Université de Montréal

Annotation syntaxico-sémantique des actants en corpus spécialisé

par
Fadila Hadouche

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Thèse présentée à la Faculté des arts et des sciences
en vue de l'obtention du grade Philosophiae Doctor (Ph.D.)
en informatique

Décembre, 2010

Université de Montréal
Faculté des études arts et des sciences

Cette thèse intitulée :

Annotation syntaxico-sémantique des actants en corpus spécialisé

Présentée par :
Fadila Hadouche

a été évaluée par un jury composé des personnes suivantes :

Jian-Yun Nie, président-rapporteur
Guy Lapalme, directeur
Marie-Claude L'Homme, co-directrice
Michel Boyer, membre du jury
Diana Inkpen, examinateur externe
Jian-Yun Nie, représentant du doyen de la FAS

SOMMAIRE

L'annotation en rôles sémantiques est une tâche qui permet d'attribuer des étiquettes de rôles telles que Agent, Patient, Instrument, Lieu, Destination etc. aux différents participants actants ou circonstants (arguments ou adjoints) d'une lexie prédicative. Cette tâche nécessite des ressources lexicales riches ou des corpus importants contenant des phrases annotées manuellement par des linguistes sur lesquels peuvent s'appuyer certaines approches d'automatisation (statistiques ou apprentissage machine).

Les travaux antérieurs dans ce domaine ont porté essentiellement sur la langue anglaise qui dispose de ressources riches, telles que PropBank, VerbNet et FrameNet, qui ont servi à alimenter les systèmes d'annotation automatisés. L'annotation dans d'autres langues, pour lesquelles on ne dispose pas d'un corpus annoté manuellement, repose souvent sur le FrameNet anglais. Une ressource telle que FrameNet de l'anglais est plus que nécessaire pour les systèmes d'annotation automatisé et l'annotation manuelle de milliers de phrases par des linguistes est une tâche fastidieuse et exigeante en temps. Nous avons proposé dans cette thèse un système automatique pour aider les linguistes dans cette tâche qui pourraient alors se limiter à la validation des annotations proposées par le système.

Dans notre travail, nous ne considérons que les verbes qui sont plus susceptibles que les noms d'être accompagnés par des actants réalisés dans les phrases. Ces verbes concernent les termes de spécialité d'informatique et d'Internet (ex. accéder, configurer, naviguer, télécharger) dont la structure actancielle est enrichie manuellement par des rôles sémantiques. La structure actancielle des lexies verbales est décrite selon les principes de la Lexicologie Explicative et Combinatoire, LEC de Mel'čuk et fait appel partiellement (en ce qui concerne les rôles sémantiques) à la notion de *Frame Element* tel que décrit dans la théorie *Frame Semantics* (FS) de Fillmore. Ces deux théories ont ceci de commun qu'elles mènent toutes les deux à la construction de dictionnaires différents de ceux issus des approches traditionnelles. Les lexies verbales d'informatique et d'Internet qui ont été annotées manuellement dans plusieurs contextes constituent notre corpus spécialisé.

Notre système qui attribue automatiquement des rôles sémantiques aux actants est basé sur des règles ou classificateurs entraînés sur plus de 2300 contextes. Nous sommes limités à une liste de rôles restreinte car certains rôles dans notre corpus n'ont pas assez d'exemples annotés manuellement. Dans notre système, nous n'avons traité que les rôles Patient, Agent et Destination dont le nombre d'exemple est supérieur à 300. Nous avons crée une classe que nous avons nommé Autre où nous avons rassemblé les autres rôles dont le nombre d'exemples annotés est inférieur à 100.

Nous avons subdivisé la tâche d'annotation en sous-tâches : identifier les participants actants et circonstants et attribuer des rôles sémantiques uniquement aux actants qui contribuent au sens de la lexie verbale. Nous avons soumis les phrases de notre corpus à l'analyseur syntaxique Syntex afin d'extraire les informations syntaxiques qui décrivent les différents participants d'une lexie verbale dans une phrase. Ces informations ont servi de traits (*features*) dans notre modèle d'apprentissage. Nous avons proposé deux techniques pour l'identification des participants : une technique à base de règles où nous avons extrait une trentaine de règles et une autre technique basée sur l'apprentissage machine. Ces mêmes techniques ont été utilisées pour la tâche de distinguer les actants des circonstants. Nous avons proposé pour la tâche d'attribuer des rôles sémantiques aux actants, une méthode de partitionnement (clustering) semi supervisé des instances que nous avons comparée à la méthode de classification de rôles sémantiques. Nous avons utilisé CHAMÉLÉON, un algorithme hiérarchique ascendant.

Mots-clés : actant, circonstant, rôles sémantiques, traits syntaxiques, classification, clustering, algorithme CHAMÉLÉON, Lexicologie Explicative et Combinatoire (LEC), *Frame semantics (FS)*, DicoInfo, FrameNet.

SUMMARY

Semantic role annotation is a process that aims to assign labels such as Agent, Patient, Instrument, Location, etc. to actants or circumstants (also called arguments or adjuncts) of predicative lexical units. This process often requires the use of rich lexical resources or corpora in which sentences are annotated manually by linguists. The automatic approaches (statistical or machine learning) are based on corpora.

Previous work was performed for the most part in English which has rich resources, such as PropBank, VerbNet and FrameNet. These resources were used to serve the automated annotation systems. This type of annotation in other languages for which no corpora of annotated sentences are available often use FrameNet by projection. Although a resource such as FrameNet is necessary for the automated annotation systems and the manual annotation by linguists of a large number of sentences is a tedious and time consuming work. We have proposed an automated system to help linguists in this task so that they have only to validate annotations proposed.

Our work focuses on verbs that are more likely than other predicative units (adjectives and nouns) to be accompanied by actants realized in sentences. These verbs are specialized terms of the computer science and Internet domains (ie. access, configure, browse, download) whose actantial structures have been annotated manually with semantic roles. The actantial structure is based on principles of Explanatory and Combinatory Lexicology, LEC of Mel'čuk and appeal in part (with regard to semantic roles) to the notion of Frame Element as described in the theory of frame semantics (FS) of Fillmore. What these two theories have in common is that they lead to the construction of dictionaries different from those resulting from the traditional theories. These manually annotated verbal units in several contexts constitute the specialized corpus that our work will use.

Our system designed to assign automatically semantic roles to actants is based on rules and classifiers trained on more than 2300 contexts. We are limited to a restricted list of roles for certain roles in our corpus have not enough examples manually annotated. In our system, we addressed the roles Patient, Agent and destination that the number of

examples is greater than 300. We have created a class that we called Autre which we bring to gether the other roles that the number of annotated examples is less than 100.

We subdivided the annotation task in the identification of participant actants and circumstants and the assignment of semantic roles to actants that contribute to the sense of the verbal lexical unit. We parsed, with Syntex, the sentences of the corpus to extract syntactic informations that describe the participants of the verbal lexical unit in the sentence. These informations are used as features in our learning model. We have proposed two techniques for the task of participant detection: the technique based in rules and machine learning. These same techniques are used for the task of classification of these participants into actants and circumstants. We proposed to the task of assigning semantic roles to the actants, a partitioning method (clustering) semi supervised of instances that we have compared to the method of semantic role classification. We used CHAMELEON, an ascending hierarchical algorithm.

Key-words: actant, circumstant, semantic roles, syntactic features, classification, clustering, CHAMÉLÉON algorithm, Explanatory and Combinatory Lexicology (LEC), *Frame semantics (FS)*, DicoInfo, FrameNet.

Table des matières

Liste des tableaux

Liste des figures

Liste des sigles

FS : Frame Semantics (une théorie)

SF : Semantic Frame (structure Frames)

FE : Frame Elements

LEC : Lexicologie Explicative et Combinatoire (une théorie)

DEC : Dictionnaire Explicatif et Combinatoire

DicoInfo : Dictionnaire d'Informatique et D'internet

Les conventions typographiques

Dans ce document, nous avons utilisé des polices de caractères pour certaines entités pour les mettre en évidence.

Entités	Polices
Arguments, actants sémantiques, frames éléments	Book antica
Annotation XML	Courier new
Citation	Normal en taille plus petite
Fonctions grammaticales	Calibri
Fonctions lexicales paradigmatiques et syntagmatiques	Gautami
Frames	Arial Narrow
Rôles sémantiques	Perpetua
Unité lexicale ou lexie	TAHOMA SMALL CAP
Exemples	Trebuchet Ms
Mots anglais	*Italique*

À

Aris, Elouize, Damya et Rayane

Remerciements

Je tiens vivement à remercier mon directeur de recherche, Guy Lapalme, pour son aide des plus précieuses, ses conseils, sa disponibilité, ses encouragements et son soutien tout au long de ma scolarité à l'université de Montréal et de ma recherche.

Je remercie également ma co-directrice, Marie-Claude L'Homme, pour son aide, ses conseils ses encouragements et sa patience tout au long de mon apprentissage en linguistique.

Je remercie Jian-Yun Nie, Michel Boyer, Diana Inkpen, membres du jury, d'avoir bien voulu accepter d'évaluer cette thèse.

Je remercie les étudiants, les chercheurs et les professeurs du laboratoire RALI pour m'avoir donné un environnement de travail stimulant

Je remercie l'équipe de L'OLST qui a annoté le corpus d'informatique et de l'Internet, particulièrement Janine Pimentel, Suzanne DesGroseilliers, Annaïch Le Serrec et Marie-Eve Laneville.

Je remercie le CRSH pour le support financier.

Mes remerciements vont également à Didier Bourigault qui nous a permis d'utiliser l'analyseur syntaxique Syntex. Et je remercie Patrick Drouin qui a rendu cet outil disponible à tout moment.

Je remercie ma mère et mon père qui m'ont encouragée à continuer mes études.

Un gros merci à mes sœurs Malika et Nacera pour leur soutien moral et leurs encouragements.

Chapitre 1 Introduction

L'objectif de notre travail est de proposer une méthode pour identifier automatiquement les actants (également appelés arguments) des unités lexicales ou lexies[1]. Il consiste à annoter les unités lexicales prédicatives verbales du français et leurs actants apparaissant dans des contextes, c.-à-d. des phrases tirées d'un corpus de textes français des domaines de l'informatique et de l'Internet. Cette annotation consiste à assigner aux actants des étiquettes de rôles sémantiques comme Agent, Patient, Destination, Instrument, Source, Lieu, Moyen, etc. aux actants. Certains rôles sémantiques ne sont pas couverts par un nombre d'exemples important. Ils ont moins de 100 exemples et aussi certains ont moins de 20 exemples. Dans notre travail, nous nous sommes limités qu'à certains rôles sémantiques dont le nombre d'exemples dépasse au moins 300. Dans notre étude, nous avons considéré le rôle Patient avec 1992 exemples, le rôle Agent avec 756 exemples, le rôle Destination avec 370 exemples. Pour les autres rôles, nous avons proposé de les prendre dans une même classe que nous avons appelée. Les exemples sont tirés du Dictionnaire de l'informatique et de l'Internet (DicoInfo) qui comporte les annotations vérifiées et validées de 104 lexies, 2311 contextes ou phrases, 3512 actants avec 22 rôles sémantiques et 1115 circonstants avec 11 rôles sémantiques. L'assignation de rôles sémantiques peut aussi concerner les circonstants[2] mais, dans notre travail, nous nous focalisons sur les actants. Les actants se distinguent des circonstants en ce sens qu'ils participent au sens d'une lexie prédicative. Leur annotation au moyen de rôles sémantiques permet de définir la relation de sens qui existe entre eux et la lexie prédicative. Cette dernière a un lien de nature différente avec chacun de ses actants. Les circonstants ne contribuent pas au sens de la lexie; mais peuvent ajouter des informations à la phrase.

Par exemple, la lexie ABANDONNER signifiant, dans le domaine de l'informatique, « 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 Agent ou le rôle Patient est un actant. Avec ces actants le sens de

[1] Dans tout le document, **lexie** et **unité lexicale** sont utilisées dans le même sens.
[2] Certains auteurs désignent les actants par arguments et les circonstants par adjoints.

la lexie est complet. Les autres participants de cette même lexie dans d'autres contextes qui ne jouent pas les rôles Agent ou Patient seront considérés comme des circonstants. Dans l'exemple :

Le programmeur ABANDONNE l'opération à ce stade.

Le programmeur et l'opération sont des actants (respectivement Agent et Patient) et à ce stade est un circonstant de Temps.

Dans certains cas, une lexie peut réaliser un actant dans une phrase et un circonstant dans une autre. Par exemple, une unité lexicale qui indique le lieu peut être un actant ou un circonstant selon la lexie verbale avec laquelle est employée. Dans la phrase :

Une personne CONSULTE de l'information sur le Web

Le participant sur le Web joue le rôle d'un circonstant de rôle sémantique Lieu de la lexie verbale CONSULTER. Cependant, dans la phrase :

Une personne NAVIGUE sur le Web

le participant sur le Web réalise un actant ayant comme rôle sémantique Lieu de la lexie verbale NAVIGUER. 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.

L'annotation de rôles sémantiques permet de désambiguïser les sens des verbes. Elle permet également de décrire les alternances, ainsi que les liens qui existent entre les dérivés morphologiques avec sens apparentés et les sens voisins. Elle présente un intérêt dans des applications TAL nécessitant des informations sémantiques telles que la traduction automatique, le résumé automatique, l'extraction d'informations et les systèmes de questions-réponses (voir chapitre 3).

Notre travail consiste à distinguer automatiquement ces deux formes de participants (actants et circonstants) et attribuer aux premiers des rôles sémantiques. Nous avons divisé cette tâche en trois étapes :

1) Identification des participants de la lexie verbale : étant donnée une lexie verbale dans une phrase, nous cherchons à trouver les unités lexicales réalisant les participants de cette lexie. Exemple :

[Une personne] CONSULTE de [l'information] sur [le Web]

Les participants de la lexie CONSULTER dans cette phrase sont une personne, l'information et le Web.

2) Distinction des actants et circonstants : les participants identifiés d'une lexie peuvent être de deux types, actants ou circonstants. Étant donné une lexie dans une phrase dont les participants sont identifiés en 1), nous cherchons à distinguer ceux qui réalisent des actants de ceux qui réalisent des circonstants. En nous inspirant de la notation de Palmer et Gildea dans leur ouvrage sur les rôles sémantiques [56], nous représentons chaque participant entre crochet et à sa gauche le type (actant ou circonstant) et le rôle sémantique. Exemple :

[$_{\text{Actant}}$ Une personne] CONSULTE de [$_{\text{Actant}}$ l'information] sur [$_{\text{Circonstant}}$ le Web]

Les participants une personne et l'information réalisent des actants et le participant le Web réalise un circonstant pour la lexie CONSULTER.

3) Distinction des rôles sémantiques : nous annotons les actants identifiés en 2) en leur attribuant des rôles sémantiques. Exemple :

[$_{\text{Agent}}$ Une personne] CONSULTE de [$_{\text{Patient}}$ l'information] sur le Web

Les actants une personne et l'information sont annotés par des étiquettes de rôles sémantiques, respectivement Agent et Patient.

Notre travail s'est déroulé dans le cadre du projet « Sémantique lexicale et terminologie : Application de deux cadres théoriques lexicaux à la description des termes spécialisés » financé par le Conseil de recherche en sciences humaines (CRSH), dont les

responsables étaient les directeurs de thèse. Ce projet impliquait également d'autres chercheurs étudiants de l'Observatoire de linguistique Sens-Texte (OLST). L'objectif de ce projet était de décrire des termes appartenant à un domaine de spécialité en appliquant deux cadres théoriques lexicaux : *Frame Semantics* (FS) [17] et Lexicologie Explicative et Combinatoire (LEC) [46].

Notre objectif informatique est l'élaboration de méthodes d'attribution automatique de rôles sémantiques aux actants des termes spécialisés. Nous nous sommes basés sur des descriptions actancielles existantes réalisées à la main et qui forment ce que nous appelons notre corpus. Nous les avons enrichies par des rôles sémantiques en proposant des méthodes automatiques d'annotation de rôle sémantique inspiré du modèle de *Frame Semantics*. Les unités lexicales sont tirées du dictionnaire spécialisé dans le domaine de l'informatique, le Dictionnaire fondamental de l'informatique (DicoInfo) (http://olst.ling.umontreal.ca/dicoinfo). Nous avons utilisé cette ressource comme un corpus. Nous cherchons à déterminer les rôles des actants de nouvelles lexies verbales qui ne sont pas déjà présentes dans le corpus de départ.

Nous voulons identifier automatiquement les actants et leur attribuer des rôles sémantiques sans avoir recours au FrameNet de l'anglais. Nous avons attribué des rôles sémantiques aux actants à partir d'un corpus de phrases du français annotées manuellement et tirées d'un dictionnaire déjà construit selon les principes de la LEC. L'annotation manuelle de phrases reste une tâche fastidieuse pour les annotateurs et très exigeante en temps; en moyenne 2 heures par lexie. Nous croyons que l'automatisation de cette tâche accélérerait le temps d'annotation et faciliterait la tâche de l'annotateur. Cette tâche d'annotation deviendrait une validation des annotations faites par notre système automatique. En s'évitant une grande partie du travail routinier pour la majorité des cas simples, le linguiste pourra se concentrer sur les cas plus difficiles.

Le corpus annoté et vérifié par des linguistes a été un grand atout pour nous par rapport à des études similaires effectuées par d'autres équipes pour d'autres langues que l'anglais. Nous n'avons pas eu à recourir au FrameNet de l'anglais pour annoter les lexies prédicatives du français. C'est ce qu'ont dû faire Padò et Lapata en 2005 [53] qui ont proposé une approche de projection d'annotations en exploitant les ressources parallèles

entre des langues. Après avoir expérimenté avec les projections anglais-allemand, ils ont testé en 2007 une projection anglais-français [54].

Nous avons effectué plusieurs expériences en utilisant le corpus de l'informatique et de l'Internet. Nous avons testé deux approches pour l'identification des participants actants et circonstants.

Notre première approche était une approche à base de règles. Nous avons extrait une trentaine de règles en nous basant sur les sorties d'un analyseur syntaxique du français, appelé Syntex[3]. Nous avons mis en correspondance les liens entre : a) la lexie verbale à l'étude avec d'autres unités lexicales dans la phrase retournées par l'analyseur syntaxique Syntex et; b) les liens de participant actant ou circonstant retrouvés dans notre corpus annoté manuellement. Ces correspondances nous ont permis de construire des règles dont la partie gauche représente les informations syntaxiques et la partie droite le résultat (c-à-d : participant). Dans cette approche d'identification de participants au moyen de règles, nous avons obtenu des taux de précision de 69 %, de rappel de 80 % et de F-mesure de 74%.

Notre deuxième approche était basée sur l'apprentissage machine sur un corpus annoté manuellement sur lequel ont été entraînés des classificateurs de Weka[4]. Les règles extraites dans la première approche nous ont permis d'identifier des traits utilisés par les classificateurs. Ces traits basés sur des informations syntaxiques ressemblent à ceux proposés par Gildea [25] dans son approche d'annotation de rôles sémantiques des verbes de l'anglais basée sur la ressource FrameNet[5]. Pour la tâche d'identification des participants, notre meilleur classificateur, RandomForestTree, obtient une F-mesure de 84,8 % si nous utilisons un analyseur syntaxique et de 74 % sans le recours à l'analyseur. Pour la tâche de distinction des participants en actants et circonstants, le classificateur RandomForestTree obtient une F-mesure de 96 % si nous utilisons l'analyseur syntaxique et de 94,6 % sinon. Les résultats sont meilleurs que ceux de la littérature; cela est attribuable en partie au fait que nous disposions d'une bonne quantité de contextes en

[3] Analyseur en dépendances, Bourigault et al., (http://w3.erss.univ-tlse2.fr/textes/pagespersos/bourigault/syntex.html)
[4] Weka fournit des implémentations des algorithmes d'apprentissage. Il a été développé à l'université de Waikato en Nouvelle-Zélande. www.cs.waikato.ac.nz/ml/weka
[5] http://framenet.icsi.berkeley.edu

français annotés et vérifiés. Ces bons résultats peuvent également être liés au fait que nous travaillons dans un domaine restreint.

Pour la tâche d'attribution de rôles sémantiques aux actants, nous avons d'abord testé le classificateur RandomFrest sur nos données. Nous avons obtenu de bons résultats en F-mesure de 93%, de 90%, de 85% et de 76% respectivement de Patient, de Agent, de Destination et de Autre. Pour pouvoir classifier de nouveaux rôles non déjà vus ou de nouvelles lexies, nous avons proposée une perspective d'utiliser le partitionnement semi supervisé des instances qui nécessite le feedback ou les contraintes de l'annotateur, au lieu de la classification des rôles sémantiques qui ne permet pas de donner le rôle naturel d'une nouvelle instance. Si une nouvelle instance dont le rôle sémantique naturel n'est pas pris en compte pendant la classification, alors le rôle prédit à cette nouvelle instance est le rôle le plus proche pas ce rôle naturel. Notre perspective de partitionnement est de permettre à l'annotateur de donner un nouveau rôle et de l'intégrer dans des groupes les plus proches. Et ce rôle sera pris la suite en considération. Notre proposition peut être une baseline d'une approche différente qui nécessite dans le futur plus d'étude pour chercher des modèles adéquats pour prendre en compte le feedback de l'annotateur. Nous avons proposé des traits de descriptions de nos actants de nature catégorielles. Nous avons essayé l'algorithme EM de partitionnement qui nous a donné un taux d'instances incorrectement regroupées de 66%. Nous avons proposé d'utiliser un algorithme hiérarchique CHAMÉLÉON qui manipule mieux les traits de nature catégorielle. C'est un algorithme hiérarchique évolué et performant testé dans plusieurs domaines tels que datamining, bioinformatique, etc. Cet algorithme définit deux mesures (inter-connectivité relative et proximité relative) entre les groupes (clusters). Ces deux mesures permettent d'avoir des groupes bien formés. Cet algorithme de clustering (partitionnement), testé pour la tâche d'annotation de rôles sémantiques, a donné les résultats en F-mesure de 88%, de 81%, de 58% et de 46% respectivement de Patient, Agent, Destination et Autre.

1.1 Organisation de la thèse

Au chapitre 2, nous présentons les deux cadres théoriques sur lesquels nous nous sommes appuyés. La LEC utilisée dans la construction du DicoInfo a servi de base théorique pour

notre corpus. La FS est utilisée comme une source d'inspiration pour l'annotation de rôles sémantiques de la structure actancielle des lexies verbales.

Dans le chapitre 3, nous définissons les rôles sémantiques, ainsi que les liens entre ces rôles et les fonctions grammaticales. Nous présentons leur apport aux applications de TAL nécessitant ce type d'informations. Le chapitre 4 présente un état de l'art des approches d'automatisation de l'annotation en rôles sémantiques.

Le chapitre 5 présente une analyse linguistique des données. Ces dernières sont extraites du DicoInfo, décrites selon la LEC et FS et forment notre corpus. L'unité de base décrite est une lexie prédicative verbale française dont les contextes sont tirés de textes liés aux domaines de l'informatique et de l'Internet. La structure actancielle des lexies verbales est enrichie et annotée de rôles sémantiques. Cette annotation est faite manuellement par les membres de L'OLST et elle est réalisée selon un modèle fortement inspiré de FrameNet. Les données du corpus sont décrites sous forme de schéma XML. Les types de participants actants ou circonstants, les rôles sémantiques, ainsi que les fonctions et les groupes syntaxiques sont décrites dans ce schéma. Ce corpus constitue la base d'entraînement et de test pour notre modèle d'automatisation de l'annotation, subdivisée en sous-tâches : identification des participants, distinction actants et circonstants et attribution de rôles sémantiques aux actants.

Le chapitre 6 décrit la première approche basée sur les règles pour identifier les participants actants et circonstants. La seconde approche d'identification par apprentissage machine est présentée au chapitre 7. Dans ce chapitre, un ensemble de traits a été proposé pour tenir compte des caractéristiques syntaxiques des participants candidats. Nous présentons les résultats de plusieurs classificateurs testés sur nos données. Le chapitre 8 présente le modèle automatique d'attribution de rôles sémantiques aux actants identifiés. Dans ce cas, nous avons proposé un algorithme de partitionnement basé sur des traits appropriés que nous comparé à la classification.

Nous concluons notre travail en identifiant nos principales contributions et nous proposons certaines perspectives.

Chapitre 2 Description des cadres théoriques

Dans ce chapitre nous présentons les deux cadres théoriques, Lexicologie Explicative et Combinatoire (LEC) et *Frame Semantics* (FS), sur lesquels se basent les ressources lexicales et terminologiques non traditionnelles, notamment des ressources dans lesquelles de nombreux renseignements sur la structure actancielle des unités lexicales. L'annotation manuelle de nos données est basée sur ces deux théories. La description de la structure actancielle d'une lexie verbale prédicative est faite selon la théorie de la LEC. L'annotation des rôles sémantiques est inspirée de la théorie FS.

2.1 La Lexicologie explicative et combinatoire (LEC)

La LEC est le volet lexical d'une théorie linguistique plus vaste qu'est la Théorie Sens-Texte. La LEC nous offre une méthode de description formelle de l'unité lexicale (ou lexie). Cette dernière est définie comme : 1) un mot pris dans une acception bien spécifique qui peut être égale à un lexème; 2) une locution prise dans une acception bien spécifique qui peut être égale à un phrasème [46]. La LEC décrit d'une manière formelle les lexies de la langue générale.

Dans cette théorie, qui a donné lieu au Dictionnaire Explicatif et Combinatoire (DEC), « Explicatif » signifie que tout élément lexical est accompagné d'une explication sémantique formelle, tandis que « Combinatoire » indique l'importance particulière portée à la représentation de la combinatoire lexicale [46]. Le DEC prend en compte deux axes : l'axe syntagmatique, l'enchaînement des unités dans le texte; et l'axe paradigmatique, oppositions sémantiques et sélection sémantique d'unités [46]. Le DEC traite chaque sens séparément. Un article du DEC correspond à la description d'une lexie, contrairement au dictionnaire issu de la lexicographie traditionnelle qui prend comme unité de base un mot polysémique.

La LEC propose des définitions pour des notions très importantes auxquelles nous ferons appel dans le présent travail, à savoir « prédicat sémantique », « actant sémantique » et « définition lexicographique ».

> Les prédicats sémantiques désignent des actions, des événements, des processus, des états, des propriétés, des relations, etc. En un mot, des faits qui impliquent nécessairement des participants [46]

P(A1,A2,A3) dénote un prédicat (sémantique) P représentant un fait ayant, par exemple, trois arguments (A1,A2,A3) de P correspondant aux participants de ce fait.

Exemples : [46]

> Le sens de DONNER est un prédicat à trois arguments : quelqu'un [1] donne quelque chose [2] à quelqu'un [3]
>
> Le sens de SOMMEIl ou de DORMIR est un prédicat à un argument : quelqu'un [1] dort.

Alors qu'en logique on parle d'un prédicat et de ses arguments, en lexicologie, on préfère parfois parler de prédicat sémantique et d'actant sémantique. On appelle lexie *à sens* prédicatif une lexie dont le sens correspond à un prédicat sémantique. Pour la décrire, on a besoin du concept fondamental « actant sémantique » qui correspond à l'argument d'un prédicat utilisé en logique tel qu'illustré dans la définition suivante [46] :

> Nous appelons actant sémantique [=ASém] de la lexie L une expression qui correspond à un argument du prédicat L(A1,A2,…,An); cette expression est soit un sens, soit une variable dans la définition de L.

Reprenant les exemples ci-dessus, DONNER a trois actants sémantiques (ASem) :

- A1 celui qui donne,
- A2 : celui qui reçoit, et
- A3 : ce qui passe de A1 à A2.

DORMIR quant à lui, a un seul ASem :

- A1 : celui qui dort.

Une définition lexicographique en LEC présente le sens de la lexie formellement et elle fait ressortir le défini et le définissant. Le défini est la lexie elle-même et le définissant est une description du sens de cette lexie. Dans [46], la définition lexicographique est présentée comme une expression de la forme propositionnelle *A=B* où *A* est le défini et *B* est le définissant. Exemple

X **assiste** à Y = Personne X est présente à un évènement Y en tant que spectateur ou participant

La LEC définit les lexies à sens prédicatif en explicitant les actants sémantiques (ASém) sous forme de variables (X,Y,Z…). Ces variables peuvent être accompagnées d'étiquettes sémantiques telles que personne et fait.

Dans le DEC, une unité lexicale est décrite par sa structure actancielle, ses fonctions lexicales et un schéma de régime. Les actants ont des réalisations linguistiques qui sont décrites dans le schéma de régime indiquant leur fonctionnement en syntaxe profonde [45]. Ces actants syntaxiques profonds sont numérotés par des chiffres romains I, II, III, … Dans le cas d'un verbe transitif, le sujet de la lexie considérée correspond à I, le complément d'objet direct correspond à II, les autres compléments à III, et ainsi de suite [46]. Dans les fonctions lexicales, on trouve les liens lexicaux partagés par l'unité lexicale décrite avec d'autres unités du lexique.

Les principes de la LEC ont été utilisés dans certains domaines spécialisés, notamment celui de l'informatique, et ont été implémentés dans une ressource appelée le DicoInfo. Par exemple, dans le DicoInfo, on trouve la description de la structure actancielle des lexies et d'autres catégories d'informations inspirées de la LEC telles que les liens lexicaux. Pour chaque actant de la lexie, on énumère également les réalisations linguistiques qui lui correspondent (voir section 2.1.1.1).

2.1.1 Les ressources

En LEC, les lexies sont regroupées avec d'autres lexies au sein d'un vocable. Chaque vocable est un regroupement de lexies qui se trouvent en relation de polysémie et qui ont des signifiants identiques et des signifiés qui sont liés entre eux. Cependant, on peut trouver des lexies qui ont des signifiants identiques mais dont les signifiés ne présentent pas de lien entre eux, dans ce cas ces lexies entretiennent une relation d'homonymie et ces lexies appartiennent à deux vocables différents. Pour distinguer ces diverses lexies, on utilise un système de numérotation [45] :

> *Deux lexies homonymes (appartenant à deux vocables différents) sont munies d'exposants numériques différents.*

> *À l'intérieur du même vocable, les numéros des lexies tentent de refléter les distances sémantiques entre ces derniers (cette distance est mesurée par l'importance des composantes sémantiques communes et par les régularités de la différence sémantique observée). Trois niveaux de numérotation sont utilisés : les chiffres romains signifient la distance maximale, les chiffres arabes la distance moyenne et les lettres minuscules la distance minimale.*

Différents dictionnaires basés sur la LEC ont été construits à l'OLST. Ces dictionnaires sont :

- le Dictionnaire Explicatif et Combinatoire (DEC) [45], en version papier et
- deux autres dictionnaires en version électronique : le Dico [35] (version base de données FileMaker, convertie en base de données relationnelle et appelé dans le web le Dicouèbe[6]) et le DicoInfo[7] en version XML.

Nous ne décrivons ici que la ressource DicoInfo sur laquelle est basée l'annotation.

2.1.1.1 **DicoInfo**

Le DicoInfo est un dictionnaire basé sur les principes de la LEC qui décrit des termes spécialisés en particulier les sens reliés aux domaines de l'informatique et de l'Internet. Il se focalise sur le fonctionnement linguistique des termes : il décrit les termes en explicitant leur structure actancielle, les liens paradigmatiques (synonymie, antonymie) et syntagmatiques qu'ils entretiennent avec d'autres termes du domaine. Le DicoInfo est publié sur le site de l'OLST à l'adresse http://olst.ling.umontreal.ca/dicoinfo.

Plusieurs rubriques participent à la description d'un terme dans le DicoInfo telles que l'entrée, les informations grammaticales, le statut, la structure actancielle, la définition, les synonymes et tous les liens lexicaux. Prenons l'exemple de ABANDONNER, sa description dans le DicoInfo est donnée dans les rubriques suivantes (figures 1 à 7).

1. Entrée

Chaque entrée est présentée par son nom et le numéro d'acception en indice même si l'unité en question a un seul sens[8]. Les acceptions dans le DicoInfo ne sont pas numérotées

[6] Le Dicouèbe est un dictionnaire compilé dans une base de données relationnelle. Il décrit les lexies dans les différents axes de la LEC à savoir la définition du sens, le schéma de régime et les fonctions lexicales. Documentation en ligne (http://olst.ling.umontreal.ca/dicouebe)
[7] Documentation en ligne (http://olst.ling.umontreal.ca/dicoinfo)

selon un protocole très strict comme dans le Dicouèbe sauf dans le cas des nominalisations des verbes qui portent le même numéro d'acception que le verbe. Les lettres minuscules (a, b, etc.) sont utilisées pour les sens voisins. Les entrées sont aussi accompagnées d'autres informations grammaticales telles que : féminin, masculin pour les noms, verbe transitif ou intransitif pour les verbes, etc. Par exemple, les informations grammaticales de ABANDONNER sont verbe transitif « v.tr. » (Figure 1).

abandonner[1] , v. tr.

Figure 1 L'entrée ABANDONNER

2. Structure actancielle et actants

La structure actancielle nous permet d'identifier les actants sémantiques d'un terme et donne la position de ces actants par rapport au terme décrit. Elle décrit aussi le rôle actanciel de chaque actant. Les rôles actanciels sont définis par des étiquettes telles que : Agent, Patient, Destination, Source, Instrument, Lieu, etc. En plus de ces rôles actanciels attribués aux actants, ces derniers sont accompagnés d'une mention indiquant le terme typique qui précise la réalisation d'un rôle actanciel. Dans le DicoInfo, cette mention est écrite entre accolades après le rôle actanciel correspondant. Bien sûr, plusieurs réalisations peuvent être observées mais on choisit comme actant typique le terme fréquemment utilisé avec le terme décrit ou le terme qui est générique (celui qui englobe les autres réalisations). Par exemple, la structure actancielle de ABANDONNER est donnée à la Figure 2.

Structure actancielle : abandonner : Agent{utilisateur 1} ~ Patient{tâche 1}

Figure 2 Structure actancielle de ABANDONNER

Par cette structure on comprend que l'action de ABANDONNER est effectuée par l'actant typique utilisateur 1 dont le rôle actanciel est Agent sur un actant typique tâche 1 dont le rôle actantiel est Patient. D'autres réalisations linguistiques peuvent être observées (Figure 3).

[8] L'unité ABANDONNER a un seul sens dans le DicoInfo mais on l'écrit abandonner$_1$ tel que dans la Figure 1

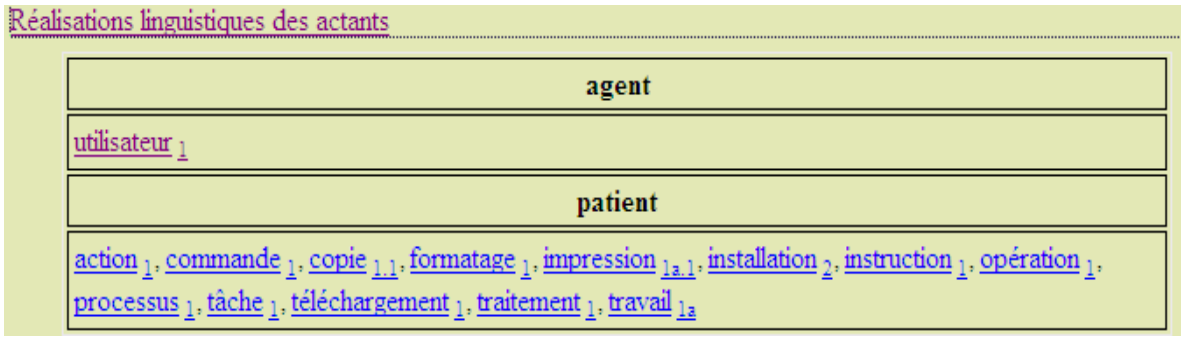


Figure 3 Les réalisations linguistiques des actants d'ABANDONNER

On pourra accéder aux autres réalisations des actants dans la rubrique réalisations linguistiques des actants. Cette rubrique est donnée après la structure actancielle. Si ces réalisations sont déjà décrites dans le DicoInfo, on peut accéder à leur description au moyen d'un lien hypertexte.

3. Définition

Certains termes sont accompagnés d'une définition. La définition est construite à partir de la structure actancielle. Elle est donnée seulement au moyen des actants typiques. Le rôle des actants n'y figure pas. Voir la définition de ABANDONNER à la Figure 4.

Définition : Un UTILISATEUR met fin à l'exécution d'une TÂCHE, souvent en raison de l'apparition d'un problème.

Figure 4 Définition de ABANDONNER

4. Contextes

Les contextes indiquent, par des exemples dans les textes du corpus, la manière dont le terme est utilisé concrètement. La Figure 5 nous donne un exemple de contextes de l'unité ABANDONNER.

Contextes
Chaque travail en attente porte un numéro, grâce auquel on peut le manipuler. Entre autres opérations, il est possible de l'abandonner à l'aide de la commande lprm. (Source : LINUX3P2)

Figure 5 Contextes de ABANDONNER

5. Les liens lexicaux

Lorsqu'un terme présente des variantes graphiques telles que 'antivirus' et 'anti-virus', ou des liens de synonymie, ces informations sont données à la suite des réalisations linguistiques ou de la définition.

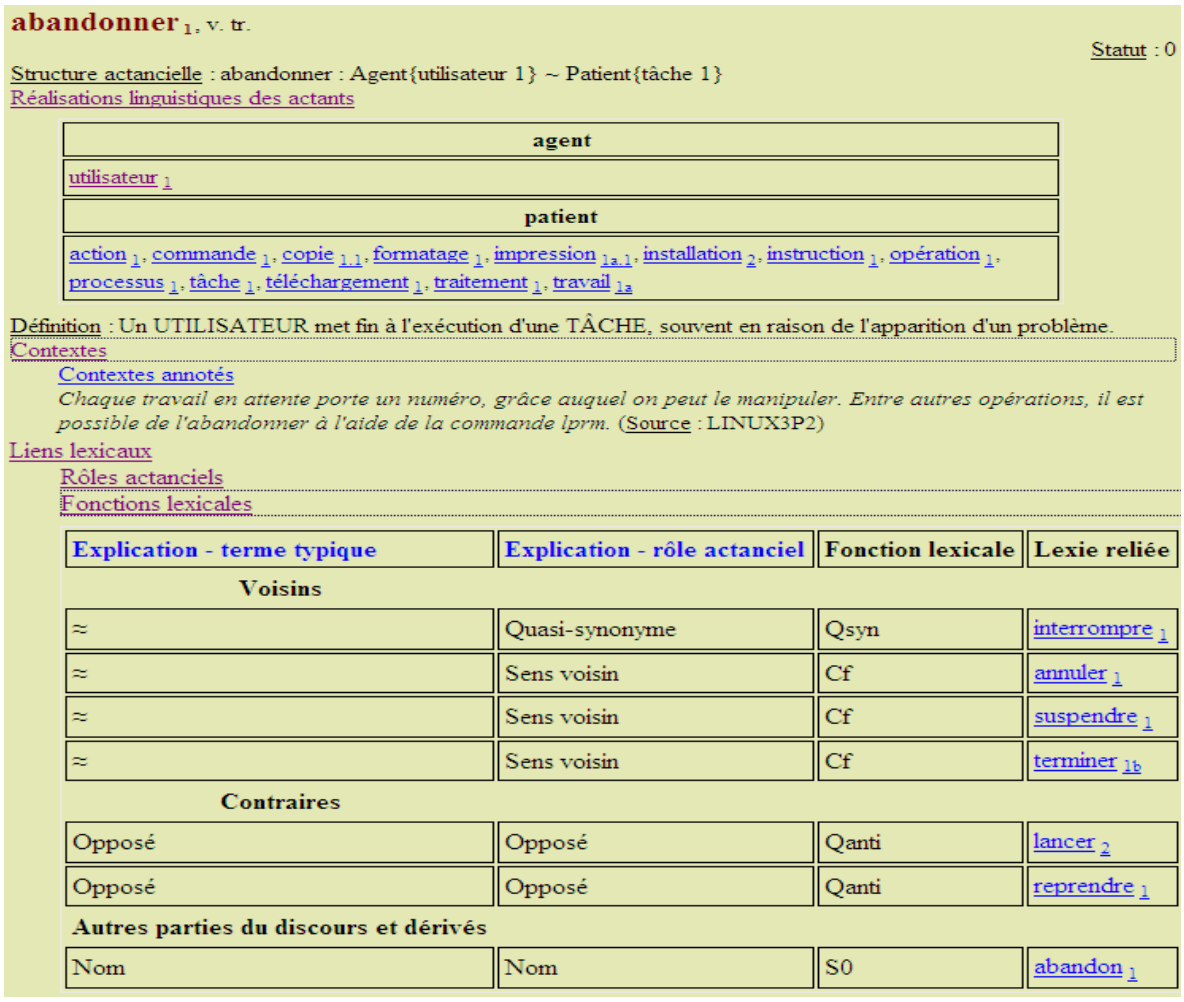
D'autres liens lexicaux paradigmatiques et syntagmatiques qu'entretient un terme avec les autres termes de DicoInfo sont aussi décrits. Dans le DicoInfo, ces liens sont affichés dans un tableau. Par exemple, les liens lexicaux de l'unité ABANDONNER sont donnés dans la Figure 6. La deuxième colonne de ce même tableau nous donne l'explication de rôle actanciel, la colonne 3, la fonction lexicale et la colonne 4 le lexie reliée correspondante.

Liens lexicaux
Rôles actanciels
Fonctions lexicales

Explication - terme typique	Explication - rôle actanciel	Fonction lexicale	Lexie reliée
Voisins			
≈	Quasi-synonyme	Qsyn	interrompre 1
≈	Sens voisin	Cf	annuler 1
≈	Sens voisin	Cf	suspendre 1
≈	Sens voisin	Cf	terminer 1b
Contraires			
Opposé	Opposé	Qanti	lancer 2
Opposé	Opposé	Qanti	reprendre 1
Autres parties du discours et dérivés			
Nom	Nom	S0	abandon 1

Figure 6 Les liens lexicaux de ABANDONNER

La Figure 7 donne la description globale de ABANDONNER.



abandonner $_{1}$, v. tr.

Statut : 0

Structure actancielle : abandonner : Agent{utilisateur 1} ~ Patient{tâche 1}

Réalisations linguistiques des actants

agent
utilisateur $_{1}$
patient
action $_{1}$, commande $_{1}$, copie $_{1.1}$, formatage $_{1}$, impression $_{1a.1}$, installation $_{2}$, instruction $_{1}$, opération $_{1}$, processus $_{1}$, tâche $_{1}$, téléchargement $_{1}$, traitement $_{1}$, travail $_{1a}$

Définition : Un UTILISATEUR met fin à l'exécution d'une TÂCHE, souvent en raison de l'apparition d'un problème.

Contextes

Contextes annotés

Chaque travail en attente porte un numéro, grâce auquel on peut le manipuler. Entre autres opérations, il est possible de l'abandonner à l'aide de la commande lprm. (Source : LINUX3P2)

Liens lexicaux

Rôles actanciels

Fonctions lexicales

Explication - terme typique	Explication - rôle actanciel	Fonction lexicale	Lexie reliée
Voisins			
≈	Quasi-synonyme	Qsyn	interrompre $_{1}$
≈	Sens voisin	Cf	annuler $_{1}$
≈	Sens voisin	Cf	suspendre $_{1}$
≈	Sens voisin	Cf	terminer $_{1b}$
Contraires			
Opposé	Opposé	Qanti	lancer $_{2}$
Opposé	Opposé	Qanti	reprendre $_{1}$
Autres parties du discours et dérivés			
Nom	Nom	S0	abandon $_{1}$

Figure 7 Description globale de l'unité ABANDONNER

2.2 La théorie des *Frames Semantics* (FS)

La théorie linguistique *Frame Semantics* (FS) proposée par Fillmore [16], connaît un intérêt croissant dans les projets de construction de bases de données lexicales. Fillmore proposa d'abord une grammaire de cas (case grammar) en 1968 [14] et il en dériva la théorie FS dans laquelle les Frames sont considérés comme des descriptions linguistiques de scénarios conceptuels (voir section 2.2.1). Il suppose que pour comprendre le sens d'une unité lexicale, nous devons d'abord connaître tous les *Semantic Frames* (SF) ou la structure conceptuelle à laquelle elle appartient [17]. Le modèle de SF caractérise les propriétés sémantiques et syntaxiques des unités lexicales en les rattachant à des Frames. Ce sont des représentations de situations qui impliquent des participants appelés *Frame Elements* (FE).

Autrement dit, en *FS*, le sens d'une unité lexicale est défini par référence à une structure d'expérience, de croyances, ou de pratiques :

> *A word's meaning can be understood only with reference to a structured background of experience, beliefs, or practices, constituting a kind of conceptual prerequisite for understanding the meaning [18].*

Avec cette approche, les sens des unités lexicales ne sont pas liés l'un à l'autre directement unité à unité, mais seulement en termes de leur relation aux Frames [18].

L'approche FS se focalise sur l'élément principal qui est le *Semantic Frame*. Ce frame est défini comme une structure cohérente de concepts liés. Ils sont liés parce que sans la connaissance de chacun d'entre eux, on n'a pas la connaissance complète des concepts.

2.2.1 Le *Semantic Frame*

La description des unités lexicales en SF permet à un lexicographe d'utiliser le même cadre sémantique pour décrire des unités lexicales appartenant à différentes parties du discours, mais partageant des éléments sémantiques communs.

Un *frame* est une construction intuitive qui formalise les liens entre la sémantique et la syntaxe dans les résultats de l'analyse lexicale [15]. Elle représente des situations dans laquelle apparaissent les différents *Frame Elements* (FE) et les différents types de relations existant entre eux. Une partie des FE sont les participants obligatoires à la situation et correspondent grosso modo aux actants sémantiques des lexies prédicatives de la LEC.

Un exemple souvent utilisé dans la littérature portant sur ce cadre théorique est celui du frame *commercial_transaction* impliquant quatre *FE* : *buyer*, *seller*, *goods*, *money*. A ce frame sont rattachés indirectement les verbes tels que *BUY*, *SELL*, *COST* ou *CHARGE,* les noms tels que *price, goods*, *money* et les adjectifs tels que *cheap*, *expensive.*

Chaque événement qui implique le fait d'échanger de l'argent contre de la marchandise qui constitue la transaction commerciale est considéré comme un exemple de la transaction commerciale. Considérons ces différentes phrases qui illustrent le frame *commercial_transaction* :

John SOLD the car to Peter for 2000$.

Peter BOUGHT the car from John for 2000$.

Peter PAID John 2000$ for the car.

John CHARGED Peter 2000$ for the car.

The car COST Peter 2000$.

De ces exemples, on fait ressortir les *FE* avec leurs réalisations linguistiques cités dans le tableau 1

FE	**Réalisations linguistiques**
Buyer	*Peter*
Seller	*John*
Goods	*Car*
Money	*2000$*

Tableau 1 *Frames éléments* (*FE*) du frame *commercial_transaction*

Les *FE* peuvent être de deux types :

- obligatoires (core) leur présence est nécessaire et fait partie du sens de l'unité lexicale ;
- optionnels (non-core) leur présence n'est pas fondamentale et ne participent pas au sens de l'unité lexicale.

Dans l'exemple 2, le verbe *BUY* prend *buyer* et *goods* comme *FE* obligatoires et *seller* et *money* comme *FE* optionnels apparaissant dans des syntagmes prépositionnels introduits respectivement par les prépositions *from* et *for*. D'une unité lexicale à une autre, ces *FE* peuvent apparaître dans différentes positions et remplir différents types de fonctions syntaxiques (Sujet, Objet, etc.). Ainsi, dans l'exemple (2), le *FE buyer* est réalisé comme sujet du verbe *BUY* et le *FE goods* est réalisé comme objet direct de ce même verbe. Dans le même exemple, les autres *FE* tels que *seller* et *money* sont réalisés comme complément. Dans le cas du verbe *CHARGE* de l'exemple (4), les *FE seller* et *money* sont réalisés respectivement

comme Sujet et Objet direct et les *FE buyer* et *goods* sont réalisés respectivement comme Objet indirect et complément.

Selon le *SF*, ces phrases sont similaires du fait qu'elles dérivent d'un *frame* commun appelé *Commercial_Transaction* et elles expriment le même événement de ce *frame* ; toutefois elles l'évoquent selon différentes perspectives. Dans l'exemple (2) ci-dessus, le cas du verbe *BUY* se focalise sur l'acheteur (*buyer*) et sur la marchandise (*goods*). Par contre, dans l'exemple (1), le verbe *SELL* se focalise sur le vendeur (*seller*) et sur l'argent (*money*). Selon Fillmore (1992), les unités lexicales non seulement sélectionnent des concepts individuels, mais indiquent aussi une perspective selon laquelle le *frame* est évoqué. Le verbe *SELL* affiche la situation depuis la perspective de *seller* et *BUY* affiche la situation depuis la perspective de *buyer*.

Les *frames* décrivant les situations dans la description présentent ce qu'on appelle des prototypes tels que : humain, objets concrets, etc. Par exemple le *frame* *Commercial_Transaction* comporte les prototypes suivants :

vendeur (*buyer*) ou acheteur (*seller*) : humain

marchandises (*goods*) : substances ou objets concrets

paiement (*money*) : argent comptant

2.2.2 FrameNet

FrameNet[9] est une base de données lexicale qui décrit le lexique anglais utilisant le modèle *Frame Semantics* [2]. FrameNet décrit les *Semantic Frames* des unités prédicatives comme les verbes, les noms et les adjectifs. Des versions sont développées pour d'autres langues, tels que FrameNet espagnol développé à l'Université autonome de Barcelone, le FrameNet allemand dans le projet Salsa est développé à l'Université de Saarbrücken [12], le FrameNet japonais développé à l'Université Keio Tokyo. Certains de ces FrameNet sont issus du transfert de FrameNet anglais en utilisant une technique de projection d'informations sémantiques dans des corpus alignés. En chinois, on a fait la projection BiFrameNet avec Hownet une ressource déjà existante pour le chinois [9].

[9] http://framenet.icsi.berkeley.edu/

Dans FrameNet, on définit des *frames* auxquels sont liées les unités lexicales en énumérant les *Frame Elements* (*FE*) de cette unité lexicale. Dans FrameNet, chaque sens d'une unité polysémique appartient à des *Semantic Frames* différents. Par exemple le verbe *ABANDON* dans FrameNet appartient à deux *frames* différents : *Activity-stop* et *Abandonment*.

La définition du *frame Activity_stop* fait intervenir les rôles Agent et Activity. Les *Frame Elements* (*FE*) sont identifiés et définis selon qu'ils sont obligatoires ou optionnels. Agent et Activity sont considérés comme des éléments obligatoires (*core*) du *frame*. Dans le cas du *frame Abandonment,* sa définition fait intervenir les *Frame Elements* (*FE*) *Agent* et *Theme* comme éléments obligatoires. D'autres participants sont considérés optionnels (*non-core*) : on retrouve *Manner*, *Means*, *Time*, *Result*, *Place*, *explanation* etc.

Dans FrameNet, la description d'une unité lexicale est faite au moyen de la description des *frames* auxquels cette unité appartient. Nous décrivons ici l'unité lexicale anglaise verbale *ABANDON*. Le vocable verbal anglais *ABANDON*, en fonction du sens qu'il véhicule, peut appartenir à deux *frames* cités précédemment. Nous présenterons, dans les figures 8 à 14, le traitement de *ABANDON* tel qu'il est décrit dans le *frame Activity_stop*. Nous présenterons d'abord la caractérisation du *frame Activity_stop* (figures 8 à 10), ensuite nous définissons l'unité lexicale *ABANDON* (figures 11 à 14).

2.2.2.1 **Description du frame *Activity_stop***

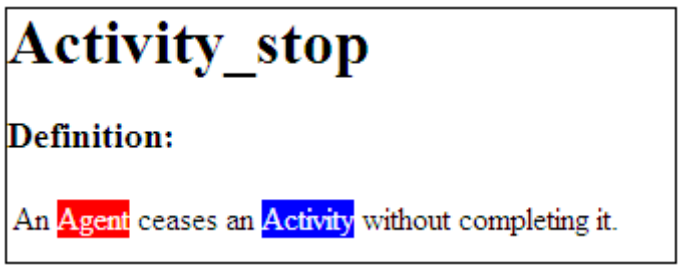


Figure 8 Définition du *frame Activity_stop* Le *FE* Agent porte la couleur rouge et le *FE* Activity porte la couleur bleu. (http://framenet.icsi.berkeley.edu/index.php)

La Figure 8 montre comment est défini le *frame Activity_stop*. Dans cette définition, on constate la mise en évidence des rôles des participants. Le *Frame Element* (*FE*) *Agent* indique le participant qui réalise l'action *Activity_stop* (celui qui stoppe) et le *FE Activity*

indique l'élément qui subit cette action (l'élément stoppé). Dans FrameNet, les *Frame Elements* (*FE*) identifiés sont visualisés et distingués par des couleurs (Figure 8). Des étiquettes sont aussi définis dans FrameNet pour remplacer des les FE qui n'apparaissent pas explicitement et porte les même couleurs que les rôles qu'elles remplacent. L'étiquette *DNI* : *Definite Null Instantiation* indique que le *FE* n'apparaît pas explicitement dans la phrase et qu'il est porté par le discours ou le contexte. FrameNet utilise deux autres types d'étiquettes que DNI, pour annoter les anaphores et les ellipses. Elles sont *CNI* : *Constructional Null Instanciation*, et *INI* : *Indefinite Null Instantiation*

Ces types de FE sont décrits dans la partie des *Frame Elements* (FE), juste au-dessous de la partie de la définition de la Figure 8. Les *FE* obligatoires sont décrits à la (

Figure 9 Description des *FE* obligatoire).

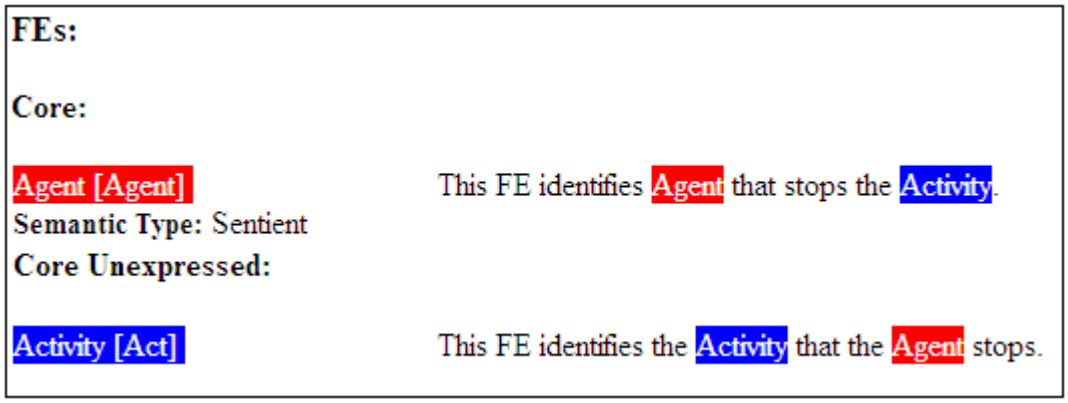


Figure 9 Description des *FE* obligatoires

Dans la partie gauche de la Figure 9, on cite le nom du rôle des *FE* qui participent à la réalisation du frame tels que *Agent* et *Activity*. Certains FE sont accompagnés de leurs types sémantiques, par exemple *Sentient* pour *Agent*. Tandis que, dans la partie droite, on explique ce que signifient ces rôles en donnant une explication explicite et parfois accompagnée d'exemples.

La description de ces *FE Core* est suivie par la description des *FE Non-Core* à la Figure 10. Dans cette dernière, nous reprenons de FrameNet seulement la description d'une partie des *FE* donnée pour ce *Frame*.

Dans la partie gauche de la Figure 10, on décrit les *FE* optionnels dont les noms sont *Area*, *Cothem*, *Distance*, etc. Certains rôles sont accompagnés d'un type sémantique tel que State_of_affairs pour le rôle *Means*, Speed pour le rôle *Speed* et Time pour le rôle *Time*. Tandis que, dans la partie droite, pour chaque rôle on définit son sens et on présente des exemples d'utilisation où on explicite les réalisations de ces *FE* (Figure 10).

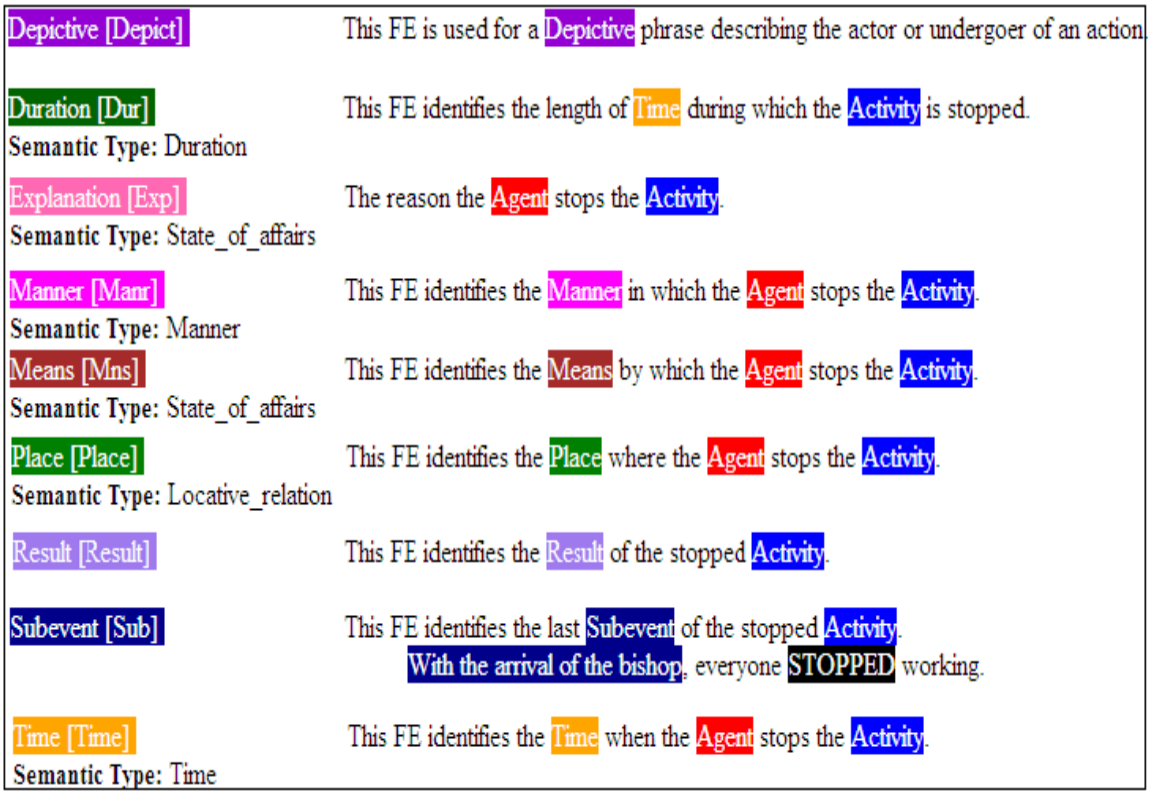


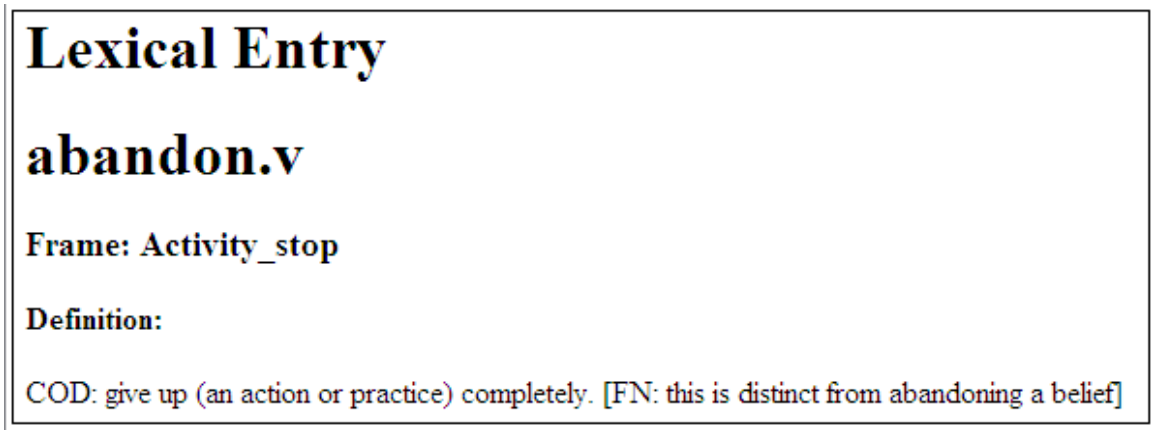
Figure 10 Description de certains *FE* optionnels du *frame Activity_stop* FrameNet utilise des couleurs pour visualiser les rôles identifiés dans leurs définitions. Toutes les réalisations de ces *FE* portent la même couleur qu'eux ou de leurs rôles correspondants. Par exemple la réalisation qwith the arrival of the bishop porte la couleur bleu nuit de son rôle Subevent. (http://framenet.icsi.berkeley.edu/index.php)

2.2.2.2 **Description de l'unité lexicale ABANDON dans le *frame Activity_stop***

Comme nous l'avons vu précédemment, l'unité lexicale verbale *ABANDON* appartient à deux *frames* différents. Nous donnerons en exemple sa description dans le *frame* Activity_stop qui a été décrit à la section précédente. Nous expliquons au fur et à mesure sa description dans ce frame par les figures ci-dessous (figures 11 à 14).

1. Définition

Chaque unité lexicale est donnée par son nom suivi de sa partie du discours : ABANDON.V dans la Figure 11 est une unité lexicale verbale. Comme une unité lexicale peut appartenir à plusieurs *frames*, on doit préciser le nom du *frame* par rapport auquel on fait la description, ce qui est indiqué dans la Figure 11 par « *frame* : *Activity_stop* ». Ensuite, on définit le sens de l'unité lexicale en question. Dans FrameNet, on utilise soit une définition tirée du Dictionnaire Oxford (notée par COD : Concise Oxford Dictionary) ou soit celle donnée par l'équipe de FrameNet (notée par FN).



Lexical Entry

abandon.v

Frame: Activity_stop

Definition:

COD: give up (an action or practice) completely. [FN: this is distinct from abandoning a belief]

Figure 11 Définition de l'entrée ABANDON

2. FE et leurs réalisations syntaxiques

On a vu précédemment que la description d'une unité dans un *frame* suit la définition de ce dernier et les rôles qui lui appartiennent. Donc, les *FE* pour le sens de l'unité lexicale *ABANDON* du *frame Activity_stop* sont donnés dans la Figure 12.

Frame Element	Number Annotated	Realization(s)
Activity	(6)	NP.Ext (3) NP.Obj (3)
Agent	(6)	CNI.-- (3) NP.Ext (3)
Explanation	(1)	2nd.-- (1)
Time	(2)	Sub.Dep (1) PP[before].Dep (1)

Figure 12 FE de ABANDON et leurs réalisations syntaxiques

Tel que décrit précédemment, le *frame* *Activity_stop* renferme les *Frame Elements* obligatoires *Activity* et *Agent* et les *Frames Elements* optionnels *Explanation* et *Time*. Ainsi, dans la table de la Figure 12, on donne les informations syntaxiques de ces *FE* indiqués par *Realizations* de la dernière colonne de la table. Ces informations sont : NP.Ext et NP.Obj pour le rôle *Activity* et CNI, et NP.Ext pour le rôle *Agent*. La réalisation syntaxique notée par NP.Obj indique que le *FE* est de type Nominal et qu'il joue le rôle d'Objet tandis que NP.Ext indique que le *FE* est de type nominal et qu'il joue le rôle de Sujet. CNI indique que le *FE* *Agent* est omis et qu'il est porté par la construction syntaxique. PP[before].Dep pour le FE optionnel *Time* indique que la réalisation syntaxique est de type syntagme prépositionnel et la préposition est *before*, On constate aussi dans cette dernière colonne des chiffres ou nombres indiquant le nombre d'apparitions des *FE*. Ainsi le *FE* dont le rôle *Activity* est réalisé par le type syntaxique NP.Ext est réalisé 3 fois et NP.Obj est réalisé 3 fois[10]. Le *FE* *Agent* est rencontré 3 fois par la réalisation NP.Ext et 3 fois par CNI.

3. Tableau de valence des FE Cores

Dans le tableau de valence on revoit les informations syntaxiques et la valence des patrons *Activity* et *Agent* dans des exemples d'emplois. Ainsi le tableau de la Figure 13 indique que : le FE *Activity* avec le type syntaxique NP.Ext apparait 1 fois avec le FE *Agent* omis et indiqué par CNI. *Activity* avec le type NP.Obj apparait 3 fois avec *Agent* de type NP.Ext. Au total on a annoté 4 exemples d'emplois où on a extrait ces occurrences de patrons. On a annoté aussi les occurrences de ces FE avec les FE optionnels *Explanation* et *Time*.

[10] Ces chiffres proviennent des phrases annotées par les lexicographes

Number Annotated	Patterns			
4 TOTAL	Activity	Agent		
(1)	NP Ext	CNI --		
(3)	NP Obj	NP Ext		
1 TOTAL	Activity	Agent	Explanation	Time
(1)	NP Ext	CNI --	2nd --	Sub Dep
1 TOTAL	Activity	Agent	Time	
(1)	NP Ext	CNI --	PP[before] Dep	

Figure 13 Valence des FE

4. Exemples annotés

Toutes les descriptions fournies dans FN sont accompagnées de phrases annotées. Nous donnons certains de ces exemples annotés dans FrameNet pour l'unité lexicale *ABANDON* (Figure 14).

- T-Wplan,project,commitment-(1)
  1. He also gave an assurance that he was ABANDONING plans to alter the Constitution so that the executive would be prime ministerial rather than presidential .
  2. The party has ABANDONED policies which made it unelectable in the 1980s .
  3. We have already seen that this may simply have been a way of saying that Wulfstan had secured an undertaking that they were going to ABANDON practices which he found displeasing .
- Wplan,project,commitment-T-(1)
  1. The project was ABANDONED when it proved too complicated and instead they drove to Spain at Easter in 1953 .CNI when it proved too complicated
  2. The commitments have been ABANDONED before they have been voted on .CNI
  3. Approaching Heligoland the weather was obviously unsuitable and so the mission was ABANDONED and the aircraft went home .CNI

Figure 14 Exemples d'emploi d'*ABANDON* dans le *frame* *Activity_stop*

Dans ces phrases, on trouve différents contextes d'utilisation de l'unité lexicale *ABANDON* et les réalisations de ses *Frame Elements* correspondant au schéma du *frame Activity_stop* dont les *Frame Elements core* sont *Agent* et *Activity*. Ces FE se retrouvent dans tous les exemples et dans le cas d'une omission ils sont présentés par l'étiquette du FrameNet « *CNI* ». Par exemple dans les trois dernières phrases, la réalisation de *Agent* est omise et elle est précisée par l'étiquette *CNI* (Figure 14). On peut aussi avoir des FE non obligatoires qui figurent dans l'annotation des exemples. Le FE figure *Time* dans les deux phrases du dernier groupe de la Figure 14.

2.3 Comparaison LEC et *FS*

Les modèles LEC et FS se focalisent tous deux sur la structure actancielle des unités lexicales et en décrivent les propriétés syntaxiques et sémantiques. Par contre, ils présentent des différences quant à la manière d'aborder la notion de structure lexicale. En *Frame Semantics* (FS), la notion de « structure lexicale » passe par celle des « *Frames* », c'est-à-dire des scénarios conceptuels qui fédèrent les unités du lexique. Dans la LEC, la structure lexicale repose sur l'appréhension de tous les liens lexicaux partagés par l'unité lexicale avec les autres unités lexicales du lexique; ces liens sont représentés par le formalisme des fonctions lexicales [37].

Bien que ces deux modèles aient des objectifs linguistiques distincts, Barque (2003) a montré une certaine convergence concernant la représentation du sens lexical. Ainsi, FrameNet, construit sur les principes de *Frame Semantics*, offre une représentation des trois aspects d'une unité lexicale traités par la LEC à savoir : 1) la description du sens; 2) la description de la combinatoire syntaxique; et 3) la description de la combinatoire lexicale [4] (Voir Tableau 2).

Aspects de LEC	Représentation de ces aspects dans FrameNet
Sens	Structures conceptuelles décomposées en *Frame Elements* (*FE*) auxquelles l'unité lexicale est liée.
Combinatoire syntaxique	Réalisation des *FE* dans les phrases dans lesquelles l'unité lexicale apparaît
Combinatoire lexicale	Manière dont les différents *FE* peuvent apparaître dans une même phrase

Tableau 2 FrameNet et correspondances avec certains aspects de la LEC [3]

On remarque dans le Tableau 2 que la représentation FrameNet utilise et s'appuie sur les structures conceptuelles ou ce qu'on appelle *Frame Element* (*FE*) dans les trois aspects de description de sens d'une unité lexicale. La notion « *Frame Element* » est présentée à la section 2.2.

2.4 Conclusion

Dans ce chapitre, nous avons présenté deux modèles de description d'unités lexicales, la LEC et les *FS*, en nous concentrant sur la manière dont ils prennent en compte les liens entre les unités lexicales prédicatives et leurs actants. Nous avons montré l'application de chaque modèle par un exemple et comparé ainsi leur optique dans la définition des éléments participant à la description de l'unité lexicale. Nous avons distingué les *Frame Elements* (*FE*) dans le modèle *FS* et les actants sémantiques dans la LEC qui nous intéressent particulièrement dans notre thèse et qui sont présentés dans le prochain chapitre. Nous y discuterons des rôles sémantiques, de leur définition et de leur importance dans des applications de TAL. Nous présenterons un état de l'art sur les différentes approches d'automatisation de l'annotation de ces rôles.

Chapitre 3 Les actants et les rôles sémantiques

Dans le chapitre précédent, nous avons vu les deux théories linguistiques qui proposent des modes de représentation des éléments d'actants, de circonstants ou de rôles sémantiques. Dans ce chapitre, nous présentons la notion d'actant par opposition à celle de circonstant et la notion de rôle sémantique.

3.1 La notion d'actant par opposition à celle de circonstant

La distinction entre "actant" et "circonstant" que nous présentons ici 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 [29].

Les notions d'actant et de circonstant ont aussi été décrites par Tesnière [65] qui a défini des fonctions subordonnées en actant, circonstant et épithète. La fonction épithète est la fonction principale de l'adjectif qui est à côté du nom (un homme **heureux**). La fonction actant peut être réalisée par un élément sujet, objet ou parfois par un groupe prépositionnel. La fonction circonstant est réalisée par un adverbe ou par un groupe prépositionnel (ex. à ce stade de la phrase ci-dessous) ou par le sujet d'autres verbes en liaison avec la lexie (ex. X permettre à Y de lexie, où X est un circonstant de la lexie). Cette division est compatible avec Tesnière. Exemple :

[Actant, Suj Vous] [Verbe ABANDONNEZ][Actant, Obj l'installation] [Circonstant, Compl à ce stade].

On peut expliciter, au niveau des participants, les liens (actant, circonstant) que ces participants entretiennent avec le verbe prédicatif ABANDONNER. Dans la phrase ci-dessus, la lexie ABANDONNER est accompagnée des réalisations de deux actants sémantiques :

- A1 (Vous) : celui qui abandonne et
- A2 (l'installation) : ce qui est abandonné

Elle apparaît également avec un circonstant :

- A3 (à ce stade) : le moment où A1 abandonne A2.

Au niveau syntaxique, les participants sont réalisés sous forme de Sujet, Objet, Complément, etc. Ainsi dans l'exemple ci-dessus, A1 est Sujet, A2 est Objet et A3 est un Complément.

3.2 Les rôles sémantiques

3.2.1 Définition

Certaines théories linguistiques proposent de rendre compte des actants et circonstants au moyen d'étiquettes de rôles sémantiques. Un rôle sémantique est une relation de sens entre un participant actant ou circonstant avec la lexie prédicative. Un prédicat détermine un rôle sémantique à chaque actant (argument ou participant essentiel) qui assiste la réalisation du sens du prédicat. Par exemple :

ANNULER (N_{Agent}, $N_{Patient}$)

AFFECTER (N_{Agent}, $N_{Patient}$, à $N_{Récipient}$)

STOCKER ($N_{Destination}$, $N_{Patient}$)

D'autres participants périphériques (circonstants) peuvent apparaître dans les contextes de ces prédicats, mais ils sont complémentaires. Ils rajoutent certaines informations qui ne participent pas au sens du prédicat. 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 les unités prédicatives différentes [29]. Comme nous le verrons à la section 3.3, ces participants occupent des positions syntaxiques dans une phrase. À ces positions, correspondent des fonctions grammaticales qui peuvent avoir le même rôle sémantique ou des rôles sémantiques différents.

3.2.2 Théories case grammar et FS

Les travaux fondateurs de la grammaire de cas (« *case grammar* ») [14] ont soulevé un intérêt considérable pour l'étude de la nature des rôles sémantiques. Cette théorie a été

modifiée par Fillmore pour remédier à certaines de ses limites. Il a construit ainsi la théorie *Frame Semantics* et la ressource lexicale FrameNet (2.2.2) basée sur cette théorie. Les rôles sémantiques définis y sont plus spécifiques que ceux définis dans la grammaire des cas. Les rôles originaux définis dans la grammaire des cas, sont : cas Agentif (cas A), cas Objectif (cas O), cas Instrumental (cas I), cas Datif (cas D) et cas locatif (cas L). Par exemple[11] [19]

[cas Agentif John] opened [cas Objectif the door] with [cas Instrumntal a chisel]

Dans la grammaire des cas, la liste de rôles sémantiques est restreinte, causant ainsi des limites dans la caractérisation des prédicats. Exemples[12] :

[Agent The teacher] gave [Theme a book] [Recipient to the student]

[Recipient The student] received [Theme a book] [Source from the teacher]

La grammaire des cas ne se rend pas compte que le donneur The teacher a les propriétés à la fois de Agent et de Source. Fillmore, dans ses travaux ultérieurs sur la sémantique lexicale, confirme sa conviction qu'une liste restreinte de cas est insuffisante pour caractériser les propriétés de la complémentation des unités lexicales [19]. Il explicite les détails des relations du verbe avec ses participants sans réduire le vocabulaire de rôles sémantiques à un ensemble restreint, car son objectif est d'élaborer la richesse et la diversité des relations sémantiques des verbes pour clarifier la nature de leurs actants [56]. Dans cet article les chercheurs débattent de l'existence de rôles sémantiques et ainsi que sur une liste définitive de rôles. En *Frame Semantics*, les rôles sémantiques essentiels sont appelés *Frame Elements core* et ceux qui sont optionnels sont appelés *Frame Elements non-core*.

Il existe différentes listes de rôles sémantiques (aucune ne semblant s'imposer comme norme), mais il ressort, que peu importe le modèle utilisé, l'annotation en rôles sémantiques des actants et des circonstants permet d'interpréter sémantiquement les contextes d'une lexie prédicative en identifiant « qui » fait « quoi » à « qui », « où », « quand », « comment », « pourquoi ». Ceci implique que, dans une phrase, il faut d'abord

[11] Exemple extrait du livre « the case for case » de Fillmore et autres
[12] Exemples extraits de l'article Background to FrameNet de Fillmore et les autres [19].

identifier les participants porteurs des rôles ensuite leur assigner des étiquettes de rôles sémantiques. Par exemple dans :

Ce garçon MANGE chaque matin une pomme dans le bus

le prédicat verbal est MANGER, ses participants identifiés sont : ce garçon qui répond à la question « qui? », chaque matin à « quand? », une pomme à « quoi? » et dans le bus à « où ?». Une fois ces participants identifiés, nous leur assignons une étiquette sémantique comme Agent, Patient, Lieu, Temps, etc. Dans cette même phrase, nous avons : ce garçon qui porte l'étiquette Agent, pomme, l'étiquette Patient, chaque matin, l'étiquette Temps et dans le bus, l'étiquette Lieu. Nous représentons cette phrase avec toutes ces descriptions, en nous inspirant de la notation de Palmer et Gildea dans leur ouvrage sur les rôles sémantiques [56], comme suit :

[$_{\text{Agent}}$ Ce garçon] [$_{\text{Prédicat}}$ MANGE] [$_{\text{Temps}}$ chaque matin] [$_{\text{Patient}}$ une pomme] [$_{\text{Lieu}}$ dans le bus]

Chaque participant est mis entre crochet avec à sa gauche ses caractéristiques sémantiques. Dans notre document, nous intégrant aussi les informations syntaxiques s'il y a lieu.

3.3 Fonctions grammaticales et rôles sémantiques

Comme nous l'avons déjà souligné, les participants en plus d'avoir un lien sémantique avec la lexie prédicative entretiennent également un lien de nature syntaxique. La correspondance entre les fonctions grammaticales et les rôles sémantiques est illustrée dans des exemples[13] des différents cas suivants, cités par Habert [28] :

1) Une même fonction grammaticale peut correspondre à plusieurs rôles sémantiques :

[$_{\text{Agent}}$ Jean] reçoit son ami

[$_{\text{Patient}}$ Jean] reçoit un livre sur la tête

[$_{\text{Bénéficiaire}}$ Jean] reçoit un livre par la poste

[13] Exemples donnés par Benoît Habert dans son cours « de la linguistique à la sémantique » [28]

Dans ces exemples, Jean occupe la fonction grammaticale sujet, mais il est associé à des rôles sémantiques différents (Agent, Patient, Bénéficiaire) selon la relation sémantique entre le prédicat verbal RECEVOIR.

2) Un même rôle sémantique peut être réalisé par des fonctions grammaticales différentes. Par exemple :

La cuisson [$_{\text{Prédicat}}$ dore] [$_{\text{Patient}}$ le gâteau]

[$_{\text{Patient}}$ Le gâteau] [$_{\text{Prédicat}}$ dore]

Dans la première phrase, le gâteau est objet ; dans la deuxième phrase, il est sujet. Mais dans les deux phrases, il joue le rôle de Patient.

3) Les participants gardent les relations qu'ils entretiennent avec le prédicat même s'ils changent de position syntaxique dans leurs emplois avec ce même prédicat. Exemples :

[$_{\text{Agent}}$ Jean] ouvre [$_{\text{Patient}}$ la porte] [$_{\text{Instrument}}$ avec la clé]

[$_{\text{Instrument}}$ La clé] ouvre [$_{\text{Patient}}$ la porte]

[$_{\text{Patient}}$ La porte] s'ouvre

ou en anglais [$_{\text{Patient}}$ the door] opens. Le verbe ne présente pas de signe de réflexion.

4) Dans les emplois de la voix active et passive la relation qu'entretiennent les participants avec leurs prédicats verbaux n'est pas modifiée. Leurs fonctions syntaxiques changent mais pas leurs rôles sémantiques. Exemples :

[$_{\text{Agent}}$ Jean] a écrit [$_{\text{Résultat}}$ ce livre]

[$_{\text{Résultat}}$ Ce livre] a été écrit par [$_{\text{Agent}}$ Jean]

5) Dans d'autres cas, un participant dans une même phrase peut relever de deux rôles sémantiques différents. Ce qui rend la phrase ambiguë. Exemple :

[$_{\text{Agent ou Instrument}}$ Ce guide] nous a renseignés

6) Deux prédicats différents peuvent mettre en relation différemment leurs participants. C'est le cas du prédicat VENDRE et du prédicat ACHETER. Le sujet et l'objet de l'un deviennent respectivement l'objet et le sujet de l'autre. Ainsi le sujet vendeur de VENDRE devient vendeur objet de ACHETER et objet acheteur de VENDRE devient acheteur sujet de ACHETER. Pour mieux expliciter cette description, nous allons avec des rôles plus spécifiques : vendeur et acheteur. Exemple :

[$_{\text{Vendeur, sujet}}$ Jean] a vendu sa voiture à [$_{\text{Acheteur, objet}}$ Pierre]

[$_{\text{Acheteur, sujet}}$ Pierre] a acheté sa voiture à [$_{\text{Vendeur, objet}}$ Jean]

Cette variété dans les différentes fonctions syntaxiques que peuvent occuper les actants viendra compliquer notre tâche d'annotation automatique.

3.4 Exemples d'annotation dans le corpus de l'informatique et de l'Internet

Nous avons déjà présenté des exemples de rôles sémantiques des unités prédicatives de la langue générale. Ici, nous présentons des exemples de phrases dans le domaine de spécialité de l'informatique et de l'Internet sur lesquelles notre projet est basé. Les annotations sont articulées autour des rôles sémantiques des participants (actants et circonstants) d'un terme prédicatif de l'informatique. Le jeu d'étiquettes de rôles sémantiques (en construction) vise à rendre compte de la manière la plus générale possible du lien sémantique entre un terme d'informatique prédicatif et un participant. Ils servent également à mettre en évidence les alternances, les dérivés morphologiques avec sens apparentés et les sens voisins [29].

Alternances

Dans ce cas, quelle que soit la position syntaxique d'un participant d'une lexie donnée, son rôle sémantique reste inchangé. Exemple :

[$_{\text{Instrument}}$ Une imprimante] imprime [$_{\text{Patient}}$ un fichier]

[$_{\text{Agent}}$ Un utilisateur] imprime [$_{\text{Patient}}$ un fichier][$_{\text{Instrument}}$ avec une imprimante]

Dans le DicoInfo, le sens de IMPRIMER dans la première phrase est représenté par IMPRIMER1a décrit par « Instrument ~ Patient » et son sens dans la deuxième phrase est donné par IMPRIMER1b décrit par « Agent ~ Patient avec Instrument».

Dérivés morphologiques avec sens apparentés

Les participants réalisés en actants par une lexie verbale, exemple « programmer », sont réalisés en actants ayant les mêmes rôles sémantiques que ceux d'un dérivé, exemple « programmation ». Cette lexie et son dérivé ont des sens apparentés. Exemple :

[$_{\text{Agent}}$ Un informaticien] programme [$_{\text{Matériau}}$ en Java]

Programmation [$_{\text{Patient}}$ du logiciel] [$_{\text{Matériau}}$ en Java] par [$_{\text{Agent}}$ un informaticien]

Dans le DicoInfo, la structure actancielle de la lexie verbale PROGRAMMER 1, correspondant au sens employé dans la première phrase, est décrite par « Agent ~ Patient en Matériau». La structure actancielle de la lexie nominale PROGRAMMATION 1, est décrite par « ~ de Patient en Matériau par Agent».

Sens voisins

Des lexies différentes ayant des sens voisins peuvent également partager des actants ayant les mêmes rôles sémantiques. Exemple :

[$_{\text{Assaillant}}$ Un pirate] s'ATTAQUE [$_{\text{Destination}}$ aux Macs]

[$_{\text{Assaillant}}$ Les vers] INFECTENT [$_{\text{Destination}}$ les fichiers]

[$_{\text{Assaillant}}$ Les virus] s'INTRODUISENT [$_{\text{Destination}}$ dans l'ordinateur]

Dans ces trois phrases, nous avons des termes prédicatifs qui ont des sens voisins décrits par la structure actancielle «Assaillant ~ Destination ». Cette dernière correspond à un sens de ATTAQUER (ATTAQUER1). Elle correspond également au sens 1 de INFECTER (INFECTER1) et au sens 1 de INTRODUIRE (INTRODUIRE1).

Les étiquettes de rôles sémantiques auxquelles nous avons recours dans le corpus de l'informatique et de l'Internet 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*).

L'assignation de rôles sémantiques aux participants permet aussi de capturer la similarité sémantique entre les langues même si leur structure syntaxique est différente [1]. Exemple de PRINT (anglais), IMPRIMER (français) et 인쇄하다 (coréen) dans :

PRINT 1b (Agent PRINTS Patient with Instrument)

인쇄하다 1b (Agent - 이 Patient - 을 Instrument - 로 인쇄하다)

IMPRIMER 1b (Agent IMPRIME Patient avec Instrument)

Les lexies verbales PRINT, 인쇄하다 et IMPRIMER ont une même structure actancielle même si leur comportement syntaxique est différent. Elles ont le même nombre d'actants annotés par les mêmes rôles sémantiques. Donc les rôles sémantiques peuvent être un outil efficace de capturer la similarité entre les langues indépendamment des phénomènes de surface qui peuvent les différencier [1].

3.5 Rôles sémantiques et applications TAL

L'annotation automatique des structures actantielles joue un rôle important dans les applications du traitement automatique des langues naturelles telles que la recherche d'informations, l'extraction d'informations, les questions/réponses et le résumé automatique. Ainsi Gerhard Fliedner a montré que l'utilisation d'annotations automatiques améliore les performances de ces applications et propose des outils de construction d'un lexique d'annotations sémantiques [22]. Certaines applications sont présentées brièvement ci-dessous.

3.5.1 Rôles sémantiques et traduction automatique

Boas [5] montre l'utilité du projet lexicographique de FrameNet de l'allemand construit en se basant sur le FrameNet de l'anglais. L'objectif du FrameNet allemand est de décrire des milliers d'unités lexicales de l'allemand au moyen des Frames sémantiques utilisés pour les unités lexicales correspondantes de l'anglais. Les informations de la combinatoire syntaxique et sémantique ainsi que les frames sémantiques des unités lexicales de l'allemand et de l'anglais permettent de déterminer la traduction équivalente d'une unité lexicale d'une langue à une autre. Par exemple, dans le *frame Communication-Statement*, l'unité lexicale verbale *ANNOUNCED* a plus d'un équivalent en allemand dans le même frame : *bekanntgeben*, *bekanntmachen*, *ankündigen*, *anzigen*, *ansagen* et *durchsagen*. Nous constatons la résolution de la polysémie, par exemple, en observant la combinatoire syntaxique et sémantique.

[$_{\text{Speaker}}$ They] ANNOUNCED [$_{\text{Message}}$ the birth of their child]

[$_{\text{Medium}}$ The document] ANNOUNCED [$_{\text{Message}}$ that the war had begun]

[$_{\text{Speaker}}$ The conductor] ANNOUNCED [$_{\text{Message}}$ the train's departure] [$_{\text{Medium}}$ over the intercom]

Dans le cas où le verbe anglais *ANNOUNCE* apparaît avec seulement le Speaker et Message comme dans le premier exemple, les verbes *ansagen* et *durchsagen* ne sont pas considérés comme ses traductions équivalentes en allemand (parce que ces deux verbes allemands sont utilisés avec Medium qui représente un équipement électronique servant à transmettre un message au destinataire, comme dans le troisième exemple). Ainsi, on peut dire que le sens général des mots n'est pas suffisant pour chercher les équivalents d'une langue à une autre. La spécification de la combinatoire syntaxique et sémantique est nécessaire pour distinguer finement les traductions équivalentes.

On peut aussi avoir des traductions équivalentes à travers des *frames* multiples. Par exemple les différents sens du verbe anglais *walk* se trouvent dans deux *frames* différents : *self-motion* et *cotheme-motion*. Ces deux sens sont exprimés des unités lexicales différentes en

allemand. Le verbe *walk* dans le frame *Self-motion* est exprimé en allemand par son équivalent *gehen* et dans le frame *Cotheme-motion* par son équivalent *begleiten*. Exemples :

[$_{\text{Self-mover}}$ Kim] WALKED [$_{\text{Goal}}$ to the store]

[$_{\text{Self-mover Kim}}$] WALKED [$_{\text{Cotheme}}$ Pat] [$_{\text{Goal}}$ to the store]

Les traductions équivalentes en allemande de ces deux phrases sont :

[$_{\text{Self-mover}}$ Kim] GING [$_{\text{Goal}}$ zum Geschäft]

[$_{\text{Self-mover}}$ Kim] BEGLEITETE [$_{\text{Cotheme}}$ Pat] [$_{\text{Goal}}$ zum Geschäft]

Dans ces phrases, nous constatons qu'en anglais la lexie *walk* exprime deux sens différents. La distinction de ces deux sens se fait par son appartenance à deux *frames* différents (*Self-mover* et *Cotheme*). Dans le frame *Self-mover*, on trouve les verbes tels que march, parade, promenade etc. et dans le frame *Cotheme* on retrouve accompagny, guide, follow, etc. en d'autres langues, les deux sens du verbe walk peut se faire en utilisant deux verbes différent. Par exemple walk dans la première phrase qui veut dire que kim va à pied au magasin, en allemand le verbe gehen est utilisé. Quand à la deuxième phrase où le verbe walk est employé dans le sens d'accompagner quelqu'un est traduite en allemand en utilisant le verbe begleten qui signifie « accompagner ». La connaissance du *frame* et de la structure actancielle accompagnée de rôles sémantiques d'un verbe permet de choisir le verbe correspondant à sa traduction dans une autre langue.

3.5.2 Rôles sémantiques et résumé automatique

Des techniques proposées par les chercheurs dans le domaine des résumés automatiques sont de deux types : extraction et abstraction [61]. L'extraction consiste à sélectionner des phrases du texte d'origine ayant des scores élevés et à les assembler pour former un texte plus court. L'abstraction utilise des méthodes linguistiques pour interpréter des textes. Suanmali [61] opte pour la méthode d'extraction en proposant d'utiliser les rôles sémantiques. Ils proposent une méthode de résumé qui se base sur les méthodes statistiques et le calcul de similarité entre les phrases en utilisant les rôles sémantiques. Rouge-(1,2 et L) montre une augmentation de F-mesure de 0.05 avec la combinaison des méthodes basées

sur les statistiques et les rôles sémantiques. En 2005, Melli et al. ont aussi utilisé les rôles sémantiques dans leur système de résumé automatique (SQUASH, proposé à DUC-2005 (*Document Understanding Conference* 2005)). Ils ont proposé d'intégrer le module de rôles sémantiques dans la sélection et la compression de phrase [47].

3.5.3 Rôles sémantiques et systèmes question/réponse

Dans le système question/réponse de Narayanan [51], les questions sont analysées et les réponses sont générées par l'identification de la structure prédicative (prédicat-argument) et les frames sémantiques de l'entrée et l'amélioration de la structure d'inférence probabiliste en utilisant les relations extraites dans le contexte du domaine et le modèle des scénarios. L'innovation de leur système est une représentation évolutive et expressive d'actions et d'événements basés sur la coordination probabiliste des modèles relationnels.

Narayanan donne un exemple de transformation de la question Q en structure argumentale prédicative et en structure de frame. Dans Q, les différents arguments sont annotés par les étiquettes de PropBank tels que Arg1, Arg2 et par des rôles sémantiques de FrameNet tels que Goods et Victim de l'exemple :

> Q : What [$_{Arg1,Goods}$ kind of nuclear materials] were [$_{Predicat}$ STOLEN] [$_{Arg2,\ Victim}$ from the Russian navy]?

Les réponses candidates à la question Q sont annotées de la même manière que Q par des étiquettes argumentales et de rôles sémantiques. L'exemple montre la réponse A correspondante à Q annotée par des informations sémantiques ; A1 la réponse annotée par des étiquettes de PropBank ; et A2 la réponse annotée par des rôles sémantiques de FrameNet.

> A: Russian's Pacific Fleet has also FALLEN prey to nuclear theft; in 1/96, approximately 7 kg of HEU was STOLEN from a naval base in Sovetskaya Gavan.

> A1 :[$_{Arg1(P1)}$ Russian's Pacific Fleet] has [$_{ArgM-Dis(P1)}$ also] [$_{Predicat(P1)}$ fallen] [$_{Arg1(P1)}$ prey to nuclear theft]; [$_{ArgM-TMP}$ in 1/96], [$_{Arg1(P2)}$ approximately 7 kg of HEU] was [$_{ArgM-ADV(P2)}$ reportedly] [$_{Predicat(P2)}$ stolen] [$_{Arg2(P2)}$ from a naval base] [$_{Arg3(P2)}$ in Sovetskaya Gavan]

A2:[$_{\text{Victim}}$ Russian's Pacific Fleet] has also [$_{\text{Predicat(P1)}}$ fallen] prey to [$_{\text{Goods}}$ nuclear] [$_{\text{Predicat(P1)}}$ theft]; in 1/96, [$_{\text{Goods(P2)}}$ approximately 7 kg of HEU] was [$_{\text{Predicat(P2)}}$ stolen] [$_{\text{Victim(P2)}}$ from a naval base] [$_{\text{Source(P2)}}$ in Sovetskaya Gavan]

La réponse exacte à Q est ***approximately 7 kg of HEU***. Narayanan montre, par sa proposition de l'architecture de QA en incluant les informations sémantiques, que le pourcentage d'identification du thème de la question est augmenté.

3.6 Ressources lexicales et rôles sémantiques

Des ressources lexicales, comportant des informations sur la structure actancielle et les rôles sémantiques, ont été créées et ont attiré l'attention des linguistes et des chercheurs en TAL. La ressource FrameNet [20] est un lexique électronique et aussi un corpus annoté par des rôles sémantiques (voir 2.2.2). Les scénarios conceptuels sont capturés et sont définis dans des *frames*. Ces *frames* comportent des éléments, appelés *frames Elements*, dont les étiquettes évoquent des rôles sémantiques. Les unités lexicales sont identifiées et classées dans des *frames* et des exemples d'emplois de ces unités lexicales, tirés du corpus « *British National Corpus* » sont annotés par ces rôles sémantiques définis dans les *frames* comportant ces unités. Plusieurs chercheurs ont utilisé cette ressource pour développer des approches automatiques d'annotation de rôles sémantiques tels que Palmer, Gildea, Pradhan etc. [56, 24, 57]. D'autres chercheurs ont utilisé cette ressource pour créer une ressource équivalente dans d'autres langues telles que Salsa pour l'allemand, FrameNet pour l'espagnole, pour le chinois etc.

Les verbes du corpus Treebank ont été annotés par des rôles sémantiques par le projet de PropBank [55]. La difficulté principale rencontrée par les chercheurs est la définition d'une liste de rôles [56]. Dans PropBank, des arguments sont définis notés Arg0, Arg1…, et des autres de temps et de location notés ArgTMP, ArgLoc. PropBank est inspiré des recherches sur le lexique VerbNet (Kipper, Dang, et Palmer 2000). VerbNet prolonge les classes de Levin's (1993). Il définit des classes de verbe qui partagent les mêmes structures argumentales et les rôles qui sont assignés à cette structure. En 2004,

cette ressource a été utilisée dans l'annotation de rôles sémantiques par Swier et Stevenson [63].

Pour le français, des ressources de ce type sont rares ou ne sont pas actuellement disponibles. Le problème s'accroît dans les domaines de spécialité.

3.7 Conclusion

Dans ce chapitre, nous avons présenté et discuté la notion de rôle sémantique. La structure actancielle peut avoir une variété de réalisations syntaxiques et l'annotation de cette structure en rôles sémantiques permet d'unifier la représentation de relations sémantiques entre les unités lexicales prédicatives et leurs participants (actants et circonstants). Cette annotation met en évidence les alternances, les dérivés morphologiques avec sens apparentés et les sens voisins. Nous avons montré des exemples d'intégration d'informations de rôles sémantiques pour améliorer la performance de systèmes de TAL (voir 3.5).

Dans les ressources faisant appel aux rôles sémantiques, l'intégration des informations sémantiques a été faite manuellement. Dans FrameNet, plusieurs *frames* et un nombre important de contextes ou d'exemples ont été annotés manuellement avec des rôles sémantiques. Cette annotation manuelle exige énormément d'effort humain et de temps, d'où l'intérêt de l'automatiser. Plusieurs travaux ont été réalisés dans cette optique d'automatisation de rôles sémantiques, en se basant sur les ressources citées en section 3.6 comme corpus de données. Le chapitre suivant sera consacré à l'ensemble de ces travaux.

Chapitre 4 État de l'art en identification et annotation d'actants

Plusieurs travaux antérieurs, particulièrement pour l'anglais, se sont intéressés à l'automatisation de l'identification des participants actants et leur annotation en rôles sémantiques. Dans ce chapitre, nous présentons les travaux réalisés dans le cadre de chaque tâche.

4.1 Identification des actants

Plusieurs critères d'identification des actants ont été suggérés en linguistique théorique [44] :

1) Le critère obligatoire et optionnel

Selon ce premier critère, il ne peut y avoir dans une phrase un verbe sans ses actants qui l'accompagnent. Ces actants doivent nécessairement être présents avec tout verbe contrairement aux circonstants qui peuvent être absents. Les actants sont dits obligatoires et les circonstants sont dits optionnels.

Le critère obligatoire et optionnel n'est pas toujours facile à appliquer lors de l'analyse de phrases réelles. Il y a bien des cas où l'actant n'est pas réalisé pour un verbe. Il peut être absent. Par exemple :

[Actant, Sujet Jean] [Verbe joue] [Actant, Complément absent] [Circonstant Complément dans la cour]

Dans cette phrase, l'actant qui occupe la fonction grammaticale complément est absent. Il correspond « à quoi jean joue » (ex. au ballon). Le sens de jouer est rempli malgré que son actant complément ne soit pas réalisé.

2) Le critère de la mobilité de position

Ce critère veut que seuls les circonstants puissent changer de position dans la phrase. Or, on trouve des cas où les actants peuvent être déplacés, comme le remarque Anne Lacheret-Dujour [39] dans son exemple :

[Actant, Objet Le chocolat] , [Actant, Sujet j'] [Verbe adore].

3) Les sujets et compléments d'objet direct correspondent généralement à des actants, tandis que certains compléments introduits par des prépositions peuvent correspondre soit à des actants soit à des circonstants. Exemples :

[Actant Le processus] [Verbe ABANDONNE] [Actant cette tâche] [Prep à] [Circonstant ce moment là]

[Actant On] [Verbe AFFECTE] [Actant un poste] [Prep à] [Actant une personne]

Les critères ci-dessus sont extrêmement difficiles à automatiser. Donc, nous avons recours aux corpus et aux calculs de fréquences dans ces corpus qui peuvent jouer un rôle important dans la distinction entre actants et circonstants [66]. Cécile Fabre et Cécile Frérot proposent une combinaison de deux mesures de productivité : productivité recteur-préposition qui calcul le nombre de régis différents que le couple (verbe, préposition) gouverne et productivité préposition-régi qui calcule le nombre de verbes différents qui gouvernent le couple (préposition, régi) pour distinguer sur corpus, au sein des groupes prépositionnels rattachés au verbe, des types de compléments différents [13].

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[14] tels que recensés par Dicovalence [52].

Certaines méthodes proposées pour l'annotation des unités lexicale en anglais s'appuient sur les informations syntaxiques pour identifier la structure argumentale d'un prédicat verbal. Dans ce cadre des travaux, Gildea et Jurafsky ont proposé l'extraction de features[15]syntaxiques qui permettent d'identifier les participants d'un prédicat [24]. Richard Johansson et Pierre Nugues ont proposé d'utiliser les relations de dépendance syntaxique comme un trait [33]. D'autres travaux se basent sur des informations syntaxiques : Surdeanu [62] a utilisé le corpus TreeBank et PropBank ainsi que Li [41] , Che [8],

[14] 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.

[15] 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.

Täckström [64] et Maria Liakata et al. [42]. Ces travaux proposent des modèles pour identifier les participants d'un prédicat verbal.

4.2 Approches automatiques d'annotation de rôles sémantiques

Plusieurs techniques ont été explorées pour induire automatiquement des rôles sémantiques : méthodes d'apprentissage non supervisé employant des informations lexicales pour développer le classificateur ou méthodes d'apprentissage supervisé se basant sur les données d'entraînement extraites de PropBank, utilisé comme corpus, dont les arguments sont annotés (Arg0,..,Arg5,ArgLoc, etc.). La ressource FrameNet a aussi servi comme corpus de données car les *Frame Elements*, équivalents aux rôles sémantiques, avec leur valence et leurs réalisations syntaxiques sont décrits. Les approches d'étiquetage de rôles sémantiques sont principalement basées sur deux tâches fondamentales : identification des arguments des prédicats et leur classification [43]. Dans tous ces travaux, les auteurs n'ont pas distingué entre actants et circonstants. Ils parlent d'arguments des prédicats, terme que nous utilisons dans ce chapitre portant sur l'état de l'art. Pour réaliser ces deux tâches d'identification et de classification des arguments, l'analyse syntaxique est importante pour calculer et déduire les rôles sémantiques [59].

L'approche d'apprentissage supervisé est étudiée dans la plupart des systèmes d'annotation sémantique. Plusieurs auteurs ont assigné des rôles sémantiques aux constituants donnés par le parseur et en ont extrait des traits sur lesquels se base le classificateur. Les traits utilisés sont ceux définis par Gildea et Jurafsky; néanmoins, le mécanisme de classification diffère d'un auteur à l'autre. Le travail de Gildea & Jurafsky utilise FrameNet pour construire un classifieur sémantique basé sur un modèle statistique. Ils divisent le problème en deux sous-tâches : identification de *Frame Elements* et classification de *Frame Elements*. Une étude similaire à celle de Gildea [69] a été faite avec le modèle « *log-linear* ». Une autre étude semblable [21] a fait appel au modèle de Maximum d'Entropy et les deux tâches d'identification et de classification de Frame Elements (*FE*) ne sont pas séparées.

Surdeanu et al. [62] ont construit un système utilisant un classifieur basé sur un arbre de décision en ajoutant certains traits additionnels aux traits déjà définis par Gildea. Un autre système est construit en se basant sur la grammaire catégorielle combinatoire (CCG) pour extraire les traits [23]. D'autres utilisent la classification par « *support vector machine* » [57] et [50]. Tous ces travaux se basent sur l'emploi d'un arbre syntaxique produit par un parseur. Gildea et Palmer ont montré que l'utilisation d'arbres syntaxiques améliore les performances et que ces arbres syntaxiques sont nécessaires pour l'identification de prédicat-arguments [25]. Le système décrit par Johansson et Nugues est basé sur les dépendances syntaxiques [32]. Xue utilise des parseurs pour étiqueter les prédicats en chinois en attribuant à leurs arguments des rôles sémantiques [68]. Un modèle de réseau de neurones a été proposé dans l'extraction sémantique qui permet d'apprendre à partir de la phrase source les étiquettes sémantiques sans avoir à utiliser aucun outil d'analyse syntaxique ni d'arbres syntaxiques sauf les catégories grammaticales [10].

Tous ces travaux utilisent des ressources, telles que FrameNet ou PropBank[16] afin d'annoter sémantiquement les phrases. Par exemple, Moschitti et Adrian ont construit un classificateur de prédicats-arguments basé sur les arbres syntaxiques annotés par l'information prédicat-argument du PropBank [49].

En SemEval-2007[17], les tâches 17 et 19 ont été consacrées aux rôles sémantiques. Une sous tâche de la tâche 17 [58] consistait à annoter les rôles sémantiques des unités lexicales en anglais. L'entraînement et le test comportent les mêmes unités lexicales mais avec des instances différentes. Des ressources ont été fournies aux participants à cette compétition : les annotations de PropBank, les classes correspondantes aux unités lexicales en étude et leurs rôles thématiques de VerbNet et le parser «Charniak ». La tâche 19 [3] a été consacrée à l'extraction des *frames*. La compétition consistait à identifier des mots et des phrases qui évoquent des *frames* et leurs assignés les *frames* adéquats. Dans le cas où le *frame* n'existe pas dans les données d'entraînement, le *frame* assigné à une unité lexicale est le *frame* le plus proche. Dans cette compétition, la ressource FrameNet est utilisée ainsi

[16] PropBank est un corpus annoté des propositions verbales et leurs arguments, ces derniers étant étiquetés au moyen de rôles sémantiques.
[17] 4th International workshop on Semantic Evaluations (http://nlp.cs.swarthmore.edu/semeval/tasks/)

qu'un ensemble de textes annotés. Cette tâche consistait aussi à identifier les parties de la phrase qui évoquent des rôles sémantiques et leurs assignés les rôles correspondants. Dans cette compétition, plusieurs systèmes ont été proposés et la plus part se sont basés sur l'apprentissage machine. En 2009, Martin Scaiano a testé son système sur les phrases de FrameNet et les données fournies à SemEval-2007 tâche 19. Il a réussit à avoir une bonne précision que les systèmes qui ont participé à la compétition de SemEval-2007. Sa méthode utilise l'apprentissage machine appliquée à un arbre de dépendance pour identifier les éléments porteurs des rôles sémantiques et leur assigner des rôles sémantiques adéquats.

L'optique de la construction de la ressource FrameNet pour d'autres langues varie d'une langue à une autre mais la plupart du temps, on se base sur le FrameNet de l'anglais et des corpus alignés ou parallèles.

Dans notre projet, l'objectif est le même que les différents travaux faits pour les autres langues, c'est-à-dire d'arriver à identifier automatiquement les actants et leur rôles sémantiques mais nous voulons y arriver sans avoir recours au FrameNet de l'anglais. Nous attribuons des rôles sémantiques aux actants à partir d'un corpus de phrases annotées manuellement et tirées d'un dictionnaire déjà construit selon les principes de la LEC.

Néanmoins, nous gardons de ces travaux l'idée d'utiliser des traits basés sur des caractéristiques syntaxiques dans le modèle d'automatisation que nous avons proposé. Dans notre cas, comme pour Johanson et Nugues [34], nous avons utilisé un analyseur syntaxique en dépendances. Cet analyseur, conçu pour le français, retourne toutes les différentes dépendances syntaxiques entre les différentes unités de la phrase. Nous avons subdivisé notre approche automatique d'annotation en deux tâches : 1) identification des participants (actants et circonstants) des lexies verbales, 2) annotation de ces actants au moyen de rôles sémantiques. La division en tâches d'identification des participants et d'annotation de rôles sémantiques est aussi proposée par d'autres auteurs et même dans des compétitions d'évaluation sémantique cette division est proposée [3,58]. Son avantage est de ne pas essayer d'annoter en rôles sémantiques des éléments alors qu'ils ne sont même pas porteurs de rôles. Donc l'idée est d'extraire les unités porteuses de rôles sémantiques ensuite assigner à ces unités des rôles. Quant à la tâche de distinguer les actants des circonstants est de notre proposition. Les linguistiques qui travaillent en collaboration avec

nous, suggèrent d'attribuer des rôles sémantiques aux actants d'abord. Ces derniers assistent à la définition d'une lexie qui est un de leur objectif. Et aussi les annotations manuelles sont au début. Il y a environ 3411 actants dont l'annotation est vérifiée et validée. Quand aux circonstant, il n'y a qu'environ 1100 circonstants dont l'annotation est vérifiée et validée.

Dans la tâche d'identification des participants actants et circonstants, nous avons récupéré les dépendances syntaxiques, retournées par l'analyseur, entre la lexie verbale faisant l'objet de l'annotation et les autres unités de la phrase afin de construire des règles d'identification des participants de cette lexie ou d'intégrer ces dépendances syntaxiques dans les traits de classification par apprentissage machine. Dans l'approche par apprentissage machine, nous avons testé la stratégie d'utiliser les dépendances syntaxiques avec les classificateurs de Weka. En nous inspirant de la proposition de Collobert et Weston [10], qui consiste à utiliser uniquement les catégories grammaticales, nous avons aussi testé cette autre stratégie avec ces mêmes classificateurs. Pour la tâche d'assignation des rôles sémantiques, nous effectuons un partitionnement, une approche différente de la classification proposée dans la plupart des travaux. La difficulté rencontrée est que la langue traitée, le français, ne dispose pas de ressources sémantiques accessibles. Le problème est accru du fait que nous travaillons sur les lexies d'un domaine de spécialité.

Une différence avec ces travaux en plus de la langue traitée, le français plutôt que l'anglais, est le corpus de données de spécialité utilisé. Nous disposons d'un corpus des domaines de l'informatique et de l'internet dans lequel les lexies verbales annotées ne sont pas regroupées dans des *Frames*, elles ne sont donc pas directement hiérarchisées entre elles. Cependant, les structures actancielles de ces unités du domaine de spécialité peuvent être enrichies d'étiquettes de rôles sémantiques dont le nombre peut être limité. Contrairement à FrameNet qui propose des hiérarchies indiquant la parenté sémantique entre les unités lexicales appartenant à des *Frames* distincts, dans FrameNet, les rôles sémantiques sont assez spécifiques qu'il n'est pas possible de limiter leur nombre.

4.3 Conclusion

Dans ce chapitre, nous avons fait référence à certains travaux connexes à notre travail. Nous avons décrit certains éléments et modèles afin d'automatiser la tâche d'annotation de rôles sémantiques. La plupart des travaux présentés sont faits pour l'anglais; par conséquent, ils se sont basés sur des ressources sémantiques riches disponibles dans cette langue. Nous avons présenté notre approche par rapport à celles de ces travaux et les éléments sur lesquels nous nous sommes basés, à savoir les caractéristiques syntaxiques et le corpus de données utilisé.

Dans le chapitre suivant, nous présentons notre corpus dans lequel les lexies verbales sont des termes du domaine de l'informatique et de l'Internet qui ont été annotées manuellement par les linguistes de l'OLST. Nous parlerons de la constitution de notre corpus de données du domaine spécialisé et l'annotation manuelle des actants dans le format XML. Ce corpus, formé par ces lexies annotées manuellement par des rôles sémantiques, constitue le noyau sur lequel est basée notre approche d'automatisation. Cette dernière est décrite dans les chapitres 5 et 6.

Chapitre 5 Annotation manuelle

Nous sommes appuyés sur le dictionnaire informatique DicoInfo (section 2.1.1.1), basé sur le modèle de la LEC pour construire notre corpus. Les annotations sont réalisées sur les participants des lexies prédicatives. Elles sont réalisées selon un modèle et une méthodologie fortement inspirés de FrameNet. L'annotation manuelle de ces données dans des contextes français est effectuée par une partie des membres de l'OLST.

L'annotation consiste à enrichir l'information donnée sur la structure actancielle des verbes appartenant au domaine de l'informatique en explicitant les propriétés syntaxico-sémantiques des verbes et de leurs participants (type actant ou circonstant, rôle sémantique, fonction syntaxique et groupe syntaxique). Les contextes annotés dans lesquels apparaissent ces verbes sont rédigés en français. L'intérêt de cette annotation syntaxico-sémantique des verbes est de mettre en évidence les constructions syntaxiques qu'ils admettent et de décrire le comportement de leurs participants, à savoir :

leur type actant ou circonstant;
leur rôle sémantique;
leur combinatoire avec le verbe.

5.1 Annotation d'une lexie verbale

Étant donné des lexies verbales dans des contextes, les linguistes enrichissent leur structure actancielle en explicitant les propriétés syntaxico-sémantiques de ces termes et de leurs participants. Cette explicitation prend la forme d'annotations insérées formellement dans des contextes où apparaissent les termes prédicatifs à décrire. Par exemple :

[$_{\text{Act Suj Agent}}$ Vous] ABANDONNEZ [$_{\text{Act Obj Patient}}$ l'installation] [$_{\text{Circ Compl Temps}}$ à ce moment là].

Dans cet exemple, ABANDONNER est le terme prédicatif faisant l'objet de l'annotation. Cette dernière met en évidence les différents participants du terme qui sont explicités avec leurs propriétés syntaxico-sémantiques (le type actant ou circonstant, le rôle sémantique, la fonction et le groupe syntaxiques).

Ces participants, partiellement équivalents aux *FE* dans FrameNet, sont annotés par des rôles sémantiques tels que : Agent, Patient, Instrument et Temps suite à l'annotation de plusieurs contextes contenant le verbe ABANDONNER (Figure 15). Les rôles sémantiques sont mis en évidence directement dans le contexte en utilisant des codes de couleurs associées aux rôles des participants.

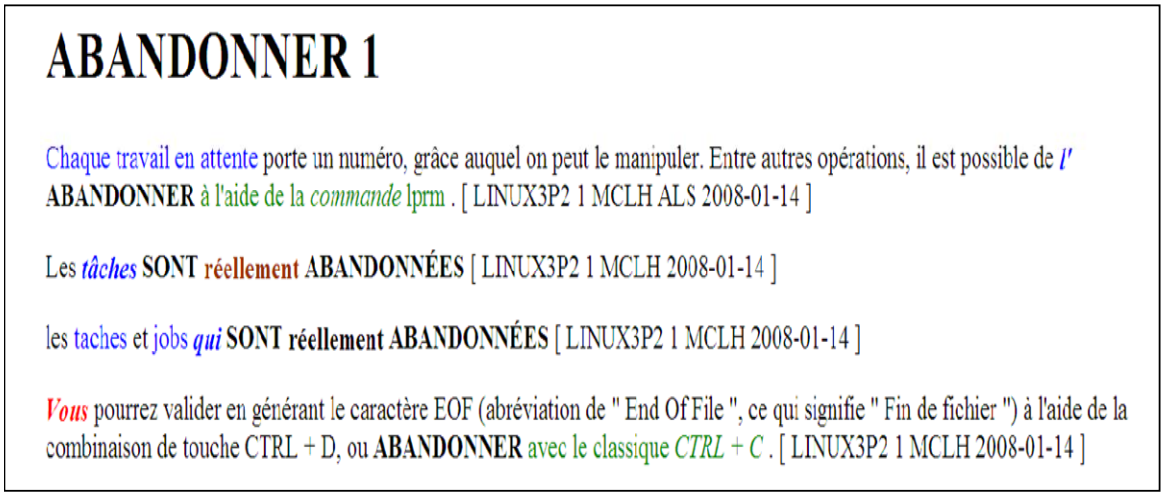


Figure 15 Contextes annotés de abandonner. Les rôles des participants sont mis en évidence par des codes de couleurs. Le rôle Agent est indiqué par la couleur rouge, le rôle Patient est signalé par la couleur bleue, le rôle Instrument est indiqué par la couleur verte et le Temps est donné par la couleur marron.

Les propriétés syntaxico-sémantiques des participants sont ensuite résumées dans un tableau récapitulatif illustré par la Figure 16. Ce tableau, décrit les actants dans la partie supérieure et les circonstants dans la partie inférieure. La colonne de gauche du tableau décrit le rôle sémantique du participant, la colonne du milieu décrit les informations syntaxiques, telles que la fonction syntaxique Sujet, Objet et Complément exprimée par le groupe syntaxique SN pour Syntagme Nominal, SP pour Syntagme Prépositionnel, SAdv pour Syntagme Adverbial, etc., suivies des statistiques sur le nombre d'occurrences des actants ou des circonstants et leurs réalisations syntaxiques sur plusieurs contextes. La colonne de droite illustre les différentes unités lexicales réalisant ces participants actants ou autres.

ABANDONNER 1		
Actants		
Patient	Objet (SN) (12) Objet (Pro) Sujet (SN)	opération (4) processus (2) le{chaque travail en attente} installation copie mise à jour chargement travail procédure traitement
Agent	Sujet (SN) (6) Lien indirect (SN)	vous (4) internaute système serveur
Autres		
Moyen	Complement (SP -avec) (2) Complement (SP -à l'aide de)	commande ctrl + c option
Temps	Complement (SP -à) (3) Complement (Prop)	stade (2) lorsque le temps de chargement leur semble excessif. heure de pointe
Cause	Complement (SP -en raison de) Complement (SP -à)	arrêt demande
But	Complement (Prop)	pour poursuivre

Figure 16 Tableau récapitulatif des participants de ABANDONNER

Cette manière de présenter l'annotation d'une lexie, illustrée par les figures 15 et 16 permet de faciliter la visualisation de l'annotation manuelle XML. Elle permet aussi aux annotateurs de valider plus facilement leur travail, car le code source XML est assez complexe et de consulter les statistiques sur les occurrences des participants.

5.2 Processus d'annotation

L'équipe de L'OLST a mis au point un modèle d'annotation des propriétés syntaxico-sémantiques des termes prédicatifs à partir d'une analyse manuelle de contextes extraits de corpus de textes spécialisé rédigés en français. Le logiciel d'acquisition de termes TermoStat (Drouin 2003) est utilisé pour extraire les termes candidats, parmi lesquels figurent les termes verbaux sur les lesquels nous travaillons dans cette thèse. Ces derniers

sont analysés et validés pour ne laisser que ceux de la spécialité d'informatique. Entre 15 à 20 contextes sont choisis pour un terme prédicatif verbal sélectionné. Ces contextes sont des phrases qui sont prises au complet pour inclure un plus grand nombre de participants. Si un participant du verbe est exprimé sous forme d'une anaphore, alors la phrase qui fait référence à ce participant est aussi prélevée pour pouvoir relier l'anaphore et le mot qu'elle renvoie dans l'annotation [29]. Ces termes verbaux sont répertoriés avec la description de leur structure actancielle dans le dictionnaire de l'informatique et de l'Internet DicoInfo (section 2.1.1.1).

Les exemples de description de ces unités verbales (appelées aussi **lexies verbales**) dans le DicoInfo constituent notre corpus de données. Cette ressource permet aux annotateurs d'identifier et de dégager les participants qui entourent la lexie verbale attestée afin de les annoter manuellement par la suite par des rôles. L'annotation manuelle de ces lexies verbales est implantée en utilisant un éditeur XML pour conserver ces annotations et pour pouvoir les traiter par la suite. L'annotation XML est validée par le schéma RNC de la l'annexe 1. Ce schéma permet de faciliter la saisie des éléments par l'annotateur.

Par exemple, les différents participants de la lexie ABANDONNER, dans le contexte de l'exemple cité en section 5.1, sont annotés en XML comme indiqué dans la Figure 17.

```
<participant type="Act" role="Agent">
  <fonction-syntaxique nom="Sujet">
    <groupe-syntaxique nom="SN">
       <realisation>Vous</realisation>
    </groupe-syntaxique>
   </fonction-syntaxique>
 </participant>
<participant type="Act" role="Patient">
  <fonction-syntaxique nom="Objet">
    <groupe-syntaxique nom="SN">l'
      <realisation>installation</realisation>
    </groupe-syntaxique>
   </fonction-syntaxique>
</participant>
<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>
```

Figure 17 Annotation en xml de la lexie abandonner dans le contexte vous pouvez ABANDONNER l'installation à ce moment là.

5.3 Conclusion

Dans ce chapitre, nous avons présenté les données utilisées dans notre corpus et leur annotation manuelle. Nous notons que l'étude est faite sur des lexies en français dans le domaine de l'informatique et de l'Internet. Les éléments annotés dans notre corpus sont : la lexie prédicative verbale, les participants actants ou circonstants de la lexie à l'étude, la fonction syntaxique et la catégorie grammaticale de ces participants et enfin les rôles sémantiques assignés aux participants.

Il ressort que cette tâche d'annotation manuelle de lexies verbales du français s'avère fastidieuse et exigeante en temps. Les chercheurs de l'OLST estiment que l'annotation manuelle d'une lexie exige 2 heures en moyenne. Notre corpus pour l'instant est constitué de 104 lexies et 2311 contextes annotés manuellement dont l'annotation a nécessité environ 4 mois de travail. Afin de faciliter la tâche de l'annotateur et d'accélérer le temps d'annotation, nous proposons une méthode. Pour y arriver, nous nous appuyons sur les données actancielles annotées et vérifiées décrites ci-dessus. Ce corpus de données annotées est utilisé par notre modèle d'automatisation comme corpus de développement, ou d'entraînement ou de test. Le 2/3 du corpus est utilisé pour le développement dans l'approche par règles ou bien pour l'entraînement dans l'approche par apprentissage et le 1/3 pour le test.

Dans notre travail, les deux types de participants, actant et circonstant, sont identifiés, mais l'assignation automatique des rôles sémantiques porte d'abord sur les actants car la distinction entre les rôles attribués aux circonstants est plus subtile. Les actants sont importants dans la réalisation du sens de la lexie contrairement aux circonstants qui ne sont qu'optionnels et qui ne contribuent pas à son sens.

Cependant, avant d'annoter les participants partageant un lien sémantique avec une lexie prédicative par des rôles sémantiques, il faut d'abord les identifier. Les deux chapitres suivants, consacrés à la tâche d'identification des actants et des circonstants présentent d'abord une approche par extraction de règles basées sur les sorties d'un analyseur syntaxique et ensuite une approche par apprentissage machine.

Chapitre 6 Identification des participants par des règles sur les sorties d'un analyseur syntaxique

Les travaux, cités à la section 4.2, réalisés dans l'objectif d'annoter automatiquement des unités lexicales de l'anglais et d'autres langues par des rôles sémantiques étaient basés sur les caractéristiques et arbres syntaxiques et certains aussi sur les dépendances syntaxiques. Dans notre tâche d'identification des participants actants et circonstants des lexies verbales, nous avons proposé d'utiliser les dépendances syntaxiques trouvées par l'analyseur syntaxique du français, « Syntex » [6, 7]. Syntex est un analyseur qui a eu une bonne évaluation en 2007 sur les textes français (évaluation EASY)[18].

Nous avons soumis à l'analyseur syntaxique les contextes du corpus, dans lesquelles apparaissent les lexies verbales à annoter, afin de récupérer les liens de dépendances entre ces lexies et les autres unités. Nous avons extrait des règles en nous basant sur les dépendances retournées par Syntex sur des contextes annotés manuellement; ces données ont été décrites au Chapitre 1. Ce corpus nous a servi à la fois d'inspiration pour développer des règles d'identification de participants et nous a permis d'évaluer la qualité de nos résultats sur des exemples n'ayant pas servi au développement. Les expérimentations liées à ce chapitre ont été présentées lors des communications acceptées à des conférences internationales [29, 30]

6.1 Les données et Syntex

Nous avons soumis les contextes ou les phrases de notre corpus à Syntex qui calcule les liens de dépendance entre les mots de la phrase. La Figure 18 illustre les participants de la lexie ACCEPTER identifiés dans le corpus annoté manuellement. Ces participants sont Bios et souris comme actants et directement comme participant circonstant.

[18] http://www.limsi.fr/Recherche/CORVAL/easy
Les résultats de l'évaluation sont donnés sur http://w3.erss.univ-tlse2.fr/membres/bourigault/syntex.html

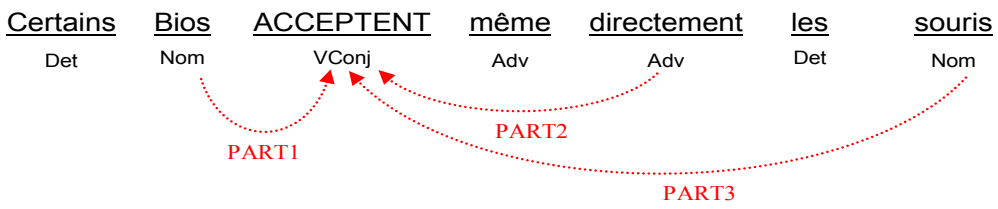


Figure 18 Les participants de la lexie ACCEPTER annotée manuellement **(PART1, PART2, PART3 étiquettes pour les participants 1, 2 et 3) et la lexie est écrite en majuscule**

Cette phrase, sans annotation,

Certains BIOS ACCEPTENT même directement les souris

est soumise à l'analyseur Syntex qui renvoie les différents liens entre les différentes unités qui la composent. Cette analyse est fournie en format XML (annexe 2) correspondant au schéma de la Figure 19.

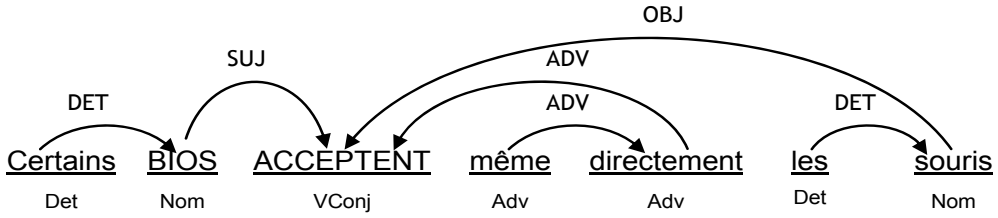


Figure 19 Schéma de l'analyse de Syntex où la lexie à étudier est indiquée en majuscules, les liens de dépendance entre les mots sont indiqués par des flèches étiquetées par les éléments du Tableau 4 Les étiquettes syntaxiques choisies parmi celles du Tableau 3 sont indiquées sous les mots.

Les étiquettes de Syntex utilisées qui figurent sur la Figure 19 sont décrites dans les tableaux 3 et 4

Étiquettes	Libellés	Exemples
Adv	Adverbe	Directement, simultanément,…
Adj	Adjectif	Difficile, excessif, nombreux,…
Det	Déterminant	Le, la, les, certains,…
Nom	Nom commun ou nom propre	Disquette, message, opération,…
Prep	Préposition	Sur, dans, en, à,…
Pro	Pronom personnel	Je, il, vous,…
ProRel	Pronom Relatif	Qui, que, lequel, lesquelles,…
VConj	Verbe Conjugué	Envoyez, classons,…
VInf	Verbe Infinitif	Zoomer, imprimer,…
VPpa	Verbe au participe passé	Classé, zoomé, imprimé,…
VPpr	Verbe au participe présent	Affectant, procédant,…

Tableau 3 Étiquettes morpho-syntaxiques des mots

Étiquettes	Libellés	Exemples
ADV	Lien avec un Adverbe	Même$_{(Adv)}$ directement$_{(Adv)}$ (adverbe modifie adverbe) Envoyer$_{(VInf)}$ directement$_{(Adv)}$ (adverbe modifie verbe)
ATTS	Lien attribut	Est$_{(VConj)}$ disponible$_{(Adj)}$
AUX	Lien avec un auxiliaire	Est$_{(VConj)}$ terminé$_{(VPpa)}$
CC	Lien entre la conjonction et ses coordonnés	La copie$_{(Nom)}$ ou la mise à jour$_{(Nom)}$ (lien CC entre copie et ou, et entre mise à jour et ou)
EPI	Épithète	La commande$_{(Nom)}$ lprm$_{(Nom)}$
NOMPREP	Lien entre un mot qui se trouve après la préposition et la préposition elle-même	à$_{(Prep)}$ une$_{(Det)}$ variable$_{(Nom)}$, dans$_{(Prep)}$ lequel$_{(ProRel)}$
PREP	Lien entre un mot et une Préposition	Enregistrer$_{(VInf)}$ sur$_{(Prep)}$ une disquette$_{(Nom)}$ Lien PREP entre le verbe enregistrer et la préposition sur.
OBJ	Lien Objet entre un verbe et un autre mot de la phrase	…acceptent$_{(VConj)}$ les souris$_{(Nom)}$ (souris est le nom Objet du verbe accepter)
REL	Lien entre un pronom relatif et le constituant auquel il réfère	Une unité$_{(Nom)}$ qui$_{(ProRel)}$ accepte$_{(VConj)}$…….. (lien REL entre le pronom relatif qui et le nom unité)
SUJ	Lien Sujet entre un verbe et un autre mot de la phrase	Les Bios$_{(Nom)}$ acceptent$_{(VConj)}$ … (Bios est le nom sujet du verbe accepter)

Tableau 4 Étiquettes des liens syntaxiques

Nous prenons comme hypothèse que les liens de dépendance donnés par Syntex permettront d'identifier les participants de la lexie. Nous nous concentrons sur les liens gauche et droit de cette dernière avec les autres mots de la phrase. Certains mots de la phrase ne possèdent pas de chemin de liens (c'est-à-dire l'ensemble de liens allant de la lexie au mot en passant éventuellement par d'autres mots) avec la lexie, mais, dans le corpus annoté, ils sont identifiés comme des participants. Pour ne pas perdre ces mots, nous proposons d'autres règles permettant de les considérer (voir section 6.4).

6.2 Méthodologie d'identification des participants

Les liens que l'analyse syntaxique nous donne entre la lexie verbale et les autres mots de la phrase vont servir à identifier les participants de cette lexie. Tel que montré à la Figure 18 et la Figure 19, nous jugeons qu'un mot est un participant de la lexie s'il entretient un lien gauche ou droit avec cette dernière. Est-ce que tous les mots qui entretiennent un lien syntaxique doivent être considérés comme participants? Pour répondre à cette question,

nous nous inspirons des exemples annotés manuellement où les participants sont identifiés et annotés par des types et des rôles. Nous comparons ces participants aux mots liés à la lexie verbale retournés par l'analyse syntaxique de Syntex. Cette comparaison est réalisée en combinant, dans la Figure 20, le schéma de dépendances donné par Syntex (Figure 19) avec le schéma de participants annotés manuellement (Figure 18).

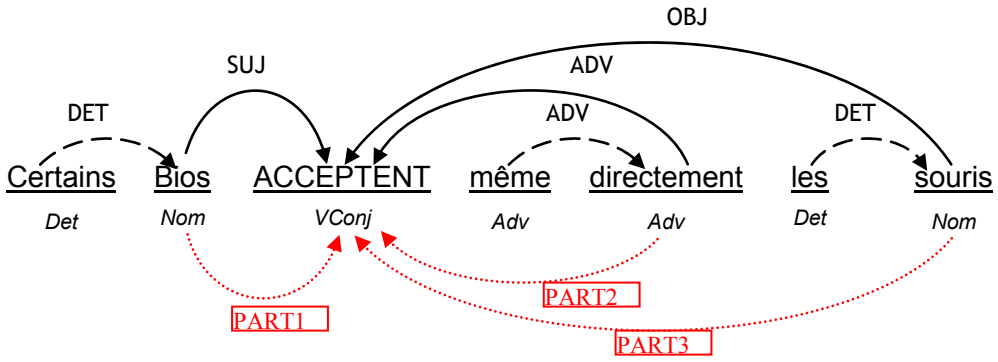


Figure 20 Schéma de combinaison les flèches au-dessus des mots correspondent aux liens donnés par Syntex. Nous indiquons en pointillé les liens qui ne seront pas considérées. Les flèches sous les mots correspondent aux participants annotés manuellement**.**

À partir du schéma de combinaison, c.-à-d. la sortie xml de l'analyse de Syntex et l'annotation xml des exemples du corpus annotés manuellement, on extrait des règles d'identification des participants ainsi que leurs traits syntaxiques. Ces derniers nous permettront d'affecter à ces participants le type Actant ou Circonstant avant de leur attribuer des rôles sémantiques. Une règle appliquée sur la sortie de Syntex est une expression de la forme « Si <condition> Alors <conclusion> ». Elle peut être expliquée comme ceci « S'il existe un lien syntaxique entre la lexie et un mot candidat de la phrase et que ce mot dans le corpus est considéré comme un participant alors ce mot candidat est un participant. S'il existe un lien syntaxique entre la lexie et un mot candidat de la phrase et que ce mot dans le corpus n'est pas considéré comme un participant alors ce mot candidat n'est pas un participant ». Une règle a deux parties : la partie gauche, qui spécifie les conditions d'application, est composée des mots de la phrase avec les liens syntaxiques qui se trouvent au-dessus des mots (Figure 20) et la partie droite, qui spécifie le résultat de l'application, ou les participants indiqués sous les mots (Figure 20). Dans une règle, un mot est décrit par ses traits tels que la catégorie grammaticale et la fonction syntaxique. Un trait est écrit <mot, Cat, Role>, où Cat est la catégorie grammaticale du mot et Role est sa

fonction syntaxique. Dans une règle, un mot peut avoir plusieurs catégories. Par exemple, un mot de catégorie Nom ou un autre de catégorie Pronom personnel peuvent réaliser une même règle. Dans ce cas, Cat est écrite « etiq1|etiq2|…|etiqn », chaque étiquette correspondant aux différentes étiquettes syntaxiques qu'un mot peut prendre dans une règle. Par exemple, le trait du mot de catégorie Nom ou Pronom est écrit <mot, Nom|Pro, Role>. On a remplacé Cat dans le trait du mot par « Nom|Pro ».

La lexie est écrite de la même manière qu'un mot en précisant « Lexie » précédant son trait et elle est écrite en majuscule. Les mots de la partie gauche d'une règle, qui forment les conditions, sont séparés par le symbole « + ». Les mots de la partie droite qui forment les participants doivent appartenir à des mots de la partie gauche. Ils sont écrits de la même manière que ces mots. La partie gauche et droite de la règle sont séparées par le symbole « → », qui ne signifie pas une réécriture mais indique la conclusion qui découle des conditions de la partie gauche.

Par exemple, à partir du schéma de combinaison de la Figure 20, on a extrait les règles d'identification pour chaque participant c'est-à-dire PART1, PART2, et PART3.

Règle pour identifier PART1

Le mot Bios, de type Nom et qui entretient le lien SUJ avec la lexie ACCEPTER, est considéré comme participant, étiqueté PART1. Alors la règle s'écrit :

<Bios, Nom, SUJ> + Lexie<ACCEPTENT, VConj, SUJ>
→ <Bios, Nom, SUJ>

Règle pour identifier PART2

Le mot directement, de type Adv et qui entretient le lien ADV avec la lexie ACCEPTENT, est considéré comme participant, étiqueté PART2. Alors la règle s'écrit :

Lexie<ACCEPTENT, VConj, ADV> + <directement, Adv, ADV)
→ <directement, Adv, ADV>

Règle pour identifier PART3

Le mot souris, de type Nom et qui entretient le lien OBJ avec la lexie ACCEPTER, est considéré comme participant, étiqueté PART3. Alors la règle s'écrit :

Lexie<ACCEPTENT, VConj, OBJ> + <souris, Nom, OBJ)
→ <souris, Nom, OBJ>

Dans le cas d'un mot de catégorie grammaticale Nom ou Pro tableau 3 et lié à la lexie par le lien SUJ ou OBJ tableau 4 donné par l'analyse de Syntex, ce mot est toujours identifié dans le corpus annoté manuellement comme un participant. Cette correspondance entre ces annotations syntaxiques et les participants, étant systématiquement vérifiée pour toutes les phrases du corpus, on peut donc dire qu'un nom ou pronom qui entretient avec la lexie verbale la fonction syntaxique SUJ ou OBJ est toujours un participant de cette lexie. Dans ce cas la règle générale sera écrite pour le mot :

SUJET

Lexie<MOT, VConj, SUJ> + <mot1, Nom|Pro, SUJ>
→ <mot1, Nom|Pro, SUJ>

OBJET

Lexie<MOT, VConj, OBJ> + <mot1, Nom, OBJ>
→ <mot1, Nom, OBJ>

Nous constatons donc un certain parallèle entre les fonctions syntaxiques et leurs rôles comme participants. Si pour ces liens directs entre la lexie et ces mots l'analogie est évidente, la situation n'est malheureusement pas toujours aussi facile et directe pour les autres liens. La situation se complique pour les relations indirectes telles que relation dans Syntex entre la lexie et les verbes, relation entre la lexie et certaines prépositions, etc. Dans la suite du présent chapitre, nous donnerons des exemples de combinaison de schémas de Syntex et participants annotés manuellement ainsi que les règles générales et leur application sur ces exemples correspondant aux différentes dépendances entretenues par la lexie et les autres mots de la phrase.

La couverture de cas de ces dépendances est faite en prenant en considération tous les contextes de notre corpus qu'on a soumis à Syntex. À partir de là, on extrait les liens gauches et droits entre la lexie à l'étude et les autres mots de la phrase. Tous les mots liés à la lexie sont considérés comme ses participants même s'ils ne le sont pas en réalité. Notre méthode est totalement empirique (voir section 6.6.1). Nous avons extrait une trentaine de règles correspondant à chaque lien identifié. Ces règles ont été validées sur notre corpus annoté manuellement en comparant les participants trouvés par les liens de l'analyse

syntaxique et ceux qui sont annotés manuellement. Nous avons ainsi modifié certaines règles et nous en avons ajouté d'autres.

D'autres dépendances entre les mots ne découlent pas de la lexie, mais ces mots peuvent être des participants de la lexie. Pour ces cas, nous avons ajouté à la partie gauche de la règle un autre champ représentant le chemin constitué de mots et de leurs catégories grammaticales qui se trouvent entre la lexie et le mot à considérer comme participant même s'il n'est pas lié syntaxiquement à la lexie. Ce chemin est noté dans la règle par $<mot_1, Cat_1><mot_2, Cat_2>\ldots<mot_n, Cat_n>$.

Ces règles sont présentées dans les sous-sections suivantes regroupées selon les dépendances syntaxiques entre la lexie et les autres mots de la phrase pour les dépendances qui mettent en jeu la lexie. Nous distinguons les autres types de dépendances ne mettant pas en jeu la lexie.

6.3 Types de dépendances mettant en jeu la lexie

La lexie possède des dépendances ou liens syntaxiques à gauche ou à droite vers les autres mots de la phrase. Nous traduisons ces dépendances en liens sémantiques, c'est-à-dire que ces mots liés à la lexie sont considérés comme des participants auxquels nous attribuons des rôles sémantiques. Ces mots peuvent être de type Nom, Adverbe, Préposition, Verbe, Adjectif, Coordination, Relatif, etc. Dans l'analyse Syntex, ces mots sont étiquetés par les étiquettes syntaxiques du Tableau 3 et les dépendances syntaxiques peuvent porter les étiquettes du Tableau 4.

Les règles de traduction des dépendances syntaxiques ou du chemin de liens syntaxiques sont données en fonction des types des mots de la phrase liés à la lexie (d'où l'écriture dépendance Lexie/type du mot lié dans ce qui suit). Elles sont écrites en suivant le chemin de lien syntaxique allant de la lexie vers le mot. Ainsi, l'ordre des mots dans la règle suit ce chemin. Les règles sont écrites dans ce document sur plusieurs lignes : des lignes pour le chemin de liens constituent la partie gauche de la règle et une autre ligne pour le participant identifié correspond à la partie droite de la règle. Les exemples associés à chaque type et règle sont présentés sous forme d'un graphe et ils sont accompagnés de l'application de la règle correspondante. Ces exemples sont donnés en annexe 3 dans des

figures et les règles correspondantes sont données dans le tableau récapitulatif de règles section 6.5.

6.3.1 Dépendance Lexie/Nom ou Lexie/Pro

Dans ce cas, nous étudions la dépendance directe entre la lexie et un mot de la phrase de type Nom ou Pronom personnel. La lexie peut être de type Verbe conjugué, au participe passé, au participe présent ou à l'infinitif. Il y a deux types de dépendances : Sujet et Objet.

Dans le cas du Sujet, si la lexie porte l'étiquette VConj, c'est-à-dire que la lexie est conjuguée, alors le mot qui lui est lié est identifié comme participant avec la fonction syntaxique Sujet (Figure 42, règle 1). Si la lexie porte l'étiquette VPpa, c'est-à-dire qu'elle est au participe passé, alors ce mot est identifié comme participant avec la fonction syntaxique Objet (Figure 43, règle 2). Cette dépendance existe aussi pour le cas de la lexie au participe présent VPpr.

Dans le cas d'objet, le mot qui est lié à la lexie par ce type de dépendance est identifié comme participant. Ce mot peut apparaître après la lexie (Figure 44) ou avant la lexie (Figure 45). Il est identifié avec la fonction syntaxique objet.

6.3.2 Dépendance Lexie|Adverbe

S'il existe un lien syntaxique direct entre la lexie verbale et un adverbe, alors nous récupérons cet adverbe comme un participant de cette lexie (Figure 46, règle 4).

6.3.3 Dépendance Lexie/Préposition

Deux types de liens entre la lexie et la préposition : le type PREP et le type NOMPREP. Quand on trouve le lien Nomprep entre la lexie et la préposition « à », on cherche le mot ayant le lien PREP avec cette préposition (Figure 47, règle 5). Si ce mot est de type Nom, on l'identifie comme participant. Si ce mot est un Verbe alors on cherche les liens de ce Verbe par les règles dégagées dans la section 6.3.6. Et Quand on trouve dans l'analyse de Syntex une relation de type PREP entre la lexie verbale et une préposition ou entre l'auxiliaire lié à la lexie et la préposition, on prend comme participant le mot lié à cette préposition par la relation NOMPREP de l'analyse Syntex (Figure 48, règle 6). Ce mot dont le rôle syntaxique

est NOMPREP avec la préposition peut être de type Nom par exemple à ce stade ou de type pronom relatif (ProRel) par exemple à laquelle (Figures 49 et 50, règles 7 et 8). Dans le cas du mot de type Nom, on récupère ce mot comme participant et dans le cas du mot ProRel, on prend le pronom relatif comme participant avec une référence au mot auquel elle réfère, à savoir son antécédent.

6.3.4 Dépendance Lexie/Conjonction

Si la relation est identifiée entre la lexie verbale et une conjonction de coordination, on suit le chemin de dépendance entre cette conjonction et les autres mots de la phrase. Ces mots liés à la conjonction peuvent être de catégories syntaxiques différentes telles que Nom, Pronom personnel, Verbe, Préposition ou Pronom relatif. Selon ces catégories des mots liés, l'analyse Syntex attribue des étiquettes syntaxiques différentes dans l'étape de la segmentation. On peut trouver les tokens d'étiquettes CCoordNom qui relient plusieurs noms (figure 51, règle 9), CCoordV qui met en liaison plusieurs verbes (figures 52, 53 et 54, règle 10, 11 et 12), CCoordCSub conjonction reliée à un constituant relatif, CCoordPrep (Figure 55, règle 13) qui est reliée à la préposition suivie d'un nom de la forme <Prep + Nom> ou bien à la préposition précédée d'un verbe de la forme <Verbe + Prep> et CCoordPrepde (Figure 56, règle 14) qui est reliée à la préposition de. On présentera les chemins de liens qui construiront nos règles selon ces différents types de conjonction.

6.3.5 Dépendance Lexie/Relative

Si la relation est identifiée entre la lexie verbale et un pronom relatif (que, qui, lequel, etc.), on récupère comme participant de cette lexie les mots en dépendance avec ce pronom relatif.

Plusieurs cas de relatives sont à distinguer. On repère les pronoms relatifs qui nous intéressent pour l'identification des participants en s'appuyant sur l'étiquette « ProRel » attribuée par l'analyse de Syntex aux pronoms relatifs. Les mots en dépendance avec le pronom relatif ont des rôles syntaxiques « REL » avec la relative. Cette dernière liée à la lexie par le rôle de SUJ, OBJ, etc selon le rôle exact que peuvent jouer les constituants liés à la relative s'ils étaient reliés directement à cette lexie (figures 57 et 58, règle 15).

6.3.6 Dépendance Lexie/Verbe

Si la lexie verbale se trouve en relation avec un autre verbe la précédant dans la phrase, alors les participants de ce mot seront identifiés comme participants de la lexie en plus des participants directs identifiés pour cette dernière. Le mot peut être un verbe, un verbe et un auxiliaire ou un verbe suivi d'une préposition. Cependant, si verbe lié est conjugué alors la lexie se trouve à l'infinitif, dont le lien en analyse de Syntex est noté OBJ (figures 59 et 60, règle 16) et si le verbe lié est un auxiliaire alors la lexie se trouve au participe passé dont le lien est AUX dans l'analyse de Syntex (figures 61 et 62, règles 17 et 18). Si le verbe lié est suivi de la préposition alors la lexie porte le lien NOMPREP. Plusieurs cas avec la préposition de (figures 63 à 70, règles 19 à 24), la préposition à (figures 71 à 76, règles 25 à 28) ou la préposition pour (figures 77 et 78, règles 29 et 30).

6.4 Autres types de dépendances ne mettant pas en jeu la lexie

Il s'agit de dépendances syntaxiques entre les mots de la phrase qui ne mettent pas en jeu la lexie à annoter ; c'est-à-dire que certains mots peuvent être identifiés comme participants alors que, lors de l'analyse Syntex de la phrase, ces mots ne présentent aucun chemin de lien syntaxique avec la lexie. Pour pouvoir les identifier comme participants, nous avons besoin de ces caractéristiques :

- chemins de liens de ces mots avec d'autres mots de la phrase;
- chemins de liens avec la lexie, s'il y a lieu;
- les mots séparant la lexie de ces mots;

Les règles qui peuvent identifier ces participants seront composées de ces trois caractéristiques dans leur partie gauche. Ces trois caractéristiques sont séparées par une virgule « , ». Une telle règle peut être exprimée comme suit :

<mot M> + <chemins de liens du mot M> , <Lexie>+<chemins de liens avec la lexie>
<mots séparant la lexie du mot M>
→ <mot M>

On peut lire cette règle de la manière suivante : un mot M est identifié participant étant donné son chemin de liens, le chemin de liens de la lexie s'il y a lieu et les mots qui se trouvent entre le mot M et la lexie.

La convention d'écriture des règles est respectée telle que définie précédemment dans la section 6.2. Les mots de chaque caractéristique sont séparé par « + » et sont décrits par leurs catégories grammaticales, leurs fonctions syntaxiques ainsi que leur position dans la phrase par rapport à la lexie qui indique si le mot apparaît avant ou après la lexie. Nous considérons ici le cas de l'adverbe puis et le cas de la préposition en.

6.4.1 Cas de l'adverbe puis

L'adverbe puis peut être lié à la lexie par le lien Syntaxique ADV trouvé par Syntex. Mais, nous ne considérons pas cet adverbe comme un participant de la lexie. Mais on peut avoir dans la même phrase une unité lexicale d'un autre verbe précédant la lexie susceptible d'être son participant et que Syntex n'a pas détecté de lien syntaxique entre elle et la lexie. La règle précédente est appliquée pour identifier ce participant (figure 79, règle 31).

6.4.2 Cas de la préposition en

En ce qui concerne le traitement de la préposition en, deux cas doivent être résolus:

1. Le cas où la préposition en apparaît avant la lexie et cette dernière se trouve au participe présent;
2. Le cas où la préposition en apparaît après la lexie.

Dans le premier cas, Syntex donne le lien NOMPREP entre la lexie et en. Ce lien représente la caractéristique chemin de lien de la lexie dans la règle. Mais ne détecte aucun lien entre la lexie et les autres mots la précédant dans la phrase susceptibles d'être des participants. Par exemple un mot Sujet d'un verbe précédant la lexie, peut être son participant Sujet (figures 80 et 81, règles 32 et 33).

Dans le deuxième cas, la lexie est suivie de la préposition en, elle-même suivie du verbe au participe présent, de la forme <en+VPpr>. L'analyse de Syntex n'identifie pas de lien entre cette préposition en suivie du verbe au participe présent et la lexie. Dans notre corpus la proposition formée par cette préposition et son verbe au participe présent est

considérée comme un participant circonstant de la lexie. Donc, si on trouve la préposition en suivie d'un verbe au participe présent apparaissant après la lexie, on prend comme participant toute la proposition formée de en+VPpr+les participants directs de ce verbe VPpr : en partant de la préposition en, suivre successivement tous ses liens syntaxiques avec d'autres mots de la phrase se trouvant après elle.

6.5 Récapitulation des règles par catégories

Toutes les règles sont récapitulées dans le Tableau 5. Dans les 2ème et 3ième colonnes, nous présentons respectivement les parties gauche et droite des règles. Dans la 1ère colonne, nous retrouvons le numéro de la règle citée à la section 6.3 et, dans la dernière, nous retrouvons le numéro de la figure correspondante, qui présente un exemple avec son schéma Syntex et l'application de la règle. Ces figures se trouvent en annexe 3.

Règle	Conditions des règles	Participants	Figure
Dépendances Lexie | NOM ou Lexie | Pro (section 6.3.1)			
1	Lexie<MOT, VConj, SUJ><mot1,Nom|Pro,SUJ>	<mot1, Nom|Pro, SUJ>	Figure 42
2	Lexie<MOT, VPpa, SUJ><mot1,Nom|Pro, SUJ>	<mot1, Nom|Pro, OBJ>	Figure 43
3	Lexie<MOT, VConj|VInf|VPpa|VPpr, OBJ> <mot1, Nom|Pro, OBJ>	<mot1, Nom|Pro, OBJ>	Figure 44 Figure 45
Dépendances Lexie | Adverbe (section 6.3.2)			
4	Lexie<MOT, VConj, ADV><mot1, Adv, ADV>	<mot1, Adv, ADV>	Figure 46
Dépendances Lexie | Préposition (section 6.3.3)			
5	Lexie<MOT, VInf, NOMPREP> <à,Prep,PREP, NOMPREP><mot1,Nom, PRep>	<mot1, Nom, Tête>	Figure 47
6	Lexie<MOT, VConj|VInf|VPpa, PREP> <mot1, Prep, PREP, NOMPREP> <mot2, Nom, NOMPREP>	<mot2,Nom,NOMPREP>	Figure 48
7	Lexie<MOT, VConj|VInf, PREP> <mot1, Prep,PREP, NOMPREP> <mot2, ProRel, NOMPREP, REL> <mot3, Nom, REL>	<mot2, ProRel, NOMPREP> avec antécédent <mot3, Nom, REL>	Figure 49

Règle	Conditions des règles	Participants	Figure
8	Lexie<MOT, VPpa, AUX> <mot1, VConj, AUX, PREP> <mot2, Prep, PREP, NOMPREP> <mot3, ProRel, NOMPREP, REL> <mot4, Nom, REL>	<mot3, ProRel, NOMPREP> avec antécédent <mot4, Nom, REL>	Figure 50
Dépendances Lexie | Conjonction (6.3.4)			
9	Lexie<MOT, VConj|VInf, OBJ|SUJ> <mot1, CCoordNom, CC, OBJ|SUJ> <mot2, Nom, CC>	<mot2, Nom, OBJ|SUJ>	Figure 51
10	Lexie<MOT, VConj, CC> <mot1, CCoordV, CC, SUJ> <mot2, Nom|Pro, SUJ>	<mot2, Nom|Pro, SUJ>	Figure 52
11	Lexie<MOT, VPpa, CC> <mot1, CCoordV, CC, Nomprep> <mot2, Prep, nomprep, prep> <mot3, Nom, Prep,obj> <mot4, VConj, obj, Suj> <mot5, Nom, suj>	<mot5, Nom, suj>	Figure 53
12	Lexie<MOT, VPpa, CC> <mot1, CCoordV, CC, prep> <mot2, Prep, prep, nomprep> <mot3, Nom, nomprep>	<mot3,Nom,nomprep>	Figure 54
13	Lexie<MOT, VConj|VInf, PREP> <mot1, CCoordPrep, prep, cc> <mot2, Prep, CC, NOMPREP> <mot2, Nom, NOMPREP>	<mot2,Nom,NOMPREP>	Figure 55
14	Lexie<MOT, VConj|VInf, nomprep> <mot1, Prep,NOMPREP, CC> <mot2, CCoordPrep,cc, rep> <mot3, VConj|VInf, PREP, suj> <mot4, Nom, suj>	<mot4, Nom, suj>	Figure 56
Dépendances Lexie | Relative (section 6.3.5)			
15	Lexie<MOT, VConj, SUJ|obj> <mot1, ProRel, REL, SUJ|obj> <mot2, Nom, Rel>	<mot1, ProRel, SUJ> avec antécédent <mot2, Nom, REL>	Figure 57 Figure 58
Dépendances Lexie | Verbe (section 6.3.6)			
16	Lexie<MOT, VInf, OBJ> <mot1, VConj, OBJ, SUJ|ADV> <mot2, Nom|Pro, SUJ|ADV>	<mot2,Nom|Pro,SUJ|ADV	Figure 59, Figure 60

Règle	Conditions des règles	Participants	Figure
17	Lexie<MOT, VPpa, AUX> <mot2, VConj, AUX, SUJ> <mot1, Nom|Pro, SUJ>	<mot1, Nom|Pro, OBJ>	Figure 61
18	Lexie<MOT, VPpa, AUX><mot1,VInf,AUX, OBJ> <mot2, VConj, OBJ, SUJ> <mot3, Nom|Pro, SUJ>	<mot1, Nom|Pro, OBJ>	Figure 62
19	Lexie<MOT, VInf, NOMPREP> <de, Prep, NOMPREP, PREP> <mot1, VConj, PREP, SUJ> <mot2, Nom|Pro, SUJ>	<mot2, Nom|Pro, SUJ>	Figure 63 Figure 64 Figure 65
20	Lexie<MOT, VInf, NOMPREP> <de, Prep,NOMPREP, PREP> <mot1, VPpa, PREP, AUX> <mot2, VConj, AUX, SUJ> <mot3, Nom|Pro, SUJ>	<mot3,Nom|Pro,SUJ,lien indirect>	Figure 66
21	Lexie<MOT, VInf, NOMPREP> <de, Prep, NOMPREP, PREP> <mot1, Adj, PREP, ATTS> <mot2, VConj, ATTS, SUJ> <mot3, Nom|Pro, SUJ>	<mot3,Nom|Pro,SUJ,lien indirect>	Figure 67
22	Lexie<MOT, VInf, NOMPREP> <de, Prep, NOMPREP, PREP> <mot1, VConj, PREP, OBJ> <mot2, Nom, OBJ>	<mot2, Nom, SUJ>	Figure 68
23	Lexie<MOT, VInf, NOMPREP> <de, Prep, NOMPREP, PREP> <mot1, VConj, PREP, PREP> <à, Prep, PREP, NOMPREP> <mot2, Nom, NOMPREP>	<mot2, Nom, lien indirect>	Figure 69
24	Lexie<MOT,VInf, NOMPREP> <de,Prep, NOMPREP, PREP> <permettre, VConj|VInf, PREP, SUJ|prep> (<à, Prep, PREP, NOMPREP>)* <mot2, Nom, SUJ|nomprep>	<mot1, Nom, SUJ, lien indirect>	Figure 70
25	Lexie<MOT, VInf, NOMPREP> <à, Prep, NOMPREP, PREP> <mot1, VConj, PREP, SUJ> <mot2, Nom|Pro, SUJ>	<mot2, Nom|Pro, SUJ, lien indirect>	Figure 71 Figure 72 Figure 73

Règle	Conditions des règles	Participants	Figure
26	Lexie<MOT, VInf, NOMPREP> <à, Prep, NOMPREP, PREP> <mot1, VPpa, PREP, AUX> <mot2, VConj, AUX, SUJ> <mot3, Nom|Pro, SUJ>	<mot1, Nom|Pro, SUJ, lien indirect>	Figure 74
27	Lexie<MOT, VInf, NOMPREP> <à, Prep, NOMPREP, PREP> <mot1, VConj|VInf, PREP, OBJ> <mot2, Nom, OBJ>	<mot2, Nom|Tête, OBJ>	Figure 75
28	Lexie<MOT, VInf, NOMPREP> <à, Prep, NOMPREP, PREP> <mot1, Adj, PREP, ATTO> <mot2, VConj|VInf, ATTO, OBJ> <mot3, Nom, OBJ>	<mot2, Nom|Tête, OBJ>	Figure 76
29	Lexie<MOT, VInf, NOMPREP> <pour, Prep, NOMPREP, PREP> <mot3, VPpa, PREP, AUX> <mot2, VConj, AUX, SUJ> <mot1, Nom, SUJ>	<mot1, Nom, Lien indirect>	Figure 77
30	Lexie<MOT, VInf, NOMPREP> <pour, Prep, NOMPREP, PREP> <mot1, VConj|VInf, PREP, OBJ> <mot2, Nom, OBJ>	<mot2, Nom, Lien indirect>	Figure 78
Autres types de dépendances (section 6.4)			
31	<mot1, Nom|Pro, SUJ><mot2, VConj, SUJ, OBJ> , Lexie<MOT, VInf, ADV> <puis, Adv, ADV> <moti, Cati>	<mot1, Nom, SUJ>	Figure 79
32	<mot1, Nom|Pro, SUJ><mot2, VConj, SUJ> <moti, Cati> , Lexie<MOT, VPpr, Nomprep> <en, Prep, nomprep>	<mot1, Nom, SUJ>	Figure 80
33	Lexie<MOT, VConj|VInf|VPpa> <mot1, Nom, OBJ> , <en, Prep, NOMPREP> <mot1, VPpr, NOMPREP, Ri><moti, Ri>	<en mot1 moti>	Figure 81

Tableau 5 Tableau de récapitulation des règles

6.6 Implémentation

Pour construire les règles décrites à la section précédente, nous avons développé un ensemble de programmes pour extraire les règles à partir des dépendances trouvées par l'analyse Syntex et des participants annotés manuellement. Un autre programme est utilisé pour l'application de ces règles sur de nouveaux exemples. Nous présentons ici l'extraction de règles et leur application.

6.6.1 Extraction des règles

Les règles ont été construites à partir de relations syntaxiques calculées sur des exemples de notre corpus annoté manuellement. Notre corpus contient 104 lexies et 2311 contextes. Ce corpus est subdivisé comme suit :

1) corpus de développement 2/3 du corpus global (1548 contextes avec 74 lexies);
2) corpus de test 1/3 (761 contextes et 30 lexies).

La construction des règles a été obtenue par un programme XSL appliqué sur la sortie XML de l'analyse Syntex et le corpus de développement. Les mots de la phrase retrouvés par Syntex avec leurs chemins de liens sont candidats à être identifiés comme participants. L'appariement entre ces mots avec leurs chemins de lien et les exemples du corpus de développement, nous a permis de sélectionner ceux qui sont réellement considérés comme participants. Les chemins de lien de ces mots nous permettent de construire les règles d'identification de participants. Les règles sont récupérées sous forme d'un fichier XML (Voir Figure 21).

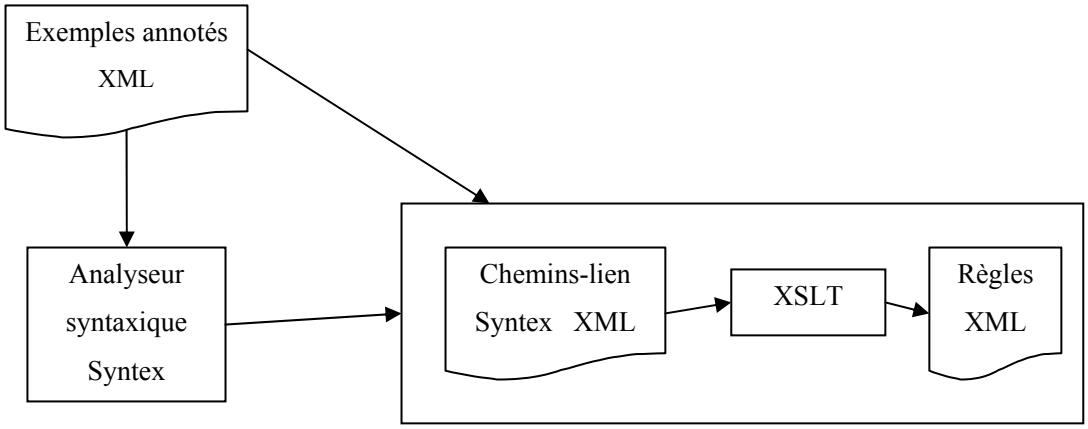


Figure 21 Schéma de construction de règles

Les sections précédentes ont illustré des types de dépendances liées à la lexie et d'autres dépendances qui n'y sont pas liées. Pour ces deux types, nous considérons deux approches d'extraction :

1) On extrait les liens qui mettent en jeu la lexie qu'on veut annoter.
2) On extrait d'autres liens qui ne mettent pas en jeu la lexie.

6.6.1.1 **Liens qui mettent en jeu la lexie**

Les liens sont sous forme de chemins entre la lexie et un autre mot qu'on notera chemin lien. Par exemple si un chemin lien entre une lexie et un mot est Objet, alors ce mot est un participant. On traduit alors ce chemin lien en une règle (voir règle 3).

Format d'une règle

Une règle est présentée sous le format XML dont la partie gauche est balisée par feature et la partie droite est balisée par participant. (Voir des extraits du fichier règles des figures 22 et 23).

```
<regle>
    <feature>
       <Lexie>Lexie(Mot;VINF,OBJ,Après)</Lexie>
       <mot>Mot;Pro,OBJ,Après</mot>
    </feature>
    <participant>Mot;Pro,OBJ,Après</participant>
</regle>
```

Figure 22 Règle XML de lien Objet correspondant à la règle 3

```
<regle>
   <feature>
      <Lexie>Lexie(Mot;VCON,PREP,Après)</Lexie>
      <mot>Mot;Nom,NOMPREP,Après</mot>
   </feature>
   <participant>Mot;Nom,NOMPREP,Après+</participant>
</regle>
```

Figure 23 Règle XML de lien prépositionnel correspondant à la règle 6

6.6.1.2 **Liens qui ne mettent pas en jeu la lexie**

Certains mots de la phrase sont identifiés comme participants dans le corpus annoté manuellement, mais Syntex ne renvoie aucun chemin lien entre eux et la lexie (voir section 6.4).

Exemple 1

ABANDONNER l'opération en répondant non

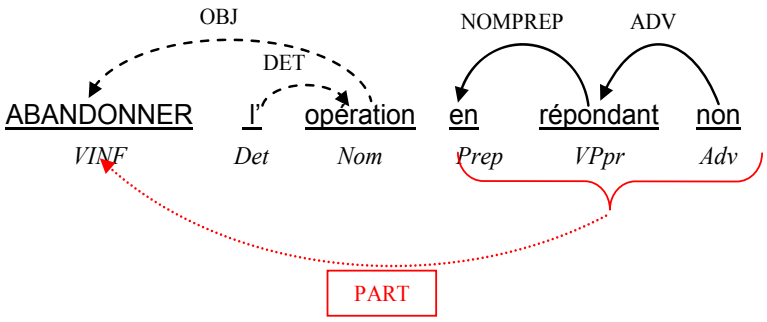


Figure 24 Lexie non liée syntaxiquement au participant prépositionnel

Dans ce cas, après appariement du corpus avec la sortie de Syntex, on l'identifie comme participant. Si ce participant est composé, alors on récupère tous les éléments qui lui sont liés retournés par Syntex, c'est le cas du participant propositionnel construit par la tête en (voir Figure 24). Sinon, on le récupère comme seul participant tel que le participant Nom-Sujet de l'exemple 2, de la Figure 25

Exemple 2 la partie gauche est imprimée puis ANNULÉE

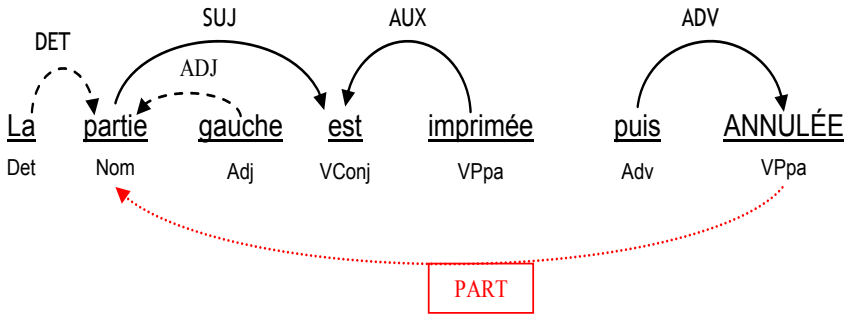


Figure 25 Lexie non liée au participant Nom-Sujet

Dans les figures 24 et 25, la lexie n'est pas en relation avec ces mots, alors on rajoute dans la règle les différentes catégories grammaticales des mots qui séparent la lexie du mot à considérer comme participant. Ces catégories sont présentées dans la règle XML par la balise *path*.

La règle XML correspondant à l'exemple 1 de la Figure 24 est présentée dans la Figure 26.

```
<regle>
   <lexie>lexie;cat</lexie>
   <feature>
      <mot>Mot1;Prep,NOMPREP,Après</mot>
      <mot>Mot2;VPpr,NOMPREP,Après)</mot>
      <mot>Mot3;Adv,ADV,Après</mot>
   </feature>
   <path>
      <Cat>(Det)</Cat>
      <Cat>(Nom)</Cat>
   </path>
   <participant>(Mot1;Prep,NOMPREP,Après)(Mot2;VPpr,NOMPREP,Après)(Mot3;Adv,ADV,Après)
```

Figure 26 Règle xml avec path correspondant à la règle 33 la valeur « Après » correspond à la position du mot par rapport à la lexie dans la phrase.

La règle XML de la Figure 26 peut être paraphrasée par : Si la proposition constitué des mots du chemin lien entre un mot prépositionnel, un verbe au participe présent et un adverbe, dont les mots qui les séparent de la lexie sont des mots de catégories Det et Nom, alors cette proposition est identifiée comme un participant.

La règle XML, correspondant à l'exemple 2 de la Figure 25 et à la règle 37 est présentée dans la Figure 27**.**

```
<regle>
   <feature>
      <mot>Mot1;Nom,SUJ, Avant</mot>
      <mot>Mot2;VConj,AUX,Avant</mot>
      <mot>Mot3;VPpa,AUX,Avant</mot>
   </feature>
   <feature>
      <Lexie>Lexie(Mot;VPpa,ADV)</Lexie>
      <mot>Mot4;Adv,ADV,Avant</mot>
   </feature>
   <path>
      <Cat>(VConj)</Cat>
      <Cat>(VPpa)</Cat>
      <Cat>(Adv)</Cat>
   </path>
   <participant>Mot1;Nom,SUJ,Avant</participant>
</regle>
```

Figure 27 Règle xml avec path correspondant à la règle 31 la valeur « Avant » correspond à la position du mot par rapport à la lexie dans la phrase.

La règle XML de la Figure 27 peut être paraphrasée par : Si un mot Nom lié à un verbe par le lien Sujet dont les mots qui le séparent de la lexie sont des mots de catégories VConj, VPpa, et Adv puis qui est lié à la lexie par le lien ADV, alors ce mot Nom est identifié comme un participant Sujet de la lexie.

6.6.2 Application des règles

Une fois la base des 33 règles construite à partir des exemples du corpus de développement, nous avons appliqué ces règles sur les exemples du corpus de test. Les phrases contenant les lexies à annoter sont analysées par Syntex pour en extraire les liens syntaxiques afin d'identifier les participants de la lexie. Ces liens ont été comparés à ceux des 33 règles d'identification de la section 6.5. S'il y a appariement entre les chemins de lien et les règles alors on applique la règle et le participant est identifié. Par exemple :

ABANDONNER l'opération en répondant non.

Nous analysons par Syntex cette phrase contenant la lexie ABANDONNER et les chemins de liens en sont extraits (voir Figure 28)

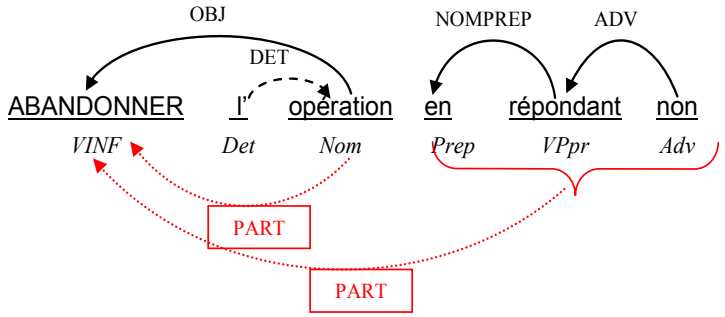


Figure 28 La lexie ABANDONNER et ses différents types de liens

Ainsi pour cet exemple, le chemin lien Objet entre la lexie ABANDONNER et le mot opération s'apparie à la règle 3 correspondant à la règle XML de la Figure 22. Cette règle alors est appliquée et le mot opération est identifié comme un participant.

Un autre chemin est présent dans la phrase, mais il ne fait pas intervenir la lexie ABANDONNER. C'est le chemin entre les mots en, répondant et non de la proposition en répondant non. On lui rajoute le chemin de catégories des mots qui séparent la lexie ABANDONNER et cette proposition. Ce chemin dans cet exemple est constitué de <l', Det, DET>+<opération, Nom, OBJ>. Les deux sont combinés et sont comparés aux règles. On retrouve alors le chemin de la règle 39 (ou la règle en format XML de la Figure 26) qui s'y apparie. La proposition en répondant non est donc identifiée comme participant.

6.6.3 Quelques exemples d'identification de participants

La Figure 29 donne les résultats de l'identification de participant à base de règles.

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 29 Participants des unités lexicales identifiés en appliquant les règles. Les unités lexicales sont en majuscules. Les mots en gras italique et soulignés sont leurs participants (Actants et Circonstants)

6.7 Évaluation de l'identification des participants

Nous avons évalué nos résultats en termes de taux de rappel et précision. La précision est définie par le nombre de participants pertinents retrouvés par rapport au nombre de participants total retrouvés (pertinents et non pertinents). La précision est donnée par la formule suivante :

$$Précision = \frac{nombre\ de\ participants\ pertinents\ retrouvés}{nombre\ total\ de\ participants\ retrouvés}$$

Le rappel est défini par le nombre de participants pertinents retrouvés par rapport au nombre de participants pertinents que possède le corpus de données. Le rappel est donné par la formule suivante :

$$Rappel = \frac{nombre\ de\ participants\ pertinents\ retrouvés}{nombre\ de\ participants\ pertinents\ dans\ le\ corpus}$$

Pour chaque lexie, on a calculé le taux de précision et de rappel (voir le Tableau 6). On a trouvé pour certaines lexies telles que CONFIGURER, SAISIR, PUBLIER, DÉVELOPPER, etc. un

bon taux et, pour d'autres, telles que DÉCONNECTER (41% de précision et 86% de rappel) et ARCHIVER (48% et 82% respectivement de précision et rappel) un taux plus faible.

Lexie	Préc.[19]	Rapp.[20]	Lexie	Préc.	Rapp.	Lexie	Préc.	Rapp.
CONFIGURER	83	76	VISITER	70	89	ACCEPTER	62	87
SAISIR	78	79	ACCÉDER	69	83	RECHARGER	61	75
PUBLIER	77	88	INSÉRER	67	83	PERSONNALISER	61	83
PLANTER	74	94	VISUALISER	66	80	REPROGRAMMER	61	70
DÉVELOPPER	74	79	TOURNER	66	84	REDÉMARRER	60	81
INSTALLER	73	90	DÉMARRER	65	88	NUMÉRISER	58	86
CALCULER	73	86	ZOOMER	64	84	SORTIR	58	82
ENREGISTRER	72	89	LANCER	63	82	INTERPRÉTER	57	85
ENTRER	71	68	MASQUER	63	87	ARCHIVER	48	82
FILTRER	71	89	LIRE	62	83	DÉCONNECTER	41	86

Tableau 6 Tableau d'évaluation d'application des règles selon l'ordre décroissant de la précision

Le taux global de précision est de 69 % et le taux de rappel est de 83% sur le corpus de test. Nous obtenons de bons résultats en se situant dans l'intervalle de mesure d'autres travaux ayant utilisé les dépendances syntaxiques. Nous sommes conscients que nous ne pouvons pas se comparer vu que les données utilisées sont différentes, mais ça peut nous servir de repère en regardant les autres systèmes. Johansson et Nugues, à SemEval-2007 [34], qui ont proposé un système en utilisant les classificateurs SVM dont les dépendances syntaxiques comme l'un des traits de classification, ont eu des résultats de 60 % pour la précision et 41 % pour le rappel. Li, Zhang et Yu [40] ont proposé une approche automatique d'annotation de textes de pages web avec des labels sémantiques en se basant sur les approches d'apprentissage machine et ont adopté la grammaire de dépendances. Dans leur travail, ils conceptualisent les mots des phrases, c.-à-d. annoter les mots pleins comme des concepts et formant ainsi des nœuds du graphe RDF qu'ils utilisent et ensuite ils attribuent des rôles sémantiques aux mots qui sont en liaison syntaxique. Dans leur travail, la tâche de relations sémantiques a eu 58 % de précision en utilisant les relations syntaxiques.

6.8 Distinction actants et circonstants en utilisant les règles

Dans les expérimentations décrites dans la section précédente, nous avons déterminé des participants d'une lexie en utilisant des règles basées sur les chemins de l'analyse de

[19] Préc. : précision
[20] Rapp. : rappel

Syntex. Les participants d'une lexie sont de type obligatoire s'ils font partie de son sens. Dans ce cas, ce participant est appelé un actant de la lexie. Les participants facultatifs sont appelés circonstants.

L'analyseur Syntex nous renvoie les chemins de liens entre les mots (voir en annexe la Figure 42) de la phrase

Le système Océ 31x5E ABANDONNE le travail d'impression

Le chemin lien Sujet s'apparie avec la règle 1, nous permet de trouver système comme un participant, qu'on a noté dans la règle par : <système, Nom, SUJ> et le chemin lien objet s'apparie avec la règle 2, nous permet de trouver travail comme un participant, qu'on a noté dans la règle par <travail, Nom, Obj>.

Nous avons constaté dans notre corpus que les participants identifiés dont la fonction syntaxique est Sujet ou Objet sont considérés comme des Actants dans tous les cas. Ce qui implique que les participants de trait <Nom, SUJ> ou de trait <Nom, OBJ> sont des Actants. Ainsi pour l'exemple ci dessus on obtient :

<système, Nom, SUJ> → Actant

<travail, Nom, OBJ> → Actant

Le schéma de la Figure 46 en annexe nous montre les chemins trouvés par l'analyse Syntex de la phrase.

Certains BIOS ACCEPTENT même directement les souris

Le chemin lien ADV entre la lexie ACCEPTER et le mot directement est extrait. La règle 4 s'applique à ce lien et nous retourne l'adverbe directement comme un participant de la lexie ACCEPTER. Ce participant est noté dans la règle par : <directement, Adv, ADV>. Le participant adverbial directement de trait <Adv, ADV> est considéré comme un circonstant dans tous les exemples de notre corpus.

<directement, Adv, ADV> → circonstant

Dans certains cas, les traits « ensemble de dépendances syntaxiques » ne sont pas suffisants pour distinguer un actant d'un circonstant. C'est le cas du trait <à + NOMPREP> qui peut servir à identifier un actant pour certains verbes et un circonstant pour d'autres. Le participant de ce trait dépend du trait lui-même et du verbe avec lequel il est employé. Par exemple :

Abandonner le processus [$_{Prep}$ à] ce [$_{NOMPRE}$ P stade]

Accéder [$_{Prep}$ à] une [$_{NOMPREP}$ information]

Affecter une [$_{OBJ}$ valeur] [$_{Prep}$ à] une [$_{NOMPREP}$ variable].

Dans ces trois exemples, stade, information et variable sont des participants. Stade est un circonstant de ABANDONNER, information est un actant COMPLÉMENT d'ACCÉDER (qui compte deux actants) et variable est un actant complément de AFFECTER (qui compte trois actants).

On constate que certains traits ne sont pas suffisants pour détecter le type du participant, il faut tenir compte aussi de la lexie verbale avec laquelle est employé ce trait. Nous proposerons de désambiguïser ce cas en regardant si la lexie verbale n'admet pas de participant objet, par exemple la lexie ACCÉDER ci-dessus. Mais cette solution naïve pose un problème parce que certains verbes se réalisent en trois actants, par exemple AFFECTER.

Selon le verbe et le trait, on peut trouver le type du participant. Pour cela, nous utiliserons le calcul de la fréquence relative de la lexie verbale et du trait de son participant, sachant que s'il apparaît dans presque tous les contextes de cette lexie, on peut en déduire le taux d'importance de sa présence avec cette lexie. La plupart des lexies devraient apparaître dans de nombreux contextes avec une grande partie de leurs actants (dans des phrases où le verbe est utilisé à l'actif). Nous pourrons faire l'hypothèse que les actants apparaissent plus souvent dans les contextes que les circonstants. On a parfois des verbes qui ont leurs participants actants omis syntaxiquement. Cette notion de fréquence relative est utilisée par Messiant, Korhonen et Poibeau [48] dans la réalisation du lexique « LexSchem » présentant la sous catégorisation des verbes français. Ils se sont aussi basés pour ce travail de sous-catégorisation sur l'analyseur Syntex pour dégager les relations

entre prédicat et ses arguments. La fréquence relative est utilisée pour filtrer les arguments extraits par le processus d'extraction.

La fréquence relative, telle que définie par Messiant et al. [48], est utilisée entre la lexie et le trait qui nous permettra de classifier son participant prépositionnel de trait <Prep, NOMPREP> ou d'un autre trait difficile à distinguer en actant ou en circonstant est donnée par la formule suivante :

$$frequence\_relative = \frac{\#(Lexie, trait)}{\#(lexie)}$$

La fréquence relative est le rapport entre le nombre de contextes où est apparue la lexie avec le trait et le nombre total de contextes où la lexie est apparue. Ainsi, #(lexie, trait) est le nombre de fois que la lexie et trait sont apparus ensemble et #(lexie) est le nombre total de fois que la lexie est apparue (avec le trait ou sans lui). Ainsi, cette fréquence nous permet de désambiguïser les cas où le participant peut être actant ou circonstant selon le verbe avec lequel est utilisé.

Dans notre corpus, nous avons calculé la fréquence relative de tous les participants prépositionnels rencontrés actant ou circonstant. Pour chaque valeur de fréquence trouvée, nous avons noté 1 si elle correspond à un actant dans le corpus et 0 si elle correspond à un circonstant. La montre Figure 30 l'intervalle de valeurs de la fréquence relative où sont concentrés les actants (valeur 1 sur l'axe des ordonnées) et les non actants (valeur 0) qui correspondent aux circonstants.

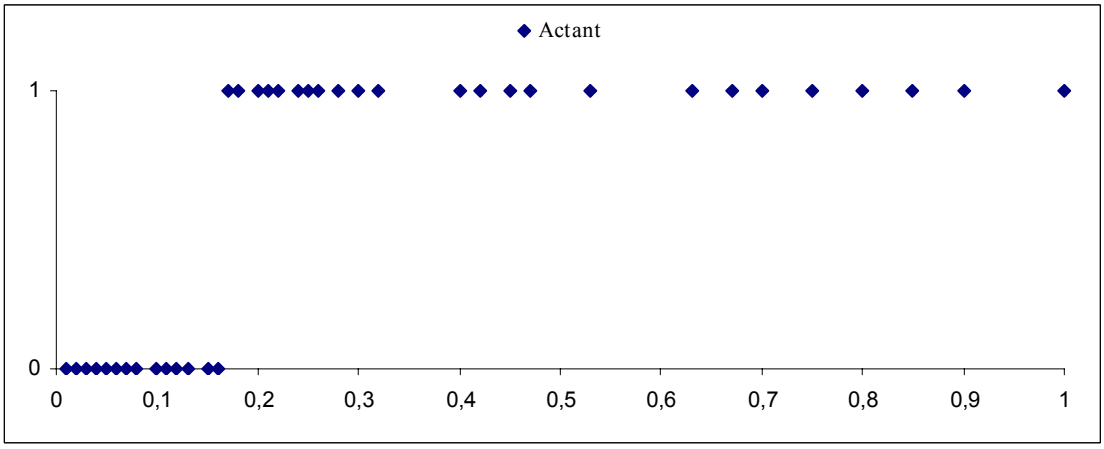


Figure 30 Fréquences relatives correspondantes à actant ou non actant

Nous constatons sur la Figure 30 que les fréquences au dessous de 0,2 correspondent à la valeur 0 c'est-à-dire non actant et celles au dessus de 0,2 correspondent à la valeur 1 c'est-à-dire actant. Nous avons déterminé le seuil=0.2 auquel nous comparons la fréquence relative de nouvelles instances de lexies. Nous avons soit le participant prépositionnel est un actant ou un circonstant. Si la fréquence relative de ce participant est supérieure à ce seuil alors ce participant est considéré comme un actant. Si cette fréquence est inférieure à ce seuil alors ce participant est considéré comme un circonstant.
En calculant la fréquence relative, nous avons évalué la distinction de participant identifiés en actants et circonstants à 80% en termes de F-mesure pour les actants et 70% pour les circonstants.

6.9 Conclusion

Nous avons exposé la méthode par extraction de règles que nous avons utilisée pour identifier les participants actants et circonstants. Nous nous sommes basés sur les dépendances syntaxiques identifiées par Syntex pour construire ces règles. Les résultats sont meilleurs que d'autres travaux qui ont utilisé aussi la caractéristique de dépendances syntaxiques. Nous avons l'avantage de disposer d'un corpus annoté manuellement, ce qui nous a permis d'extraire plus de règles d'identification à partir des chemins de liens syntaxiques de Syntex. Certains participants importants peuvent apparaître dans le contexte sans partager un lien syntaxique avec la lexie annotée (par exemple « X permet de + lexie annotée », « X est capable de + lexie annotée », « X empêche Y de + lexie annotée »). Ils sont alors annotés et considérés comme des liens indirects. Notre corpus nous a permis de rajouter certaines règles liées aux cas indirects que l'analyseur Syntex ne détecte pas.

Le chapitre suivant présente une approche d'identification des actants et circonstants par apprentissage machine.

Chapitre 7 Identification des actants et circonstants par apprentissage machine

Dans notre travail, 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 traits pour la classification des participants. Les règles que nous avons extraites au Chapitre 1 ont permis d'identifier des traits sur lesquels sont testés les classificateurs de Weka. Nous avons proposé des traits basés sur deux approches :

- les relations de dépendance trouvées par un analyseur syntaxique, en considérant les mots dépendants de la lexie comme participants candidats;
- les catégories grammaticales des mots de la phrase où apparaît la lexie verbale susceptibles d'être des participants de cette lexie.

Aucune de ces stratégies ne permettant d'effectuer une distinction parfaite entre les deux types de participants, nous souhaitons évaluer jusqu'à quel point elles peuvent y contribuer.

Dans le chapitre précédent, nous avons testé une approche à base de règles en interprétant les liens syntaxiques entre la lexie et les mots de la phrase. 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 dont nous aurons besoin. Nous proposons l'utilisation de l'apprentissage machine pour une classification supervisée basée sur ces informations syntaxiques. Les expérimentations liées à ce chapitre ont été présentées lors des communications acceptées à des conférences internationales [31]

7.1 Classification des participants par Weka[21]

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 traits construits à partir des informations et liens syntaxiques pour

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

retrouver les participants. Ces informations et liens syntaxiques sont illustrés dans la Figure 31.

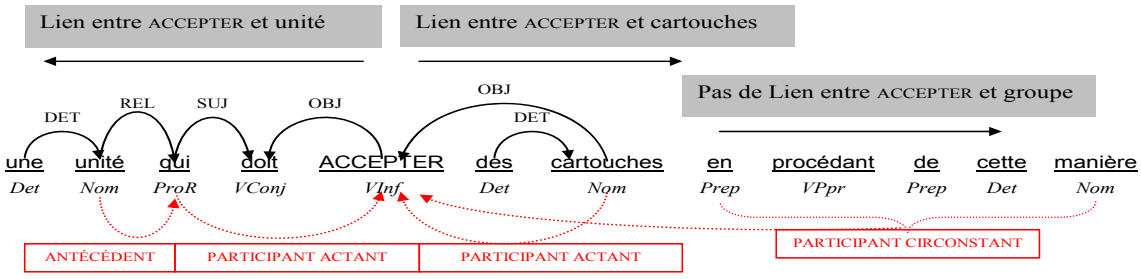


Figure 31 Exemple de contexte de la lexie ACCEPTER et ses participants avec les liens syntaxiques

Le Tableau 7 montre la description des participants de la lexie ACCEPTER dans la phrase de la Figure 31 par des informations syntaxiques que nous considérons comme des traits de classification.

Traits / Participants	Lexie	Cat-Lexie	Mot	Cat-Mot	Posi-tion	Dist-ance	Lien-syntaxique-cg	Lien-syntaxique-fn
unité	ACCEPTER	Vinf	CN	Nom	avant	2	doit, ProRel	OBJ, SUJ, REL
qui	ACCEPTER	VInf	Qui	ProRel	avant	1	doit	OBJ, SUJ
doit	ACCEPTER	VInf	Doit	VConj	avant	0	nul	OBJ
cartouches	ACCEPTER	VInf	CN	Nom	après	1	Nul	OBJ

Tableau 7 Participants de l'exemple de la Figure 31 décrits par les traits : 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 traits sont décrits au Tableau 8

7.2 Identification des participants

Les participants d'une lexie verbale annotée manuellement dans le corpus prennent la forme d'un groupe nominal dont la réalisation est un nom ou pronom, d'un groupe prépositionnel dont la réalisation est un nom ou d'un groupe adverbial dont la réalisation est un adverbe. Nous répartissons les mots dans les catégories classiques de « mots pleins »

et « mots vides » [65]. Dans notre travail, nous avons considéré comme participants candidats aux lexies verbales les groupes nominaux, les adverbes et les pronoms relatifs. Nous avons traité 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 31. Nous avons aussi considéré certaines prépositions telles que « en » et les coordinations de substantif telles que « si ».

Pour réaliser cette tâche d'identification, nous avons extrait des traits décrits dans Tableau 8 en nous inspirant de ceux de base définis par Gildea & Jurafsky [24].

Nom traits	Signification
Lexie	Unité lexicale verbale à l'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 reliée à 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 8 Traits de base de classification

Le trait **Mot** du Tableau 8 peut prendre des valeurs réelles comme contenu dans les cas de verbes modaux comme permettre ou les cas de certains adverbes. Dans d'autres cas, nous pouvons prendre la classe correspondant à la catégorie grammaticale 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 trait **Mot**.
- Si le mot est un 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 correspondent à 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 pourtant identifiés adverbes 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 traits à ajouter à ceux du Tableau 8 qui diffèrent selon leur utilisation des relations de dépendance syntaxique entre la lexie et les autres mots de la phrase.

7.2.1 Traits de classification basés sur les relations de dépendance de Syntex

Nous avons suggéré 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 les relations de dépendance. Ces dernières constituent un autre critère de classification. Deux autres traits, définis dans le Tableau 9, sont ajoutés à ceux du Tableau 8.

Nom traits	Signification
Lien-syntaxique-cg	Chemin de **Lexie** à **Mot**. Ce chemin est un ensemble de liens syntaxiques trouvés 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 9 Traits de la classification en se basant sur Syntex

Les valeurs possibles du trait 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 prenons pour certains verbes leurs valeurs réelles. Ce sont les verbes dits *modaux* dont nous avons défini une liste à partir de 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 pas sa catégorie. La préposition à et de n'identifie pas les mêmes participants.

Nous avons testé les classificateurs Weka sur les traits proposés ci-dessus individuellement pour savoir leur contribution à la tâche d'identification des participants. Les résultats de cette classification sont donnés par la F-mesure de la Figure 32

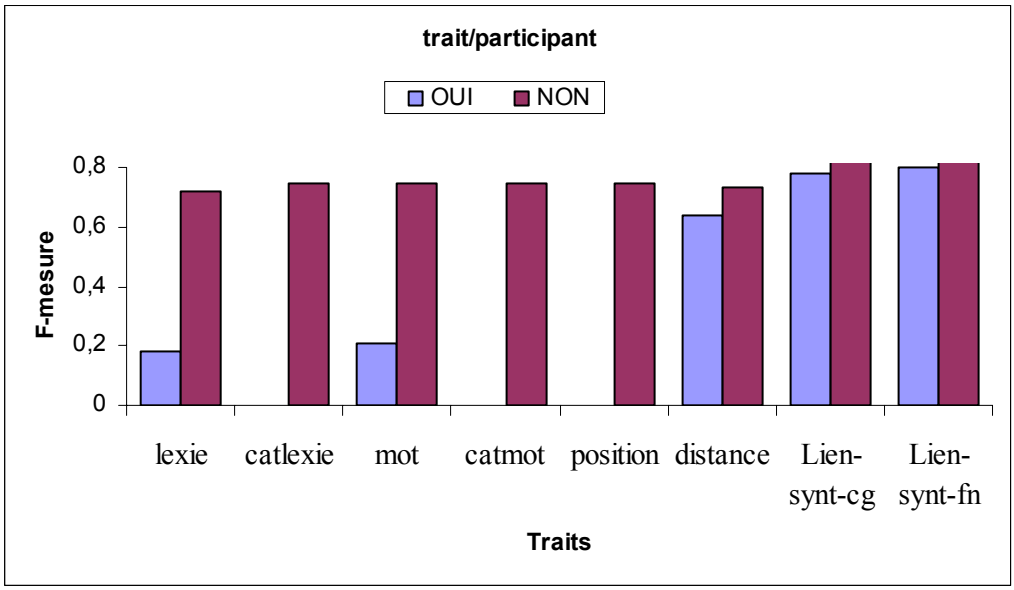


Figure 32 Contribution de chaque trait à la classification des participants

Nous constatons que la les traits catégorie grammaticale de la lexie (catlexie), catégorie grammaticale du mot candidat (catmot) et position ne contribuent pas à la classification des participants. Sachant qu'avec Syntex, nous avons pris que les mots candidats qui ont un lien syntaxique avec la lexie, la fonction syntaxique (lien-synt-fn) et les catégories grammaticales des mots (Lien-synt-cg) qui séparent la lexie et le mot candidat donne la meilleure contribution. Ces traits donnent une F-mesure respectivement d'environ 80% et 78%. Nous avons combinés les traits qui ont pu contribuer à la classification tels que lien-synt-fn, lien-synt-cg, distance, mot et lexie et nous avons enlevé les traits catlexie, catmot et position qui n'ont pas été des traits discriminants. Le meilleur résultat de F-mesure sur la combinaison des traits est de 0,86% pour le classificateur RandomForest. Les résultats de la classification sont donnés à la section 7.2.4.

Il arrive également que Syntex omette certains liens entre des éléments de la phrase et la lexie. Comme dans l'exemple de Figure 31, Syntex ne détecte aucun lien entre la lexie et la proposition « en procédant de cette manière ». Cette dernière est considérée comme un participant circonstant qui exprime 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. Nous avons proposé d'autres traits de classification sans cet analyseur.

7.2.2 Traits de classification sans l'analyseur Syntex

Dans l'approche précédente faisant appel à Syntex, nous avons remarqué 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 traits 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 traits du Tableau 8, d'autres traits 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.

Nous avons également testé la contribution individuelle des traits 1) à 5) déterminés ci-dessus dans la classification des participants. La Figure 33 donne la F-mesure de chacun des traits

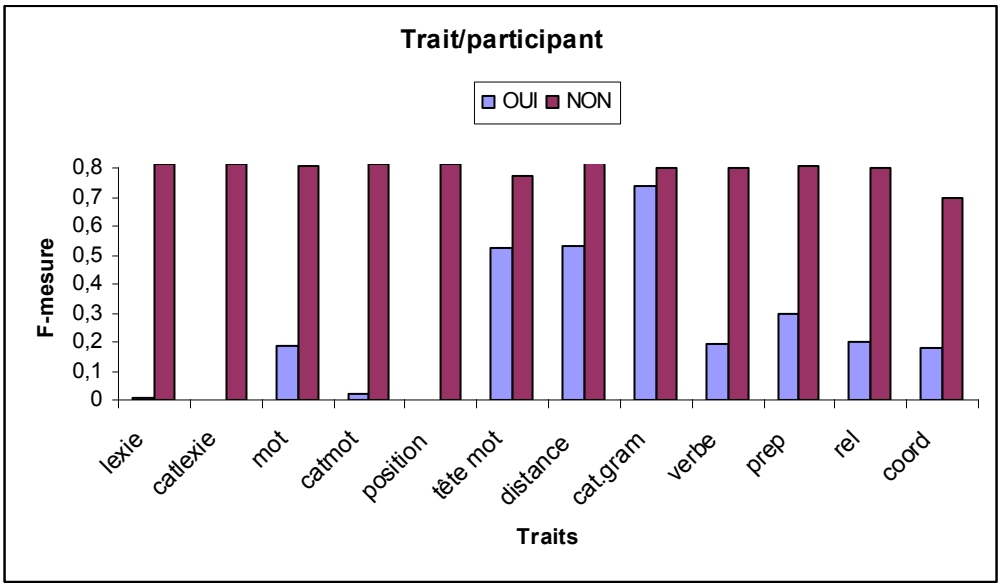


Figure 33 Contribution de chaque trait dans la classification des participants sans syntex

À la section 7.2.4, nous avons présenté les résultats de la classification des classificateurs Weka en utilisant la combinaison des traits qui ont contribué individuellement à la classification tels que les traits mot, tête mot (ce qui précède le mot candidat), distance, catégories grammaticales entre la lexie et le mot candidat (cat.gram), verbe modal, préposition (prep), pronom relatif (rel) et la coordination (coord) qui se trouvent entre la lexie et le mot candidat, et en enlever ceux qui ne sont pas discriminants tels que lexie, catégorie de la lexie (catlexie), catégorie du mot (catmot) et position. Le meilleur résultat de F-mesure en combinant les trais est de 0,76%. (voir section 7.2.4).

7.2.3 Participants propositionnels et les participants composés

Les participants ayant la forme de propositions ne sont pas identifiés en entier : uniquement la tête de la proposition est retournée par notre système. Par exemple :

> 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 normalement considérée comme un participant circonstant de la lexie verbale ABANDONNER. Avec notre système toutefois, nous n'identifions qu'une seule unité de toute la proposition, ici si, car le vecteur de traits que nous avons défini ne correspond qu'à une seule unité à la fois. Dans ce cas, nous soumettons le verbe de la proposition en question à notre système pour en identifier les participants. Nous générons la proposition en nous basant sur l'unité tête renvoyée par le système, le verbe et ses participants.

D'autres participants se réalisent sous forme de mots composés de la forme NN (nom suivi d'un nom) ou de NA (nom suivi d'un adjectif). Dans ce cas aussi, le vecteur trait, qui détecte une seule unité à la fois, ne peut pas détecter le mot composé. Cependant, notre système détecte les unités composant le mot séparément comme des participants différents et non pas comme une seule entité correspondant à un seul participant. Nous utilisons le DicoInfo (voir section 2.1.1.1) pour vérifier si ces entités sont réellement des entités composées par exemple traitement de texte, disque dur, etc. Nous vérifions si le mot composé constitue une seule entrée du DicoInfo, alors nous pourrons prendre ce mot

comme un seul participant. S'il ne constitue pas une entrée du DicoInfo, alors nous considérons les unités composant ce mot comme des participants différents.

7.2.4 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 37, nous présentons les classificateurs avec un taux de classification élevé, supérieur à 78%. Le tableau de gauche de la Figure 34 montre le taux de classification sur les 8376 instances en utilisant l'analyseur syntaxique Syntex comparé à celui obtenu 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-mesure 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	89,70	86,53
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	86,8	76,00
RTree	78,6	69,60
BFTree	82,4	71,00

Figure 34 : Résultats des classificateurs : IBk avec k=5 plus proches voisins qui utilise la distance euclidienne convertie en poids ; RFTree : RandomForestTree 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 DecisionTree. 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.

Les résultats de la Figure 34 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 76 %, ne propose pas d'aussi bons résultats que celle faisant appel à un analyseur en dépendance (avec Syntex), qui se situe autour de 86,8 %. Mais un taux de 76

% reste intéressant, et laisse supposer que les traits que nous avons proposés pour cette approche sont prometteurs. Le trait **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.

7.3 Distinction entre actant et circonstant

Syntex ne permet pas de distinguer les actants des circonstants directement. Il permet d'affecter des fonctions syntaxiques telles que sujet ou objet qui correspondent à des actants. Par contre, le problème se pose pour des mots introduits par des prépositions et les liens indirects. Syntex affecte dans tous les cas la fonction NOMPREP. Par exemple :

On ABANDONNE l'opération à ce stade

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 considéré comme un actant de la lexie ACCÉDER. Les participants identifiés à la section 7.2 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 correspondent généralement à des circonstants.
- Les noms régis par des verbes modaux, voir section 7.2.1, 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.

Nous avons ajouté la fréquence relative définie au chapitre précédent à la section 6.8 aux autres traits de classification définis dans les sections ci-dessus dans ce chapitre.

7.3.1 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 35 : 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 excellent. 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.

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 35 Tableaux de résultats des classificateurs sur la classe actant et 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 37

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. Ceux-ci varient entre 80 % et 90 % et font appel à une approche similaire de traits 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 [13] qui utilise une autre approche sur le français pour distinguer les arguments des adjoints.

7.4 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 traits, 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 traits 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 (le cas des liens indirects), l'approche sans Syntex vient y remédier. Avec cette dernière, plusieurs participants non détectés par Syntex sont bien repérés.

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, qui ne sont pas vues dans le corpus. La validation des résultats a donné un taux de 80%. Les participants actants et circonstants sont identifiés avec cette approche, nous passons maintenant à annoter les actants identifiés par des rôles sémantiques. L'approche d'automatisation de cette annotation est présentée dans le Chapitre 1.

Chapitre 8 Attribution de rôles sémantiques à des actants

Dans le chapitre précédent, nous avons identifié les actants d'une lexie prédicative verbale qui peuvent jouer des rôles sémantiques différents d'une lexie à une autre. Ils sont alors annotés différemment. Une liste d'étiquettes de rôles (Agent, Assaillant, Patient, Source, Destination, Instrument, Lieu, Cause, Récipient, Résultat, Matériau, Environnement, Substitut, etc.) a été définie pour les termes de l'informatique par le groupe de L'OLST. Cette liste toujours en construction, énumère les critères sémantiques et syntaxiques sur lesquels les linguistes se basent pour attribuer un rôle à un participant. Les exemples suivants, tirés du corpus annoté manuellement de l'informatique et d'Internet, illustrent quelques rôles :

[$_{\text{Assaillant}}$ Des virus] ne s'attaquent pas [$_{\text{Destination}}$ aux Mac]

Pour désinstaller [$_{\text{Patient}}$ le logiciel][$_{\text{Source}}$ de votre PC][$_{\text{Instrument}}$ Un logiciel] permet de naviguer [$_{\text{Lieu}}$ sur le Web]

[$_{\text{Agent}}$ Des virus] comme Tchernobyl ou [$_{\text{Cause}}$ certains échecs] de flashage corrompent [$_{\text{Patient}}$ le boot block]

[$_{\text{Agent}}$ Vous] voulez affecter [$_{\text{Patient}}$ une valeur] [$_{\text{Récipient}}$ à une variable]

[$_{\text{Patient}}$ Les codes] envoyés sont convertis [$_{\text{Résultat}}$ en Unicode]

[$_{\text{Patient}}$ Les commentaires] écrits [$_{\text{Matériau}}$ avec la syntaxe]

[$_{\text{Patient}}$ Des programmes] s'exécutent [$_{\text{Environnement}}$ sur une machine]

Ces rôles sémantiques sont accompagnés d'informations syntaxico-sémantiques, de définitions et d'exemples. Le Tableau 10, illustre certains rôles sémantiques (Agent, Assaillant et Cause) tels qu'ils sont définis dans le guide d'annotation [38] de l'OLST.

<table>
<tr><th>Rôle sémantique</th><th colspan="2">Définition</th><th>Actant ou circonstant</th><th>Classe sémantique</th></tr>
<tr><td rowspan="2">Agent</td><td colspan="2">Le participant à l'origine de l'action.
Le participant à l'origine de la création ou de l'utilisation de qqch.
Le participant jouant le rôle de qqch.</td><td>Toujours actant</td><td>Animé
Machine
Logiciel</td></tr>
<tr><td>Exemples</td><td colspan="3">L'ordinateur (agent) traite des données (patient)
Un client est une machine (agent) qui adresse une requête (patient) au serveur (destinataire).</td></tr>
<tr><td rowspan="2">Assaillant</td><td colspan="2">Participant réalisant une action nuisible.
Participant dont on veut se débarrasser.</td><td>Toujours actant</td><td>Virus
Pirate</td></tr>
<tr><td>Exemples</td><td colspan="3">80 % des virus (Assaillant) ne s'attaquent pas aux Mac.
Un hacker (Assaillant) est parvenu à s'introduire dans le réseau (Destination)</td></tr>
<tr><td rowspan="2">Cause</td><td colspan="2">Participant réalisant une action évoquant une action réalisée par un agent</td><td>Actant ou circonstant</td><td>Activité
Erreur</td></tr>
<tr><td>Exemples</td><td colspan="3">L'abandon (Cause) durant la procédure de gravure peut corrompre le disque (Patient) ou le rendre inutilisable.
La moindre erreur (Cause) ferait échouer le démarrage du serveur X (Patient).</td></tr>
</table>

Tableau 10 informations syntaxico-sémantiques des rôles sémantiques Agent, Assaillant et Cause

8.1 Analyse des participants actants du corpus

Dans notre corpus, chaque participant est décrit par sa fonction syntaxique, son groupe syntaxique, son type et son rôle. Dans le Tableau 11, nous récapitulons pour chaque rôle, certaines caractéristiques syntaxiques correspondantes. La colonne 1 indique le nombre de participants ayant un rôle donné, cité dans la colonne 2, dans tout le corpus. Les colonnes 3 et 4 indiquent les fonctions syntaxiques et groupes syntaxiques possibles pour des actants portant ces rôles sémantiques. La colonne 5 indique la valeur de la préposition dans le cas où le groupe syntaxique est prépositionnel.

nombre	Rôles	Fonctions syntaxiques[22]					Groupes syntaxiques	Préposition
		suj	obj	compl	LI	tête		
1992	Patient	x	x	x	x	x	SN, SP, Pro, Prop	sur, entre, de, dans
756	Agent	x	x	x	x		SN, SP, Pro	par
370	Destination	x	x	x	x	x	SN, Pro, SP, SAdv	dans, sur, à, au sein de, vers, etc.
72	Source	x	x	x			SP, SN, Pro	à partir de, dans, depuis, de, sur
72	Lieu			x	x		SP, SN, SAdv, Pro	sur, dans, de, en, à, sous
61	Instrument	x		x	x		SN, SP, Pro	à l'aide de, avec, par, grâce, à, sur
31	Assaillant	x		x	x		SP, Pro, SN	par
26	Matériau			x	x		SP, SN, Pro	dans, en, avec, sur, selon, à
20	Récipient			x	x		SP, Pro, SN	à, pour
18	Résultat			x			SP	en, dans
15	Cause	x			x		Prop, SN	
12	Environnement			x			SP	sur, sous, dans

Tableau 11 Caractéristiques syntaxiques des unités associées à des rôles sémantiques dans le corpus

Nous constatons que tous les sujets ne jouent pas toujours le rôle d'agent et que tous les objets ne sont pas toujours des patients (voir section 3.3). Les fonctions syntaxiques et groupes syntaxiques ne déterminent donc pas un rôle d'une manière unique. Nous avons défini d'autres caractéristiques complémentaires décrites au Tableau 12 : les caractéristiques concernant la lexie, celles concernant l'actant et celles concernant la lexie et l'actant. Deux colonnes sont utilisées. La première colonne indique le nom de la caractéristique et la deuxième colonne, sa signification.

Nous remarquons aussi que certains rôles de notre corpus, cités dans le Tableau 11 tels que Assaillant, Lieu, Instrument, Source, Récipient, Cause, Environnement et Matériau ne sont pas suffisamment représentés par des exemples. Ceci nous a conduit à rassembler ces rôles dans une seule classe appelée Autre. Nous avons ainsi restreint la liste des rôles à prendre en compte dans l'apprentissage machine. La liste finale des rôles étudiés ici est donc Agent, patient, Destination et Autre.

[22] Les abréviations des fonctions syntaxiques dans le Tableau 11 sont : (suj : sujet, obj : objet, compl : complément et LI : Lien Indirect)

La lexie	
Caractéristique	**Signification**
Lexie	Une lexie peut affecter un rôle sémantique à ses actants : ex. une lexie L1 peut affecter un rôle R1 pour son actant sujet alors qu'une autre lexie L2 lui affectera un rôle R2.
Tête de la lexie	La tête est sous forme d'une préposition quand la lexie est précédée par une préposition telle que de ou à. Un actant sujet d'une lexie précédée par la préposition à est patient.
Catégorie lexie	La catégorie grammaticale de la lexie (conjugué, participe passé, participe présent, etc.)

L'actant	
Caractéristique	**signification**
Mot actant	La réalisation de l'actant de la lexie.
Tête de l'actant	Les actants prépositionnels jouent des rôles sémantiques selon la tête (préposition utilisée) : à, avec, dans, sur, pour, etc.
Fonct-synt-act	Ce sont les fonctions syntaxiques (sujet, objet, lien-indirect, modificateur, tête, complément) rencontrées dans notre corpus
Group-synt-act	Sont telles que SN (syntagme nominal), SP (syntagme prépositionnel), etc.

Relations entre la lexie et son actant	
caractéristique	**signification**
Position	Position où apparaît l'actant par rapport à la lexie. « avant ou après »
Distance	Le nombre de mots plein qui séparent un actant de sa lexie
Ordre	Ordre de l'actant par rapport aux autres actants de la lexie dans une même phrase. C.-à-d., s'il est 1er actant, 2ème actant, etc.
Nombre_actant	Nombre d'actants que possède la lexie dans le contexte en question
Verbe1 et verb2	Si un actant sujet est suivi du verbe être alors cet actant ne peut être celui qui fait l'action mais plutôt celui qui subit l'action. il faut spécifier les verbes être et avoir s'ils y sont. Si présence d'un verbe, autre que avoir et être, entre la lexie et son actant on donne la valeur du verbe. Particulièrement pour les liens indirects, employés par les verbes modaux tels que permettre, demander, pouvoir, vouloir etc. affectent des rôles sémantiques spécifiques.
Nombre_verbe	Nombre de verbes qui apparaissent entre la lexie et son actant

Tableau 12 Caractéristiques décrivant un actant

8.2 Description des données

Les éléments de la première colonne du Tableau 12 constituent les traits décrivant un actant. Chaque actant de lexies différentes et contextes différents est décrit par des valeurs attribuées à chacun de ces traits. Par exemple

[Actant, Agent Vous] CLIQUEZ sur [Actant, Patient le bouton]

Les actants Vous et le bouton de la lexie CLIQUER sont décrits par les traits suivants :

Vous													
Traits de la lexie			Traits de l'actant				Traits entre la lexie et son actant						
lexie	Cat.	Têt	Act	Cat.	têt	FS	V1	V2	Pos	Dist	NbrV	Ordr	NbrA
cliquer	Vconj	?	vous	Pro	?	Sujet	?	?	avant	0	0	1	2

bouton													
Traits de la lexie			Traits de l'actant				Traits entre la lexie et son actant						
lexie	Cat.	Têt	Act	Cat.	têt	FS	V1	V2	Pos	Dist	NbrV	Ordr	NbrA
cliquer	Vconj	?	bouton	Nom	?	Objet	?	?	après	1	0	2	2

Chaque ligne de ce type décrit un actant d'une lexie. Nous disposons de 3445 lignes ou instances correspondant à chacun des actants de différentes lexies dans des contextes différents rencontrés dans notre corpus. Les valeurs représentées par « ? » sont des valeurs absentes ou indiquent que ce trait ne concerne pas l'actant en question dans ce contexte. Nous avons proposé en premier lieu une classification des rôles sémantiques. Nous avons utilisé le package de Weka [67] pour classifier ces rôles.

8.3 Classification des rôles sémantiques par Weka

Nous avons testé le classificateur RandomForest, le classificateur qui a donné les meilleurs résultats pour les tâches précédentes, sur nos données décrites par les traits du Tableau 12. Nous avons d'abord testé ce classificateur en tenant compte des traits individuellement. La contribution de ces traits à la classification des rôles sémantiques est donnée par la F_mesure de la Figure 36

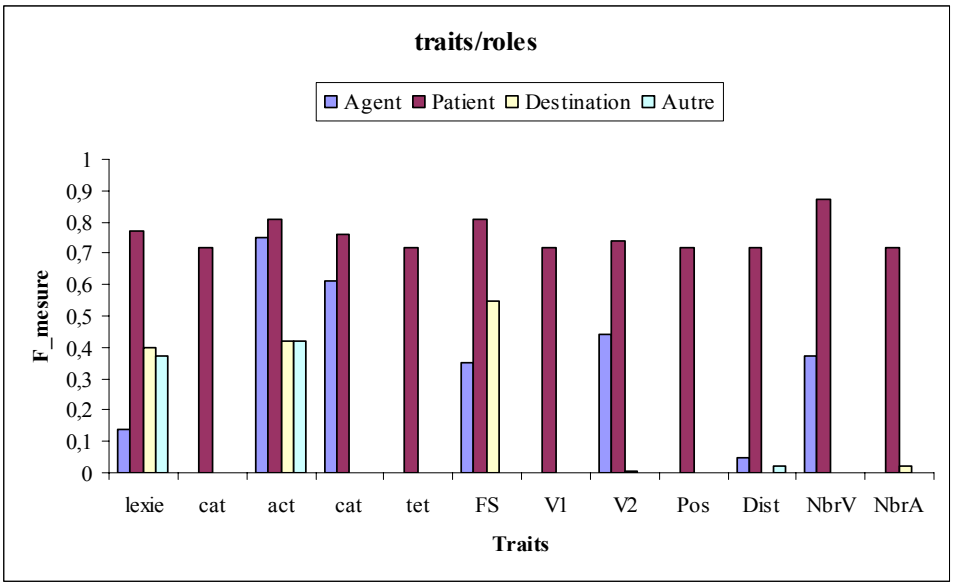


Figure 36 Contribution de chaque trait à la classification des rôles sémantiques

On constate à Figure 36, que s'il est assez facile d'identifier le rôle Patient, les autres rôles sont beaucoup plus difficiles à départager. La combinaison de ces traits permet de distinguer ou de démarquer ces rôles sémantiques plus efficacement que de les prendre individuellement. Dans le Tableau 13, nous donnons les résultats en précision, rappel et F_mesure, du classificateur RandomForest en considérant tous les traits. Ces résultats sont calculés sur la validation croisée (10 folds) du corpus d'entraînement composé de 3445 actants.

	RandomForest		
	Précision	Rappel	F_mesure
Agent	0,94	0,86	0,90
Patient	0,90	0,97	0,93
Destination	0,90	0,82	0,85
Autre	0,85	0,69	0,76

Tableau 13 Combinaison de tous les traits dans la classification des rôles

La classification des rôles sémantiques donne de très bons résultats sur les rôles proposés pour les 3301 actants rencontrés dans le corpus. Nous savons que les rôles sémantiques ne constituent pas une liste finie. Ces derniers sont découverts au fur et à

mesure que l'annotation de nouvelles lexies ou données est réalisée. De nouveaux rôles peuvent apparaître et que la classification n'a pas pris en compte auparavant. L'objectif premier de notre travail est d'accompagner le linguiste annotateur pendant l'annotation des rôles sémantiques. Notre système devrait donc être capable de donner à l'annotateur la possibilité non seulement de choisir le rôle proposé mais aussi d'en définir un nouveau. Pour intégrer cette façon de faire dans le système d'annotation manuel, nous proposons donc de classifier les instances d'actants décrites à l'aide des traits définis ci-dessus, au lieu de classifier les rôles sémantiques tel que vu dans les travaux antérieurs (voir chapitre 4). Nous regroupons toutes les instances proches ou qui partagent les mêmes valeurs des traits suggérés dans les mêmes groupes. Nous proposons de faire un partitionnement semi supervisé sur l'ensemble de nos données. C'est-à-dire nous regroupons les instances dans des partitions selon leur similarité et ensuite classifions les rôles sémantiques dans ces groupes. Pendant le regroupement, les rôles sémantiques sont ignorés.

8.4 Partitionnement semi supervisé

Le partitionnement est un processus de regroupement des données dans des groupes afin que des objets dans un même groupe aient une similarité plus élevée entre eux qu'avec les instances d'un autre groupe. Dans notre cas, nous procédons au regroupement semi supervisé dont les instances à regrouper possèdent des étiquettes qui sont les rôles sémantiques. Dans les regroupements et la mesure de similarité, nous ne tenons pas compte de ces étiquettes. Une fois les groupes sont formés, nous affectons aux instances leurs étiquettes ainsi nous classifions les étiquettes dans chaque groupe. Donc dans un groupe, on peut avoir plusieurs étiquettes représentées.
Dans le processus de partitionnement, deux conditions doivent être vérifiées : maximiser la similarité intra-cluster et minimiser la similarité inter-clusters.

Plusieurs types d'algorithmes de partitionnement existent : le clustering par partitionnement k-means est performant pour des données numériques. Le clustering hiérarchique ascendant (CURE [27], ROCK [26], CHAMÉLÉON [36], etc.) pour les données non numériques ou avec des traits catégoriels.

Nous avons testé les algorithmes de partitionnement de Weka. Nous avons trouvé un taux de 66% d'instances incorrectement classifiées. Sachant que nos données ont des traits de descriptions de type catégoriels, nous avons pensé d'utiliser un algorithme hiérarchique, En s'inspirant des autres domaines pour lesquels ça a donné de meilleurs résultats dans le cas de manipulation de données catégorielles ou non numériques.

L'algorithme CHAMÉLÉON [36] est le plus évolué des algorithmes de partitionnement hiérarchique ascendant. Il a été aussi prouvé par ses auteurs George Karypis et al. [36] que CHAMÉLÉON, par sa mesure d'inter-connectivité relative et de proximité, vient remédier aux limites des autres algorithmes hiérarchiques existants.

CHAMÉLÉON est utilisé dans plusieurs domaines, notamment en bioinformatique, pour identifier les fonctions des gènes dans une séquence. Ayant remarqué une certaine similarité entre notre problème et celui de la détermination de la fonction d'un gène étant donné sa position dans une séquence, nous avons choisi CHAMÉLÉON.

8.5 Partitionnement en utilisant l'algorithme CHAMÉLÉON

Nous regroupons nos données selon leurs mesures de similarité. Comme nous utilisons des caractéristiques catégorielles ou non numériques, la mesure de similarité entre les instances actants utilisée est celle définie en utilisant le coefficient Jaccard. Cette notion de similarité est basée sur les caractéristiques ou traits communs entre les instances. Elle est calculée pour toutes les instances deux à deux.

Si nous considérons les actants $A_1,A_2,\ldots,A_n$. Nous définissons la similarité entre un actant A_i et un actant A_j par le rapport entre le nombre de traits en commun de Ai et de Aj et l'union de ces traits. Cette similarité, notée Sim (A_i,A_j), est donnée par :

$$Sim(A_i,A_j)=|Ai \cap Aj| / |Ai \cup Aj| \qquad (1)$$

La similarité n'est pas suffisante pour fusionner des groupes car elle peut trouver des groupes avec des instances très similaires et la contrainte de partitionnement de maximiser la similarité à l'intérieur d'un même groupe est assurée. Par contre qu'en est-il de la contrainte de minimiser la similarité entre les groupes? Avec juste la similarité cette contrainte n'est pas assurée. Cette contrainte est aussi importante que la première

contrainte. Donc il faudra déterminer des mesures concernant les groupes pour minimiser les similarités entre les groupes et maximiser les similarités dans un même groupe. On cherche à obtenir dans chaque groupe un grand degré d'inter-connectivité. Dans CHAMÉLÉON, on maximise deux mesures : l'inter-connectivité et la proximité.

8.5.1 L'algorithme CHAMÉLÉON

CHAMÉLÉON (Hierarchical Clustering Using Dynamic Modeling) [Karipys 99] manipule des données modélisées sous forme de graphe. Le graphe représentant les données est basé sur l'approche des k plus proches voisins (k-pp). L'algorithme CHAMÉLÉON opère en deux phases : 1) partitionnement du graphe initial des k-pp voisins en sous-partitions ou groupes et 2) fusion basée sur un modèle dynamique de ces différents groupes en mesurant la similarité entre eux. Deux groupes sont fusionnés si l'inter-connectivité et la proximité entre eux est plus grande que celles à l'intérieur de chaque groupe. Ceci est interprété par le calcul de deux mesures entre deux clusters C_i et C_j :

1) Inter-connectivité relative, notée RI, définie comme une inter-connectivité absolue entre C_i et C_j normalisée par l'inter-connectivité interne de ces deux clusters C_i et C_j. La RI entre deux clusters C_i et C_j est donnée par

$$RI(C_i, C_j) = \frac{2 * |EC(C_i, C_j)|}{|EC(C_i)| + |EC(C_j)|} \qquad (2)$$

Où l'inter-connectivité absolue $|EC(C_i, C_j)|$ est l'ensemble des arcs qui relient les nœuds de C_i et de C_j. Elle représente l'inter-connectivité absolue entre deux clusters. $|EC(C_i)|$ et $|EC(C_j)|$ sont la somme des poids des arêtes qui divisent un cluster C_i respectivement C_j en deux clusters à peu près égaux, (minimum 25% pour chaque cluster). Elle représente l'inter-connectivité interne d'un cluster.

2) Proximité relative « Relative Closeness » notée RC, définie comme une proximité absolue entre C_i et C_j normalisée par la proximité interne du cluster C_i respectivement du cluster C_j. La proximité relative est donnée par :

$$RC(C_i, C_j) = \frac{(|C_i| + |C_j|)\,\overline{EC}(C_i, C_j)}{|C_i|\,\overline{EC}(C_i) + |C_j|\,\overline{EC}(C_j)} \qquad (3)$$

où la proximité absolue $\overline{EC}(C_i, C_j)$ est la moyenne des poids des arcs qui relient les nœuds du cluster C_i aux nœuds du cluster C_j . $\overline{EC}(C_i)$ et $\overline{EC}(C_j)$ sont la moyenne des poids des arcs qui divisent en deux un cluster C_i respectivement C_j. $|C_i|$ et $|C_j|$ est le nombre de nœuds dans C_i respectivement C_j .

Il y a plusieurs façons de diviser un cluster en deux. Dans notre cas, on veut trouver une bissection minimale. C'est-à-dire minimiser les poids des arcs qui relient les deux partitions. Le problème est NP-complet. Plusieurs heuristiques ont été proposées. Dans notre cas nous nous somme inspiré de l'algorithme de Kernighan-lin-Algorithm.

À l'aide de ces deux mesures, la fusion des clusters est faite selon deux approches. La première approche est de fusionner seulement les paires de clusters dont la RI et la RC ne dépassent pas leurs seuils respectifs T_{RI} et T_{RC} spécifiés par l'utilisateur. Dans ce cas, pour chaque cluster C_i , CHAMÉLÉON vérifie si un cluster C_j parmi les clusters adjacents de Ci satisfait les deux conditions

$$RI(C_i, C_j) \geq T_{RI} \quad \text{et} \quad RC(C_i, C_j) \geq T_{RC}$$

Et si plusieurs clusters satisfont ces deux conditions, CHAMÉLÉON choisit celui qui a la plus grande inter-connectivité absolue entre lui et Ci.

La deuxième approche consiste à maximiser une fonction combinant les deux mesures par

$$RI(C_i, C_j) * RC(C_i, C_j)^{\alpha}$$

Où α indique la préférence attribuée à ces mesures. Il est spécifié par l'utilisateur. Si $\alpha > 1$ alors CHAMÉLÉON donne une grande importance à la proximité relative, et si $\alpha < 1$ alors il donne une plus grande importance à l'inter-connectivité relative.

8.6 Application de CHAMÉLÉON

Nous avons adapté la méthode CHAMÉLÉON à nos données. Nous avons calculé la matrice de similarité pour toutes les instances. Nous avons obtenu, par l'application de la

méthode des k-pp voisins, le graphe initial nécessaire dans la première phase de CHAMÉLÉON. Ce graphe, non orienté, contient des nœuds représentant les instances d'actants décrites par des traits (voir section 1.2 et section 1.3) et des arêtes pondérées dont les poids représentent la relation de similarité entre les nœuds qu'elles relient. (voir formule (1) page 88). Nous construisons un exemple de graphe avec nos données d'instances d'actants. Prenons 8 instances d'actants représentées par I1,I2,…,I8 :

I1 (écrire,VINF,non,non,y,Pro,?,avant,0,Complement,Pro,nr,0,2,2);
I2 (affecter,VINF,non,non,lui,Pro,?,avant,0,Complement,Pro,nr,0,6,6);
I3 (affecter,VINF,non,non,lui,Pro,?,avant,0,Complement,Pro,nr,0,1,2);
I4 (démarrer,VINF,non,non,vous,Pro,?,avant,1,Sujet,SN,nr,0,3,3);
I5 (interrompre,VINF,non,V,vous,Pro,?,avant,2,Sujet,SN,nr,1,2,3);
I6 (interrompre,VINF,non,pouvoir,vous,Pro,?,avant,2,Sujet,SN,nr,1,2,3);
I7 (masquer,VINF,non,V,vous,Pro,?,avant,1,Sujet,SN,nr,1,3,3);
I8 (masquer,VINF,non,non,vous,Pro,?,avant,1,Sujet,SN,nr,0,3,3);

Nous calculons entre ces instances, deux à deux, la mesure de similarité donnée dans la formule (1). Ainsi nous formons une matrice de similarité de 8 lignes et 8 colonnes. Cette matrice de similarité est interprétée par un graphe dont les nœuds sont les instances I1,I2,…,I8 et dont le poids des arêtes sont les similarités calculées entre elle. Vu que le graphe complet est dense, nous n'avons illustré dans le graphe suivant que certains liens, pour mieux visionner le fonctionnement de CHAMÉLÉON sur les exemples

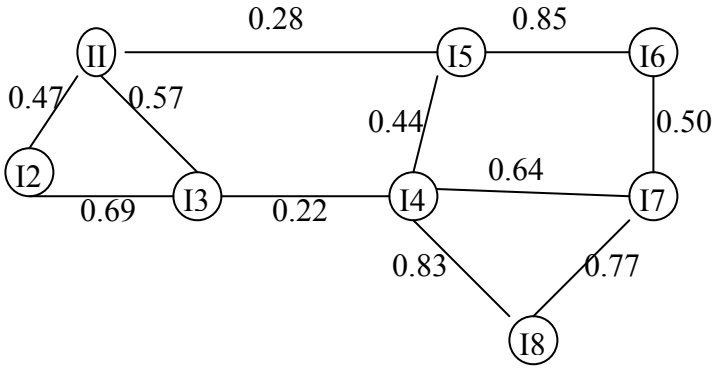


Figure 37 Graphe initial : les nœuds I1,I2,I3,I4,I5,I6,I7,I8 sont les instances et les nombres sur les arêtes sont les similarités entre ces nœuds.

Le graphe initial global de la Figure 37 est partitionné pour trouver les premières partitions. Nous utilisons un package de partitionnement de graphe «Metis » [Georg Karypis 98], le mieux adapté à notre graphe, en nous basant sur les poids de chaque arête. Nous arrivons ainsi au graphe de la Figure 38.

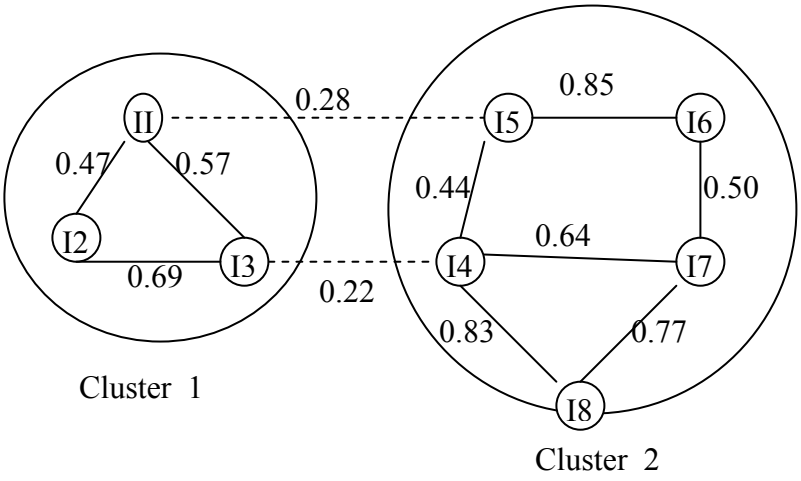


Figure 38 Division du cluster global en sous clusters. Arêtes symbolisées par (------)

Metis partitionne le graphe en choisissant les arcs dont la somme de leurs poids est minimisée et dont la taille des partitions ou clusters résultants ne peut être moins de 25% de la taille du cluster ou de la partition globale.

CHAMÉLÉON utilise le résultat de Metis, c'est-à-dire les différentes partitions trouvées et les arêtes éliminées pour calculer la mesure de l'inter-connectivité absolue entre le cluster 1 et le cluster 2 : $EC(C_1,C_2)$.

Chaque cluster est divisé en 2 clusters de taille plus au moins égales et dont la somme des poids est minimisée. Les arêtes éliminées pendant la division seront utilisées pour calculer l'inter-connectivité interne d'un cluster. La Figure 39 illustre une telle configuration. Cet exemple est juste pour expliquer le principe de l'algorithme CHAMÉLÉON. La division d'un cluster en deux est aussi complexe qu'on essaie pas de minimiser cette tâche avec cet exemple qui ne peut pas couvrir la complexité de la tâche de la division d'un cluster.

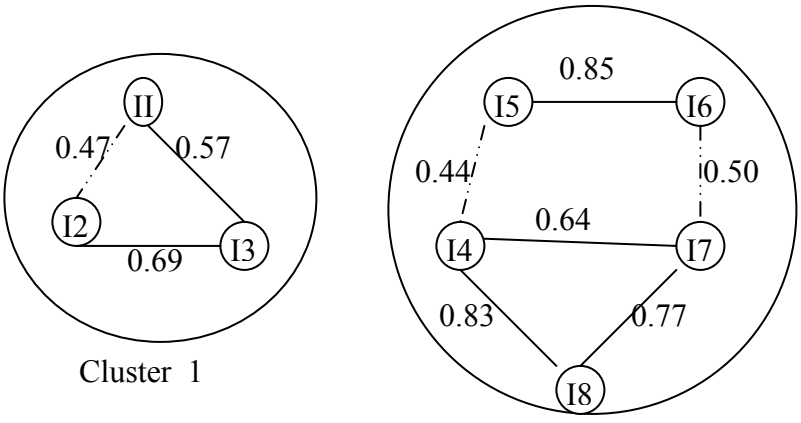


Figure 39 un même cluster subdivisé en deux : arêtes symbolisées par (- - -)

Dans le cluster 1, l'arête entre I1 et I2 est éliminée. Et dans le cluster 2, l'arête entre I4 et I5 et l'arête entre I6 et I7 sont éliminées. À partir des poids de ces arêtes éliminées noté P, la mesure de l'inter-connectivité interne du cluster 1 $EC(C_1)$ respectivement du cluster 2 $EC(C_2)$ est calculée. La mesure d'inter-connectivité RI est donnée dans le Tableau 14

$EC(C_1, C_2)$	$P_{I1\text{-}I5} + P_{I3\text{-}I4}$	0.28+0.22=0.50
$EC(C_1)$	$P_{I1\text{-}I2}$	0.47
$EC(C_2)$	$P_{I5\text{-}I4} + P_{I6\text{-}I7}$	0.44+050=0.94
$RI(C_i, C_j)$	$2*(EC(C_1, C_2))/EC(C_1)+EC(C_2)$	0.70

Tableau 14 Calcul de l'inter-connectivité relative RI pour l'exemple

Quant à la proximité absolue, les mêmes concepts sont utilisés. L'inter-connectivité absolue est alors la somme des poids des arêtes alors que la proximité absolue est la moyenne des poids des arêtes. La mesure de la proximité RC est donnée dans le Tableau 15.

$\overline{\overline{EC}}(C_1, C_2)$	$(P_{I1\text{-}I5} + P_{I3\text{-}I4}) / 2$	0.25
$\overline{\overline{EC}}(C_1)$	$(P_{I1\text{-}I2}) / 1$	0.47
$\overline{\overline{EC}}(C_2)$	$P_{I5\text{-}I4} + P_{I6\text{-}I7} / 2$	0.47
$\|C_1\|$	I1+I0+I3	3
$\|C_2\|$	I4+I5+I6+I7+I8	5
$RC(C_1, C_2)$	$\overline{\overline{EC}}(C_1, C_2)/(\|C1\|/\|C1\|+\|C2\|)*\overline{\overline{EC}}(C_1)+(\|C1\|/\|C1\|+\|C2\|)*\overline{\overline{EC}}(C_2)$	0.35

Tableau 15 Calcul de la proximité relative RC pour l'exemple

8.7 Expérimentation

Dans notre expérimentation, nous avons divisé notre ensemble de données (3301 instances d'actant) en deux parties. Nous en avons utilisé les 2/3 (2200 instances) pour la formation des clusters, et 1/3 (1101 instances) pour tester si les clusters sont bien formés. Nous avons essayé deux expériences : 1) une représentation naïve où chaque instance est considérée comme un cluster et 2) regrouper les instances similaires dans les mêmes groupes en utilisant l'algorithme CHAMÉLÉON

8.7.1 Représentation naïve

Nous avons considéré une représentation naïve, où chaque instance d'actant forme un groupement. Nous avons ainsi 2200 groupements. Nous avons classifié les instances de test dans ces groupements « naïfs » en utilisant la mesure de similarité (formule (1) (page 106)). Nous comparons chaque instance de test avec chaque groupement formé par une seule instance en nous basant sur la mesure de similarité entre instances. Nous sélectionnons le groupe le plus proche ou dont la similarité est maximale. Nous avons calculé la précision et le rappel pour estimer le taux de classification dans les groupes formés par toutes ces instances. La précision est le rapport entre le nombre d'instances de test correctement classifiées et la somme du nombre d'instances correctement classifiées et du nombre d'instances incorrectement classifiées. Le rappel est le rapport entre le nombre d'instances de test correctement classifiées et le nombre total d'instances de chaque classe. Ces deux mesures sont calculées pour chaque rôle. Le Tableau 15 présente les résultats de la formation des groupes avec la représentation naïve en se basant sur ce calcul de précision et rappel.

Rôles	Précision	Rappel
Agent	0,84	0,82
Patient	0,95	0,80
Destination	0,50	0,69
Autre	0,51	0,52

Tableau 16 résultats de la représentation naïve

Les résultats de la représentation naïve nous permettent de voir si la classification d'instances en faisant un partitionnement peut donner des résultats acceptables pour l'annotation de nouvelles instances en rôles sémantiques. À comparer à la classification des rôles sémantiques, nous constatons que ces résultats sont moins bons. Au lieu d'utiliser toutes les instances comme des groupes, ce qui n'est pas l'objectif du partitionnement, nous avons proposé de faire un partitionnement en rassemblant les instances similaires dans les mêmes groupes en sorte que les deux contraintes de partitionnement soient assurées. Nous avons proposé d'expérimenté le partitionnement avec la méthode CHAMÉLÉON présentée ci-dessus.

8.7.2 Partitionnement par l'algorithme CHAMÉLÉON

Nous avons pris les 2200 instances à partitionner selon cet algorithme. Tel que dit ci-dessus, le partitionnement que nous avons proposé est semi supervisé. Nous ignorons les rôles sémantiques pendant la phase de partitionnement. Une fois que les groupes sont formés, on associe à chaque instance du groupe son rôle sémantique original. Ainsi dans les groupes formés on peut trouver des groupes avec seulement un rôle sémantique comme on peut trouver un groupe avec plusieurs rôles sémantiques.

La méthode CHAMÉLÉON comporte plusieurs paramètres à déterminer par l'utilisateur. Pour trouver les valeurs de ces paramètres, nous avons attribué plusieurs valeurs à chacun des paramètres. Avec chacune de ces valeurs, nous avons appliqué l'algorithme CHAMÉLÉON pour former les groupes. Une fois les groupes sont formés avec ces valeurs, nous avons essayé de classifier les 1101 instances de test dans ces groupes. Nous avons évalué le résultat en termes de précision, rappel définis précédemment et de F-mesure. Nous avons pris les valeurs des paramètres qui ont donné une F-mesure maximale.

Par exemple, nous avons testé, dans la phase I de l'algorithme, des valeurs pour le paramètre k pour calculer le graphe des k-pp voisins. Nous avons testé des valeurs plus petites k=2 et k=3 et nous avons testé le graphe complet avec la valeur k=8. Nous avons constaté que si nous prenons k=2 ou bien le graphe de similarité entre les instances en entier (correspondant à k=8), nous arrivons aux mêmes résultats (voir Figure 40).

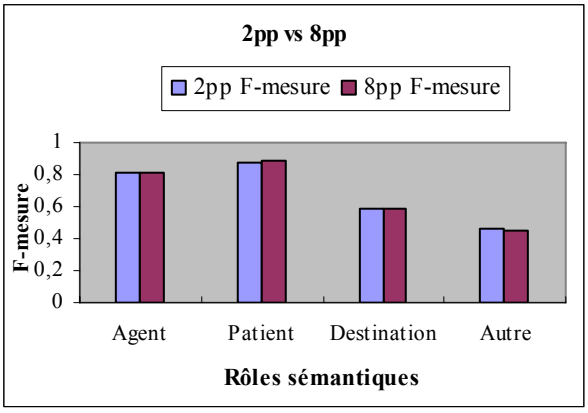


Figure 40 F-mesure obtenue à partir d'un graphe de 2pp voisin et de graphe entier

Dans ce cas, nous avons choisi le graphe le moins dense et moins coûteux en temps de partitionnement. Nous avons donc choisi k=2 dont le graphe comporte 2200 sommets et environ 106500 arcs que le graphe entier qui comporte 2200 sommets et environ 557400 arcs. Le graphe de 2pp voisin constitue le graphe initial permettant à CHAMÉLÉON de former des groupes homogènes.

Ce graphe initial doit être partitionné en sous graphes qui constituent les sous partitions initiales pour CHAMÉLÉON. Nous avons testé sur le nombre de partitions initiales à retrouver par Metis. Nous avons testé plusieurs valeurs 5,10,15,20. Nous avons choisi un nombre de partitions égal à 20. Ce qui permet à CHAMÉLÉON de fusionner les partitions qui se ressemblent. Si ce nombre est faible, CHAMÉLÉON risque de ne pas trouver les groupes naturels.

La deuxième phase de CHAMÉLÉON fusionne les groupes en se basant sur le calcul de RI et RC. Dans notre cas, nous avons utilisé la deuxième approche : $RI * RC^{\alpha}$. Nous avons testé plusieurs valeurs de α (0.5, 1, 1.5, 2, 2.5, 3). Le α=1 que nous avons choisi, d'après les tests, est celui qui donne une même importance aux deux mesures RI et RC.

Une fois les groupes trouvés, nous cherchons pour chaque groupe des représentants qui peuvent mieux décrire l'ensemble des éléments pouvant appartenir à ce groupe. Dans un groupe, on peut trouver des instances similaires avec des rôles différents. Nous avons

proposé de calculer un seuil qui permet de sélectionner les instances représentantes du groupe. Nous avons calculé pour chaque instance son importance ou son poids dans le groupe. Ce poids est calculé en utilisant la mesure de similarité (1) (page 88).

Le poids d'une instance i noté P_i dans un groupe C est donné par la somme des similarités de cette instance i avec les autres instances noté j du groupe C divisé par la somme des similarités entre toutes les paires d'instances k,j du groupe C. Le poids P_i est donné par :

$$P_i = \frac{\sum_{j \in C} Sim(i,j)}{\sum_{k,j \in C} Sim(k,j)} \qquad (4)$$

Le seuil est définit par la moyenne des poids de toutes les instances du groupe C. Nous sélectionnons toutes les instances ayant un poids supérieur ou égal à ce seuil comme instances représentante du groupe C. les groupes obtenus sont formés de ces instances représentantes.

Pour tester si les groupes obtenus sont bien formés, nous avons pris les instances de notre ensemble de test et nous avons cherché à identifier le groupe dans lequel on devrait la classer. Nous avons testé sur une centaine de tests de milles instances tirées aléatoirement. Nous avons constaté que les rôles représentés par un grand nombre d'instances, par exemple les 1992 cas de Patient et les 756 instances de Agent se retrouvent bien dans les groupes trouvés par CHAMÉLÉON. Le rôle Destination avec 370 instances et les autres rôles, représentés par la classe Autre, dont le nombre d'exemples est inférieur à 200, présentent un taux de classement acceptable malgré qu'ils ne sont pas des rôles fréquents. Avec l'algorithme CHAMÉLÉON, ces rôles eux aussi peuvent être des représentants des groupes.

Le Tableau 17 montre les résultats obtenus par CHAMÉLÉON vs la représentation naïve en utilisant les mêmes mesures de précision et de rappel définies précédemment et de F-mesure.

	Naive			Chaméléon		
	Préc.[23]	Rapp.[24]	F-mes[25]	Préc.	Rapp.	F-mes
Agent	0,84	0,82	0.82	0,75	0,88	0,81
Patient	0,95	0,80	0,86	0,88	0,89	0,88
Destination	0,50	0,69	0,57	0,66	0,53	0,58
Autre	0,51	0,52	0,51	0,62	0,37	0,46

Tableau 17 résultats de comparaison de mesures de la naïve et CHAMÉLÉON

L'algorithme CHAMÉLÉON donne également dans certains cas ambigus un choix entre deux rôles sémantiques. Ces réponses de choix ne sont pas prises dans le calcul de la précision et rappel du Tableau 17.

Un pourcentage correspondant à ces réponses des cas ambigus est calculé pour chaque classe. Le Tableau 18 montre le nombre d'instances pertinentes dans chaque classe, le nombre d'instances indécises ou à choix de deux rôles et le ratio de réponses indécises par rapport au nombre d'instances pertinentes.

Rôles sémantiques	Chaméléon		
	Nombre instances	Instances indécises	Ratio
Agent	212	5	2 %
Patient	549	5	1 %
Destination	107	4	4 %
Autre	100	3	3 %

Tableau 18 Pourcentage de réponses indécises retournées par CHAMÉLÉON

Ces réponses indécises, nous permettent de proposer diverses hypothèses de rôles aux linguistes dans les cas ambigus. Ces hypothèses faciliteront la tâche au linguiste qui sélectionnera un rôle dans une liste restreinte de rôles sémantiques.

[23] Précision
[24] Rappel
[25] F-mesure

Au terme de la F-mesure du Tableau 17, nous constatons que les résultats sont très proches entre la représentation naïve et CHAMÉLÉON. Nous constatons aussi qu'on perd dans la précision et on gagne dans le rappel pour les rôles ayant un grand nombre d'instances et le contraire pour les rôles de faible nombre d'instances. Nous pouvons dire que la contrainte d'avoir des rôles qui ne sont pas représentés par un nombre insuffisant d'instances est considérable qui rend un très faible rappel de 37%. Pour les rôles Patient et Agent les résultats ne sont pas mauvais.

Le Tableau 19 montre la comparaison de la F-mesure de CHAMÉLÉON avec celle de classification

	RandomForest	Chaméléon
	F-mesure	F-mesure
Agent	0,90	0,81
Patient	0,93	0,88
Destination	0,85	0,58
Autre	0,76	0,46

Tableau 19 Comparaison de F-mesure de RandomForest à celle de CHAMÉLÉON

La classification donne de bons et de meilleurs résultats que CHAMÉLÉON. La classification est la meilleure pour classifier les rôles sémantiques et prédire des rôles, déjà existants pour de nouvelles instances. Par contre, elle ne donne aucune initiative pour prendre en compte de nouveaux rôles et de nouvelles lexies et leur prédiction est juste de choisir un rôle le plus proche. Nous avons proposé un partitionnement semi supervisé d'instances, une perspective d'intégrer ces nouveaux rôles. Un tel processus de partitionnement sera basé sur le *feedback* ou des contraintes de l'annotateur pour lesquels dans le futur des modèles qui prendront en compte ce *feedback* peuvent être proposés.

8.8 Conclusion

Dans ce chapitre, nous avons testé la classification pour annoter les rôles sémantiques. La classification a donné de bons résultats avec un taux de F_mesure qui est compris entre 0,76% et 0,90%. La classification peut classifier de nouvelles instances avec un bon score

mais elle ne peut pas classifier de nouveaux rôles ou de nouvelles instances. Pour cela nous avons proposé une nouvelle perspective qui est de faire un partitionnement des instances ensuite classifier les rôles sémantiques. Ce qui permettra de classifier de nouveaux rôles.
Nous avons testé les algorithmes de partitionnement de Weka sur nos données et le résultat est de 66% d'instances mal classifiées. Nous avons adapté un algorithme de partitionnement hiérarchique ascendant, CHAMÉLÉON, qui a prouvé son efficacité dans plusieurs domaines, particulièrement en bioinformatique. Nous avons suggéré cette approche car nous utilisons des traits de description de nature catégoriels que les autres méthodes de partitionnement ne manipulent pas assez bien sachant qu'elles sont plutôt destinées à manipuler des données numériques.

Nous avons d'abord considéré toutes les instances comme des groupes. Ces groupes servent à la classification de nouvelles instances. Ces dernières sont comparées à ces groupes en utilisant la similarité définie à la page 106. Plusieurs actants ont exactement les mêmes valeurs pour les traits de description mais ne portent pas les mêmes rôles sémantiques. Dans ce cas la similarité entre ces actants, deux à deux, est de valeur 1. Ceci implique que si un rôle est représenté par un grand nombre d'instances tel que Patient, il l'emporte sur les autres rôles dont le nombre d'instances est moindre. Plusieurs instances sont similaires, il est plus intéressant de les regrouper dans les mêmes partitions qui est l'objectif du partitionnement en respectant la contrainte de maximiser la similarité à l'intérieur d'un groupe et de la minimiser entre deux groupes différents.

Nous avons alors testé l'approche de partitionnement supervisé qui permet de regrouper les instances semblables dans des groupes pour ensuite classifier les rôles sémantiques. Nous estimons avoir obtenu un taux de F-mesure acceptable en considérant cette approche comme une perspective tout au début à exploiter. On peut la considérée comme une baseline. L'algorithme CHAMÉLÉON choisi offre deux mesures performantes d'inter connectivité relative et de proximité relative pour fusionner dynamiquement les groupes. Il permet aussi de fusionner plusieurs groupes à la fois, contrairement aux autres algorithmes hiérarchiques ascendants.

Chapitre 9 Travaux Futurs

Suite à nos travaux, il serait important d'effectuer une validation linguistique de l'annotation des actants de nouvelles lexies par les quatre rôles sémantiques considérés dans notre étude comme Agent, Patient, Destination et Autre. Nous sommes limités à ces rôles car les autres rôles ont un nombre de contextes annoté très faible (voir Tableau 11) pour pouvoir les classer que nous avons d'ailleurs jumelé dans la classe Autre. Même les rôles que nous avons considérés n'ont pas un très grand nombre d'instances. Nous proposons d'annoter manuellement plus de contextes afin d'avoir assez de données à entraîner et à tester.

Nous proposons également de mesurer le gain de temps réel pour les linguistes en se servant de l'annotation automatique et de mieux intégrer les tâches d'annotation manuelle et automatique. L'annotation manuelle n'est pas complètement exemptes d'erreurs et l'automatisation du processus permettrait de systématiser certaines analyses ou, du moins, de relever des problèmes d'analyses non systématiques. Proposer des modèles d'apprentissage du *feedback* de l'annotateur pour annoter de nouvelles lexies et de nouveaux rôles

Par la suite nous proposons de structurer toutes ces annotations et de construire une base de données de toutes les lexies verbales annotées automatiquement et les mettre disponibles pour les linguistes et les applications TAL afin de les enrichir et de construire une ressource lexicale du domaine de spécialité d'informatique et d'Internet en accès libre. Nous proposons d'utiliser cette ressource et d'essayer d'annoter les actants des lexies verbales directement dans des textes du même domaine.

À présent nous disposons d'un autre corpus spécialisé dans un autre domaine qui est celui du réchauffement climatique. Ce corpus est aussi annoté manuellement par l'équipe de l'OLST. Nous proposons d'utiliser notre système sur les données de ce nouveau corpus. L'adaptation de notre système à un autre domaine de spécialité permettrait de dégager les changements à apporter d'un domaine à un autre. Nous nous attendons à une meilleure adaptation de la tâche d'identification de participants actants et circonstants basée sur les informations syntaxiques. Quant à l'attribution des rôles sémantiques, le problème résidera

dans les rôles qui ne sont pas prévus dans le corpus spécialisé d'informatique et d'Internet. Notre système arrivera à sélectionner le groupe le plus similaire aux données à classifier et le linguiste pourrait valider le nouveau rôle sémantique et mettre à jour ce groupe en tenant compte des instances du nouveau rôle.

Jusqu'ici, nous n'avons utilisé que le corpus annoté manuellement comme ressource. Les informations syntaxiques (traits) sur lesquels nous nous sommes basés s'avèrent intéressantes et utiles pour identifier les participants et leur attribuer des rôles sémantiques. Néanmoins, ces traits sont insuffisants pour la tâche d'annotation de rôles sémantiques qui nécessite d'autres informations sémantiques complémentaires à ces informations syntaxiques. Nous pouvons trouver ces informations sémantiques dans des ressources lexicales sémantiques équivalentes aux ressources VerbNet et WordNet de l'anglais par exemple, riches et accessibles. Malheureusement en français, de telles ressources sont inaccessibles ou inexistantes ce qui ralentit d'autant les recherches dans le domaine des applications de TAL pour cette langue. C'est ce qui nous a obligé à mener notre étude sans avoir recours aux ressources sémantiques. Dans le futur, nous aimerions utiliser un dictionnaire sémantique des verbes du français, qui donnerait des informations syntaxico-sémantiques sur les verbes ainsi que le sens d'un verbe selon le domaine de spécialité dans lequel est employé. Un tel dictionnaire est depuis peu disponible en accès libre, il s'agit du dictionnaire « Les Verbes français » (LVF) de Jean Dubois et Françoise Dubois-Charlier [11]. Nous avons d'ailleurs effectué une comparaison de LVF avec d'autres ressources sémantiques, FrameNet, VerbNet et WordNet de l'anglais et nous avons constaté que LVF peut rivaliser ces ressources anglaises dans la description syntaxico-sémantique des unités prédicatives verbales [31] (une copie de cet article figure dans l'annexe 4). Les schèmes syntaxiques décrits dans LVF peuvent donner la valence du verbe, la restriction de sélection (animé, humain, chose, animal, etc.) des participants actants. Ainsi les rôles sémantiques pourraient être désambiguïsés dans le cas où plusieurs rôles sont sélectionnés. Par exemple, si pour un participant actant les rôles Agent et Destination sont sélectionnés, on pourrait choisir Agent si le schème syntaxique du verbe indique que le participant du verbe est plutôt humain. Le champ opérateur décrivant le sens

du verbe dans LVF utilise des fonctions sémantiques et opérateurs qui servent à désambiguïser les sens d'un verbe.

Avec la disponibilité du dictionnaire sémantique LVF, nous pourrions annoter des actants des lexies verbales de la langue générale. Nous proposons aussi d'utiliser cette annotation de rôles sémantiques des actants dans des applications de TAL.

Enfin, nous proposons d'annoter les circonstants des lexies verbales. L'approche par classification ou par partitionnement peut être utilisée. Mais nous doutons que les résultats soient aussi bons que pour les actants. Les circonstants sont plus ambigus que les actants. Les circonstants peuvent être sous forme d'une phrase, ce qu'on appelle des propositions. Des traits doivent être définis pour représenter ces circonstants propositionnels.

Et aussi d'annoter d'autres lexies prédicatives telles que les noms et les adjectifs et construire une ressource lexicale pour le français telle que la ressource FrameNet.

Chapitre 10 Conclusion

Ce travail se situe dans la lignée des travaux portant sur l'identification des participants (arguments et adjuncts), porteur de rôles sémantiques, en se basant sur des traits, (Gildea et Jurafsky) [24] pour la langue anglaise. Les travaux antérieurs d'annotation de rôles sémantiques utilisent le corpus FrameNet de la langue générale anglaise, PropBank, Treebank, et les ressources sémantiques VerbNet et WordNet de cette même langue. Ce type de ressources est plutôt rare pour la langue française et encore plus pour un domaine de spécialité.

Dans notre travail, nous avons utilisé un corpus spécialisé, annoté manuellement par des linguistes. Ce corpus, composé de phrases écrites en langue française, était lié aux domaines de l'informatique et de l'Internet. Nous avons adapté l'approche basée sur des traits décrivant les participants des lexies verbales de notre corpus spécialisé. Notre étude se distingue de ces travaux sur ces aspects :

- la langue traitée est le français dont les ressources sémantiques en accès libre sont rares vs anglais;
- le domaine de spécialité vs langue générale;
- les actants vs les circonstants;
- La classification des instances vs classification des rôles sémantiques;

Nous avons proposé deux méthodes pour identifier les participants actants et circonstants :

Dans la première approche, nous avons proposé d'extraire des règles à l'aide d'un analyseur syntaxique. Nous avons testé l'analyseur Syntex du français sur nos données. En faisant un appariement avec les données du corpus et les liens de dépendance entre les différentes unités des phrases retournés par Syntex, nous avons extrait une trentaine de règles basées sur ces dépendances et composées d'informations syntaxiques telles que groupe syntaxique et fonction syntaxique. Nous avons constaté que Syntex identifie convenablement les unités de la phrase, de nature sujet ou objet, liées à la lexie verbale. Ces fonctions syntaxiques nous permettent aussi d'identifier le type actant de ces unités. Par contre les unités liées à la lexie verbale à travers les prépositions présentent une ambiguïté quant à leur nature actant ou circonstant. L'étiquette attribuée à ces unités prépositionnelles

est « NOMPREP », qui signifie nom de la préposition. Ceci n'indique aucune information plus approfondie, même pas la fonction syntaxique complément ou autre. Dans ce cas, pour distinguer entre une unité prépositionnelle de type actant et celle de type circonstant, nous avons recours au calcul de la fréquence relative entre cette unité et la lexie verbale avec laquelle est employée dans notre corpus spécialisé. Cette fréquence est basée sur le fait qu'un actant est obligatoire et un circonstant est optionnel. Ce dernier est un critère linguistique, difficilement automatisable (car pas toujours respecté dans les phrases réelles), ce sont les cas de lexies verbales qui apparaissent dans des contextes sans leurs actants. Selon les données de notre corpus, nous avons tenté de discriminer ces cas de participants prépositionnels en calculant la fréquence relative entre les lexies verbales et leurs participants prépositionnels.

Cette approche nous a permis d'identifier les participants avec un haut taux de F-mesure et de valider l'importance des traits syntaxiques utilisés ou l'approche basée sur les traits. Mais elle est fastidieuse; nous ne pouvons pas prévoir tous les cas possibles.

Notre deuxième approche est plutôt basée sur l'apprentissage machine en utilisant les traits syntaxiques identifiés par nos règles comme des traits de classification.

Comme Syntex n'identifie pas les participants liés à une lexie par des liens indirects, les liens de dépendance entre ces unités et la lexie sont inexistants. Généralement ce sont les cas d'unités lexicales liées à d'autres verbes dans la phrase que la lexie à l'étude. Ces verbes sont généralement des verbes modaux et ils sont liés par des liens de dépendance de Syntex à la lexie à l'étude. Mais la difficulté est que les participants de ces verbes modaux liés à la lexie ne sont pas tous nécessairement des participants de cette lexie. Ceci dépend du verbe modal utilisé. Nous avons pour cela utilisé ces verbes modaux comme traits de classification. Une liste de ces verbes a été établie par les linguistes de l'OLST qui ont annoté manuellement les différentes lexies du corpus.

À l'aide des classificateurs de Weka, nous avons testé deux approches : la première est basée sur les dépendances de Syntex, sachant que Syntex détecte presque parfaitement les participants sujet et objet et une deuxième basée uniquement sur les catégories

syntaxiques de Treetagger, ce qui nous a permis de détecter les participants que Syntex est incapable de détecter, par exemple, les participants de type propositionnel.

Les classificateurs testés sur nos données ont réalisé de bons résultats de F-mesure comparés à ceux de la littérature (Sachant que les données sont différentes de celle de la littérature, nous sous entendons par comparer si nos résultats appartiennent à l'intervalle des résultats obtenus par d'autres travaux réalisés sur d'autres données et d'autres langues). Parmi ces classificateurs, le classificateur RandomForest a donné les meilleurs résultats.

Notre contribution dans cette partie a été d'arriver à détecter les participants issus de liens indirects et ceux propositionnels que l'analyse syntaxique n'avait pu faire ressortir et ceci en se basant sur la notion de traits et de classification binaire.

Une fois les participants identifiés, nous avons tenté de distinguer les participants actants des circonstants afin de leur attribuer des rôles sémantiques. Nous avons testé les classificateurs de Weka en utilisant en plus la fréquence relative comme trait et aussi les verbes modaux car ces derniers permettent d'affecter-eux même un type aux participants.

La dernière tâche consistait à attribuer des rôles sémantiques aux actants identifiés. Nous avons proposé de classifier les instances au lieu de classifier les rôles sémantiques. Nous présentons ainsi aux linguistes d'éventuels rôles dans les cas ambigus. Nous avons proposé de regrouper les instances selon leur similarité. Cette dernière est calculée sur la base du nombre de valeurs communes aux traits syntaxiques considérés et l'union de ces valeurs. Les traits syntaxiques utilisés sont catégoriels ce qui justifie le choix de la similarité utilisée. Nous avons testé l'approche naïve en considérant toutes les instances du corpus comme groupes et testé les nouvelles instances sur ces derniers. Par la suite nous avons proposé testé une approche de partitionnement. Parmi les approches de partitionnement existantes, nous avons choisi le partitionnement hiérarchique ascendant car il manipule mieux les traits de nature catégoriels que les autres approches de partitionnement manipulant les données plutôt numériques.

Les algorithmes hiérarchiques sont utilisés dans plusieurs domaines, particulièrement en datamining et bioinformatique. Parmi ces algorithmes, CHAMÉLÉON est basé sur le calcul de deux mesures (inter-connectivité relative et proximité relative) entre les groupes.

CHAMÉLÉON, a été utilisé en bio-informatique pour détecter la fonction d'un gène dans une séquence, une problématique semblable à notre problème de détection d'un rôle sémantique d'un actant d'une lexie verbale dans une phrase. Nous avons testé la bonne formation des groupes trouvés par cet algorithme en testant de nouvelles instances de notre corpus de test.

Notre contribution dans cette partie, est de venir en aide à l'annotateur afin d'annoter des milliers de lexies non déjà traitées. Nous avons proposé de faire un partitionnement semi supervisé afin de donner la possibilité d'annoter des actants avec de nouveaux rôles qui apparaissent au fur et à mesure. Le problème majeur de l'annotation en rôle sémantique est la définition d'une liste exhaustive de rôles. Notre proposition de regrouper les instances au lieu de classifier directement les rôles sémantiques permet de retrouver le groupe adéquat pour une nouvelle instance et de lui associer un rôle déjà existant dans ce groupe ou de proposer un autre rôle par l'annotateur, si nécessaire, et de l'intégré dans ce groupe.

Dans le futur nous proposons de prendre en compte le feedback de l'annotateur. Comme en recherche d'information, une fois les documents sélectionnés, nous essayons d'enrichir la requête par le feedback de l'utilisateur pour sélectionner les documents qui répondent adéquatement à la requête. En résumé, nous pouvons dire que cette perspective de l'approche de partitionnement, en regroupant les instances similaires permet de proposer à l'annotateur un choix de rôle dans les cas ambigus et lui permet aussi de proposer un nouveau rôle et de l'intégrer dans le groupe approprié pour qui on recalcule par la suite ses nouveaux représentants. Cette approche de partitionnement semble mieux adaptée pour des instances d'un autre domaine de spécialité.

Bibliographie

1. Bae H.S., L'Homme M.-C., et Lapalme G., Semantic roles in multilingual terminological descriptions: Application to french and korean contexts. In Proceedings of Workshop on Multilingual and Comparative Perspectives in Specialized Language Ressources (LREC'08), 2008, Marrakech, Marrocco.
2. Baker C.F., fillmore C.J., et Lowe J.B., The berkeley framenet project. In Proceedings of the 17th international conference on Computational linguistics (COLING-ACL'98), 1998. 1: p. 86-90.
3. Baker C., Ellsworth M., et Erk K., SemEval-2007 task 19 : Frame Semantic Structure Extraction. Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), 2007, Prague, Czech Republic, p.99-104
4. Barque L., Opérations sémantiques sur la base de données de définitions sens-texte, Mémoire de DEA, UFR de Linguistique, Université Paris 7, 2003, Paris, France. 67 p.
5. Boas H.C., Bilingual framenet dictionaries for machine translation. In Proceedings of the Third International Conference on Language Ressources and Evaluation (LREC), 2002, Las Palmas de Gran Canaria, Spain.
6. Bourigault D., Un analyseur syntaxique opérationnel: Syntex, Mémoire d'Habilitation à Diriger les Recherches, Laboratoire CLLE-ERSS, CNRS & Université de Toulouse-le Mirail, 2007. 158 p.
7. Bourigault D. et Fabre C., Approche linguistique pour l'analyse syntaxique de corpus. Cahiers de Grammaires n° 25, 2000, Université de Toulouse - Le Mirail, p. 131-151.
8. Che W., et al., Multilingual dependency-based syntactic and semantic parsing. Proceedings of CoNLL, 2009, Boulder, Colorado, p. 49-54.
9. Chen B. et Fung P., Automatic construction of an English-Chinese bilingual framenet. In Proceedings of HLT-NAACL, 2004, Boston, Massachusetts, USA, p. 29-32.

10. Collobert R. et Weston J., Fast semantic extraction using a novel neural network architecture. Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-07), 2007, Prague, Czech Republic, p. 560-567.
11. Dubois J. et Dubois-Charlier F., Les verbes français, ed. Larousse-Bordas. 1997, Paris.
12. Erk K., et al., Towards a ressource for lexical semantics: A large german corpus with extensive semantic annotation. Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL-03), 2003, Sapporo, Japan, p. 537-544.
13. Fabre C. et Frérot C., Groupes prépositionnels arguments ou circonstants : Vers un repérage automatique en corpus. TALN, 2002, Nancy, France.
14. Fillmore C.J., The case for case. In Emmon W.Bach and Robert T. Harms, (Eds), Universals in Linguistic Theory. New York, 1968: p. 1-88.
15. Fillmore C.J., An alternative to checklist theories of meaning. Proceeding of the first Annual Meeting of the Berkeley Linguistics Society, 1975: p. 123-132.
16. Fillmore C.J., Frame semantics and the nature of language. In Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech 1976. 280 p. 20-32.
17. Fillmore C.J., Frame semantics, in Linguistic society of korea, ed. Linguistics in the Morning Calm. Seoul:Hanshin. 1982. p. 111-137.
18. Fillmore C.J. et Atkins B.T., Towards a frame-based organization of the lexicon: The semantics of risk and neighbours. New essays in semantics and lexical organization, ed. Lehrer A et E. Kittay. 1992.
19. Fillmore C.J., Johnson C.R., et Petruck M.R.L., Background to framenet. International Journal of Lexicography, 2003. 16(3): p. 235-250.
20. Fillmore C.J., Ruppenhofer J., et Baker C.F., Framenet and representing the link between semantic and syntactic relations., in Computational linguistics and beyond, Chu Ren Huang et Winifred Lenders, Editors, ed. Language and Linguistics Monographs Series B. 2004: Institute of Linguistics, Academia Sinica. p. 19-62.

21. Fleischman M. et Hovy E., A maximum entropy approch to framenet tagging usc information science institute. In Proceedings of HLT-NAACL : short papers, 2003, Edmonton, Canada, p. 22-24.
22. Fliedner G., Tools for building a lexical semantic annotation. In Proceedings of Lorraine-Saarland Workshop on Prospects and Advances in the Syntax/Semantic Interface, 2003, Loria, Nancy, France, p. 5-9.
23. Gildea D. et Hockenmaier J., Identifying semantic roles using combinatory categorial grammar. Proceedings of the conference on Empirical methods in natural language processing (EMNLP'03), 2003, Sapporo, Japan, p. 57-64.
24. Gildea D. et Jurafsky D., Automatic labeling of semantic roles. Computational Linguistics, 2002. 28(3): p. 245-288.
25. Gildea D. et Palmer M., The necessity of parsing for predicate argument recognition. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL'02) 2002, Philadelphia, Pennsylvania, p. 239-246.
26. Guha S., Rastogi R., et Shim K., Rock: A robust clustering algorithm for categorical attributes. Proceedings of 15th International Conference on Data Engineering, 1999, Sydney, Australia, p. 512-521.
27. Guha S., Rastogi R., et Shim K., Cure: An efficient clustering algorithm for large databases. Information Systems, 2001. 26(1): p. 35-58.
28. Habert B., Cours de la linguistique à la sémantique http://archives.Limsi.Fr/individu/habert/cours/deasciencescognitivesp11/deascience scognitivesp11-01-02/index.Html. 2002.
29. Hadouche F., et al., Intégration d'informations syntaxico-sémantiques dans les bases de données terminologiques: Méthodologie d'annotation et perspectives d'automatisation. First International Workshop on Terminology and Lexical Semantics (TLS'09), 2009, Montréal, Québec, Canada, p. 22-31.
30. Hadouche F., L'Homme M.C., et Lapalme G., Automatic annotation of actants in specialized corpora. eLexicography in the 21st Century: New Challenges, New Applications (eLex'09). Les Cahiers du Cental, 2009, Louvain-La-Neuve, Bruxelles, p. 393-398.

31. Hadouche F. et Lapalme G., Une version électronique du LVF comparée avec d'autres ressources lexicales. Langages n° 179 -180, 2010, Armand Colin, p.93-220.
32. Hadouche F., Lapalme G., et L'Homme M.-C. Identification des participants actants et circonstants par apprentissage machine. in TALN. 2010. Montréal, Québec, Canada.
33. Johanson R. et Nugues P., Sparse bayesian classification of predicate arguments. Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL'05), 2005, Ann Arbor, Michigan, p. 117-180.
34. Johansson R. et Nugues P., Lth: Semantic structure extraction using nonprojective dependency trees. Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval'07), 2007, Prague, Czech Republic, p. 227-230.
35. Kahane S., Polguère A., et Steinlin. J., Base dico. Http://olst.Ling.Umontreal.Ca/dicouebe/. OLST 2008.
36. Karypis G., Eui-Hong H., et Kumar V., Chameleon: Hierarchical clustering using dynamic modeling. Computer, 1999. 32(8): p. 68-75.
37. L'Homme M.-C., Projet sémantique lexicale et terminologie. 2007, Université de Montréal: Montréal, Québec, Canada (http://olst.ling.umontreal.ca/?page_id=361).
38. L'Homme M.-C., Serrec A.L., et Laneville M.-È., Le guide d'annotation. 2010: Université de Montréal. 69 p.
39. Lacheret-Dujour A. et François J., Circonstance et prédication verbale en français parlé : Contraintes sémantico-pragmatiques et filtrage prosodique. Syntaxe & sémantique n° 6, 2005, Presses universitaire de Caen, p. 35-56.
40. Li J., Zhang L., et Yu Y. 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). 2001. Victoria, Canada: ACM New York, USA.
41. Li L., et al., Discriminative learning of syntactic and semantic dependencies. Proceedings of the Twelfth Conference on Computational Natural Language Learning (CoNLL'08), 2008, Manchester, United Kingdom, p. 218-222.
42. Liakata M. et Pulman S., Using predicate-argument structures for information extraction. Proceedings of ACL, 2003, p. 8-15.

43. Màrquez L., Carreras x., et Srevenson S., Special issue on semantic role labeling. Computational Linguistics, 2008. 34(2): p. 145-159.
44. Mel'čuk I., Actants in semantic and syntax. Actants in Semantics. Linguistics, 2004. 42(2): p. 1-66.
45. Mel'čuk I.A., Dictionnaire explicatif et combinatoire du français contemporain, recherche lexico-sémantiques i. Presses de l'université de montréal. 1984, Montréal, Québec, Canada. 176 p.
46. Mel'čuk I.A., Polguère A., et Clas A., Introduction à la lexicologie explicative et combinatoire. AUPELF-UREF et Duculot ed. 1995, Louvain la Neuve. 256 p.
47. Melli G., et al. Description of squash, the sfu question answering summary handler for the duc-2005 summarization task. in Proceedings of the HLT/EMNLP Document Understanding Workshop (DUC). 2005. Vancouver, Canada.
48. Messiant C., Korhonen A., et Poibeau T. Lexschem: A large subcategorization lexicon for french verbs. in LREC Proceedings. 2008. Marrakech, Maroc.
49. Moschitti A. et Bejan C.A., A semantic kernel for predicate argument classification. Proceeding of CoNLL, 2004, Boston, MA, USA, p. 17-24.
50. Moschitti A., Pighin D., et Basili R., Tree kernels for semantic role labeling. Computational Linguistics, 2008. 34(2): p. 193-224.
51. Narayanan S. et Harabagiu S. Question answering based on semantic structures. In Procedings of the 20th International Conference on Computational Linguistics (COLING). 2004. Genève, Suisse.
52. Nasr A. et Béchet F., Analyse syntaxique en dépendances de l'oral spontané. Actes de la conférence Traitement Automatique des Langues Naturelles (TALN), 2009, Senlis, France.
53. Padò S. et Lapata M., Cross-lingual projection of role-semantic information. Actes de HLT/EMNLP, 2005, Vancouver, Canada.
54. Padò S. et Pitel G., Annotation précise du français en sémantique de rôles par projection cross-linguistique. Actes de la conférence Traitement Automatique des Langues Naturelles (TALN), 2007, Toulouse, France.
55. Palmer M., Gildea D., et Kingsbury P., The proposition bank: An annotated corpus of semantic roles. Computational Linguistics, 2005. 31: p. 71-105.

56. Palmer M., Gildea D., et Xue N., Semantic role labeling. Synthesis Lectures on Human Language Technologies, 2010: p. 103.
57. Pradhan S., et al., Shallow semantic parsing using support vector machines. Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL), 2004, Boston, USA, p. 233-240.
58. Pradhan S., et al., SemEval-2007 task 17: English lexical sample, SRL and all words. In Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), 2007, Prague, Czech Republic, p. 87-92
59. Punyakanok V., Roth D., et Yih W.-t., The importance of syntactic parsing and inference in semantic role labeling. Computational Linguistics, 2008. 34(2): p. 257-287.
60. Scaiano M., et Inkpen D., Automatic frame extraction from sentences. In Proceedings of the 22th Canadian Conference on Artificial Intelligence (AI 2009), 2009, Kelowna, BC, Canada, p. 110-120.
61. Suanmali L., Salim N., et Binwahlan M.S., A hybrid approach based on semantic role labeling and general statistic method for test summarization. Applied Sciences, 2010. 10(3): p. 166-173.
62. Surdeanu M., et al., Using predicate-argument structures for information extraction. Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL '03), 2003, Sapporo, Japan, p. 8-15.
63. Swier R.S. et Stevenson S., Unsupervised semantic role labeling. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2004, Barcelona, Spain, p. 95-102.
64. Täckström O., Multilingual semantic parsing with a pipeline of linear classifiers Proceedings of the 13th Conference on Computational Natural Language Learning (CoNLL): Shared Task, 2009, Boulder, Colorado, USA, p. 103-108.
65. Tesnière L., Éléments de syntaxe structurale, ed. Klincksieck. 1959, Paris. 674 p.
66. Vázquez G. et Montraveta A.F., Annotation de corpus : Sur la délimitation des arguments et des adjoints. SKY Journal of Linguistics 2008. 21: p. 243-269.

67. Witten I. H., et Frank E., Data mining: Practical machine learning tools and techniques, 2nd Edition. Morgan Kaufmann (ed.), 2005

68. Xue N., Labeling chinese predicates with semantic roles. Computational Linguistics, 2008. 34(2): p. 225-255.

69. Xue N. et Palmer M., Calibrating features for semantic role labeling. Proceedings of Empirical Methods in Natural Language (EMNLP) 2004, Barcelona, Spain, 2004, p. 88-94.

Annexe 1

```
start = element vocables{element-vocable*}[26]
element-vocable = element vocable {
    attribute identificateur {text},
    element-lexie*
}
element-lexie = element lexie {
    attribute numero-acception {xsd:positiveInteger},
    element-contexte*
}
element-contexte = element contexte {
    attribute source {text},
    attribute statut {xsd:nonNegativeInteger},
    attribute annotateur {list{TypeAnnotateur*}},
    attribute mise-a-jour {xsd:date}?,
    element-contexte-texte,
    mixed {(element-lexie-att | element-participant | element-
antecedent)*}
}
element-contexte-texte = element contexte-texte {text}

element-lexie-att = element lexie-att {
   attribute auxiliaire{Auxiliaire}?,
    mixed {element-reference?,element-lemme?},
    text
}
element-reference=element reference{
attribute xml:id {xsd:ID}?,text
}
element-lemme=element lemme {
   attribute lem {text},text
}
element-participant = element participant {
   attribute type {TypeParticipant},
   attribute role {RoleParticipant},
   element-fonction-syntaxique
}
element-antecedent = element antecedent {
   attribute xml:id {xsd:ID},
   mixed {element-valeur-antecedent*},
   text
}
element-valeur-antecedent= element valeur-antecedent{
attribute lemme{text}?,text
}
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}?,
```

[26] Dans ce schéma le symbole « * » suivant un élément signifie une répétition d'un ou de plusieurs éléments et le symbole « ? » signifie zéro ou une occurrence d'un élément.

```
   mixed {element-realisation}
}
element-realisation = element realisation{
   attribute lemme {text}?,
   attribute etiquette {text}?,
   attribute ref {xsd:IDREF}?,
   attribute reflex {xsd:IDREF}?,
   text
}
```

Figure 41 Schéma RNC validant la forme XML du corpus des annotations

Annexe 2

La sortie en format xml de l'analyse syntaxique de Syntex de la phrase :

« Certains Bios acceptent même directement les souris » de la Figure 19.

```
<SEQ id="T_1944">
  <TXT>Certains BIOS acceptent même directement les souris.</TXT>
  <tokens>
   <t i="1" l="certain" f="Certains" c="Det" p="D"/>
   <t i="2" l="Bios" f="BIOS" c="NomPrXXInc" p="NP"/>
   <t i="3" l="accepter" f="acceptent" c="VCONJP" p="V"/>
   <t i="4" l="même" f="même" c="Adv" p="R"/>
   <t i="5" l="directement" f="directement" c="Adv" p="R"/>
   <t i="6" l="le" f="les" c="Det??" p="D"/>
   <t i="7" l="souris" f="souris" c="Nom?P" p="N"/>
   <t i="8" l="." f="." c="Typo" p="T"/>
  </tokens>
  <dependances>
   <d r="DET" s="1" c="2"/>
   <g r="DET" s="2" c="1"/>
   <d r="SUJ" s="2" c="3"/>
   <g r="SUJ" s="3" c="2"/>
   <g r="ADV" s="3" c="5"/>
   <g r="OBJ" s="3" c="7"/>
   <d r="ADV" s="4" c="5"/>
   <g r="ADV" s="5" c="4"/>
   <d r="ADV" s="5" c="3"/>
   <d r="DET" s="6" c="7"/>
   <g r="DET" s="7" c="6"/>
   <d r="OBJ" s="7" c="3"/>
  </dependances>
 </SEQ>
```

Annexe 3

(Figures correspondantes aux règles du Chapitre 1)

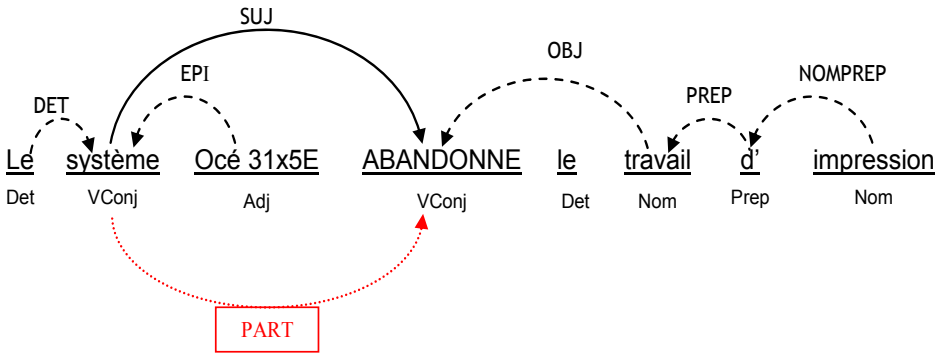


Figure 42 Dépendance Sujet (SUJ) (Règle 1)
Lexie<ABANDONNE, VConj, SUJ> + <système, Nom, SUJ>
→ <système, Nom, SUJ>

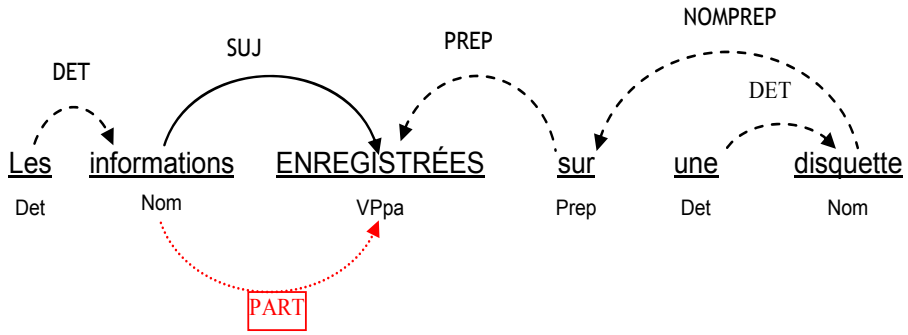


Figure 43 Dépendance Sujet (SUJ) avec la lexie au participe passé (Règle 2)
Lexie<ENREGISTRÉES, VPpa, SUJ> + <informations, Nom, SUJ>
→ <informations, Nom, OBJ>

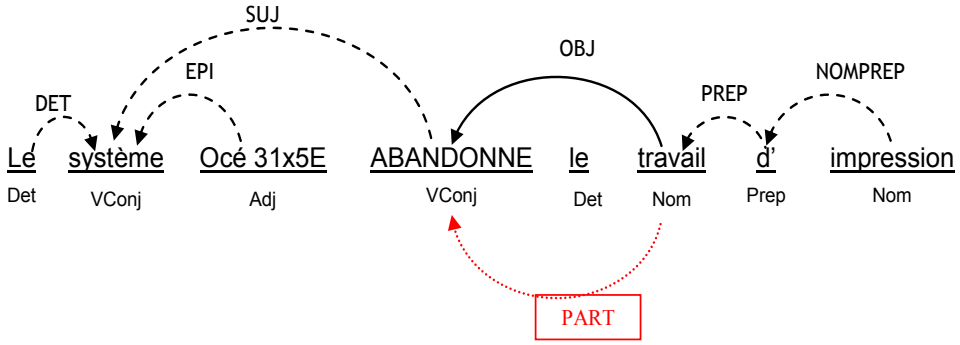


Figure 44 Dépendance Objet (OBJ) (Règle 3)
Lexie<ABANDONNE, VConj, OBJ> + <travail, Nom, OBJ>
→ <travail, Nom, OBJ>

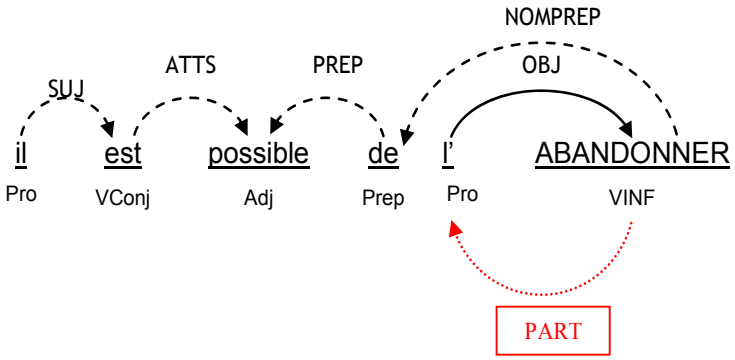


Figure 45 Dépendance Objet d'un pronom sujet (Règle 3)
<l', Pro, OBJ> + Lexie<ABANDONNER, VInf, OBJ>
→ <l', Pro, OBJ>

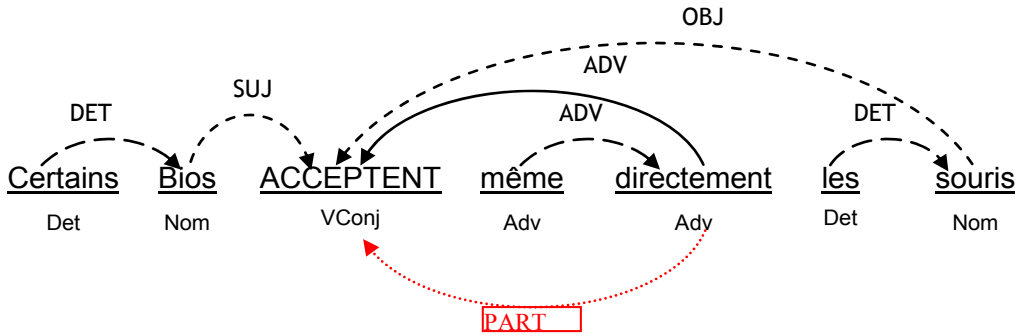


Figure 46 Dépendance Lexie| Adverbe (Règle 4)
Lexie<ACCEPTENT, VConj, ADV> + <directement, Adv, ADV>
→ <directement, Adv, ADV>

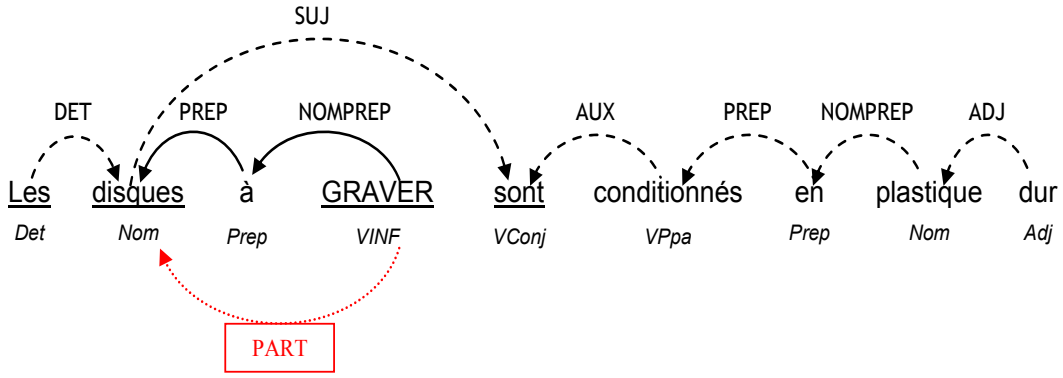


Figure 47 Dépendance Lexie| Préposition à avant la lexie (Règle 5)
Lexie<GRAVER, VInf, NOMPREP> + <à, Prep, PREP, NOMPREP> + <disques, Nom, PREP>
→ <disques, Nom, TÊTE>

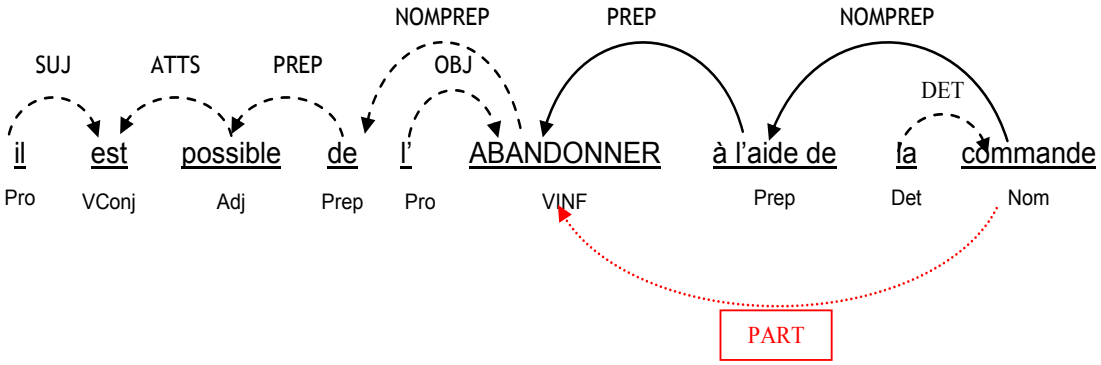


Figure 48 Dépendance Lexie | Préposition à l'aide de (Règle 6)
Lexie<ABANDONNER, VInf, PREP> + <à l'aide de, Prep, PREP, NOMPREP> + <commande, Nom, NOMPREP>
→ <commande, Nom, NOMPREP>

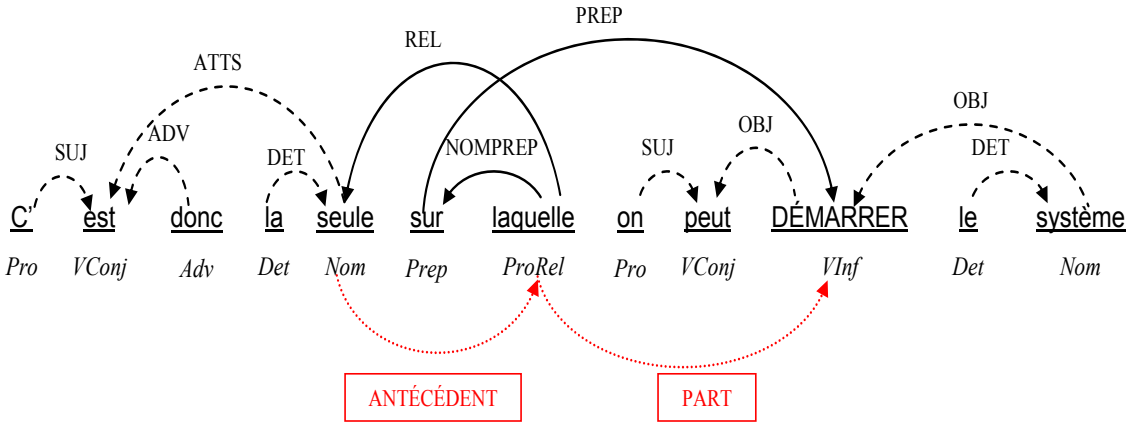


Figure 49 Dépendance Lexie | Préposition sur liée à un relatif (Règle 7)

Lexie<DÉMARRER, VInf, PREP> + <sur, Prep, PREP, NOMPREP> + <laquelle, ProRel, NOMPREP, REL> + <seule, Nom, REL>

→ <laquelle, ProRel, NOMPREP> avec antécédent <seule, Nom, REL>

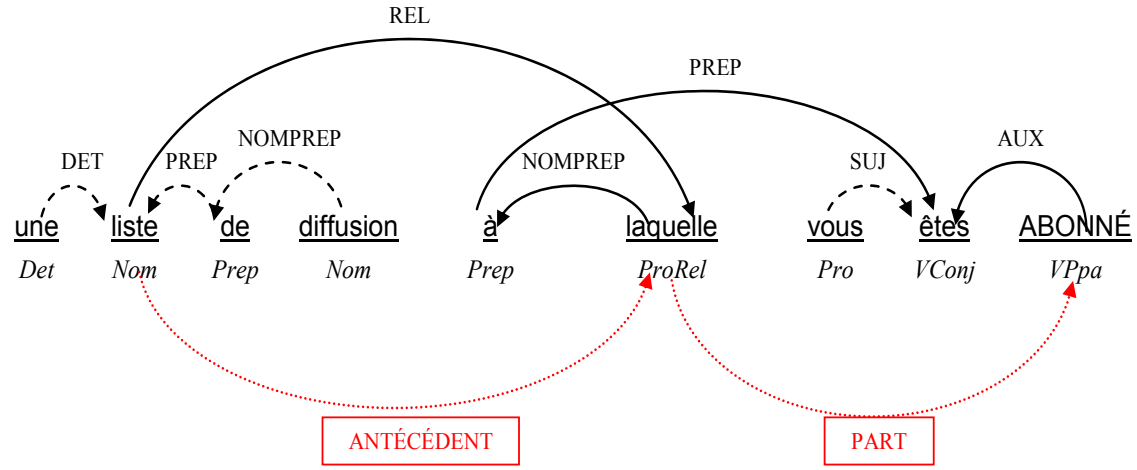


Figure 50 Dépendance Lexie | Préposition (Règle 8)

Lexie<ABONNÉ, VPpa, AUX> + <êtes, VConj, AUX, PREP> + <à, Prep, PREP, NOMPREP> + <laquelle, ProRel, NOMPREP, REL> + <liste, Nom, REL>

→ <laquelle, ProRel, NOMPREP> avec antécédent <liste, Nom, REL>

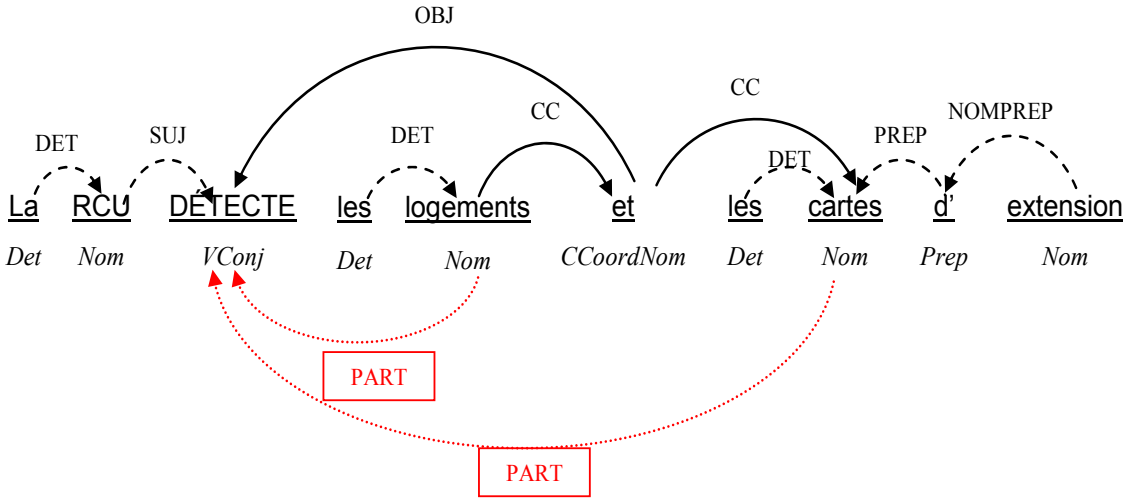


Figure 51 Dépendance Lexie | Conjonction entre deux syntagmes nominaux (Règle 9)
Lexie<DÉTECTE, VConj, OBJ> + <et, CCoordNom, CC, OBJ> + <logements, Nom, CC>
→<logements, Nom, OBJ>
Lexie<DÉTECTE, VConj, OBJ> + <et, CCoordNom, CC, OBJ> + <cartes, Nom, CC>
→<cartes, Nom, OBJ>

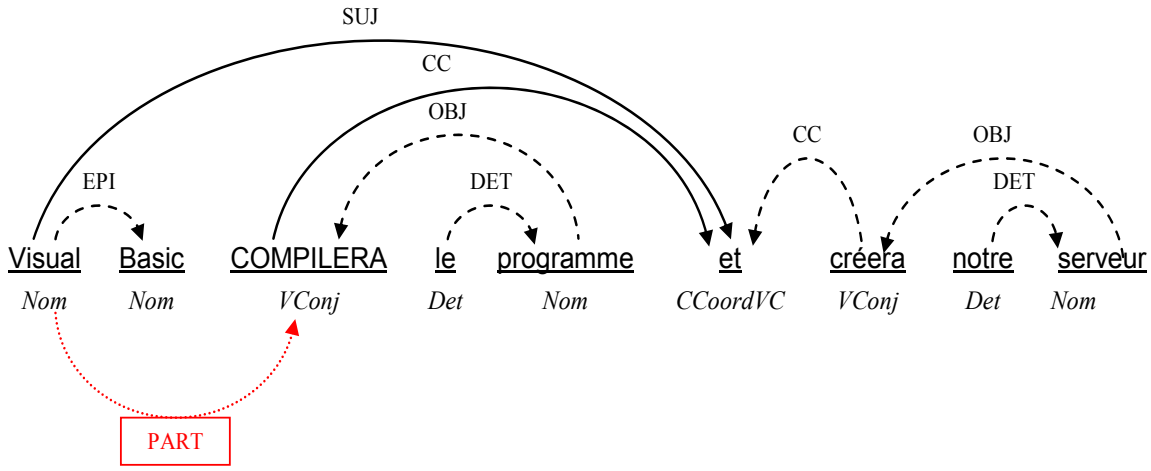


Figure 52 Dépendance Lexie | Conjonction entre deux syntagmes verbaux (Règle 10)
Lexie<COMPILERA, VConj, CC> + <et, CCoordV, CC, SUJ> + <Visual, Nom, SUJ>
→ <Visual, Nom, SUJ>

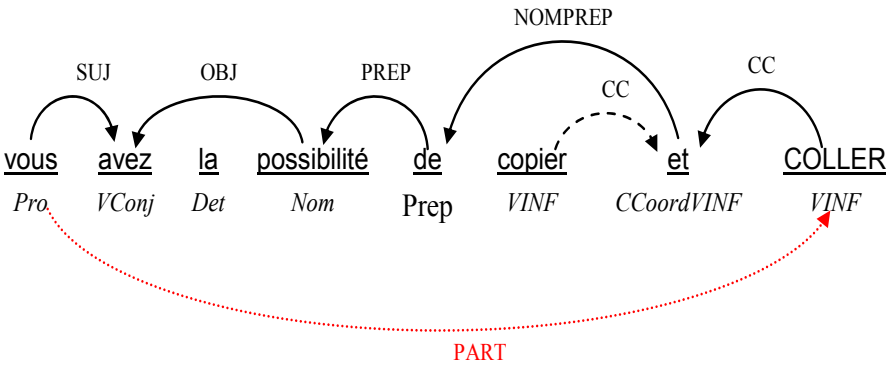


Figure 53 Dépendance Lexie | Conjonction (lien indirect avec le sujet) (Règle 11)

Lexie<COLLER, VPpa, CC> + <et, CCoordV, CC, NOMPREP> + <de, Prep, NOMPREP, PREP> +
<possibilité, Nom, PREP, OBJ> + <avez, VConj, OBJ, SUJ> + <vous, Nom, SUJ>
→ <vous, Nom, SUJ>

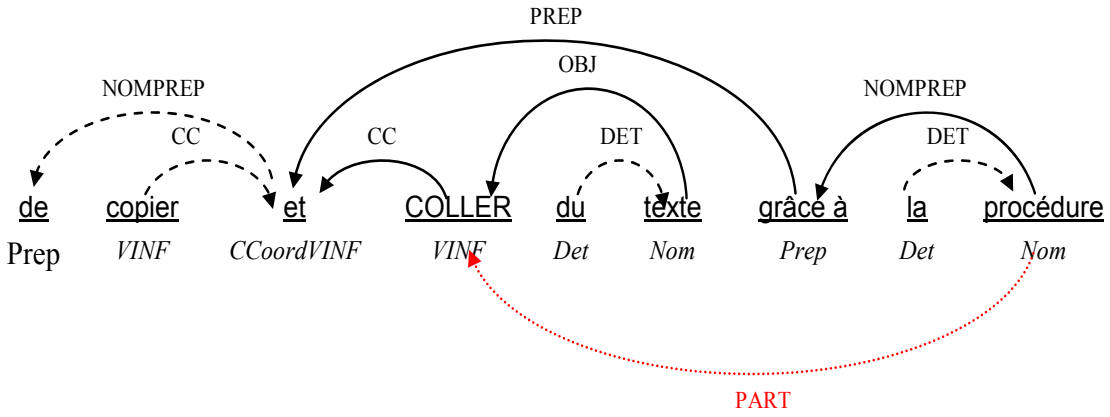


Figure 54 Dépendance Lexie | Conjonction liée à une préposition (Règle 12)

Lexie<COLLER, VPpa, CC> + <et, CCoordV, CC, PREP> + <grâce à, Prep, PREP, NOMPREP> +
<procédure, Nom, NOMPREP>
→ <procédure, Nom, NOMPREP>

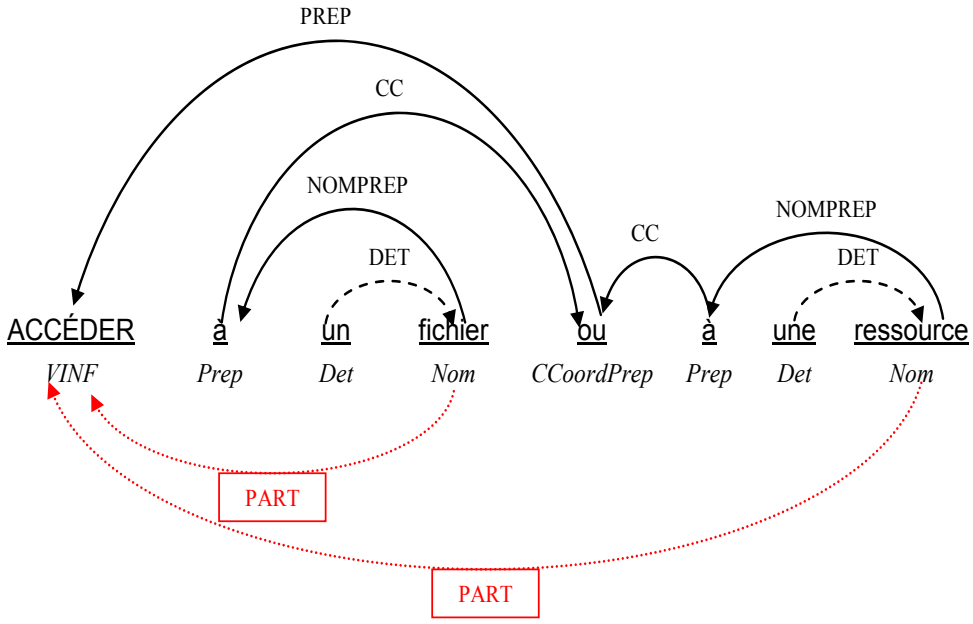


Figure 55 Dépendance Lexie | Conjonction avec le lien PREP (Règle 13)

Lexie<ACCÉDER, VInf, PREP> + <ou, CCoordPrep, PREP, CC> + <à, Prep, CC, NOMPREP > +
<fichier, Nom, NOMPREP>
→ <fichier, Nom, NOMPREP>
Lexie<ACCÉDER, VInf, PREP> + <ou, CCoordPrep, PREP, CC> + <à, Prep, CC, NOMPREP> +
<ressource, Nom, NOMPREP>
→ <ressource, Nom, NOMPREP>

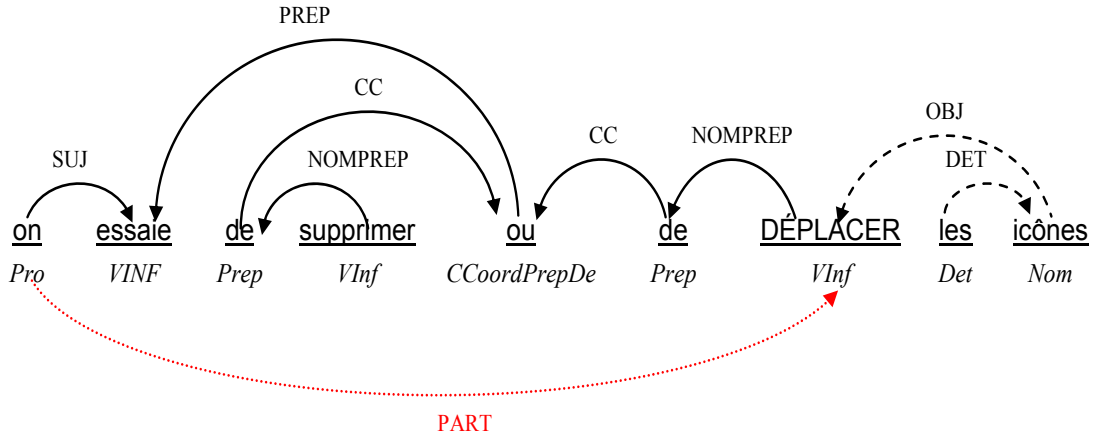


Figure 56 Dépendance Lexie | Conjonction (verbes introduits par une préposition) (Règle 14)

Lexie<DÉPLACER, VInf, NOMPREP> + <de, Prep, NOMPREP, CC> + <ou, CCoordPrep, CC, PREP> +
<essaie, VConj, PREP, SUJ> + <on, Nom, SUJ>
→ <on, Nom, SUJ>

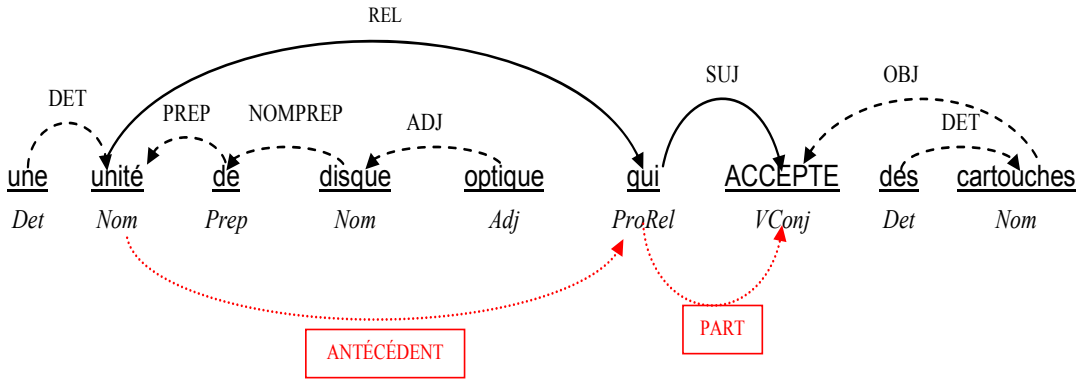


Figure 57 Dépendance Lexie| Pronom relatif sujet (Règle 15)
Lexie<ACCEPTE, VConj, SUJ> + <qui, ProRel, REL, SUJ> + <unité, Nom, REL>
→ <qui, ProRel, SUJ> avec antécédent <unité, Nom, REL>

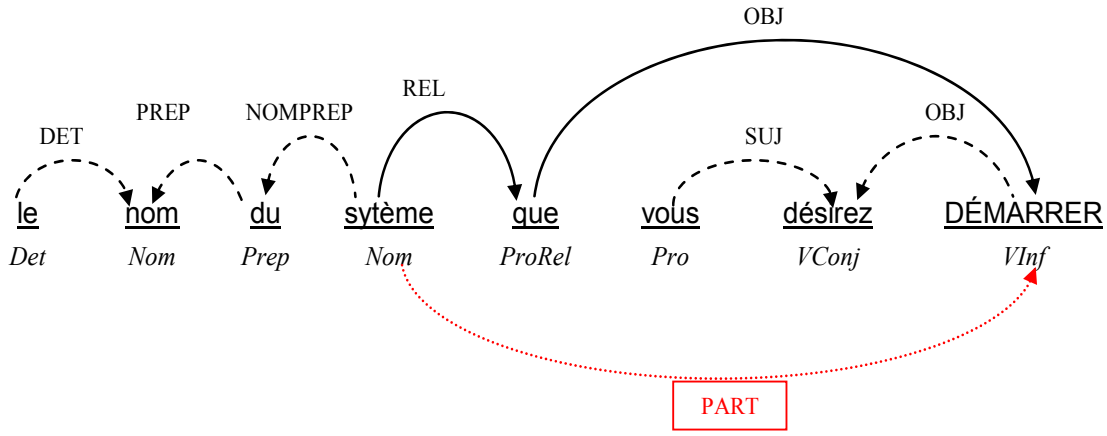


Figure 58 Dépendance Lexie| Pronom relatif objet (Règle 15)
Lexie<DÉMARRER, VConj, OBJ> + <que, ProRel, REL, OBJ> + <sytème, Nom, REL>
→ <que, ProRel, OBJ> avec antécédent <système, Nom, REL>

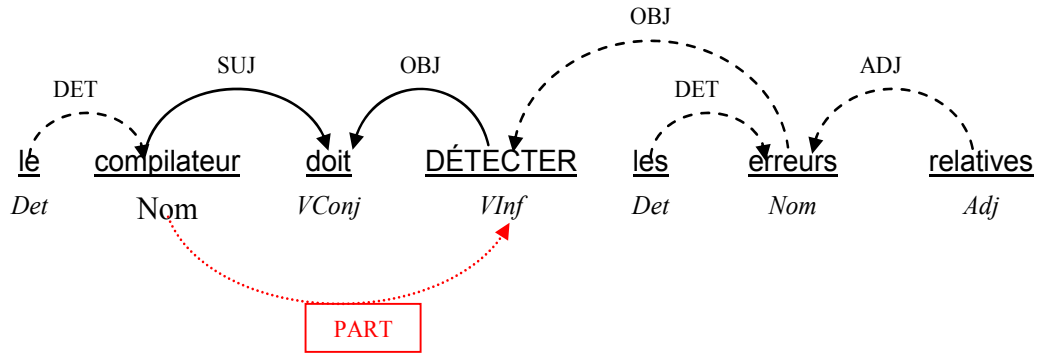


Figure 59 Dépendance Lexie| devoir (Règle 16)
Lexie<DÉTECTER, VInf, OBJ> + <doit, VConj, OBJ, SUJ> + <compilateur, Nom, SUJ>
→ <compilateur, Nom, SUJ>

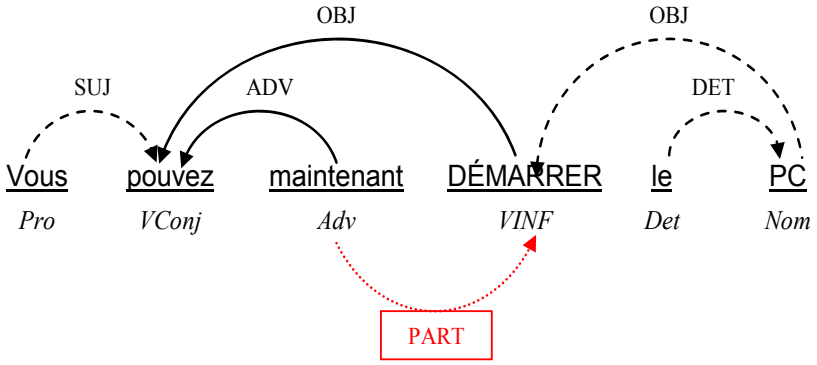


Figure 60 Dépendance Lexie| pouvoir (Règle 16)
Lexie<DÉMARRER, VInf, OBJ> + <pouvez, VConj, OBJ, ADV> + <maintenant, Nom, ADV>
→ <maintenant, Nom, ADV>

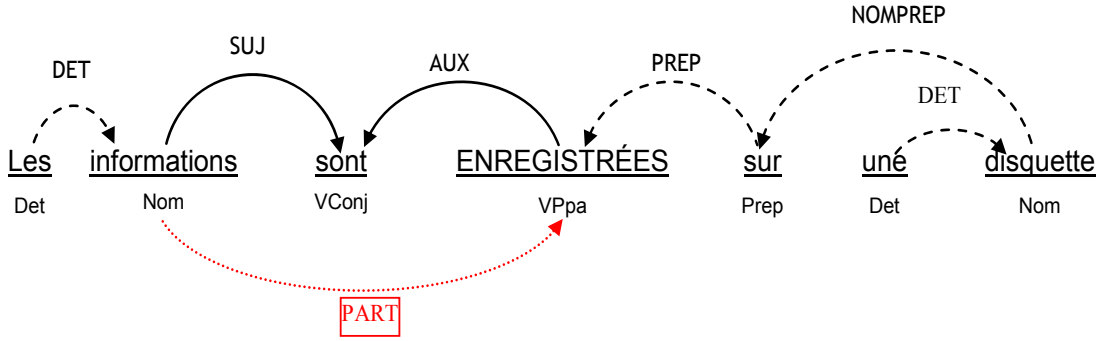


Figure 61 Dépendance Lexie|Auxiliaire (Règle 17)
Lexie<ENREGISTRÉES, VPpa, AUX> + <sont, VConj, AUX, SUJ> + <informations, Nom, SUJ>
→ <informations, Nom|Pro, OBJ>

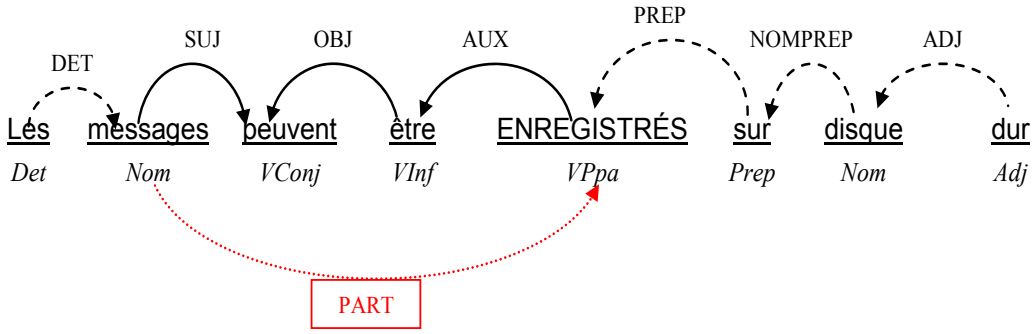


Figure 62 Dépendance Lexie| pouvoir+auxiliaire (Règle 18)
Lexie<ENREGISTRÉS, VPpa, AUX> + <être, VInf, AUX, OBJ> + <peuvent, VConj, OBJ, SUJ> +
<messages, Nom, SUJ>
→<messages, Nom, OBJ>

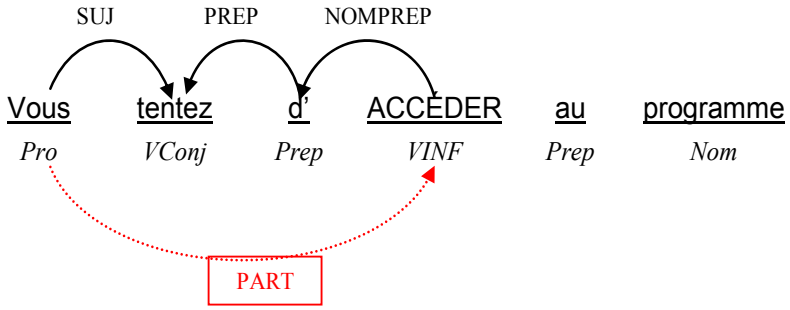


Figure 63 Dépendance Lexie| tenter de (Règle 19)

Lexie<ACCEDER, VInf, NOMPREP> +<de, Prep, NOMPREP, PREP> + <tentez, VConj, PREP, SUJ> + <vous, Pro, SUJ>
→ <vous, Pro, SUJ>

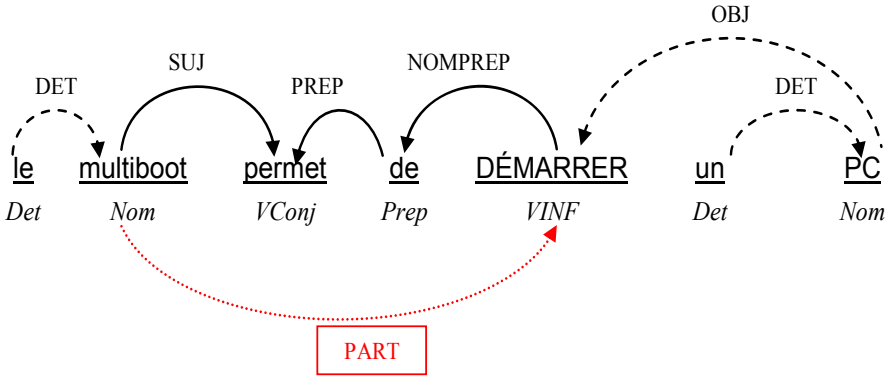


Figure 64 Dépendance Lexie| permettre de (Règle 19)

Lexie<DÉMARRER, VInf, NOMPREP> +<de, Prep, NOMPREP, PREP> + <permet, VConj, PREP, SUJ> + <multiboot, Nom, SUJ>
→ <multiboot, Nom, SUJ>

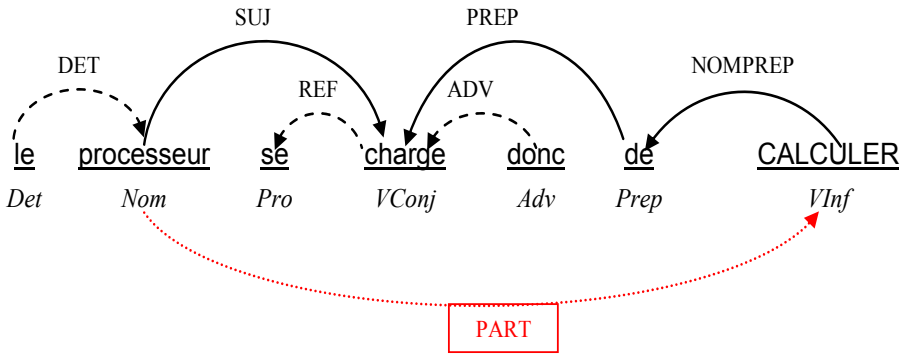


Figure 65 Dépendance Lexie| se charger de (Règle 19)

Lexie<CALCULER, VInf, NOMPREP> +<de, Prep, NOMPREP, PREP> + <charge, VConj, PREP, SUJ> + <processeur, Nom, SUJ>
→ <processeur, Nom, SUJ>

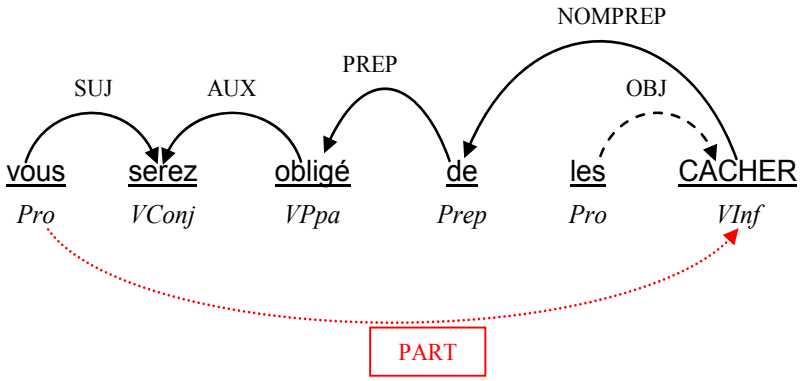


Figure 66 Dépendance Lexie | être obligé de (Règle 20)

Lexie<CACHER, VInf, NOMPREP> + <de, Prep, NOMPREP, PREP> + <obligé, VPpa, PREP, AUX> + <serez, VConj, AUX, SUJ> + <vous, Pro, SUJ>
→ <vous, Pro, SUJ, lien indirect>

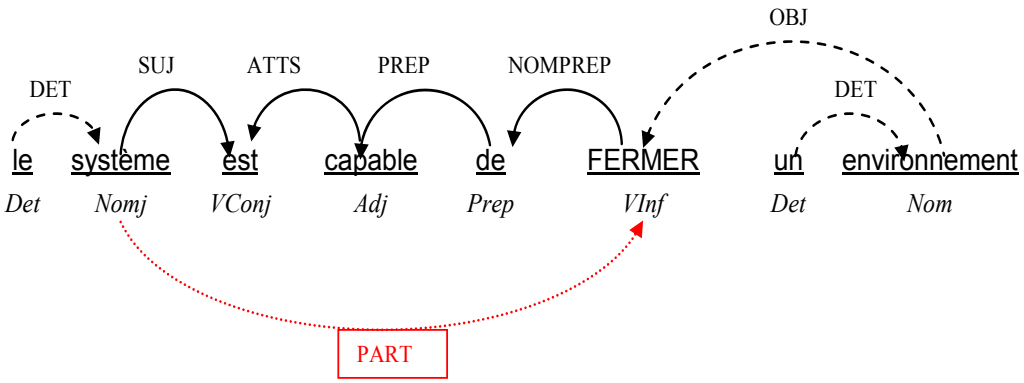


Figure 67 Dépendance Lexie| être capable de (Règle 21)

Lexie<FERMER, VInf, NOMPREP> + <de, Prep, NOMPREP, PREP> + <capable, Adj, PREP, ATTS> + <est, VConj, ATTS, SUJ> + <système, Nom, SUJ>
→ <système, Nom, SUJ, lien indirect>

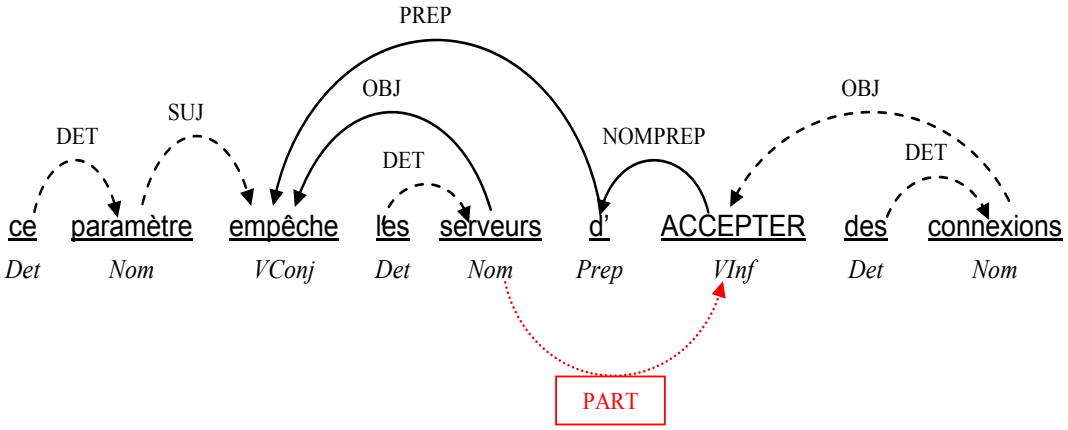


Figure 68 Dépendance Lexie| empêcher de (Règle 22)

Lexie<ACCEPTER, VInf, NOMPREP> + <de, Prep, NOMPREP, PREP> + <empêche, VConj, PREP, OBJ> + <serveurs, Nom, OBJ>
→ <serveurs, Nom, SUJ>

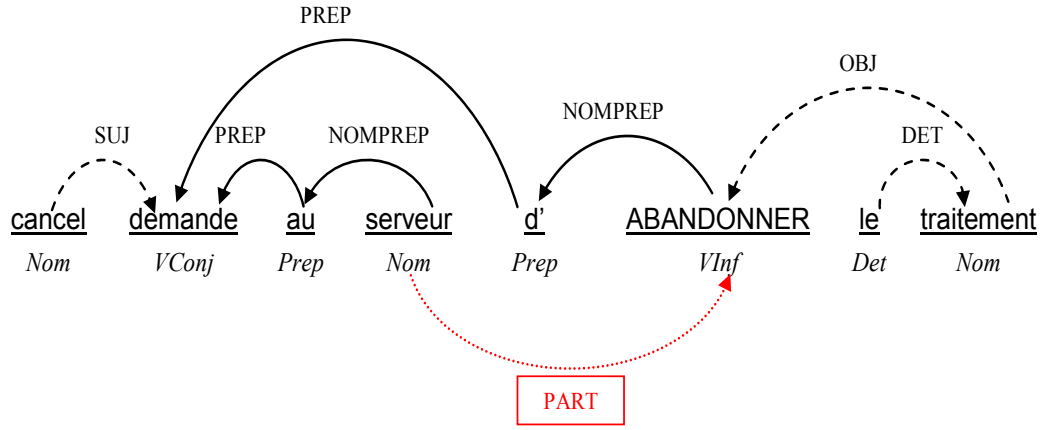


Figure 69 Dépendance Lexie| demander de (Règle 23)

Lexie<ABANDONNER, VInf, NOMPREP> + <de, Prep, NOMPREP, PREP> + <demande, VConj, PREP, PREP> + <à, Prep, PREP, NOMPREP> + <serveur, Nom, NOMPREP>
→ <serveur, Nom, lien indirect>

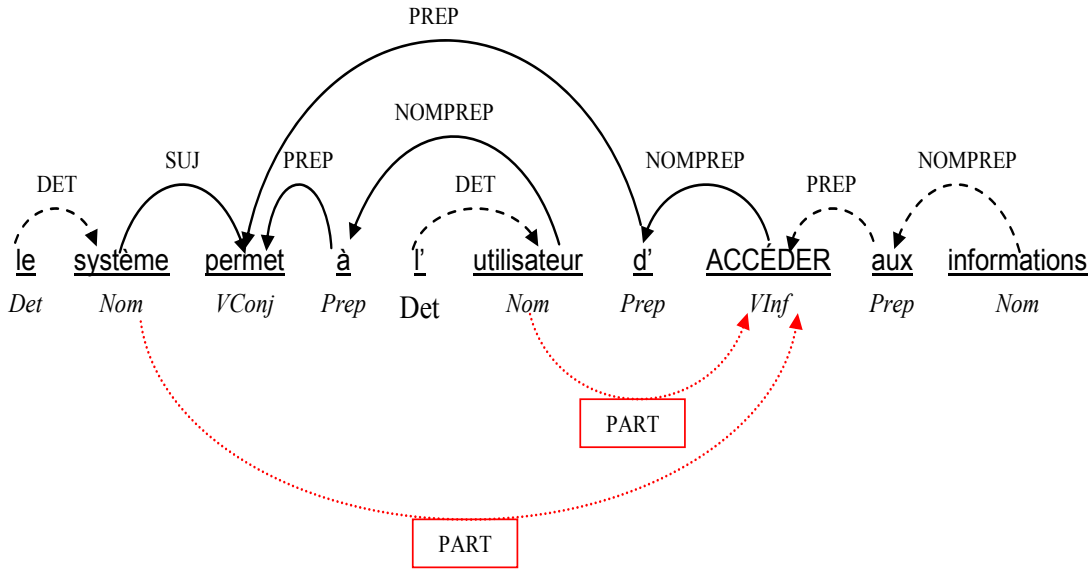


Figure 70 Dépendance Lexie| permetre à … de … (Règle 28)(Règle 24)

Lexie<ACCEDER, VInf, NOMPREP> + <de, Prep, NOMPREP, PREP> + <permet, VConj, PREP, SUJ> + <système, Nom, SUJ>
→<système, Nom, SUJ, lien indirect>

Lexie<ACCEPTER, VInf, NOMPREP> + <de, Prep, NOMPREP, PREP> + <permet, VConj, PREP, PREP> + <à, Prep, PREP, NOMPREP> + <utilisateur, Nom, NOMPREP>
→<utilisateur, Nom, SUJ, lien indirect>

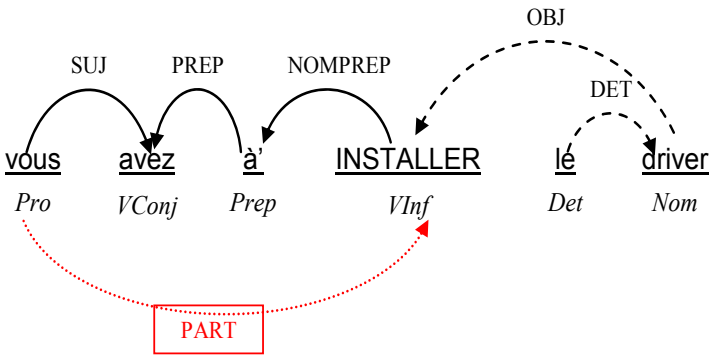


Figure 71 Dépendance Lexie| avoir à (Règle 25)

Lexie<INSTALLER, VInf, NOMPREP> + <à, Prep, NOMPREP, PREP> + <avez, VConj, PREP, SUJ> + <vous, Pro, SUJ>
→<vous, Pro, SUJ, lien indirect>

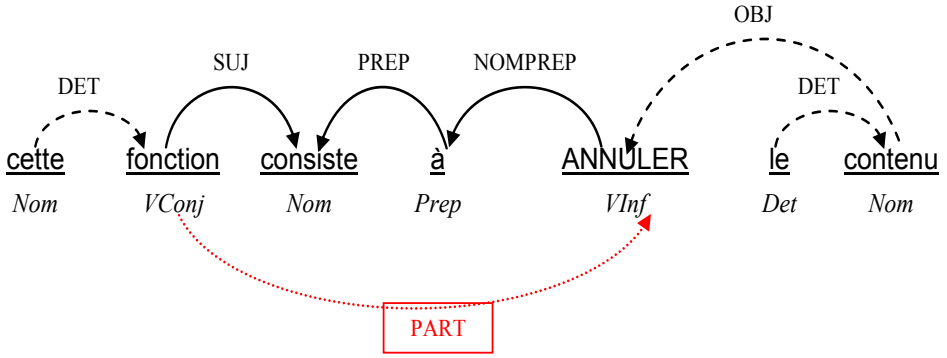


Figure 72 Dépendance Lexie| consiste à (Règle 25)

Lexie<ANNULER, VInf, NOMPREP> + <à, Prep, NOMPREP, PREP> + <consiste, VConj, PREP, SUJ> + <fonction, Nom, SUJ>
→<fonction, Nom, SUJ, lien indirect>

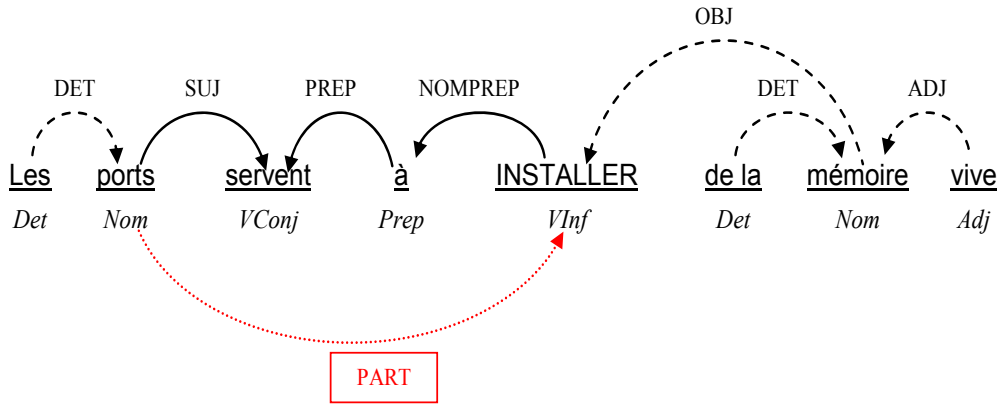


Figure 73 Dépendance Lexie| servir à (Règle 25)

Lexie<INSTALLER, VInf, NOMPREP> + <à, Prep, NOMPREP, PREP> + <servent, VConj, PREP, SUJ> + <ports, Nom, SUJ>
→<ports, Nom, SUJ, lien indirect>

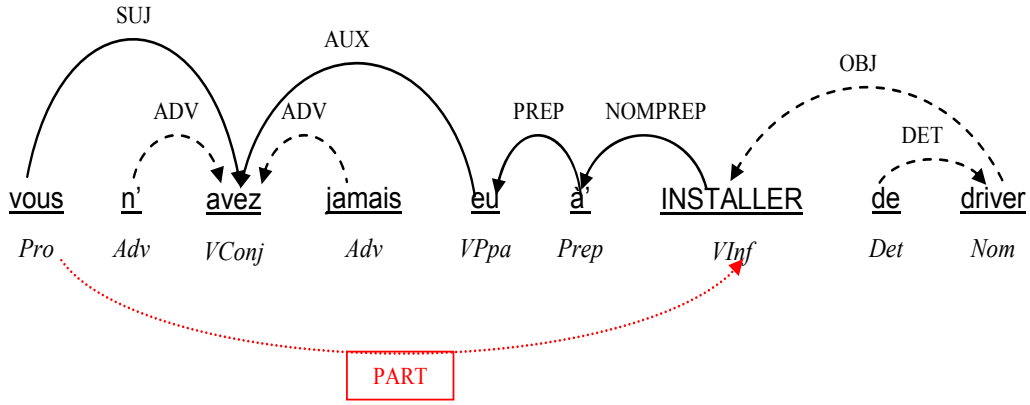


Figure 74 Dépendance Lexie| avoir+auxiliaire+à (Règle 26)

Lexie<INSTALLER, VInf, NOMPREP> + <à, Prep, NOMPREP, PREP> + <EU, VPpa, PREP, AUX> + <avez, VConj, AUX, SUJ> + <vous, Pro, SUJ>
→ <VOUS, Pro, SUJ, lien indirect>

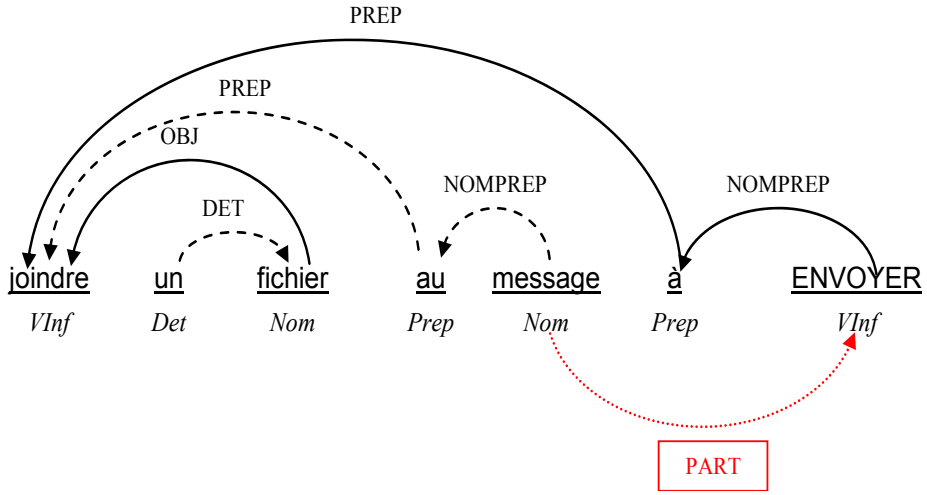


Figure 75 Dépendance Lexie|Verbe liés à la préposition à (Règle 27)

Lexie<ENVOYER, VInf, NOMPREP> + <à, Prep, NOMPREP, PREP> + <joindre, VConj, PREP, PREP> + <message, Nom, PREP>
→ <message, Nom|TÊTE, PREP>

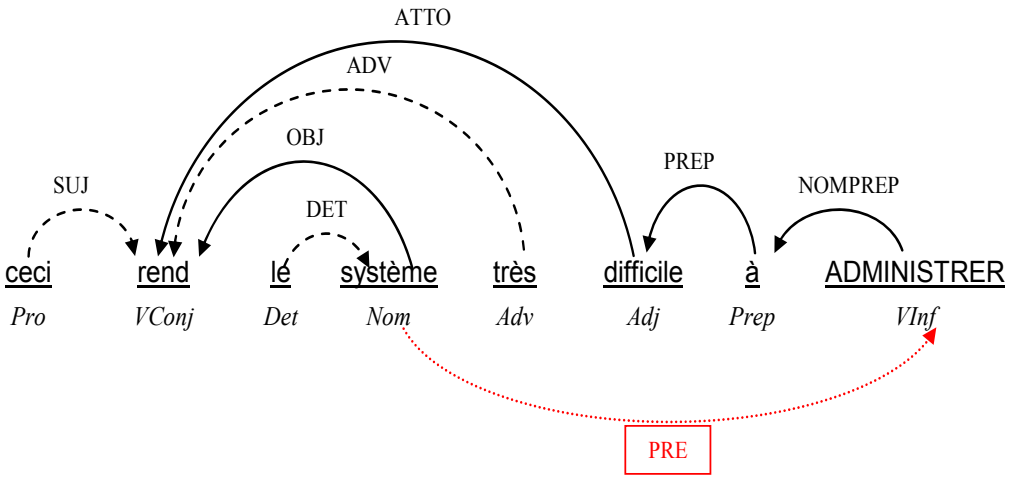


Figure 76 Dépendance Lexie| rendre+attribut à (Règle 28)

Lexie<ADMINISTRER, VInf, NOMPREP> + <à, Prep, NOMPREP, PREP> + <difficile, Adj, PREP, ATTO> + <rend, VConj, ATTO, OBJ> + <système, Nom, OBJ>
→<système, Nom|TETE, OBJ>

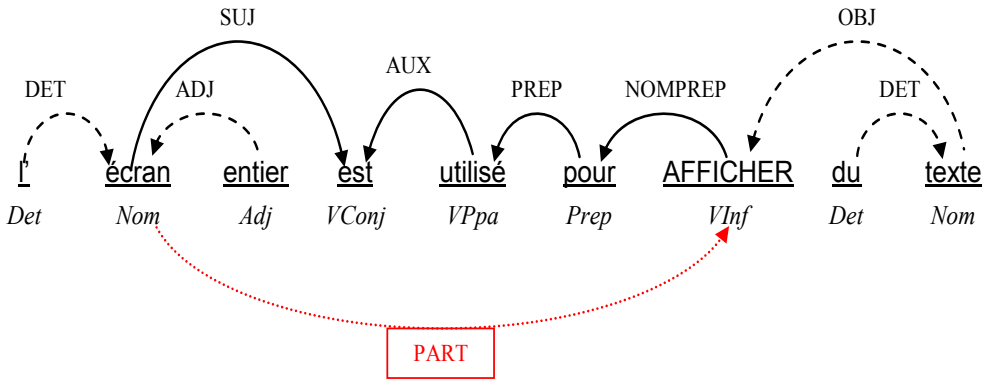


Figure 77 Dépendance Lexie| être utilisé pour (Règle 29)

Lexie<AFFICHER, VInf, NOMPREP> + <pour, Prep, NOMPREP, PREP> + <utilisé, VPpa, PREP, AUX> + <est, VConj, AUX, SUJ> + <écran, Nom, SUJ>
→ <écran, Nom, Lien indirect>

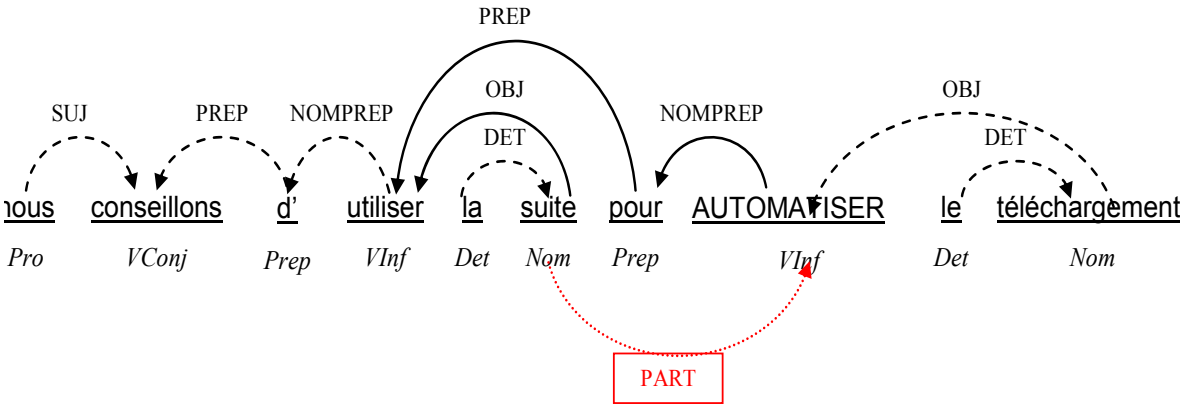


Figure 78 Dépendance Lexie|Verbe de+utiliser+pour (Règle 30)

Lexie<AUTOMATISER, VInf, NOMPREP> + <pour, Prep, NOMPREP, PREP> + <utiliser, VInf, , PREP, OBJ> + <suite, Nom, OBJ>
→ <suite, Nom, Lien indirect>

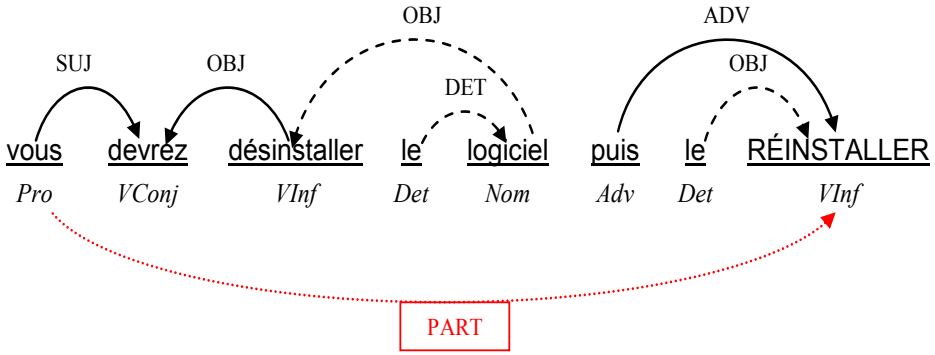


Figure 79 Autres types de dépendance (adverbe puis) (Règle 31)
<vous, Pro, SUJ> + <devrez, VConj, SUJ, OBJ> + <désinstaller, VInf, OBJ> + <logiciel, Nom, OBJ> ,
<puis, Adv, ADV> + Lexie<RÉINSTALLER, VInf, ADV>
→ <vous, Pro, SUJ>

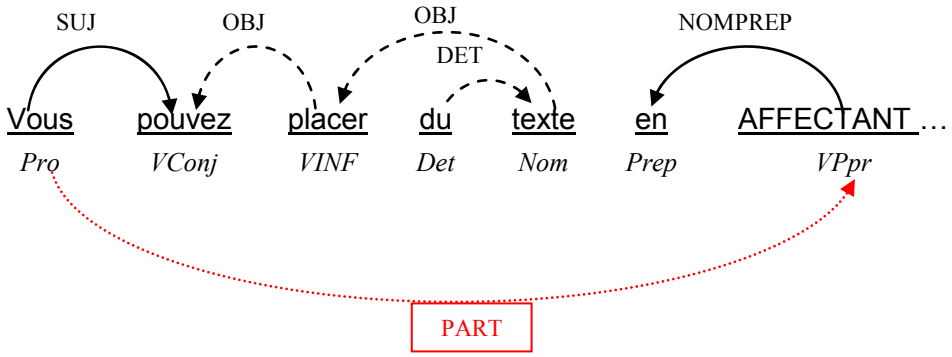


Figure 80 Autres types de dépendances (préposition en) (Règle 32)
<vous, Pro, SUJ> + <pouvez, VConj, SUJ, OBJ> + <placer, VInf, OBJ, OBJ> + <texte, Nom, OBJ> +
<en, Prep, NOMPREP> + Lexie<AFFECTANT, VPpr, NOMPREP>
→ <vous, Pro, SUJ>

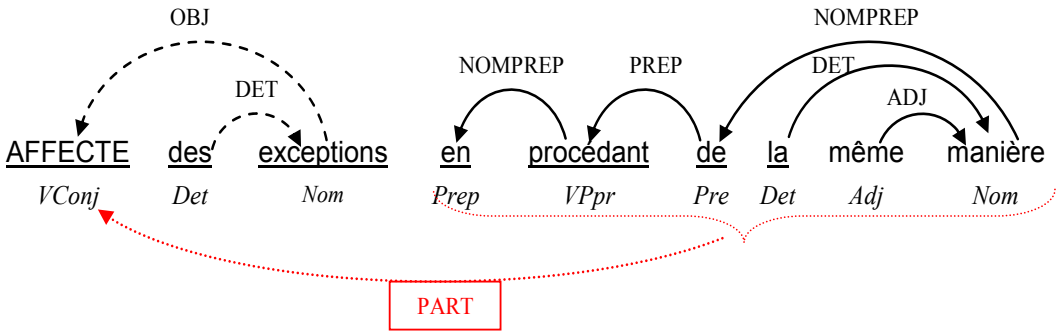


Figure 81 Autres types de dépendances (proposition) (Règle 33)

Lexie<AFFECTEZ, VConj, OBJ> + <exceptions, Nom, OBJ> , <en, Prep, NOMPREP> + <procédant, VPpr, NOMPREP, PREP> + <de, Prep, PREP, NOMPREP> + <la, Det, DET> + <même, Adj, ADJ> + <manière, Nom, NOMPREP>

→ <en procédant de la même manière>

Annexe 4

La version de l'article « LVF comparée aux autres ressources lexicales » dans la revue Langage.

Hadouche F. et Lapalme G., Une version électronique du LVF comparée avec d'autres Ressources lexicales. Langages n° 179 -180, 2010, Armand Colin, p.193-220.