Intégration d'informations syntaxico-sémantiques dans les bases de données terminologiques : méthodologie d'annotation et perspectives d'automatisation

Fadila Hadouche
Laboratoire RALI, DIRO
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal (Québec), Canada H3C 3J7
hadouchf@iro.umontreal.ca

Marie-Claude L'Homme
Observatoire de linguistique Sens-Texte
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal (Québec), Canada H3C 3J7
mc.lhomme@umontreal.ca

Guy Lapalme
Laboratoire RALI, DIRO
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal (Québec), Canada H3C 3J7
lapalme@iro.umontreal.ca

Annaïch Le Serrec
Observatoire de linguistique Sens-Texte
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal (Québec), Canada H3C 3J7
Annaich.le.serrec@umontreal.ca

Résumé

Dans le présent article, nous décrivons un modèle accompagné d'une méthodologie visant à expliciter les propriétés syntaxico-sémantiques de termes. L'explicitation est réalisée au moyen d'une annotation des propriétés de termes de nature prédicative et de leurs actants dans des contextes extraits de corpus spécialisés. Le projet comporte deux volets. Le premier consiste à définir le modèle à partir d'une annotation manuelle. Le second consiste à mettre au point une méthode d'annotation automatique. Nous décrivons les objectifs du projet ainsi que son l'état d'avancement.

1 Introduction

Les dictionnaires spécialisés et les bases de données terminologiques, quoique riches en information de nature conceptuelle, fournissent en général très peu de renseignements sur les propriétés linguistiques des termes ou sur leur comportement en langue. Quelques exceptions, toutefois, ont commencé à faire leur apparition : des dictionnaires spécialisés conçus dans une perspective d'apprentissage (par exemple, le Binon *et al.* 2000 dans le domaine des affaires) ou des bases terminologiques incorporant des rubriques sur le fonctionnement syntaxique ou sur la combinatoire des termes (par exemple, Lépinette & Olmo Cazevieille 2007; DiCoInfo dans le domaine de l'informatique). En outre, les bases de données lexicales sont de plus en plus sollicitées dans de nombreuses applications en traitement des langues naturelles (TAL). Il devient donc nécessaire de fournir sur les unités lexicales – qu'elles soient générales ou spécialisées – une description explicite et formelle de l'ensemble de leurs propriétés linguistiques.

Le présent projet s'inscrit dans cette mouvance. Il vise à proposer un modèle ainsi qu'une méthodologie permettant d'intégrer aux bases de données terminologiques une explicitation des propriétés syntaxico-sémantiques des termes. Les propriétés ainsi décrites peuvent se prêter à trois applications : 1. offrir aux utilisateurs un accès à des renseignements liés au fonctionnement des termes dans les textes; 2. permettre aux terminologues d'appuyer leurs descriptions sur des réalisations véritables

en corpus et de confirmer ou d'infirmer leurs intuitions; 3. intégrer à des applications de TAL des descriptions qui rendent compte des propriétés syntaxiques et sémantiques des termes et des liens entre ces deux ensembles de propriétés.

L'explicitation est réalisée au moyen d'une annotation formelle des dites propriétés réalisés dans des contextes dans lesquels les termes apparaissent. Le projet comprend deux volets : 1. le premier volet consiste à mettre au point le modèle d'annotation à partir d'une analyse manuelle d'un nombre important de contextes; 2. le second volet vise à mettre au point une méthode d'annotation automatique dont l'amorce repose sur un apprentissage réalisé à partir des contextes traités manuellement.

Dans les pages qui suivent, nous rendons compte de l'état d'avancement du projet. L'article est construit de la manière suivante. La section 2 décrit les objectifs généraux du projet et aborde la question de l'intérêt d'une annotation syntaxico-sémantique des contextes extraits de corpus spécialisés. La section 3 porte plus spécifiquement sur l'annotation manuelle réalisée dans deux domaines spécialisés, à savoir l'informatique et le réchauffement climatique. La section 4 porte sur la méthode d'automatisation de l'annotation envisagée et sur les étapes réalisées jusqu'à présent. Nous concluons en présentant quelques perspectives.

2 Pourquoi annoter les termes de nature prédicative ?

L'explication des propriétés syntaxico-sémantiques des termes dans les dictionnaires spécialisés et bases de données constituent une source de renseignements extrêmement utiles sur plusieurs plans. Nous nous intéressons plus spécifiquement aux termes de nature prédicative, et, dans une première étape, aux verbes.

Une description des propriétés syntaxico-sémantiques des verbes permet de mettre en évidence, en premier lieu, les constructions syntaxiques qu'ils admettent. Outre les propriétés des verbes eux-mêmes, la description s'intéresse au comportement de leurs participants, à savoir les actants (également appelés *arguments*) et les circonstants: [2] leur nombre, leur rôle sémantique, les modalités de leur combinatoire avec le verbe, etc.

2.1 Annotation syntaxico-sémantique des structures actancielles de termes

Comme nous l'avons signalé dans l'introduction, l'enrichissement des bases de données terminologiques envisagé ici prend la forme d'annotations insérées formellement dans les contextes dans lesquels apparaissent les termes. La méthodologie d'annotation s'inspire fortement de celle développée dans le cadre du projet FrameNet (Ruppenhofer 2002; FrameNet 2008),[3] mais s'en distingue sur certains plans. Une partie de ces distinctions seront évoquées plus loin dans l'article.

La figure 1 montre deux contextes dans lesquels les propriétés de termes sont explicitées : le premier illustre le terme d'informatique *affecter* et ses différents participants; le deuxième est extrait d'un corpus du changement climatique.

Les éléments explicités sont les suivants :

1. Terme prédicatif faisant l'objet de l'annotation (dans la figure 1, *affecter*);
2. Participants et leur nature (actants ou circonstants) : les participants sont mis en évidence directement dans le contexte, puis énumérés dans un tableau récapitulatif (à la figure 1, les actants sont en caractères gras dans le contexte et apparaissent dans la partie supérieure du tableau récapitulatif; les circonstants apparaissent dans la partie inférieure du tableau);
3. Rôle sémantique du participant : le rôle sémantique (**Agent**, **Patient**, **Destination**, etc.) est indiqué dans le contexte et dans le tableau récapitulatif (la question des rôles sémantiques est abordée à la section 2.3);

[2] La distinction entre "actant" et "circonstant" s'appuie sur Mel'čuk (2004). Les actants sont définis comme des participants obligatoires et contribuent au sens d'unités lexicales de sens prédicatif. Les circonstants apparaissent dans des phrases et peuvent entretenir un lien syntaxique avec l'unité lexicale, mais ne contribuent pas au sens de l'unité en question.

[3] Signalons qu'un nombre croissant de travaux terminologiques ont recours à des représentations s'inspirant des Frames sémantiques et de la méthodologie utilisée pour élaborer FrameNet (2008). Nous pouvons citer Dolbey *et al.* (2006) Faber *et al.* (2006) et Schmidt (à paraître).

4. Fonction syntaxique du participant : la fonction syntaxique (sujet, objet, complément, modificateur) du participant est indiquée dans le tableau récapitulatif;
5. Groupe syntaxique du participant : le groupe syntaxique (SN, SP, SAdv, Prop) du participant est également donné dans le tableau récapitulatif.

Les *développeurs* peuvent AFFECTER **une *valeur* à cette *variable***.

AFFECTER 1		
Actants		
Agent	Sujet (SN) (1)	développeur
Patient	Objet (SN) (1)	valeur
Récipient	Complément (SP-à) (1)	variable

L'*élévation* du niveau de la mer AFFECTERA **les *écosystèmes* des mangroves** en éliminant leurs habitats actuels et en créant des zones inondées par les marées vers lesquelles certaines espèces de mangrove pourraient se déplacer.

AFFECTER 1		
Actants		
Cause	Sujet (SN) (1)	élévation
Patient	Objet (SN) (1)	écosystème
Autres		
Mode	Complément (Prop) (2)	en éliminant leurs habitats actuels en créant des zones inondées par les marées vers lesquelles certaines espèces de mangrove pourraient se déplacer

Figure 1. Contextes annotés contenant le verbe *affecter*

2.2 Rôles sémantiques

Les annotations sont articulées autour des rôles sémantiques des participants (actants et circonstants) d'un terme prédicatif. Le rôle sémantique est défini comme le lien partagé par le participant et le terme prédicatif (**Agent**, **Patient**, **Instrument**, **Récipient**). Malgré le caractère subjectif qu'attribue parfois la littérature à la notion de « rôle sémantique », la représentation des participants par des étiquettes de rôles se révèle un outil efficace pour rendre compte des participants partageant le même lien avec des unités prédicatives différentes (Fillmore 1968; FrameNet 2008; VerbNet 2008).

L'exemple (1) montre de quelle manière les actants de termes de l'informatique sont représentés au moyen d'étiquettes de rôles. Bien que la position syntaxique des actants puisse varier, leur rôle restera le même.

(1) COMPILER$_{1a}$: **Instrument** ~ **Patient** (ex. *un compilateur compile du code*)
COMPILER$_{1b}$: **Agent** ~ **Patient** avec **Instrument** (ex. *un programmeur compile du code au moyen du compilateur x*)
COMPILABLE$_1$: **Patient** ~ (ex. *du code compilable*)
COMPILATEUR$_1$: ~ utilisé par **Agent** pour intervenir sur **Patient** (ex. *compilateur de Java*)

L'exemple (2) montre que l'annotation en rôles permet de mettre en évidence des liens de synonymie, de quasi-synonymie et d'opposition entre termes.

(2) STOCKER$_{1a}$: **Destination** ~ **Patient** (ex. *des végétaux stockent du CO_2*)
PIÉGER$_1$: **Destination** ~ **Patient** (ex. *le permafrost piège du dioxyde de carbone*)
EMPRISONNER$_1$: **Destination** ~ **Patient** (ex. *l'atmosphère emprisonne plus de chaleur*)

LIBÉRER$_{1a}$: **Source** ~ **Patient** (ex. *les eaux chaudes libèrent de la chaleur*)

Les étiquettes de rôles sémantiques auxquelles nous avons recours s'apparentent aux étiquettes utilisées pour représenter les éléments d'un Frame (FE) dans FrameNet (2008). Toutefois, elles s'en distinguent en ce sens que nous tentons de définir un nombre limité d'étiquettes qui s'appliqueront à l'ensemble des termes dans un domaine spécialisé (et non uniquement à l'intérieur d'un seul Frame).[4]

2.3 Schéma d'annotation XML

Le processus décrit en 2.1 est effectué en ajoutant des balises XML aux exemples d'utilisation et de descriptions tirés de corpus spécialisés. Afin de systématiser les différents types de balises selon les éléments décrits en 2.1 ainsi que leur imbrication, nous avons défini un schéma XML pour l'ensemble des unités lexicales que nous avons annotées. La figure 2 donne l'extrait du Schéma correspondant à l'annotation d'un participant. Chaque participant est identifié par son type (Actant ou circonstant) et son rôle (**Agent**, **Patient**, etc.) et contient une description de la fonction syntaxique (Objet, Sujet, etc.) exprimée en terme d'un groupe syntaxique (SN, SP, etc.) qui décrit la réalisation de cet actant par du texte comprenant une réalisation syntaxique (ceci est indiqué par *mixed*). À chaque réalisation est associée un identificateur *ref* qui permet de relier plusieurs occurrences (ex. une référence pronominale ou anaphorique) d'un même actant.

```
element-participant = element participant {
   attribute type {TypeParticipant},
   attribute role {RoleParticipant},
   element-fonction-syntaxique
}
element-fonction-syntaxique = element fonction-syntaxique {
   attribute nom {NomFonctionSyntaxique},
   attribute cas {CasFonctionSyntaxique}?,
   element-groupe-syntaxique
}
element-groupe-syntaxique = element groupe-syntaxique {
   attribute nom {NomGroupeSyntaxique},
   attribute preposition {text}?,
   attribute particule {text}?,
   mixed {element-realisation}
}
element-realisation = element realisation{
   attribute lemme {text}?,
   attribute etiquette {text}?,
   attribute ref {xsd:IDREF}?,
   attribute reflex {xsd:IDREF}?,
   text
}
```

Figure 2: Schéma XML (RelaxNG forme compact) des annotations d'un participant

Selon ce schéma, les participants dans le contexte **`Vous pouvez aussi, à ce moment-là, abandonner l'installation`**, sont annotés en XML comme indiqués à la figure 3. Cette annotation

[4] Ce faisant, nous nous opposons assez radicalement aux hypothèses formulées dans le cadre du modèle des Frames Sémantiques. Nous croyons effectivement qu'il est possible de rendre compte des structures actancielles de termes d'un domaine au moyen d'un nombre plus limité d'étiquettes s'apparentant davantage aux étiquettes proposées par Fillmore (1968). En outre, les unités lexicales ne sont pas regroupées dans des Frames et, par conséquent, ne sont pas hiérarchisées entre elles. Les hiérarchies proposées par FrameNet permettent de rendre compte de parentés sémantiques existant entre unités lexicales, mais appartenant à des Frames distincts.

permet une identification fine des participants: **Vous** comme **Agent**, **à ce moment-là** comme *Circonstant* avec rôle **Temps** et **l'installation** comme **Patient**. Le contenu de la balise participant donne des informations syntaxiques sur les actants. En utilisant un éditeur spécialisé pour XML, l'annotation peut être effectuée rapidement en sélectionnant les participants avec la souris et en choisissant les balises appropriées dans des menus. La définition d'un schéma autorise l'annotateur à n'afficher dans les menus que les balises qui, à chaque étape, font en sorte que l'annotation respecte le schéma.

L'utilisation de feuilles de styles appropriées offre la possibilité de générer plusieurs types de présentation des annotations : nous utilisons présentement des pages HTML qui distinguent les types d'actants par des codes de couleur (figure 1). Cette présentation permet également d'obtenir des statistiques sur le nombre d'occurrences des actants et leurs réalisations sur plusieurs contextes (figure 1). Ces informations permettent aux annotateurs de valider plus facilement leur travail car le code source XML est assez *touffu.* De plus, l'insertion de balises force une division des éléments de la phrase imposant à l'annotateur une lecture peu naturelle.

```
<participant type="Act" role="Agent">
  <fonction-syntaxique nom="Sujet">
    <groupe-syntaxique nom="SN">
       <realisation>Vous</realisation>
    </groupe-syntaxique>
   </fonction-syntaxique>
 </participant> pouvez aussi,
 <participant type="Circ" role="Temps">
  <fonction-syntaxique nom="Complement">
    <groupe-syntaxique nom="SP" preposition="à">à ce
      <realisation>moment-là</realisation>
    </groupe-syntaxique>
   </fonction-syntaxique>
 </participant>,
 <lexie-att>abandonner</lexie-att>
 <participant type="Act" role="Patient">
  <fonction-syntaxique nom="Objet">
    <groupe-syntaxique nom="SN">l'
      <realisation>installation</realisation>
    </groupe-syntaxique>
   </fonction-syntaxique>
 </participant>
```

Figure 3 : Annotation XML (respectant les schémas de la figure 2) du contexte: Vous pouvez aussi, à ce moment-là, abandonner l'installation. La balise lexie-att non définie dans cet article, entoure l'occurrence de la lexie sous étude dans ce contexte.

3 Annotation des contextes dans deux domaines spécialisés

La méthodologie d'annotation manuelle des contextes comprend les étapes suivantes :

1. Constitution d'un corpus : la construction et la gestion du corpus suivent les recommandations établies au laboratoire du groupe ÉCLECTIK (Marshman 2003).
2. Extraction automatique de candidats termes simples : les candidats termes simples (nom, adjectif, verbe, adverbe) sont extraient par TermoStat (Drouin 2003). Ce logiciel d'acquisition automatique de termes s'appuie sur une approche contrastive, c'est-à-dire qu'il exploite une méthode de mise en opposition de corpus spécialisés et non spécialisés. Dans le présent travail, nous avons opposé un corpus de référence composé de textes journalistiques (Le Monde, année 2002) à un corpus d'informatique, d'une part, et à un corpus de changement climatique, d'autre part.

3. Analyse de la liste de candidats-termes : les listes de candidats générées par l'extracteur sont ensuite étudiées par des terminologiques qui ne retiennent que les unités étant de nature terminologique. Cette analyse nécessite l'application de critères de validation de la part des terminologues. Puisque nous annotons actuellement des termes de nature verbale, nous avons choisi des verbes extraits par TermoStat, puis validés par des analystes.
4. Choix de contextes dans lesquels apparaissent les termes : pour chaque terme sélectionné, nous prélevons entre 15 et 20 contextes. Pour préserver la représentativité du corpus, nous évitons de prendre trop de contextes figurant dans un même texte. De plus, à moins que les phrases ne soient vraiment trop longues, nous les prenons au complet afin d'inclure le maximum de participants. Il arrive que dans certaines phrases, un participant soit exprimé sous une forme anaphorique. Dans ces cas-là, nous prélevons également la phrase qui fait référence à ce participant (généralement la phrase qui précède). De cette façon, au moment de l'annotation, il est possible de relier l'anaphore au mot qu'elle renvoie.
5. Annotation dans le Schéma XML décrit à la section 2.3 : l'annotation est réalisée en deux étapes (première annotation et révision par une personne différente).

Actuellement, nous annotons des contextes tirés de deux corpus spécialisés distincts, à savoir un corpus d'informatique et un corpus portant sur le changement climatique. Les deux domaines abordant des thématiques assez différentes (concepts techniques dans le corpus d'informatique et concepts scientifiques parfois abordés sous un angle social dans le corpus sur le changement climatique), il apparaît intéressant de mettre au point le modèle d'annotation en les comparant l'un à l'autre.

3.1 Annotation des contextes : informatique

Les premières annotations manuelles ont été réalisées pour des verbes appartenant au domaine de l'informatique (ex. *abandonner, cliquer, configurer*, *joindre, télécharger*). Les verbes étaient déjà répertoriés dans une base de données terminologiques décrivant leur structure actancielle (la base de données porte le nom de DiCoInfo, Dictionnaire fondamental de l'informatique et de l'Internet[5]). Les annotateurs pouvaient donc se reporter à ces descriptions afin de distinguer les actants des circonstants dans les contextes réels.

Les contextes (rappelons que de 15 à 20 contextes sont retenus pour chaque unité lexicale annotée) sont extraits d'un corpus de textes rédigés en français d'environ 2 millions de mots. Les niveaux de spécialisation des textes varient, mais de manière générale, ces textes sont de nature didactique ou de vulgarisation et sont écrits, pour la plupart, par des spécialistes.

Les annotateurs disposaient d'une liste de rôles sémantiques préalablement définis dans la base de données pour étiqueter les actants (ex. **Agent**, **Patient**, **Instrument**, **Destination**, **Source**, **Assaillant**). Toutefois, puisque l'annotation des contextes prend également en compte les circonstants, nous avons dû définir de nouveaux rôles dont nous illustrons l'application ci-dessous :

But : *Pour lancer cette commande*
Environnement : *Sous Windows*
Temps : *à ce moment-là*
Voie : *via le bus de données*

Jusqu'à présent, nous avons annoté entre 15 et 20 contextes pour approximativement 122 unités lexicales. Ces contextes ont exigé la définition d'un total 32 étiquettes de rôles sémantiques.

3.2 Annotation des contextes : changement climatique

Pour les contextes du changement climatique, contrairement aux contextes de l'informatique, l'annotation précède la création des fiches, et ce principalement pour deux raisons. Premièrement, l'élaboration du dictionnaire de l'informatique est antérieure au projet d'annotation. Deuxièmement, débuter par l'annotation semble une démarche progressive et logique. En ce sens qu'elle permet aux terminologies de

[5] Le DiCoInfo est accessible à l'adresse suivante : http://olst.ling.umontreal.ca/dicoinfo/.

définir la structure actancielle des termes en se basant sur des données contextuelles. Les fiches sont par conséquent composées d'après les faits de langues observés en corpus – les contextes ne sont pas choisis en fonction des fiches.

Le corpus sur lequel nous travaillons compte approximativement 1 020 000 mots. Pour la sélection des textes, nous nous sommes appuyés principalement sur cinq critères : 1) les auteurs des textes sources sont des spécialistes ou du moins des personnes connaissant très bien le domaine; 2) les textes proviennent de sites Internet ou de publications reconnus dans le domaine; 3) les textes sont diversifiés au niveau de la spécialisation (technique, scientifique, vulgarisé, etc.); 4) les textes proviennent de différents types de documents (manuels, périodiques, sites Internet, etc.); 5) l'équilibre entre la taille et le nombre de documents est respecté.

Parmi les verbes que nous annotons, certains ont la même forme que les unités lexicales utilisées en informatique, mais la plupart ont des acceptions distinctes : *affecter*, *analyser*, *calculer*, *emmagasiner*, *générer*, *modifier*. Ce recoupement, permettra de vérifier dans quelle mesure le domaine influence les propriétés linguistiques de ces formes.

Dès à présent, il nous est possible d'identifier des rôles sémantiques qui se manifestent avec plus de régularité dans le domaine du changement climatique que dans celui de l'informatique. Au nombre des rôles associés aux actants, **Cause** et **Destination** sont particulièrement fréquents (figures 1 et 2.2). Par ailleurs, la liste des rôles associés aux circonstants à, pour le moment, été augmentée par des rôles à caractère spatial, quantitatif et temporel :

Direction : *vers le nord*
Coût : *à un coût se situant entre 50 et 180 dollars É.-U. par véhicule*
Valeur : *de 50 à 70 %*
Durée : *durant une courte période hivernale*

4 Annotation automatique des contextes : premières étapes

La tâche d'annotation manuelle décrite à la section 3 est très fastidieuse : l'annotation des contextes d'une unité lexicale exige 2 heures en moyenne. Notre corpus pour l'instant est constitué de 105 unités lexicales et nous avons annoté manuellement environ 3321 contextes.[6] L'annotation de ces 3321 contextes a nécessité environ 4 mois de travail. Pour accélérer l'annotation, nous proposons une méthode automatique d'apprentissage des traits des actants en nous basant sur un système de classification entraîné sur notre corpus de données annotées manuellement avec les rôles sémantiques tels qu'**Agent**, **Patient**, **Destination**, **Instrument**, etc.

La tâche consiste à annoter les participants des unités lexicales en rôles sémantiques. Ces participants sont de deux types : actants et circonstants. L'annotation en rôles sémantiques portera d'abord sur les actants, car la distinction entre les circonstants est plus complexe. Dans notre travail, trois tâches sont considérées : 1) identification des participants; 2) distinction entre les actants et les circonstants; 3) assignation de rôles sémantiques aux actants. Nous utilisons les liens syntaxiques d'unité lexicale avec les autres mots de la phrase comme traits caractéristiques ou « features » pour réaliser ces tâches principales. Les liens syntaxiques entre les unités de la phrase sont identifiés à l'aide de l'analyseur syntaxique *Syntex* (Bourigault *et al.* 2005).

On interprète la combinaison de l'analyse produite par *Syntex* et les liens entre l'unité lexicale et ses participants identifiés lors de l'analyse manuelle au moyen d'un graphe (voir figure 4).

[6] La méthode d'annotation a été mise au point à partir des contextes relevés pour des verbes d'informatique et non ceux du corpus de réchauffement climatique. Le nombre total d'unités lexicales retenues est moins élevé que le nombre total d'unités annotées puisque le travail d'annotation manuelle s'est poursuivi en parallèle.

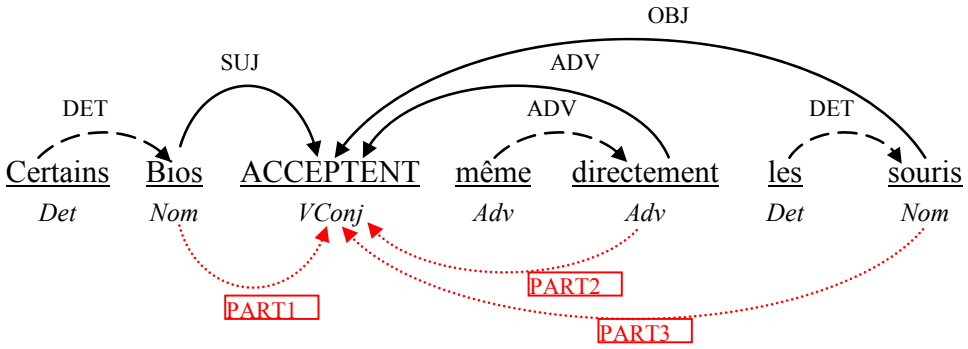

Figure 4 : Graphe de combinaison et règles déduites. Les flèches au-dessus des mots correspondent aux liens donnés par Syntex. Nous indiquons en pointillé les flèches qui ne seront pas considérées. Les flèches sous les mots correspondent aux participants annotés manuellement.

À partir d'un tel graphe, on extrait des règles d'identification de participants et leurs « features ». Ces règles affecteront le type Actant ou Circonstant à ces participants. Ces features sont aussi utilisés pour trouver les rôles sémantiques des participants identifiés. Sous le graphe, on indique une règle déduite par l'observation manuelle de plusieurs situations semblables.

Une règle a deux parties : la partie gauche, qui constitue les conditions de l'application, est composée des unités de la phrase avec les liens syntaxiques qui se trouvent au-dessus des unités et la partie droite qui représente les résultats de l'application où les participants sont indiqués sous les unités.

Dans une règle, une unité est décrite par ses traits. Un trait est écrit <*mot, Cat, Fonc*>, où *Cat* correspond à la catégorie grammaticale de l'unité et *Fonc* est sa fonction syntaxique. Dans une règle, une unité peut appartenir à plusieurs catégories grammaticales; dans ce cas, *Cat* est noté comme une alternative de plusieurs étiquettes syntaxiques, chacune d'elles séparées par un «|». Par exemple, le trait du mot de catégorie *Nom* ou *Pronom* est écrit <*mot, Nom|Pro, Role*>.

La figure 5 montre une partie des résultats obtenus lors de l'identification des participants au moyen de la stratégie décrite ci-dessus.

Lorsque la chaîne atteint la longueur maxi on recherche le premier espace en partant de la fin la ***partie*** gauche est imprimée puis ANNULÉE

Pour ABANDONNER le ***processus*** d installation à ce ***stade*** redémarrez l ordinateur et retirez la disquette d amorçage ou le CD ROM

ABANDONNEZ la ***copie*** ou la ***mise à jour***

Pour ce faire le ***système*** est capable de FERMER un ***environnement*** Classic devenu instable ou inopérant

Si l on veut protéger les ***informations*** ENREGISTRÉES sur une ***disquette*** de13 5 cm il suffit d obstruer cette encoche avec un morceau de papier adhésif

Dans les débuts de l informatique les ***programmeurs*** ÉCRIVAIENT des ***programmes*** ***en codant directement en langage machine***

Les ***disques*** à GRAVER sont conditionnés en plastique dur

La requête actuellement affichée a été modifiée et ***vous*** tentez de SORTIR sans la sauver

Certains ***BIOS*** ACCEPTENT même ***directement*** les ***souris***

Figure 5 Participants des unités lexicales identifiés automatiquement. Les unités lexicales sont en majuscules. Les mots en gras italique et soulignés sont leurs participants (Actants et Circonstants)

À partir d'un certain nombre d'exemples, nous avons dégagé une quarantaine de patrons d'identification de participants et nous les avons appliqués à l'ensemble du corpus (en dehors des

exemples que nous avions étudiés pour le développement de nos règles) soit 105 lexies sur 3321 contextes. En comparant les participants identifiés automatiquement et ceux qui avaient été identifiés manuellement, nous avons obtenu une précision de 75 % (nombre de participants pertinents retrouvés par rapport au nombre de participants total) et un rappel de 80 % (nombre de participants pertinents retrouvés par rapport au nombre de participants pertinents). Ces premiers résultats nous montrent que nous sommes sur la bonne voie et qu'il est possible d'accélérer le processus d'identification des actants au moyen de règles automatiques. Ces annotations pourraient servir de point de départ, quitte à ce qu'elles soient révisées par la suite.

Pour l'attribution des rôles sémantiques aux actants, certaines fonctions syntaxiques retournées par l'analyseur *Syntex* nous permettent d'affecter avec un bon niveau de confiance des rôles sémantiques aux actants de l'unité lexicale. Par exemple, les fonctions syntaxiques « Sujet » ou « Objet » nous permettent d'attribuer les rôles **Agent** ou **Patient** aux actants correspondants. Alors que pour les actants ayant ces fonctions, on arrive à leur affecter des rôles sémantiques, ce n'est pas toujours le cas pour d'autres fonctions. Par exemple, les actants prépositionnels occupant une fonction de Complément peuvent être identifiés à l'aide des liens syntaxiques mais l'attribution des rôles sémantiques ne pourra s'appuyer uniquement sur le type de fonction syntaxique renvoyé par Syntex. Dans ce cas, on utilise les mêmes features (dépendances syntaxiques et position du participant indiquant s'il apparaît avant ou après l'unité lexicale dans la phrase) de l'étape d'identification de participants auxquels on rajoute d'autres features tels le feature tête comme les prépositions *à*, *dans*, *sur*, *grâce à*, *avec* pour les prépositionnels ou le feature Lexie (unité lexicale) car le rôle d'un actant peut aussi dépendre de l'unité lexicale avec laquelle il est utilisé.

Dans une étape ultérieure, nous comptons utiliser des méthodes d'apprentissage machine pour nos traitements. Nous utiliserons des classificateurs automatiques qui s'appuieront sur les features cités ci-dessus. À terme, nous expérimenterons avec deux approches d'apprentissage : statistique (Gildea & Jurafsky 2002) et *instance-based learning* (Li *et al.* 2001).

5 Conclusion et perspectives

Dans cet article, nous avons voulu rendre compte d'un projet en cours visant à enrichir des bases de données terminologiques en y introduisant, plus spécifiquement, des renseignements de nature syntaxico-sémantique. Ce projet comporte un travail manuel important, mais nous espérons le réduire de manière considérable au moyen d'une méthode qui permettra de générer des annotations automatiquement, annotations qui seront par la suite révisées par des annotateurs humains.

Les résultats obtenus jusqu'à présent quant à l'identification automatique des participants sont prometteurs. Nous espérons pousser plus avant l'automatisation en procédant à l'attribution automatique de certains rôles, notamment ceux associés aux actants. Ce travail nécessite une définition rigoureuse des rôles décrits jusqu'à maintenant et une meilleure compréhension des structures syntaxiques dans lesquelles ils peuvent se retrouver. Naturellement, la démarche n'est pas linéaire, rien n'empêche une fois la rédaction des fiches commencée d'apporter des modifications aux contextes annotés. Autrement dit, le travail d'annotation automatique repose sur une analyse manuelle importante; de même, le traitement automatique des contextes – en repérant certaines erreurs – permet de systématiser des paramètres de l'annotation manuelle.

Remerciements

Les travaux de recherche présentés dans le présent article sont financés par le Conseil canadien en sciences humaines (CRSH) du Canada et par les Fonds québécois de recherche sur la société et la culture (FQRSC). Les auteurs aimeraient remercier Stéphanie Caron, Stéphanie Klebetsanis et Charlotte Tellier qui ont participé au travail d'annotation manuelle des contextes d'informatique.

Références

Binon, Jean, Serge Verlinde, Jan Van Dyck & Ann Bertels. 2000. *Dictionnaire d'apprentissage du français des affaires*. Paris: Didier.

Bourigault, Didier, Cécile Fabre, Cécile Frérot, Marie-Paule Jacques & Sylvia Ozsdowska. 2005. Syntex, analyseur syntaxique de corpus. In *Actes des 12èmes journées sur le Traitement Automatique des Langues Naturelles.* Dourdan, France.

DiCoInfo. Dictionnaire fondamental de l'informatique et de l'Internet. (http://olst.ling.umontreal.ca/dicoinfo/). Accessed 5 February 2009.

Dolbey, Andrew, Michael Ellsworth & Jan Scheffczyk. 2006. BioFrameNet: A Domain-Specific FrameNet Extension with Links to Biomedical Ontologies. In Bodenreider, Olivier (ed.). *Proceedings of KR-MED*, 87-94.

Drouin, Patrick. 2003. Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1): 99-117.

Faber, Pamela, Silvia Montero Martínez, María Rosa Castro Prieto, José Senso Ruiz, Juan Antonio Prieto Velasco, Pilar León Arauz, Carlos Márquez Linares & Miguel Vega Expósito. 2006. Process-oriented terminology management in the domain of Coastal Engineering. *Terminology*, 12(2): 189-213.

Fillmore, Charles J. 1968. The case for case. In Bach, Emmon & Robert T. Harns (eds.). *Universals in linguistic Theory*, New York: Holt, Rinehard & Winston, 1-88.

Fillmore, Charles J., Christopher R. Johnson & Miriam R.L. Petruck. 2003a. Background to FrameNet, In Fontenelle, Thierry (ed.). FrameNet and Frame Semantics. Special Issue of the *International Journal of Lexicography*, 16(3): 235-250.

Fillmore, Charles J., Miriam R.L. Petruck, Joseph Ruppenhofer & Abby Wright. 2003. FrameNet in Action: The case of attaching. *International Journal of Lexicography*, 16(3): 297-332.

FrameNet (http://framenet.icsi.berkeley.edu/). Accessed 6 February 2008.

Gildea, Daniel & Daniel Jurafsky. 2002. Automatic labelling of semantic roles. *Computational Linguistics,* 28(3): 245-288.

Lépinette, Brigitte & Françoise Olmo Cazevieille. 2007. Un dictionnaire syntaxico-sémantique bilingue de la zootechnie. *Cahier de linguistique*, 33(1): 83-102.

Li Jiamming, Lei Zhang & Yong Yu. 2001. Learning to Generate Semantic Annotation for Domain Specific Sentences. In *Workshop on Knowledge Markup and Semantic Annotation at the 1st International Conference on Knowledge Capture* (K-CAP 2001), October, Victoria, B.C., Canada

Marshman, Elizabeth. 2003. Construction et gestion des corpus : Résumé et essai d'uniformisation du processus pour la terminologie. (http://www.olst.umontreal.ca/pdf/terminotique/corpusentermino.pdf). Accessed 8 February 2009.

Mel'čuk, Igor. 2004. Actants in semantics and syntax. I: actants in semantics. *Linguistics*, 42(1): 1-66.

Ruppenhofer, Joseph, Michael Ellsworth, Miriam R.L. Petruck, Christopher Johnson & Jan Scheffczyk. 2002. *FrameNet II: Extended Theory and Practice*. (http://framenet.icsi.berkeley.edu/). Accessed 6 February 2008.

Schmidt, Thomas. forthcoming. *The Kicktionary – A Multilingual Lexical Resources of Football Language.*

VerbNet (http://verbs.colorado.edu/~mpalmer/projects/verbnet.html). Accessed 7 February 2008.