Identification des actants et circonstants par apprentissage machine

Fadila Hadouche[1] Guy Lapalme[1] Marie-Claude L'Homme[2]

(1) RALI, (2) OLST,
Université de Montréal,
C.P. 6128, succursale Centre-ville,
Montréal, Québec,
Canada H3C 3J7
hadouchf@iro.umontreal.ca, lapalme@iro.umontreal, mc.lhomme@umontreal.ca

Résumé

Dans cet article, nous traitons de l'identification automatique des participants actants et circonstants de lexies prédicatives verbales tirées d'un corpus spécialisé en langue française. Les actants contribuent à la réalisation du sens de la lexie alors que les circonstants sont optionnels : ils ajoutent une information supplémentaire qui ne fait pas partie intégrante du sémantisme de la lexie. Nous proposons une classification de ces participants par apprentissage machine basée sur un corpus de lexies verbales du domaine de l'informatique, lexies qui ont été annotées manuellement avec des rôles sémantiques. Nous présentons des features qui nous permettent d'identifier les participants et de distinguer les actants des circonstants.

Abstract

In this paper we discuss the identification of participants actants and circumstants of specialized verbal lexical units in a French specialised corpus. The actants are linguistic units that contribute to the sense of the verbal lexical unit while circumstants are optional: they add extra information that is not part of the meaning of the verbal unit. In this work, we propose a classification of participants using machine learning based on a specialized corpus of verbal lexical items in the field of computing which are annotated manually with semantic roles labels. We defined features to identify participants and distinguish actants from circumstants.

Mots-clés : Structure actancielle, actants et circonstants, features de classification
Keywords: Actantial structure, actants and circumstants, classification features

1 Introduction

Identifier la frontière entre les actants et les circonstants reste une tâche difficile en linguistique et présente des défis d'automatisation encore plus élevés. Entre autres difficultés, nous pouvons évoquer le fait que certains actants et circonstants peuvent occuper les mêmes positions syntaxiques et il arrive que les actants soient omis dans les phrases[1]. Dans la phrase suivante, Vous voulez simplement AFFECTER une valeur à une variable que vous avez déclarée: les actants sont vous, valeur et variable que vous avez déclarée ; simplement est un circonstant.

Les actants permettent de définir le sens de la lexie verbale, contrairement aux circonstants qui ne contribuent pas au sens de la lexie. Par exemple, la lexie ABANDONNER signifiant « cesser une activité sans la finir » est définie en faisant appel aux actants agent (celui à l'origine de l'action) et patient (subissant l'action exprimée par le verbe). L'unité linguistique qui joue le rôle d'agent ou le rôle de patient est un actant. On voit bien qu'avec ces actants le sens de la lexie est rempli et complet. Les autres participants de cette même lexie dans d'autres contextes qui ne jouent pas les rôles d'agent ou de patient seront considérés comme des circonstants. Dans l'exemple le programmeur ABANDONNE l'opération à ce stade, programmeur et opération sont des actants (respectivement agent et patient) et stade est un circonstant de temps.

Notre travail consiste précisément à distinguer ces deux formes de participants. Nous avons divisé la tâche en deux : identification des participants de la lexie verbale et leur classification en actant et circonstant. Dans notre travail, aucune ressource lexicale de sous-catégorisation des verbes ou dictionnaire n'est utilisé pour identifier les participants actants et circonstants des verbes. Nous nous basons sur un corpus français de lexies verbales annotées manuellement avec des rôles sémantiques. Nous nous sommes inspirés des travaux basés sur l'extraction des features pour la classification des participants. Nous avons proposé des features basés sur deux approches : 1) les relations de dépendance trouvées par un analyseur syntaxique, en considérant les mots en dépendance avec la lexie comme participants candidats; 2) les catégories grammaticales des mots de la phrase où apparaît la lexie verbale susceptibles d'être des participants de cette lexie. Aucune de ces deux stratégies ne permettant d'effectuer une distinction parfaite entre les deux types de participants, nous souhaitons évaluer jusqu'à quel point ces stratégies peuvent y contribuer. Nous cherchons à proposer à des annotateurs humains des phrases pré-annotées afin d'accélérer l'annotation manuelle.

2 Travaux connexes

Plusieurs critères d'identification des actants ont été suggérés en linguistique théorique [1] mais certains sont extrêmement difficiles à automatiser. Le critère obligatoire pour les actants et optionnel pour les circonstants n'est pas toujours facile à appliquer lors de l'analyse de phrases réelles. Un autre critère, celui de la mobilité de position, veut que seuls les circonstants peuvent changer de position dans la phrase. Or, on trouve des cas où les actants peuvent être déplacés, comme le remarque Anne Lacheret-Dujour [2] dans son exemple « le chocolat, j'adore ». Par ailleurs, les sujets et compléments d'objet direct correspondent généralement à des actants, tandis que certains compléments introduits par des prépositions peuvent avoir les deux rôles. Enfin, la fréquence peut jouer un rôle important dans la distinction entre actants et circonstants [3]. Cécile Fabre et Cécile Frérot proposent une combinaison de deux mesures de productivité[2] pour distinguer sur corpus, au sein des groupes prépositionnels rattachés au verbe, des types de compléments différents [4].

---

[1] Certains auteurs désignent les actants par *arguments* et les circonstants par *adjoints.*

[2] Productivité recteur-préposition : nombre de régis différents que le couple (verbe, préposition) gouverne. Productivité préposition-régi : nombre de verbes différents qui gouvernent le couple (préposition, régi)

Des ressources telles que Framenet[3], Verbnet[4] ou Propbank[5] sont des ressources qui fournissent une annotation des verbes et de leurs actants (et, dans certaines d'entre elles, de leur circonstants) en étiquetant ces derniers au moyen de rôles sémantiques[6]. Ces ressources ont été construites manuellement et ont servi de point de départ pour la mise au point de méthodes d'annotation automatiques des prédicats (dont certaines s'appuient sur les informations syntaxiques pour identifier la structure argumentale d'un prédicat verbal). Dans ce cadre, Gildea et Jurafsky ont proposé l'extraction de features[7] syntaxiques qui permettent d'identifier les participants d'un prédicat [5]. Richard Johansson et Pierre Nugues ont proposé d'utiliser les relations de dépendance syntaxique comme un feature [6]. D'autres travaux se basent sur des informations syntaxiques : Surdeanu [7] a utilisé le corpus TreeBank et PropBank ainsi que Li [8] , Che [9], Täckström [10] et Maria Liakata et al. [11]. Ces travaux proposent des modèles pour identifier les participants d'un prédicat verbal mais ne distinguent pas les actants des circonstants.

D'autres propositions, qui s'appuient sur les schémas de sous-catégorisation des verbes, utilisent des informations sur la transitivité du verbe. Nasr et Béchet proposent un analyseur syntaxique axé sur la détection des cadres valenciels[8] tels que recensés par Dicovalence [12].

3 Les participants et l'analyseur syntaxique

Dans notre travail, nous utilisons l'analyseur syntaxique Syntex [13] qui effectue une analyse en dépendance pour déterminer les liens entre les mots de la phrase. La Figure 1 illustre le parallèle entre l'analyse syntaxique faite par Syntex (liens au-dessus de la phrase) et l'annotation manuelle des participants de la lexie verbale (liens sous la phrase). La lexie ACCEPTER est liée syntaxiquement ou gouverne le mot cartouches. Ce dernier est lié à la lexie par le lien fonctionnel objet et il est considéré comme un participant de cette lexie. Dans d'autres cas, la lexie est elle même gouvernée par un mot de la phrase.

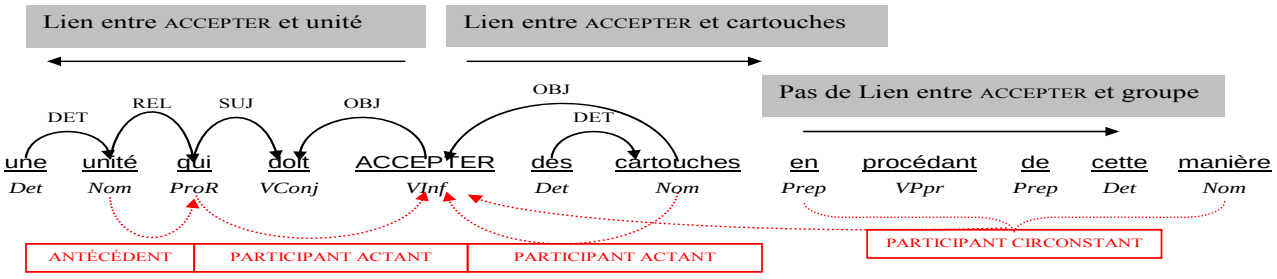

Figure 1 : Exemple de contexte de la lexie verbale ACCÉPTER et ses participants

Dans la Figure 1, la lexie ACCEPTER est gouvernée par le verbe devoir (doit dans la phrase) par le lien fonctionnel objet. Dans ce cas, nous considérons que le sujet de doit, réalisé ici sous la forme d'un pronom relatif, est un participant de la lexie ACCEPTER. Toutefois, la règle consistant à considérer comme participants tous les mots gouvernés par un gouvernant de la lexie n'est pas toujours applicable. Dans l'exemple cancel demande au serveur d'ABANDONNER le traitement, le sujet cancel du verbe demander n'est pas un participant de la lexie ABANDONNER. Suivre uniquement les liens

[3] http://framenet.icsi.berkeley.edu/
[4] http://verbs.colorado.edu/~mpalmer/projects/verbnet.html
[5] http://verbs.colorado.edu/~mpalmer/projects/ace.html
[6] À noter que la ressource FrameNet préfère l'appellation *Frame elements* (FE) (qui comprennent les FE obligatoires – core - et optionnels – non core) à celle de *rôle sémantique*. Cette préférence terminologique reflète également une distinction théorique importante.
[7] Les features définis par Gildea sont : prédicat, catégorie grammaticale du prédicat et du mot candidat, tête, arbre-syntaxique, voie passive ou active, position du mot candidat par rapport au prédicat.
[8] Un cadre valenciel d'un verbe décrit ses compléments. Les informations données sont le nombre de compléments avec leurs fonctions, que nécessite un verbe dans sa réalisation.

donnés par Syntex sans critères supplémentaires n'est pas suffisant. Il arrive également que Syntex omette certains liens entre des éléments de la phrase et la lexie. Syntex ne détecte aucun lien entre la lexie et « en procédant de cette manière » (Figure 1). Ce dernier est considéré comme le participant circonstant qui indique le mode. Dans notre corpus, 975 éléments considérés comme des participants (environ 22 %) n'ont pas été détectés par Syntex comme dépendant syntaxiquement de la lexie à l'étude.

Dans [14], nous avons testé une approche à base de règles en interprétant ces liens syntaxiques. Les résultats montrent 69 % de détermination correcte des participants. L'inconvénient de cette méthode à base de règles est l'impossibilité de prévoir toutes les règles. Nous proposons l'utilisation de l'apprentissage machine pour une classification supervisée basée sur ces informations syntaxiques dont les features correspondants sont définis dans le Tableau 2.

4 Classification des participants par Weka[9]

Nous avons utilisé Weka qui fournit des implémentations des algorithmes d'apprentissage les plus connus que nous pouvons appliquer sur les contextes des lexies verbales de notre corpus, en utilisant des features construits à partir des informations syntaxiques pour retrouver les participants. Nous avons testé plusieurs de ses classificateurs pour définir celui qui offre le meilleur résultat. Nous avons divisé notre travail en deux tâches : 1) identifier les participants de la lexie verbale ; et 2) distinguer les actants des circonstants. Cette division nous aidera à filtrer les mots candidats en tant qu'actants ou circonstants et nous évitera de traiter les nombreux mots qui ne sont pas participants et qui bruitent la performance de la classification. Deux expérimentations pour chaque tâche sont réalisées. Une première utilise les relations de dépendance syntaxique entre les mots de la phrase et la lexie verbale ; la seconde fait appel aux catégories grammaticales des mots de la phrase. Nous notons que ce sont les mots simples des phrases où apparaît la lexie qui sont des candidats à être des participants et non pas des syntagmes. Chaque mot candidat est décrit par des features dans le but de le classifier comme participant (OUI) ou non-participant (NON). Le Tableau 1 montre cette description correspondant à certains mots de la phrase de la Figure 1

Feature / Participants	Lexie	Cat-Lexie	Mot	Cat-Mot	Position	Dist-ance	Lien-syntaxique-cg	Lien-syntaxique-fn	Part
unité	ACCEPTER	Vinf	CN	Nom	avant	2	doit, ProRel	OBJ, SUJ, REL	OUI
qui	ACCEPTER	VInf	qui	ProRel	avant	1	doit	OBJ, SUJ	OUI
doit	ACCEPTER	VInf	doit	VConj	avant	0	nul	OBJ	NON
cartouches	ACCEPTER	VInf	CN	Nom	après	1	Nul	OBJ	OUI

Tableau 1 : Participants de l'exemple de la Figure 1 décrits par les features : catégorie grammaticale de la lexie, la classe du mot en traitement (à classifier), sa catégorie grammaticale, sa position et sa distance par rapport à la lexie, les catégories et les fonctions grammaticales des mots rencontrés sur les liens allant de la lexie au mot. Ces features sont décrits au Tableau 2.

4.1 Corpus

Nous avons utilisé un corpus d'informatique et d'Internet contenant une centaine de lexies verbales spécialisées de langue française (ex. *accéder, configurer, télécharger*). La structure actancielle de ces lexies est annotée manuellement au moyen d'étiquettes de rôles sémantiques (ex. agent, patient, instrument) [15]. Nous avons utilisé 2/3 du corpus pour l'entraînement des classificateurs et le 1/3

[9] Weka est développé à l'Université de Waikato en Nouvelle-Zélande. www.cs.waikato.ac.nz/ml/weka

restant pour le test. Le corpus d'entraînement est composé de 72 lexies différentes et de 1512 contextes; et le corpus de test, est composé de 36 lexies et de 756 contextes. Le corpus global est composé d'environ 3464 actants et de 1151 circonstants. Selon la terminologie de Weka, nous avons 14860 instances pour l'entrainement et 6526 instances pour le test. Chaque instance correspond à un mot pouvant être un participant.

5 Identification des participants

Bien que d'autres lexies soient de nature prédicative, nous ne considérons que les verbes dans les expérimentations décrites ci-dessous : les verbes sont plus susceptibles que les noms d'être accompagnés par des actants réalisés dans les phrases. Les participants d'une lexie verbale annotée manuellement dans le corpus prennent la forme : 1. d'un groupe nominal dont la réalisation est un nom ou pronom, 2.d'un groupe prépositionnel dont la réalisation est un nom ; 3. d'un groupe adverbial dont la réalisation est un adverbe. Nous répartissons les mots dans les catégories classiques de « mots pleins » et « mots vides » [16]. Dans notre travail, nous considérons comme participants candidats aux lexies verbales les groupes nominaux, les adverbes et les pronoms relatifs. Nous traitons les pronoms relatifs comme des participants dans notre corpus annoté manuellement et indiquons une référence à un groupe nominal annoté *antécédent*. C'est le cas d'une unité de la phrase de la Figure 1. Pour réaliser cette tâche d'identification, nous avons extrait des features décrits dans Tableau 2 en nous inspirant de ceux de base définis par Gildea & Jurafsky.

Nom Features	Signification
Lexie	Unité lexicale verbale en étude (c.-à-d. : **ACCÉDER**, **IMPRIMER**, etc.)
Catlexie	Catégorie grammaticale de **Lexie** (c.-à-d. : VInf, VConj, VPpa, etc.)
Mot	Unité lexicale de la phrase en liaison avec la lexie par des liens syntaxiques de Syntex.
Catmot	Catégorie grammaticale du **Mot** (c.-à-d. Nom, Adverbe, Pronom, etc.)
Position	Position du **Mot** par rapport à **Lexie**. sa valeur est « avant » si **Mot** apparaît avant **Lexie** et elle est « après » si **Mot** est après **Lexie**.
Distance	Le nombre de mots qui séparent **Lexie** du **Mot** en empruntant les liens syntaxiques qui existent entre eux

Tableau 2 : Features de classification

Le feature **Mot** du Tableau 2 peut prendre soit des valeurs réelles comme le verbe `permettre` on doit prendre la valeur `permettre` comme contenu du mot. Dans d'autres cas, nous pouvons prendre la classe correspondant à la catégorie grammaticale. Ces deux cas sont donnés selon les critères suivants :

- Si le mot est un nom, peu importe sa valeur, sa classification reste la même. Nous proposons dans ce cas la classe des noms notée CN comme valeur du feature **Mot**.
- Si le mot est pronom personnel dont la catégorie grammaticale est notée Pro, nous prenons sa valeur réelle, car les pronoms tels que `je`, `tu`, `il`, etc., peuvent être des participants réalisés sous forme de sujet du verbe, qui diffèrent des pronoms `le`, `les`, etc., qui peuvent être des participants réalisés sous forme d'objet direct. En outre, d'autres pronoms tels que `se` ne sont pas toujours annotés comme participants.
- Si le mot est un pronom relatif nous prenons sa valeur réelle, sachant que le pronom relatif `qui` pointe sur un participant antécédent de catégorie grammaticale Nom de fonction SUJET et le pronom relatif `que` pointe sur un participant antécédent de catégorie grammaticale Nom de fonction OBJET.
- Nous avons constaté dans notre corpus que certains adverbes (`simultanément`, `directement`, etc.) sont annotés par les linguistes comme des participants, mais d'autres ne le sont pas (`puis`, `ne pas`, `même`, etc.). Nous avons pris toutes les valeurs des adverbes

rencontrés dans notre corpus et si un adverbe n'a jamais été rencontré, nous avons défini une classe inconnue notée INCADV.

Dans notre étude, nous avons proposé d'autres features à ajouter à ceux du Tableau 2 qui diffèrent selon leur utilisation des relations de dépendance syntaxique entre la lexie et les autres mots de la phrase.

5.1 Features de classification basés sur les relations de dépendance de Syntex

Nous suggérons de prendre en compte les relations de dépendance syntaxique entre la lexie verbale et les autres mots de la phrase. Nous nous sommes basés sur l'analyseur syntaxique Syntex à partir duquel nous avons extrait ces relations de dépendance. Ces dernières constituent un autre critère de classification. Deux autres features définis dans le Tableau 3 sont ajoutés à ceux du Tableau 2.

Nom Features	Signification
Lien-syntaxique-cg	Chemin de **Lexie** à **Mot**. Ce chemin est un ensemble de liens syntaxiques trouvé par Syntex allant de **Lexie** jusqu'à **Mot**. Il est une combinaison de toutes les catégories grammaticales des mots qui se retrouvent sur ce chemin (ie : Nom, Prep, Adv, etc.).
Lien-syntaxique-fn	Dans ce cas le chemin de liens syntaxiques entre **Lexie** et **Mot** est une combinaison de toutes les fonctions syntaxiques de ces liens (SUJ, OBJ, etc.)

Tableau 3 : Features de la classification en se basant sur Syntex

Les valeurs possibles du feature Lien-syntaxique-cg dépendent de certains critères :

- Si la catégorie grammaticale du mot se trouvant sur le chemin des liens entre la lexie et le mot est un verbe, alors nous proposons de prendre pour certains verbes leurs valeurs réelles. Ce sont les verbes qu'on appelle *verbes modaux* dont nous devons tenir compte. Nous avons défini une liste de ces verbes, rencontrés aussi dans notre corpus (faire, aller, vouloir, pouvoir, permettre, demander, devoir, etc.). Ces verbes n'identifient pas les mêmes participants et ils attribuent des statuts différents, soit actant ou circonstant, aux participants. Par exemple, X tente de Lexie, X est considéré comme actant. Par contre dans X permet à Y de Lexie, X est un circonstant. Dans le cas, X demande à Y de Lexie, X n'est pas un participant. Pour d'autres verbes une classe des verbes CV est définie.
- Si la catégorie du mot se trouvant sur le chemin est une préposition, nous prenons sa valeur et non pas sa catégorie. La préposition à et de n'identifie pas les mêmes participants.

5.2 Features de classification sans l'analyseur Syntex

Dans l'approche précédente faisant appel à Syntex, nous remarquons que les noms qui sont participants de la lexie verbale peuvent être directs (ayant un lien syntaxique avéré dans la phrase) ou indirects (apparaissant dans la phrase, mais n'ayant pas de lien syntaxique avec la lexie verbale). Ils peuvent être régis par une préposition, une relative ou une coordination. Ils peuvent aussi être liés à un autre verbe précédant la lexie verbale annotée régie par une préposition. Étant donnée que, dans cette deuxième approche, nous n'utilisons pas les liens de dépendance syntaxiques, nous proposons des features qui permettent de restituer ces informations en testant la nature des mots qui apparaissent entre la lexie et le mot candidat. Nous proposons de considérer en plus des features du Tableau 2, d'autres features tels que : 1) les **catégories grammaticales** des mots séparant le mot candidat de la lexie, 2) le **nombre de verbes** entre la lexie et le mot candidat s'ils existent, 3) la **valeur du verbe ou sa catégorie**, 4) la **distance**, 5) la **position,** 6) la **valeur de la préposition,** du **pronom relatif** et de la **coordination** s'ils existent entre la lexie et le mot candidat.

5.3 Résultats des classificateurs Weka et comparaison des deux cas

Nous avons testé plusieurs classificateurs de Weka (un classificateur de chaque classe d'algorithmes présenté par Weka : bayesian, trees, rules, lazy etc.). Dans la Figure 2, nous présentons les classificateurs avec un taux de classification élevé, supérieur à 78%. Le tableau de gauche de la Figure 2 montre le taux de classification sur les 8376 instances en utilisant l'analyseur syntaxique Syntex comparé à celui sans analyseur syntaxique : c'est la proportion de participants bien classifiés tant de la classe OUI que de la classe NON. Par exemple, avec Syntex le taux 87 % bien classés (13 % mal classés) indique que, sur les 3413 de la classe OUI, 541 participants sont mal classés dans la classe NON et, sur les 4963 de la classe NON, 489 participants sont mal classés dans la classe OUI. Ainsi le tableau de droite indique la F-measure de la classe OUI dans les deux cas d'utilisation (avec ou sans les dépendances fournis par l'analyseur syntaxique). Ces taux ont été calculés en comparant les résultats des classificateurs sur le corpus de test avec les annotations manuelles.

Taux de biens classifiés		
Classifier	avec Syntex	sans Syntex
IB5	87.39	83.56
RFTree	87.70	84.53
RTree	81.72	80.70
BFTree	86.01	82.50

F-mesure de la classe OUI		
Classifier	avec Syntex	sans Syntex
IB5	85.2	73.70
RFTree	84.8	73.00
RTree	78.6	69.60
BFTree	82.4	71.00

Figure 2 : Résultats des classificateurs. IB5 : classifieur IBk avec k=5 plus proches voisins qui utilise la distance euclidienne convertie en poids ; RFTree : Random Forest Tree qui consiste en un ensemble d'arbres de décision. Dans notre cas, nous avons utilisé un nombre de 10 arbres ; il fait comme la cross validation, il prend un et teste sur les 9 restant ainsi de suite ; RTree : RandomTree est sous forme d'une arborescence ; BFTree : Best First Decision Tree. Il choisit la racine de l'arbre et pour les branches, il divise l'ensemble d'entrainement en sous ensembles et choisit les meilleurs sous ensembles à mettre sous les nœuds. Ce processus est répété pour tous les nœuds. Ce classificateur a pris énormément de temps sur notre corpus que les autres.

Les résultats de la Figure 2 montrent que la F-mesure de la tâche d'identification des participants en utilisant uniquement les catégories grammaticales (sans Syntex), qui se situe autour de 80 %, ne propose pas d'aussi bons résultats que celle faisant appel à un analyseur en dépendance (avec Syntex), qui se situe autour de 93 %. Mais un taux de 80 % reste intéressant, et laisse supposer que les features que nous avons proposés pour cette approche sont prometteurs. Le feature **distance** que nous avons défini joue un rôle particulièrement important dans cette classification. Cette notion de distance entre le mot candidat et la lexie permet de désambiguïser des mots candidats qui peuvent avoir plus ou moins les mêmes caractéristiques les séparant de la lexie.

6 Distinction entre actant et circonstant

Syntex ne permet pas de distinguer les actants des circonstants. Il permet d'affecter des fonctions syntaxiques telles que sujet ou objet qui peuvent être le plus souvent des actants. Par contre le problème se pose pour des mots introduits par des prépositions. Syntex affecte dans tous les cas la fonction NOMPREP (par exemple On ABANDONNE l'opération à ce stade et on ACCÈDE à cette information). Dans ces deux exemples, Syntex annote stade et information par NOMPREP et aucune autre information n'est ajoutée. Pourtant, stade et information ne sont pas du même type : stade est considéré comme un circonstant de la lexie ABANDONNER et information est considérée comme un actant de la lexie ACCÉDER. Les participants identifiés à la section 5 sont soit des noms, des pronoms ou des adverbes. Pour attribuer le type actant ou circonstant à ces participants, nous considérons :

- Les noms ou les pronoms occupant les fonctions de sujet ou d'objet direct sont sélectionnés ; s'ils ne sont pas régis par d'autres éléments dans la phrase, alors on peut les considérer comme des actants.

- Selon notre corpus annoté manuellement, les adverbes sont généralement des circonstants.
- Les noms régis par des verbes modaux, voir section 5.1, sont considérés actants ou circonstants selon le verbe modal employé. Pour certains verbes comme devoir, pouvoir, vouloir, etc., ces noms sont considérés comme des actants. Par contre, pour d'autres verbes tels que permettre, servir, etc. ces noms sont des circonstants. Nous tenons donc compte d'une liste de verbes modaux pour décider de la classe de ces noms.
- Les noms régis par un pronom relatif qui ou que sont des actants à condition que ces pronoms relatifs ne soient pas régis par des prépositions ou des verbes modaux.
- Les noms régis par les prépositions (introduisant un complément du verbe) peuvent être des actants pour certaines lexies verbales, et des circonstants pour d'autres. Nous entendons par là que ces noms ne dépendent pas uniquement de leur position syntaxique mais qu'ils dépendent aussi de la lexie verbale. Dans ces cas, l'annotation exige l'accès à plus d'informations sur la lexie verbale. Puisqu'on n'utilise pas un dictionnaire nous informant des schémas de sous-catégorisation de cette lexie verbale, nous proposons de calculer la fréquence relative de ces noms, régis par une préposition, avec la lexie à partir de notre connaissance des contextes dans lesquels est employée cette lexie.

6.1 Calcul de la fréquence relative

La notion de la fréquence relative est utilisée par Messiant [17] dans la réalisation du lexique « LexSchem » présentant la sous-catégorisation des verbes français. Dans notre cas, La fréquence relative constitue le rapport entre le nombre de contextes où est apparue la lexie verbale avec les noms régis par une préposition et le nombre total de contextes où la lexie est apparue. Nous faisons l'hypothèse que cette fréquence traduira l'importance du participant et son rôle quant au sémantisme de la lexie verbale. Cette fréquence relative est donnée par la formule suivante

$$frequence - relative = \frac{\#(Lexie, feature)}{\#(lexie)}$$ [10]

Si cette fréquence est élevée, elle peut indiquer que la relation sémantique entre la lexie verbale et ce participant est très étroite, et donc que ce participant est vraisemblablement un actant. Mais si elle est faible, elle indique que la relation sémantique n'est pas étroite, donc ce participant est peut être un circonstant. Dans cette tâche aussi, nous avons testé les classificateurs Weka. Dans ce cas, nous avons ajouté le feature -fréquence relative- aux features définis dans les sections précédentes.

6.2 Les résultats de la classification en actant et en circonstant

La classification en actant et circonstant fonctionne mieux lorsque nous faisons appel à Syntex comme le montrent les résultats de la Figure 3 : une F-mesure de 96 % pour les actants et de 83 % pour les circonstants. Dans le cas où Syntex n'est pas utilisé, la classe actant est classifiée avec une F-mesure pour le meilleur classificateur de 94 %, ce qui est acceptable. Nous avons remarqué, que dans le corpus d'entraînement, les circonstants n'étaient pas nombreux représentant environ 1/3 des participants actants. La F-mesure de la classe des circonstants est autour de 79 % pour le meilleur classificateur. Le calcul de la fréquence relative pour distinguer les actants des circonstants donne de bons résultats. La différence de performance avec ou sans Syntex est due au fait que Syntex fournit la fonction grammaticale sujet et objet du participant.

[10] feature représente les noms régis par une préposition. Par exemple à une variable, son feature est « préposition, nom » qui correspond à « à, variable ». #(lexie,feature) est le nombre de fois que la lexie et ce feature sont apparus ensemble et #(lexie) est le nombre total de fois que la lexie est apparue (avec ou sans ce feature).

Taux de biens classifiés		
Classifier	avec Syntex	sans Syntex
IB5	92.99	88.17
RFTree	93.55	91.38
RTree	89.48	87.24
BFTree	91.97	89.80

F-mesure de la classe ACT		
Classifier	avec Syntex	sans Syntex
IB5	95.60	92.60
RFTree	96.00	94.60
RTree	93.50	91.90
BFTree	95.00	93.50

F-mesure de la classe CIRC		
Classifier	avec Syntex	sans Syntex
IB5	82.20	70.90
RFTree	83.00	79.00
RTree	71.70	69.80
BFTree	79.00	76.10

Figure 3 : Tableaux de résultats des classificateurs sur la classe actant et la classe circonstant. À gauche, nous avons le taux des biens classifiés de la classe actant ou de la classe circonstant. Ces classificateurs sont les mêmes que ceux utilisés à la Figure 2

Les résultats de l'identification des participants qui sont pour le meilleur classificateur de 87 % et de 84 % selon qu'on utilise Syntex ou non sont analogues à ceux obtenus dans les différents travaux sur l'anglais qui varient entre 80 % et 90 % utilisant une approche similaire de features sur le Penn Treebank, plus riche et plus grand en taille. Quant à la tâche de distinction entre actant et circonstant, une tâche propre à notre objectif, les résultats sont satisfaisants. Ils sont semblables aux 76 % trouvés par Fabre [4] qui utilise une autre approche sur le français pour distinguer les arguments des adjoints pour un autre objectif que le nôtre.

Conclusion

Nous avons présenté une approche d'identification de participants, actant et circonstant, en langue française, afin de pouvoir par la suite annoter ces actants par des rôles sémantiques. Cette tâche facilitera la distinction entre les participants obligatoires et optionnels quant à leur annotation par des rôles sémantiques. Cette approche d'utilisation de features, une première pour le français dans ce contexte, est inspirée de celle réalisée en anglais dans le cadre de FrameNet pour l'annotation de rôles sémantiques. Nos résultats montrent que les features que nous avons proposés sont appropriés à la tâche d'identification et de distinction d'actants et circonstants. On peut tirer avantage d'un analyseur syntaxique automatique tel que Syntex, dont les résultats sont meilleurs sur les 78 % unités lexicales qu'il détecte pouvant avoir une relation avec la lexie verbale. Mais dans les 22 % d'unités qui restent et que Syntex ne détecte aucune relation entre eux et la lexie, l'approche sans Syntex vient y remédier. Avec cette dernière, plusieurs participants non détectés par Syntex sont bien repérés.

Nous avons rencontré certaines difficultés dans le cas de participants propositionnels comme dans si vous essayez d'installer une application qui affiche ce type d'alerte, ABANDONNEZ l'installation. Toute la proposition « si vous essayez d'installer une application qui affiche ce type d'alerte » est prise comme un participant circonstant de la lexie verbale ABANDONNER. Avec notre système, nous n'identifions qu'une seule unité de toute la proposition, par exemple si, car le vecteur de feature que nous avons défini ne correspond qu'à une seule unité. Dans ce cas, nous avons proposé de soumettre le verbe de la proposition en question à notre système pour en identifier les participants. Nous générons la proposition en nous basant sur le verbe et ses participants. Quant aux participants qui se réalisent sous forme de mots composés de la forme NN (nom suivi d'un nom) ou de NA (nom suivi d'un adjectif), nous proposons de vérifier que si ces composés sont en une seule entrée dans DicoInfo[11] alors nous pourrons les considérer comme une entité.

[11] Dictionnaire de l'Informatique et de l'Internet, base de notre corpus http://olst.ling.umontreal.ca/cgi-bin/dicoinfo/search.cgi

L'identification et la distinction entre actants et circonstants ont été testées sur des lexies verbales annotées manuellement. Nous avons également annoté de nouvelles lexies verbales et leur validation manuelle est en cours.

Références

1. MEL'ČUK I. (2004) Actants in Semantic and Syntax, in *Actants in Semantics. Linguistics 42(2)*. p. 1-66.
2. LACHERET-DUJOUR A., FRANÇOIS J. (2004) Circonstance et prédication verbale en français parlé : contraintes sémantico-pragmatiques et filtrage prosodique, *Syntaxe & sémantique* 2004. 6: p. 35-56
3. VÁZQUEZ G., MONTRAVETA A.F. (2008) Annotation de corpus : Sur la délimitation des arguments et des adjoints. *SKY Journal of Linguistics* 2008. 21: p. 243-269.
4. FABRE C., FRÉROT C. (2002) Groupes prépositionnels arguments ou circonstants : vers un repérage automatique en corpus, in *TALN*. 2002: Nancy.
5. GILDEA D., JURAFSKY D., (2002) Automatic labeling of semantic roles. *Computational Linguistics,*
6. JOHANSON R., NUGUES P. (2005) Sparse Bayesian Classification of Predicate Arguments. *In Procs of CoNLL*-2005. Ann Arbor, Michigan.
7. SURDEANU M., HARABAGIU S.,WILLIAMS J., AARSETH P. (2003) Using predicate-argument structures for information extraction. *Langage Computer Corp*. USA.
8. LI L., FAN S., WANG X., WANG XI. (2008) Discriminative Learning of Syntactic and Semantic Dependencies, *In Procs of CoNLL*-2008. p. 218-222.
9. CHE W., LI Z., LI Y., GUO Y., QIN B., LIU T. (2009) Multilingual Dependency-based Syntactic and Semantic Parsing, In *Proceedings of CoNLL*-2009: Boulder, Colorado. p. 49-54.
10. TÄCKSTRÖM O. (2009) Multilingual semantic parsing with a pipeline of linear classifiers In *Proceeding of CoNLL*-2009: Boulder, Colorado. p. 103-108.
11. LIAKATA M., PULMAN S. (2003) Using Predicate-Argument Structures for Information Extraction, *In Proceedings of ACL*-2003. p. 8-15.
12. NASR A., BÉCHET F. (2009) Analyse syntaxique en dépendances de l'oral spontané in *TALN*-2009, Senlis.
13. BOURIGAULT D., FABRE C. (2005) L'analyseur syntaxique de corpus Syntex.
14. HADOUCHE F., L'HOMME M-C., LAPALME G. (2009) Automatic annotation of actants, eLexicography in the 21st century 2009, Université catholique de Louvain, Belgique.
15. HADOUCHE F., L'HOMME M-C., LAPALME G., LE SERREC A. (2009) Intégration d'informations syntaxico-sémantiques dans les bases de données terminologiques: méthodologie d'annotation et perspectives d'automatisation, *TLS'09 : Workshop international sur la Terminologie sémantique lexicale*, Montréal, Québec
16. TESNIÈRE L. Éléments de syntaxe structurale 1966, Paris, Klincksieck.
17. MESSIANT C., KORHONEN A., POIBEAU T. (2008) LexSchem: A Large Subcategorization Lexicon for French Verbs, in *LREC Proceedings*, Marrakech, Maroc.