vue syntactico-sémantique, ils sont caractérisés en termes de classes d'objets (<aliment>,<moyen de transport>, <outil>, <voie>, etc.) et de domaines (<aéronautique>, <médecine>, <sciences>, etc.) (Buvet et Mathieu-Colas, 1999). La macrostructure du dictionnaire est donc constituée de l'ensemble des noms correspondant à des arguments élémentaires. La microstructure comporte la vedette et des informations métalinguistiques relatives aux classes et aux domaines[9].

PRED-DIC : la structuration du dictionnaire des prédicats est plus complexe. Nous décrivons successivement la macrostructure puis la microstructructure.

Macrostructure

La nomenclature du dictionnaire PRÉD-DIC est constituée des racines prédicatives correspondant à autant d'emplois prédicatifs. Nous précisons successivement les notions de racine prédicative et d'emploi prédicatif.

Racine prédicative : la notion de racine prédicative rend compte du caractère polymorphe de certains prédicats : *Il l'aime/Il est amoureux d'elle/Il éprouve de l'amour pour elle*. La parenté morphologique entre le verbe, l'adjectif et le nom en position de prédicat et la stricte équivalence entre les trois énoncés permettent d'interpréter *aimer*, *amoureux* et *amour* comme trois formes différentes d'une même racine prédicative. Toutes les racines prédicatives ne donnent pas lieu à des énoncés équivalents : *Ceci gêne Luc/Ceci est gênant/Luc ressent de la gêne*.

Emploi prédicatif : un emploi prédicatif est défini conjointement par une racine prédicative, une classe sémantique et une interprétation donnée. Deux cas de figure sont à envisager selon que l'emploi prédicatif est autonome (*grognon*) ou polymorphique (*dédain/dédaigner/ dédaigneux*).

Une racine prédicative polymorphique permet de rendre compte d'emplois prédicatifs non équivalents mais morphologiquement reliés et sémantiquement apparentés : *Luc déteste Max/Max est détesté/Max est détestable*. Les trois formes prédicatives ont une interprétation spécifique : *détester* s'interprète comme une 'propriété occasionnelle', *détesté* comme une 'propriété résultative' et *détestable* comme une 'propriété causale permanente'.

L'interprétation d'un prédicat est le produit de sa construction, de son trait et de son aspect inhérent. Si *détester*, *détesté* et *détestable* sont des prédicats qui partagent la même racine prédicative, en l'occurrence **détest-**, et appartiennent à la même classe sémantique <haine>, ils n'ont pas cependant la même interprétation.

1. celle du verbe résulte du fait qu'il a la construction **X0 V X1** avec **X0** ='humain' et **X1**='humain', le trait 'état' et l'aspect 'provisoire' ;
2. celle de l'adjectif participe tient au fait qu'il a la construction **X0 être Appé** avec **X0**= 'humain', le trait 'état' et l'aspect 'accompli' ;
3. celle de l'adjectif en *–able* s'explique parce qu'il a la construction **X0 être A** avec **X0**= 'humain', le trait 'état' et l'aspect 'permanent'.

Quels que soient les dictionnaires, les vedettes sont à l'intersection de leur macrostructure et de leur microstructure. Les vedettes de PRED-DIC sont des racines prédicatives correspondant à autant d'emplois prédicatifs. Autrement dit, une racine prédicative apparaît dans plus d'une entrée lorsqu'elle correspond à plus d'un emploi prédicatif. Par contre, un emploi prédicatif polymorphique est décrit sous la même entrée et il est spécifié dans l'article afférent quelles sont les différentes formes qu'il recouvre. Ainsi, *détestation* est dans le même article que *détester* du fait de l'équivalence des deux phrases suivantes :

Luc déteste Max

[9] Tous les noms du dictionnaire ne nécessitent pas d'être caractérisés par un domaine.

Luc a de la détestation pour Max

Nous présentons maintenant les informations métalinguistiques qui constituent le reste d'un article de PRED-DIC.

Microstructure

Les descripteurs associés à l'entrée d'un dictionnaire sont tous des propriétés linguistiques. Ils sont de trois ordres : les descripteurs de définition, les descripteurs de conditions et les descripteurs de validation

Les descripteurs de définition : il s'agit de la classe sémantique et de l'interprétation de l'emploi prédicatif correspondant à l'entrée. Elles constituent les deux informations qui sont associées à la racine prédicative lors de l'étiquetage syntactico-sémantique.

Les classes sémantiques qui caractérisent les prédicats sont en assez grand nombre (environ 2000 en l'état actuel des travaux du LLI). Par contre, les interprétations possibles sont limitées en nombre (une douzaine).

Les descripteurs de conditions : ce sont les différentes propriétés linguistiques qui permettent à l'étiqueteur sémantique de déterminer quels sont les emplois prédicatifs des racines prédicatives. Elles sont de cinq sortes.

Les propriétés morphologiques : ces propriétés sont au nombre de trois. Tout d'abord, elles indiquent les diverses formes simples possibles d'un même emploi prédicatif (le verbe *prendre* et le nom *prise* ou uniquement le verbe *prendre*). Elles font état de l'éventuel caractère complexe de l'entrée (*prendre ombrage*). Elles signalent aussi la défectivité (*prendre fin* ne s'emploie qu'à la troisième personne).

Les propriétés structurelles : elles font état du nombre des arguments et de leur mode de structuration. Le nombre de structures possibles dépend de la forme des prédicats. Par exemple, *prendre* est caractérisé par :

1. la construction **X0 V X1** en tant que synonyme de *commander* (*Luc prend une bière*) ;
2. la construction **X0 V X1 PREP2 X2** en tant que synonyme de *tenir* (*Luc prend Léa par la taille*) ;

Les propriétés distributionnelles : elles font état de la structure argumentale du prédicat en indiquant, d'une part, la nature syntaxique des arguments (groupe nominal, complétive, infinitive, etc.) et, d'autre part, la nature sémantique des prédicats (en termes de classes d'objets ou d'hyperclasses). Il est possible de la sorte d'établir que *prendre* a deux acceptions différentes selon que la position **X0** est occupée :

1. indifféremment par un groupe nominal ou une infinitive (*(Cette affaire Faire cela) prend du temps*) ;
2. seulement par un groupe nominal qui, de plus, correspond nécessairement à un humain (*Luc prend du temps*).

Les propriétés combinatoires : ces propriétés sont dissociées selon qu'elles ressortissent à la signification grammaticale (*brûler* dans *Luc brûle d'amour pour Léa*) ou bien à la signification lexicale (*intelligemment* dans *Luc a présenté le projet intelligemment*) (Blanco et Buvet, 2004).

Les propriétés paraphrastiques : il s'agit de reconstructions des phrases canoniques, typiquement le passif ou la forme pronominale. La construction standard *Luc prend Max au sérieux* donne la reconstruction du type passif *Max est pris au sérieux par Luc* et la reconstruction du type forme pronominale réfléchie *Luc se prend au sérieux*.

Elles sont symptomatiques de la polysémie des racines prédicatives dans la mesure où les reconstructions varient selon les emplois.

Les descripteurs de validation : ces descripteurs justifient les interprétations des racines constituant les entrées. Ils permettent de vérifier la cohérence de la définition proposée dans l'article. Si certains sont aussi des descripteurs de conditions (par exemple, les propriétés structurelles), les autres sont de propriétés sémantiques spécifiques qui ne participent pas directement à l'identification de l'emploi. Il s'agit du type (état, action événement) et l'aspect intrinsèque de l'emploi prédicatif (e.g. le ponctuel pour *gifler*).
La description paramétrée des différents emplois prédicatifs et des arguments donne à lieu à diverses informations explicites qu'il est possible de structurer sous un format informatisable.

4.2 Modélisation informatique

L'augmentation des performances des systèmes TAL est directement liée à celle des ressources linguistiques tant du point de vue qualitatif que du point de vue quantitatif. L'exploitation informatique de ces ressources, qu'elles soient de nature morphologique, syntaxique ou sémantique, est toujours problématique. Pour des raisons de réutilisabilité et de pérennité, il faut que les ressources respectent des normes reconnues par tous les acteurs du domaine (Francopoulo, 2006). La structuration informatique de ARGU-DIC et PRED-DIC utilise la pré-norme LMF (ISO, 2006). Celle-ci propose non pas une DTD XML toute faite mais plutôt un cadre dans lequel il possible de construire et documenter un grand nombre de ressources linguistiques décrites dans des formalismes très divers. Le risque de contraindre une représentation par le biais d'un DTD est l'abandon de celle-ci lorsqu'un modèle ne peut y trouver sa place. Nous nous appuyons également sur les recommandations de la TEI en ce qui concerne la représentations des dictionnaires « papier » et le codage des entêtes. La structuration métalinguistique de PRED-DIC est prise en charge au format XML comme suit :

```
<entry id="prendre_1" class="capture" int="operation">
    <root>prendre</root>
    <example>les enfants prennent un chat
    </example>
    <definition>
        <item name="class" val="capture"/>
        <item name="int" val="operation"/>
    </definition>
    <morphProp>
        <item val="verb"/>
    </morphProp>
    <structprop>
        <item pred="verb" val="X0_V_X1"/>
    </structProp>
    <distProp>
        <struct type="syntax">
            <item arg="0" val="np"/>
            <item arg="1" val="np"/>
        </struct>
        <struct type="semantic">
            <item arg="0" val="hum"/>
            <item arg="1" val="animal"/>
        </struct>
    </distProp>
    <semanticProp>
        <item type="feature" val="action"/>
        <item type="aspect" val="perfective"/>
    </semanticProp>
    <appropriateProp></appropriateProp>
    <paraphrasticProp>
        <item pred="verb" val="passive"/>
    </paraphrasticProp>
</entry><classFrame>
        <item id="capture"/>
        <item id="hum"/>
        <item id="animal"/>
</classFrame>

<intFrame>
        <item id="operation"/>
</intFrame>

<interpretationFrame>
        <item id="operation"/>
</interpretationFrame>

<propertyFrame typepred="verb">
        <struct id="X0_V_X1">
                <item val="0"/>
                <item pred="verb"/>
                <item val="1"/>
        </struct>
        <struct id="X0_V_X1opt">
                <item num="0"/>
                <item pred="V"/>
                <item num="1" type="opt"/>
        </struct>
</propertyFrame>
```

5 Exemple de traitement[10]

Niveau d'analyse	Infos	Les	jeunes	prennent	les	autoroutes	à	contre-sens	le	14	juillet	à	13h35	sur	la	huitième	avenue	.	Commentaires
Typographique	Sign	w	w	w	w	w	w	w	w	num	w	w	time	w	w	w	w	punct	**0**
	Type												hour					term	**1**
	Case	tc	lc	lc	lc	lc	lc	lc	lc		lc	lc		lc	lc	lc	lc		
Morphologique	morph																		
	cat	det|pro	adj|n	v	det|pro	n	prep	n	det|pro		n	prep		prep	det|pro	adj	n		
	sub_cat	def|pers			def|pers				def|pers						def|pers	num			
	Nb	plu		plu															
	Gender																		
	Pers	3		3															
	Tense			pres															
	Mode			ind|subj															
5. Analyse syntaxico-sémantique																			
a. Désambig morpho		det	n		det				det						det				**2**
b. Projection dict. Arguments	Sem		hum			voie											voie		
	Dom																		
c. Reconnaissance des Groupes		GN		GV	GN		GPREP		GN			GPREP		GPREP					**3**
	Head	Jeune		Prendre	Autoroute										Avenue				

Commentaires (exemples de règles utilisées pour générer les représentations)

0 : les signes sont reconnus via des expressions régulières UNICODE.

Exemples : `w=(\p{L})+(-)(\p{L})+; punct=\p{P}`

1 : un certain nombre d'entités nommées à base numérique sont reconnues dès la reconnaissance typographique.

Exemple : `Hour=[0-9]{2}[h:][0-9]{2}`

2 : En français, la désambiguisation morpho-syntaxique porte notamment sur les déterminants/pronoms.

Exemple de règle de désambiguisation : `morph[1][@cat in 'det,pro'] and morph[@cat in 'n,adj'] and morph[3][@cat='v']=>det+n+v`

3 :la reconnaissance des groupes se fait sur la base de grammaires locales exprimées en XPATH :

Exemples : `w[1][@cat='det'] and w[2][@cat='n'] and w[3][@cat='v'] => GN[det n] V`

4 : Reconnaissance prédicative

Exemple : `GN[1][@sem='hum'] and GV[2][@head='prendre'] and GN[3][@sem='voie']`

[10] La phrase exemple est tirée du Monde en ligne, article du 19/07/2003

6 Conclusion

Nous avons présenté un prototype d'étiqueteur syntactico-sémantique intégré dans une plate-forme d'analyse linguistique. La particularité de cet étiqueteur est d'utiliser le modèle linguistique des classes d'objets. Celui-ci offre un moyen efficace de lier la syntaxe et la sémantique au sein d'une structure. Du point de vue informatique, nous avons décrit un système qui reprend les grandes étapes « classiques » en T.A.L., en nous appuyant sur des ressources linguistiques, et en mettant au point un langage d'expression de grammaires locales très proche d'une expression abstraite et quasi linguistique des phénomènes. L'implémentation proprement dite de ce modèle a été réalisé en respectant les normes (ou futures normes) et recommandations reconnues.

7 Bibliographie

BLANCO X et BUVET P.-A. (2004), « Verbes supports et significations grammaticales. Implications pour la traduction espagnol-français » in *Lingvisticae Investigationes* 27(2), John Benjamins B.V., Amsterdam

BUVET P.-A. (à paraître), « Détermination et figement au regard de la traduction », *META*.

BUVET P.-A. et MATHIEU-COLAS M. (1999), « Les champs *domaine* et *sous-domaine* dans les dictionnaires électroniques », *Cahiers de lexicologie*, 75, Didier Erudition, Paris, pp. 173-191.

SABAH G. (1996), Le sens dans les traitements automatiques des langues — le point après 50 ans de recherches, conférence invitée, journée ATALA (un demi-siècle de traitement automatique des langues : Paris.

GROSS G. (1995), « Une sémantique nouvelle pour la traduction automatique : les classes d'objets », in *La Tribune des Industries de la Langue et l'Information électronique*, 17-18-19, Paris.

GROSS G. et VIVES R. (1986), « Les constructions nominales et l'élaboration d'un lexique-grammaire », *Langue française*, 69, Larousse, Paris, pp. 5-27.

IDE, N., VERONIS, J., (1998). The state of the art. Computational Linguistics 24, 1–40, Introduction to the Special Issue on Word Sense Disambiguation

LE PESANT D. et M. MATHIEU-COLAS (1998), « Introduction aux classes d'objets » in *Langages* 131, Larousse, Paris.

MIHALCE R., EDMOND P. (Eds.), (2004). Proceedings of SENSEVAL-3 : Third International Workshop on Evaluating Word Sense Disambiguation Systems

PREISS J., STEVENSON M., (2004). *Word Sense Disambiguation*, Computer Speech & language,Volume 18, Issue 3

PIAO S.L., ARCHER D., MUDRAYA O., RAYSON P., GARSIDE R., MCENERY T., WILSON A. (2005) A Large Semantic Lexicon for Corpus Annotation. In proceedings of the Corpus Linguistics 2005 conference, July 14-17, Birmingham, UK. Proceedings from the Corpus Linguistics Conference Series on-line e-journal, Vol. 1, no. 1, ISSN 1747-9398.

VERONIS, J. (2004). « Quels dictionnaires pour l'étiquetage sémantique ? » Le français moderne, 72(1):27-38.